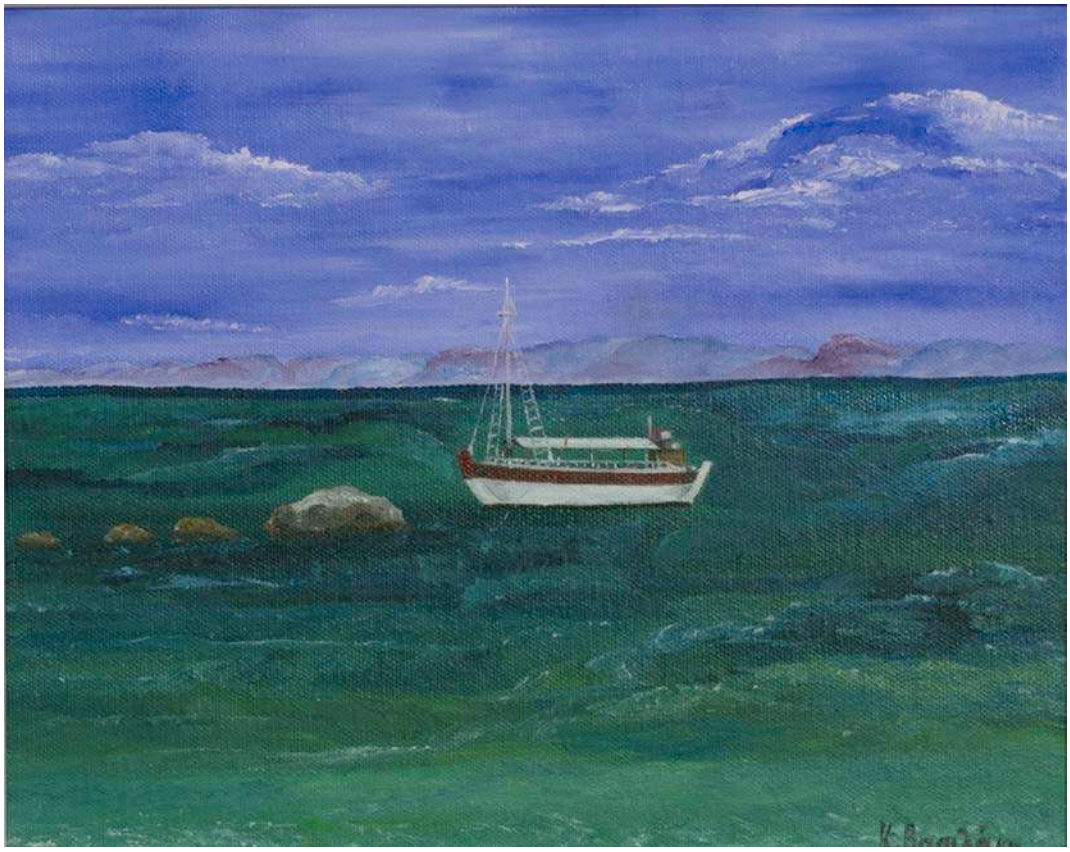
HELLENIC REPUBLIC
National and Kapodistrian University of Athens
Department of Informatics and Telecommunications

ABSTRACTS OF DOCTORAL DISSERTATIONS



Athens 2015

Volume 10

HELLENIC REPUBLIC
National and Kapodistrian
University of Athens

Department of Informatics and Telecommunications

ABSTRACTS OF DOCTORAL DISSERTATIONS

The Committee of Research and Development

A. Eleftheriadis
M. Koubarakis
E. Manolakos (Chair)
T. Theoharis

ISSN: 1791-7948

Volume 10

National and Kapodistrian University of Athens
Department of Informatics and Telecommunications
Panepistimiopolis, 15784 Athens, Greece

PREFACE

This volume includes extended abstracts of Doctoral Dissertations conducted in the Department of Informatics and Telecommunications, University of Athens, and completed from 1/2014 to 12/2014.

We publish this volume to demonstrate the breadth and quality of the original research performed by our Ph.D. students and faculty and to facilitate the dissemination of their innovative research results. We are happy to present the 10th yearly collection of this kind and expect this initiative to continue in the years to come. The submission of an extended abstract in English is required by all graduating doctoral students in our Department.

We would like to thank all graduates who contributed to this volume and hope that this was a positive experience for them. Finally, we would like to thank PhD candidate Nikos Bogdos for his help and attention to detail in putting together this volume.

The painting in the cover is called "*Journey in time*" contributed by artist *Katerina Vasilaki*.

The DiT Dept. Committee on Research and Development Activities

A. Eleftheriadis
M. Koubarakis
E. Manolakos (Chair)
T. Theoharis

Athens, May 2015

Table of Contents

Data Driven Adaptation of Heterogeneous Service Oriented Processes

Georgios Athanasopoulos*

National and Kapodistrian University of Athens
Department of Informatics and Telecommuncations
gathanas@di.uoa.gr

Abstract. Within the currently forming pervasive computing environment services and information sources thrive. Instantiations of the service oriented computing paradigm e.g. Web, Peer-to-Peer (P2P) and Grid services are continuously emerging, whilst information can be collected from several information sources e.g. materialisations of the Web 2.0 and Web 3.0 trends, Social Networking apps and Sensor Networks. Within this context the development of adaptable service oriented processes utilising heterogeneous services, in addition to available information is an emerging trend. This paper presents an approach and an enabling architecture that leverage the provision of data-driven, adaptable, heterogeneous service processes. Core within the proposed architecture is a set of interacting components that accommodate the acquisition of information, the execution of service chains and their adaptation based on collected information.

1 Dissertation Summary

Our era has been marked by a shift in the way of thinking and acting across many domains such as business, science and community. Cornerstone in this new way of thinking is the notion of service. In spite of the lack of a commonly accepted definition a service can be conceived as a function that is offered by someone and may be used by anyone else, or as Douglas Barry sets it more formally: *a function that is well-defined, self-contained, and does not depend on the context or state of other services* [1].

This shift in thinking and acting has an effect on software systems and the way they are developed. Service Oriented Computing (SOC) focuses on the use of services as system constituent parts. It is regarded as an evolution to the Component-Based development and distributed object oriented computing, and has been widely accepted as the current and future trend in distributed system development. Among its goals is to promote the loose coupling and flexible integration of the system parts in a far better way than component and object oriented technologies do.

* Dissertation Advisor: Aphrodite Tsalgatidou, Associate Professor

1.1 Service Oriented Process Adaptation

Service Oriented Processes (SOPs) constitute an indicative materialization of the loose coupling notion of SOC, as they rely on the lenient integration of comprising services. They are normally defined in higher-level languages, e.g. BPMN [2], WS-BPEL [3], and provide descriptions of coordinated flows of constituent (atomic or compound) services. One of their prime characteristics that has contributed to their proliferated usage is their easy execution by contemporary orchestration engines, e.g. Apache ODE [1].

Even though, SOP development is regarded as an evolution of Enterprise Application Integration approaches [4], it is also a new paradigm that is referred as Mega-Programming [5]. An inherently assigned characteristic, stemming from this consideration, is that processes are expected to be long-running and stateful as well as they may involve the interaction with stateless or stateful services. In the frame of our approach, SOPs are regarded as systems, which operate within a specific environment, and perform pre-specified activities, via the use of external services, in an orderly manner producing and/or consuming related information during their execution.

Nonetheless, in the currently forming Pervasive Computing environment, resources such as services and information sources emerge with an increasing pace. The Service-Oriented Computing (SOC) model has been instantiated by several distinct paradigms, e.g. Web, P2P, OGC services, and a plethora of service instances are emerging every day. In addition, the emerging Sensor Web [6], and the materializations of the Web 2.0 [6] and Web 3.0 [7] paradigms, e.g. Social Networking applications, provide new types of information sources. In this context, rendering SOPs able to tap onto these resources is becoming a necessity rather than an option. These resources could be used for the adaptation of SOPs with the goal to optimize their execution. Hence, processes should be flexible enough in order to facilitate the use of services that have not been identified at design time, irrespectively of their type, and the use of information that may stem from emerging sources.

This notion of process adaptation differs from contemporary definitions, e.g. Cassati et al [8], in that process adaptation should also support process optimization. Hence, the benefits acquired by process adaptation should be measurable in terms of specific indicators (criteria). The list of criteria that can be used for the optimization of SOPs may comprise cost, execution time, throughput, etc.

In the scope of our research we argue that process adaptation should be used for the optimization of the process execution time. An approach to achieve this optimization is via the reduction of unnecessary process activities. Considering the properties of SOPs, i.e. long-running activities, it is plausible to expect that the reduction of unnecessary process activities can lead to smaller execution times and as a result to higher throughput.

[1] `http://ode.apache.org/`

1.2 Background

Nonetheless, when looking into contemporary approaches for the provision of SOPs (dynamic or not) we see that they fail to accommodate all the aforementioned constraints, i.e. reuse of several types of services and information sources. For example process specification standards, e.g. WS-BPEL [3], or similar approaches, e.g. JOpera [9], are incapable of supporting the provision of adaptable processes. Even though they provide some form of flexibility, e.g. late-binding, or few of them the ability to accommodate heterogeneous types of services JOpera [9], they are capable neither to incorporate services, whose interface is not prescribed at the process design time, nor to use information from undefined sources.

More advanced approaches such as AI-based techniques, have been extensively applied in the provision of dynamically composed service oriented orchestrations [10]. Most of them focus on the provision of automatically constructed orchestrations (or similarly called task plans) that comprise Web services solely and neglect the information that may exist in the process environment, unless this is accessible through services. Along the same lines, the Context Aware Computing (CAC) research community has invested considerable efforts towards the use of contextual information for the adaptation of web service compositions, e.g. [11]. However, the majority of these approaches fail to address the interoperability concerns raised by the multiple instantiations of the service oriented computing paradigm [12].

Modern approaches, such as the ones based on Aspect Oriented Programming, e.g. [12], can provide the means for the adaptation of SOPs at runtime through the weavering of appropriate aspects. Nevertheless, the specification of hooks where these aspects will be integrated as well as the specification of the actual aspects themselves is an arduous task that should be performed by the process developer. In addition, when considering the vast amount of alternative services and/or information sources that could be incorporated in a process, the provision of the required aspects is becoming a daunting task.

Within this frame, the development of adaptable service processes, comprising heterogeneous services and information stemming from existing sources, is an emerging trend that calls for novel approaches.

2 Data Driven Adaptation

Our proposal to the specified requirements and research challenges is a set of middleware and tools catering for the provision of Data-Driven Adaptable Service Oriented Processes (DDA-SOP) that comprise heterogeneous types of services. The DDA-SOP approach is based on the combined use of a) the Tuplespace paradigm [13], for the collection and exchange of information with unanticipated information sources, and b) AI planning techniques [14], for the discovery of alternative execution paths that can exploit the collected information for the adaptation of a given SOP.

The prime assumption of our approach is based on the observation that *a SOP, comprising heterogeneous services, should be able to use the information available within its environment and adapt its execution accordingly.* In this frame, a process state is not solely depending on the values of its internal parameters, on the resulting outcomes from the invocation of its constituent services and/or on its internal operations but, also on information pertaining to the process environment, i.e. the process space; therefore, the latter should be taken into account during process development and execution.

The proposed approach (see Fig. 2) is implemented by a set of tools and middleware that accommodate the following three basic functional needs: Collection of contextual information; Execution of heterogeneous service processes; Process adaptation driven by collected information. More specifically these components are as follows:

- The *Semantic Context Space engine* (SCS engine) provides an open space where external information sources may place relevant information; in our context relevant information refers to semantically related information elements, which are structured and affiliated to concepts of a domain ontology.
- The *Process Optimizer* implements an AI planner that facilitates the discovery of process plans, which control the execution and adaptation of service processes at runtime.
- The *Service Orchestration engine* provides a BPEL-based engine that facilitates the execution of heterogeneous service orchestrations, whilst in parallel accommodates the monitoring and reconfiguration of processes according to the suggestions of the Process Optimizer.

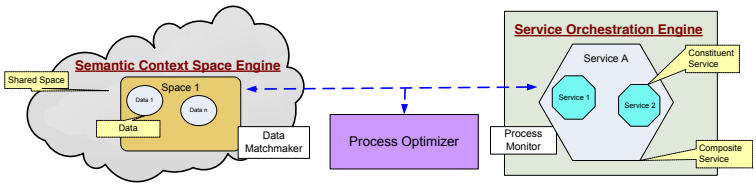


Fig. 1. Figure - High-level architecture

The contributions of this work can be briefly summarized into the following:

- An approach for mapping the problem of Data-Driven Adaptation of Service Oriented Processes to an extended, non-deterministic planning problem. To accommodate this, we use observations as the means for checking both the result of services invoked throughout the process and of the process environment, i.e. the space associated to a SOP. Details of the observed properties are used for defining queries that are executed on the associated space.

– Appropriate algorithms for extending the definition of a non-deterministic planning problem description with the inclusion of related observations and activities. The proposed approach relies on the use of ontologies and similarity measurements for the extrapolation of an original set of observations on a given SOP. In essence, our approach introduces additional related sensors for monitoring extra properties of a given SOP.
– A Generic Service Model capable of supporting the specification of common as well as distinct characteristics of various types of services. This service model enables us to use services irrespectively of their type in a seamless way. Distinct features, e.g. spatial and temporal characteristics of OGC services, provide us with required properties for customizing our approach, e.g. spatial and temporal properties are used in querying for related information as well as for discovering services.
– An environment catering for the provision of Data-Driven Adaptable Service Oriented Processes comprising heterogeneous types of services. The provided set of tools accommodates all functional needs of our approach, i.e. starting from the optimization of a given SOP model with the inclusion of appropriate adaptation steps, to the collection and exchange of related information with a SOP running on the service orchestration engine.

In addition to this list of outcomes our approach also contributed to the:

– Design and implementation of an engine facilitating the collection and exchange of information annotated with several types of meta-information. This engine is an open source, extendable implementation of a Tuplespace service that includes extensions for the collection and exchange of information, which is annotated with semantics, e.g. RDF-based, or WSML-based, as well as spatial and/or temporal meta-information.
– Specification of a Peer-to-Peer Service Definition Language, namely PSDL, with groundings for the description of JXTA services. PSDL is a WSDL-based [**?**] language with extensions for the description of Peer-to-Peer services.
– Design and implementation of a middleware catering for the dynamic invocation of Peer-to-Peer services described in PSDL. This middleware is an extension to the Web Service Invocation Framework [**?**] that uses PSDL service descriptions for the static and dynamic invocation of Peer-to-Peer services.

2.1 Process Optimization

Both the Semantic Context Space engine and the Service Orchestration engine can facilitate the provision of service-oriented processes, which are capable of exchanging information with external sources/actors. Nevertheless, the provision of processes that are able to incorporate additional behaviors, i.e. in terms of services, and information that is related to a given process, is facilitated via the Process Optimizer.

Building on the representation of the automated process composition problem as a non-deterministic, partially observable planning problem (see Def. 1) we can reuse existing planners for the discovery of policies, i.e. conditional plans, that guide the execution of a process for achieving a set of defined goals [15]. Conditions in such plans, expressed as if-then-else structures, assert the values of specific observations and decide on the execution of specific sets of actions.

Definition 1. *A non-deterministic, partially observable system (Σ) can be defined as a tuple $\Sigma = \langle S, A, R, O, X \rangle$, where:*

- *S is the finite set of states of the associated state transition system,*
- *is the finite set of actions $A = \{a_i | i \leq n \wedge i, n \in \mathbb{N}\}$*
- *$R \subset S \times A \times S$ is the transition relation,*
- *O is a finite set of observation variables $O = \{o_i | i \leq n \wedge i, n \in \mathbb{N}\}$*
- *$X_O : S \times \{\top, \bot\}$ is the relation for the evaluation of observation variables $o \in O$ on each state $s \in S$. Within our context the value of an observation variable is independent of the action that may have preceded*

In the frame of such planning problems the transition relation R maps the execution of an action $a \in A$, on a state $s \in S$ (assuming that a is applicable on s) to possibly more than one successor states i.e. $S' \subseteq S, \quad |S'| \geq 1$. An action a is applicable on a state $s \in S$, if and only if, there exists a state $s' \in S$ such that $R(s, a, s')$ stands. The set O contains the finite set of observation variables o_i whose values are evaluated at runtime. The value of each observation variable at each state is defined by the observation relation X. A simplification normally introduced to avoid the unnecessary complexities and to render the planning problem finite, is to consider observation variables as boolean variables whose values could be either true or false (i.e. $\{\top, \bot\}$). Therefore if $X_o(s, \top)$ holds at a state $s \in S$ then the value of variable o at state s is *True*. The dual holds in cases where variable o is *False*. In cases where both $X_o(s, \top)$ and $X_o(s, \bot)$ hold, variable o has an undefined value.

Nevertheless, even though this representation (Def. 1) is capable of supporting the non-deterministic and partially observable nature of services, it needs additional extensions so as to accommodate the requirements of data-driven adaptable processes [16]. Briefly the limitations of this representation are that a) it cannot consider information, which comes from sources that are not prespecified, i.e. information that is neither produced by the interacting services, nor the process at hand, and b) it fails to consider information that is related, but not exactly what is expected by the process. To accommodate these concerns we provide appropriate extensions, which consist of a) additions to the set of observations (O) and the set of actions (A) so as to leverage the consideration of related information, and b) a mechanism facilitating the valuation of observations (i.e. observation variables) based on queries executed over an open set of information elements that can be collected from external sources.

These two extensions are implemented by the Process Optimizer and the Semantic Context Space engine respectively. Although both extensions are of equal importance for the provision of data-driven adaptable processes, in the following we elaborate on the extensions of the observation and action sets.

2.2 Process Expansion

A crucial step in the provision of Data-Driven Adaptable Service-Oriented process is the introduction of extensions to the observation and action sets of the planning problem (Def. 1). This step can be regarded as *the incorporation of additional sensors for monitoring the process, along with appropriate actions for handling the accruing believes*. Nevertheless, this expansion processes should be guided so as to avoid the introduction of an overwhelming set of observations and actions that will make the planning problem practically intractable.

Starting with a WS-BPEL specification, called D hereafter, the first step is to identify the initial set of observations and actions associated to the process at hand. With regards to the initial set of observations, this should maximize the coverage of process states, whilst the initial set of actions should include the services associated to the process along with the process's internal activities, e.g. assign activities.

Similar to other existing approaches, e.g. [17], the most appropriate candidate for monitoring, i.e. assigning observations, is the set of interacting parties. Interactions with external service providers can be modeled via the exchange of messages over a set of predefined channels [17]; read and write operations performed over these channels are the means for controlling the state of the process. In this context, the initial set of observations includes observation properties, which monitor the outcomes of the services interacting with the process at hand. It is important here to note that in the frame of our work we do not consider observations for service interactions performed inside loop control structures. This is because such a consideration would entail an (explicit or implicit) ordering on the set of observed properties so as to avoid consistency errors, e.g. using observed information that is for a consequent execution step and not for the current.

Given the set of original observations (O) and actions (A) along with the related states (S), transition relations (R) and observation valuations (X), extracted from the provided process model D, the steps of the model expansion process include a) the semantic-based extension of the observation set, b) the extension of the action set with actions capable of supporting the exploitation of the introduced observations and c) the consolidation of the extended action and observation sets. In the following we provide additional details on each of the comprising steps of the expansion process.

2.3 Observation Set Expansion

The underpinning assumption which drives the observation expansion is that *instead of considering partially matching results to the performed observations we may as well look for exact matches to partially matching (i.e. similar) observations*. Hence, the set of observations $_D$ linked to D is expanded with the introduction of similar candidate observations.

Definition 2. *An observation property* $o \in O$ *can be defined as a tuple* $o = \langle name_o, vc_o, to \rangle$ *, where:*

– *$name_o$ is the name of the observation,*
– *vc_o is the semantic concept associated to the given observation and*
– *t_o is the syntactic type, i.e. specifies the related domain of values for the observation o.*

To facilitate the expansion we introduce two additional features, i.e. an expansion function (exp_o) and an expansion ratio property (r_{exp}). The r_{exp} property dictates the minimum similarity distance (i.e. the cut-off value) between the concepts of an ontology so as to consider them part of the same set (i.e. expansion set). Thus, given an ontology (V), a concept (vc) and an expansion ration (r_{exp} the exp_o function (see 1) returns all concepts of the provided ontology with similarity to vc equal to or greater than r_{exp}.

$$exp_o : \mathbb{R} \times V \rightarrow 2^V \tag{1}$$

For the calculation of the expansion function (1) several similarity distance measurements can be used. Such a measure providing an indication of similarity between two concepts, i.e. $vc_a, vc_b \in V$ based on the IS-A relation hierarchy, is Dices coefficient. We need to state here that our approach is capable of supporting additional measures through the inclusion of appropriate plug-ins.

Given the set of original observations O_D that are semantically associated to the concepts of an ontology V, applying exp_o generates a set of extra concepts. If O_V is the set of concepts assigned to the observations in O_D, i.e. $O_V = \{vc_i | vc_i = vc_{o_i}, \forall o_i \in O_D\}$, then O'_V contains the additional concepts introduced through the expansion process (2).

$$exp_o\left(r_{exp}, O_V\right) = O'_V \tag{2}$$

This set of extended concepts (2) can be regarded as a set of candidate observations to be considered by the process D. However, this set of candidate observations is partially defined, as O'_V contains solely the semantic concepts (vc_i) of the extended observations. Candidate observations should be refined so as to specify their syntactic types (t_i) and names as well. Moreover, the set of candidate observations should be pruned from observations (i.e. semantic concepts) that cannot be handled by available actions. These refinements are presented next.

2.4 Action Set Expansion

The set of actions (A_D) originally specified for a process (D) point to service operations that are already defined in the process specification. This set of actions should be extended with the inclusion of additional actions, i.e. service operations, that will be able to use the extended set of observations. To facilitate this extension we introduced a service query engine and appropriate heuristics. The prime goal of this engine is to discover alternative service chains (Sc) comprising one or multiple actions (i.e. service operations) that can use as input the extended observations and lead to the achievement of the process goals or sub-goals.

Definition 3. *Sub-goals S_{sg} in the context of this paper refer to intermediate states of the specified process D, i.e. $S_{sg} = S_D - (S_0 \cup G)$.*

The service query engine uses as input the merged set of observation concepts, i.e. $O_V \cup O'_V$, the set of process states (S_D), and the set of process goals (G). The outcome of the discovery process is a service chain Sc, which satisfies the specified search objectives; a service chain Sc (see 3) is typically defined as a finite, ordered set of actions (i.e. service operations).

$$Sc = \{a_0 \prec a_1 \prec \cdots \prec a_n, \ and \ n \in \mathbb{N}\} \tag{3}$$

In the frame of our approach we consider candidate service chains corresponding to acyclic finite automata. This is in order to comply with the rule for shorter execution paths; by introducing cyclic service chains we risk ending-up with longer paths, depending on the iterations that may incur at execution time. Moreover, we do not assign additional observations for monitoring the outcomes of the discovered service chains. This is because introducing additional observations would increase the complexity of the accruing planning problem, without modifying its nature, e.g. render it fully observable, or providing additional benefits.

To accommodate the reuse of heterogeneous types of services, i.e. not just Web services, we reuse and built upon a generic service model that facilitates the modeling of the commonly shared and distinct characteristics of several types of services [**?**]. According to this model a service is a collection of operations, which are accessed by service clients via the exchange of appropriate messages over the web at specific endpoints.

Given the set of features specified in the service model candidate service chains should have input messages with concepts defined in the merged set of observation concepts, i.e. $Oc \cup Oc'$, and their outcomes should lead to either the goal states (G) or sub-goal states (S_{sg}) of process D.

Nonetheless, the formulation of the respective service queries is primarily controlled by the discovery strategy to be used. The two most appealing ones are the forward and backward search strategies. In the backward search strategy, the goal is to define service queries, which set constraints on the expected output messages (Os) and post-conditions (C_{post}) of a candidate action, in addition to checking whether input messages contain ontology concepts related to the merged concept set $Oc \cup Oc'$. Contrary to that, in a forward search strategy the goal is to define queries that constraint the input messages (Is), i.e. starting with the ontology concepts of merged concept set $Oc \cup Oc'$, and checking if a specific process goal (G) or sub-goal (S_{sg}) is achieved. In the provided implementation we employed a forward search strategy. The description of the employed strategy is skipped for reasons of brevity.

Irrespectively of the applied search strategy and in order to avoid searching through a very large set of candidate services, we exploit the given problem description for the introduction of appropriate search bounds and heuristics. Based on the objectives set beforehand, the introduced adaptations should reduce the set of actions used for achieving the goals (G) or sub-goals (S_{sg}) of process D.

Definition 4. *We, introduce a function* $lenghtTo\,(s_a, s_b)$*, which returns the distance, i.e. the cardinality of the set of transitions, between state* s_a *and the requested state* s_b *of a process or of a given service chain (4).*

$$lenghtTo : S \times S \rightarrow \mathbb{N} \tag{4}$$

Based on (4), for a process D with a set of starting states s_0 and an intermediate or goal state $s_i \in \{S_{sg} \cup G\}$, a candidate service chain Sc with a starting state s_0 and the same goal state $s_i \in \{S_{sg} \cup G\}$ (i.e. Sc leads to the same state as in D) a search bound can be expressed as follows:

$$lenghtTo\,(s_0, s_i) > lenghtTo\,(s'_0, s_i) \tag{5}$$

Using this search bound we can avoid falling into an endless and iterative search over a large set of candidate service chains. To further enhance the search process we can also introduce additional heuristics, by exploiting knowledge from the problem domain. In conclusion, the outcome of the action set expansion process is a finite set of candidate service chains (6) for the given set of candidate observations, when searching over a finite service registry:

$$SC = \{Sc_{vc} | Sc_{vc}, vc \in O_V \cup O'_V\} \tag{6}$$

2.5 Observation and Action Set Consolidation

The calculated candidate observation (O'_V) and service chain sets (SC) contain values that do not correspond to each other, hence before proceeding with the formulation of the extended planning problem, i.e. ED, these sets should be cleaned up. Thus, starting from the set of additional service chains $Sc \in SC$ each of the candidate observations, i.e. $o' \in O'_V$ that is not used from any of the identified service chains, i.e. $\nexists Sc_{vc} \in SC \wedge vc_{Sc} = vc_{o'}$, is removed from the candidate observations set. The pruned set of candidate observations O''_V contains all semantic concepts that are used by the identified service chains.

The next step, is the specification of the corresponding observation variables. According to the definition given for an observation variable, i.e. Def. 2, apart from the associated semantic concept, the syntactic type along with a name are also required. Based on our hypothesis that each candidate observation, i.e. $o' \in O'_V$ should be used by at least one service chain, i.e. $Sc \in SC$, we can safely assume that the syntactic type of each observation variable should correspond to the one expected by the candidate service chain, Sc. Nonetheless, as a candidate observation may be used by several service chains we should introduce different observation variables for each of these chains unless they use the same syntactic type for the candidate observation; in the latter case we consider that they refer to the same variable. Each of the newly introduced variables according to the definition given in Def. 1 is expected to stand prior to the execution of the related service chains, whilst in other cases it may be undefined. The resulting observation variable set for the extended planning problem is the union of the original observation and the pruned observations O'', i.e. $O_{ED} = O_D \cup O''$.

Following the formulation of the extended observation variable set, the consolidation process continues with the merge of the extended action set A_{SC} and the accruing state sets, i.e. S_{SC}, to the original action A_D and state S_D sets. Since additional service chains $Sc \in SC$ can be modeled as independent, finite automata (i.e. STS_{Sc_i}), the resulting action and state sets can be calculated from the composition [17] of the original process automaton with each of the additional service chain automata, i.e. $STS_{ED} = STS_D \otimes STS_{Sc_0} \otimes STS_{Sc_1} \cdots \otimes STS_{Sc_n}$.

The formulation of the extended planning problem finishes with the specification of the initial and goal states of the extended planning problem. These correspond to the constraints of the original process D. The resulting planning problem can be then fed to an external planner for the identification of possible solutions. The resulting solution corresponds to the automaton of the controlling process, i.e. the extended adaptable process, that can be transformed to a WS-BPEL process model. This transformation process is not described here for reasons of brevity.

3 Conclusions

The provision of adaptable Service-Oriented Processes (SOPs) in the currently forming web is becoming a necessity as the number of untapped resources increases. Within this new Web era, services, of various types, and information are partially utilized by existing approaches tackling the provision of adaptable SOPs. To this end our approach, the Data-Driven Adaptation of Service Oriented Processes (DDA-SOP), can accommodate the provision of adaptable SOPs that exploit both existing services, irrespectively of their type, and information, i.e. semantically annotation structured data, for their adaptation. The main goal of our approach has been the increase of the overall process performance. To achieve this goal we re-design the given processes so that when related information emerges at the process environment along with the use of existing services we end up with smaller execution paths.

The evaluation of our prototype implementation has clearly shown that our approach can lead into significant performance improvements even from small adaptation rates. These improvements get even more evident in the case of long running processes, where improvements appear from very small adaptation rates, e.g. $< 2\%$. As a result we also observed an increase in the throughput of the executing process, i.e. the number of requests served per second, which is proportional to the achieved performance gains. We have to note that the cost paid for the calculation of DDA-SOPs may seem to counter the benefits accruing by our approach. A crucial factor in the calculation of these costs is the complexity and size, i.e. in terms of involved actions, of a given process model. Nevertheless, considering that these costs are paid once prior to the deployment and execution of the resulting DDA-SOP this is not a prohibiting cost.

The provided prototype implementation has unveiled a wide range of future directions that deserve further research. These directions include improvements and optimizations to our prototype implementation as well as emerging research

fields such as: a) investigation of the implicit interaction pattern as a mechanism for the loose coupling of complex processes, b) use of different planning techniques such as stochastic models, e.g. Markov based planning techniques, and c) use of different similarity measurements for the discovery of related observations.

References

1. Barry, D.K.: Service-oriented architecture definition. Web Services and Service-Oriented Architectures (2010)
2. OMG: Business process model and notation (bpmn), v. 2.0, 2011. OMG recommendation
3. Alves, A., Arkin, A., Askary, S., Barreto, C., Bloch, B., Curbera, F., Ford, M., Goland, Y., Guízar, A., Kartha, N., Liu, C.K., Khalaf, R., König, D., Marin, M., Mehta, V., Thatte, S., van der Rijn, D., Yendluri, P., Yiu, A.: Web services business process execution language version 2.0. OASIS standard (April 2007)
4. Alonso, G., Casati, F., Kuno, H., Machiraju, V.: Web services. Springer (2004)
5. Pautasso, C., Alonso, G.: From web service composition to megaprogramming. In: Technologies for E-Services. Springer (2005) 39–53
6. Botts, M., Percivall, G., Reed, C., Davidson, J.: Ogc sensor web enablement: Overview and high level architecture. In: GeoSensor networks. Springer (2008) 175–190
7. Spivack, N.: Web 3.0: The third generation web is coming (2006)
8. Casati, F., Ilnicki, S., Jin, L., Krishnamoorthy, V., Shan, M.C.: Adaptive and dynamic service composition in eflow. In: Advanced Information Systems Engineering, Springer (2000) 13–31
9. Pautasso, C., Alonso, G.: Jopera: a toolkit for efficient visual composition of web services. International Journal of Electronic Commerce (IJEC) **9** (Winter 2004/2005 2004) 107–141
10. Rao, J., Su, X.: A survey of automated web service composition methods. In: In Proceedings of the First International Workshop on Semantic Web Services and Web Process Composition, SWSWPC 2004. (2004) 43–54
11. Qiu, L., Shi, Z., Lin, F.: Context optimization of ai planning for services composition. In: (ICEBE'06) IEEE International Conference on e-Business Engineering, IEEE (2006) 610–617
12. Kongdenfha, W., Saint-Paul, R., Benatallah, B., Casati, F.: An aspect-oriented framework for service adaptation. In: Service-Oriented Computing–ICSOC 2006. Springer (2006) 15–26
13. Rossi, D., Cabri, G., Denti, E.: Tuple-based technologies for coordination. In: Coordination of Internet agents. Springer (2001) 83–109
14. Russell, S., Norvig, P., Intelligence, A.: A modern approach. Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs **25** (1995)
15. Nau, D., Ghallab, M., Traverso, P.: Automated Planning: Theory & Practice. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2004)
16. Athanasopoulos, G., Tsalgatidou, A.: An approach to data-driven adaptable service processes. In: ICSOFT (1). (2010) 139–145
17. Pistore, M., Barbon, F., Bertoli, P., Sharparau, D., Traverso, P.: Planning and monitoring web service composition. In: Artificial Intelligence: Methodology, Systems, and Applications, 11th International Conference (AIMSA 2004). (2004) 106–119

Proof Procedures for Extensional Higher-Order Logic Programming*

Angelos Charalambidis

Department of Informatics & Telecommunications
University of Athens
a.charalambidis@di.uoa.gr

Abstract. We consider an extensional higher-order logic programming language which possesses the minimum Herbrand model property. We propose an SLD-resolution proof procedure and we demonstrate that it is sound and complete with respect to this semantics. In this way, we extend the familiar proof theory of first-order logic programming to apply to the more general higher-order case. We then enhance our source language with constructive negation and extend the aforementioned proof procedure to support this new feature. We demonstrate the soundness of the resulting proof procedure and describe an actual implementation of a language that embodies the above ideas.

1 Introduction

The two most prominent declarative paradigms, namely logic and functional programming, differ radically in an important aspect: logic programming is traditionally first-order while functional programming encourages and promotes the use of higher-order functions and constructs. The initial attitude of logic programmers towards higher-order logic programming was somewhat skeptical: it was often argued that there exist ways of encoding or simulating higher-order programming inside Prolog itself (see, for example, [8]). However, ease of use is a primary criterion for a programming language, and the fact that higher-order features can be simulated or encoded does not mean that it is practical to do so.

Eventually extensions with genuine higher-order capabilities were introduced. These extensions allow predicates to be applied but also passed as parameters. The existing proposals can be placed in two main categories, namely the *intensional* and the *extensional* ones. In the former category, the two most prominent languages are λProlog [6] and HiLog [4]. The latter category is much less developed: currently there exist two main proposals for extensional higher-order logic

* Advisor: Panos Rondogiannis, Associate Professor, University of Athens. The doctoral dissertation has been co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Heracleitus II. Investing in knowledge society through the European Social Fund.

programming, namely [7] and [1]; however, apart from the work reported in this dissertation, no other actual systems have been built so far.

In an extensional language, two predicates that succeed for the same instances are considered equal. On the other hand, in an intensional language it is possible that predicates that are equal as sets will not be treated as equal. In other words, a predicate in an intensional language is more than just the set of arguments for which it is true. For example, in Hilog, two predicates are not considered equal unless their names are the same.

Example 1. Consider a program that consists only of the following rule:

```
p(Q):-Q(0),Q(1).
```

In an extensional logic language, predicate `p` can be understood in purely set-theoretic terms: `p` is the set of all those sets that contain both `0` and `1`.

It should be noted that this program is also a syntactically acceptable program of the existing intensional logic programming languages. The difference is that in an extensional language the above program has a purely set-theoretic semantics. Actually, as we are going to see, this set theoretic interpretation allows us to permit queries of the form `?-p(R)` which will get meaningful answers (the answer in this case will express the fact that `R` is any relation which is true of both `0` and `1`). Notice that an intensional language will not in general provide an answer to such a query (since there does not exist any actual predicate defined in the program that is true of both `0` and `1`).

The first work in this area was [7]. In that paper, W. W. Wadge demonstrated that there exists a modest fragment of higher-order logic programming that can be understood in purely extensional terms. More specifically, Wadge discovered a simple syntactic restriction which ensures that compliant programs have an extensional declarative reading. Roughly speaking, the restriction says that rules about predicates can state general principles but cannot pick out a particular predicate for special treatment. Wadge gave several examples of useful extensional higher-order programs and outlined the proof of a minimum-model result. Finally, Wadge conjectured that a sound and complete proof system exists for his fragment, but did not further pursue such an investigation.

2 The Proposed Approach: an Intuitive Overview

The first problem we consider is to bypass one important restriction of [7], namely the inability to handle program clauses or queries that contain uninstantiated predicate variables. The following example illustrates these ideas:

Example 2. Consider the following higher-order logic program written in an extended Prolog-like syntax.

```
p(Q):-Q(0),Q(s(0)).
nat(0).
nat(s(X)):-nat(X).
```

The Herbrand universe of the program is the set of natural numbers in successor notation. According to the semantics of [7], the least Herbrand model of the program assigns to predicate `p` a continuous relation which is true of *all* unary relations that contain at least `0` and `s(0)`. Consider now the query `?-p(R)` which asks for all relations that satisfy `p`. Such a query seems completely unreasonable, since there exist uncountably many relations that must be substituted and tested in the place of `R`.

In our example, despite the fact that there exists an infinite number of relations that satisfy `p`, all of them are supersets of the finite relation $\{0, s(0)\}$. In some sense, this finite relation *represents* all the relations that satisfy `p`. But how can we make the notion of "finiteness" more explicit? For this purpose we adopt the semantics described in [3]. The new semantics allows us to introduce a relatively simple, sound and complete proof system which applies to programs and queries that may contain uninstantiated predicate variables. The key idea can be demonstrated by continuing Example 2. Given the query: `?-p(R)` one (inefficient and tedious) approach would be to enumerate all possible finite relations of the appropriate type over the Herbrand universe. Instead of this, we use an approach which is based on what we call *basic templates*: a basic template for `R` is (intuitively) a finite set whose elements are individual variables. For example, assume that we instantiate `R` with the template $\{X, Y\}$. Then, the resolution proceeds as follows.

```
?-p(R)
?-p({X,Y})
?-{X,Y}(0),{X,Y}(s(0))
?-{0,Y}(s(0))
?-true
```

and the proof system will return the answer $R = \{0, s(0)\}$. The proof system will also return other finite solutions, such as $R = \{0, s(0), Z_1\}$, $R = \{0, s(0), Z_1, Z_2\}$, and so on. However, a slightly optimized implementation can be created that returns only the answer $R = \{0, s(0)\}$, which represents all the finite relations produced by the proof system. The intuition behind this answer is that the given query succeeds for all unary relations that contain at least `0` and `s(0)`.

One basic property of all the higher-order predicates that can be defined in the language considered so far, is that they are monotonic. Intuitively, the monotonicity property states that if a predicate is true of a relation `R` then it is also true of every superset of `R`. However, there are many natural higher-order predicates that are *nonmonotonic* and which a programmer would like to be able to write. For example, assume we want to define a predicate `disconnected(G)` which succeeds if its input argument, namely a graph `G`, is disconnected. A graph is simply a set of pairs, and therefore `disconnected` is a second-order predicate. Notice that `disconnected` is obviously nonmonotonic: given graphs `G1` and `G2` with $G1 \subseteq G2$, it is possible that `disconnected(G1)` succeeds but `disconnected(G2)` fails.

The obvious idea in order to add nonmonotonicity is to enhance the language with negation-as-failure. However, this is not as straightforward as it sounds,

because even the simpler higher-order programs with negation face the well-known problem of *floundering* [5]. In classical logic programming a computation is said to flounder if at some point a goal is reached that contains only nonground negative literals.

Fortunately, there exists an approach to negation-as-failure that bypasses the problem of floundering. This is usually called *constructive negation* [2] and its main idea can be explained by a simple example. Consider the predicate `p` that holds for `1` and `2` and consider the query `?-not(p(X))`. The original idea behind constructive negation [2] is that in order to answer a negative query that contains uninstantiated variables, the following procedure must be applied: we run the positive version of the query and we collect the solutions as a disjunction; we then return the negation of the disjunction as the answer to the original query. In our example, the corresponding positive query (namely `?-p(X)`) has the answers `X=1` and `X=2`. The negation of their disjunction is the conjunction $(\texttt{X} \neq \texttt{1}) \wedge (\texttt{X} \neq \texttt{2})$. Observe now that the procedure behind constructive negation returns as answers not only substitutions (as it happens in negationless logic programming) but also inequalities. Generalizing the above idea to the higher-order setting requires the ability to express some form of inequalities regarding the elements of sets. Intuitively, we would like to express that some element *does not belong to a set*.

Example 3. Consider the following simple program:

```
p(Q):-Q(0),not(Q(1)).
```

Intuitively, `p` is true of all relations that contain `0` but they do not contain `1`. A reasonable answer for `?-p(R)` would be $\texttt{R} = \{\texttt{0}\} \cup \{\texttt{X} \mid \texttt{X} \neq \texttt{1}\}$.

It turns out that the extension of higher-order logic programs with constructive negation offers a much greater versatility to extensional higher-order logic programming. We can extend higher-order logic programming with constructive negation. Moreover, there exists a relative simple sound proof procedure for the new language.

3 Definite Higher-Order Programs

Definition 1. *A type can either be* functional*,* argument*, or* predicate*, denoted by σ, ρ and π respectively and defined as:*

$$\begin{aligned} \sigma &:= \iota \mid (\iota \to \sigma) \\ \rho &:= \iota \mid \pi \\ \pi &:= o \mid (\rho \to \pi) \end{aligned}$$

We will use τ to denote an arbitrary type.

As usual, the binary operator $\to$ is right-associative. A functional type that is different from ι will often be written in the form $\iota^n \to \iota$, $n \geq 1$. Moreover, it can be easily seen that every predicate type π can be written in the form $\rho_1 \to \cdots \to \rho_n \to o$, $n \geq 0$ (for $n = 0$ we assume that $\pi = o$).

Definition 2. *The set of* positive expressions *of the higher-order language $\mathcal{H}$ is recursively defined as follows.*

1. *Every predicate variable (respectively, predicate constant) of type π is a positive expression of type π; every individual variable (respectively, individual constant) of type ι is a positive expression of type ι; the propositional constants* false *and* true *are positive expressions of type o.*
2. *If* f *is an n-ary function symbol and $\mathsf{E}_1, \ldots, \mathsf{E}_n$ are positive expressions of type ι, then $(\mathsf{f}\ \mathsf{E}_1 \cdots \mathsf{E}_n)$ is a positive expression of type ι.*
3. *If E_1 is a positive expression of type $\rho \to \pi$ and E_2 is a positive expression of type ρ, then $(\mathsf{E}_1\mathsf{E}_2)$ is a positive expression of type π.*
4. *If V is an argument variable of type ρ and E is a positive expression of type π, then $(\lambda\mathsf{V}.\mathsf{E})$ is a positive expression of type $\rho \to \pi$.*
5. *If $\mathsf{E}_1, \mathsf{E}_2$ are positive expressions of type π, then $(\mathsf{E}_1 \bigwedge_\pi \mathsf{E}_2)$ and $(\mathsf{E}_1 \bigvee_\pi \mathsf{E}_2)$ are positive expressions of type π.*
6. *If $\mathsf{E}_1, \mathsf{E}_2$ are positive expressions of type ι, then $(\mathsf{E}_1 \approx \mathsf{E}_2)$ is a positive expression of type o.*
7. *If E is an expression of type o and V is an argument variable of type ρ, then $(\exists_\rho \mathsf{V}\,\mathsf{E})$ is a positive expression of type o.*

The notions of *free* and *bound* variables of a positive expression are defined as usual. A positive expression is called *closed* if it does not contain any free variables.

Definition 3. *The set of* clausal expressions *of the higher-order language $\mathcal{H}$ is defined as follows.*

1. *If p is a predicate constant of type π and E is a closed positive expression of type π then $\mathsf{p} \leftarrow_\pi \mathsf{E}$ is a clausal expression of $\mathcal{H}$, also called a* program clause.
2. *If E is a positive expression of type o, then $\mathsf{false} \leftarrow_o \mathsf{E}$ (usually denoted by $\leftarrow_o \mathsf{E}$ or just $\leftarrow \mathsf{E}$) is a clausal expression of $\mathcal{H}$, also called a* goal clause.

All clausal expressions of $\mathcal{H}$ have type o. A program of $\mathcal{H}$ is a finite set of program clauses of $\mathcal{H}$.

Example 4. The following is a higher-order program that computes the closure of its input binary relation `R`. The type of `closure` is $\pi = (\iota \to \iota \to o) \to \iota \to \iota \to o$.

$$\texttt{closure} \leftarrow_\pi \lambda\texttt{R}.\lambda\texttt{X}.\lambda\texttt{Y}.(\texttt{R X Y})$$
$$\texttt{closure} \leftarrow_\pi \lambda\texttt{R}.\lambda\texttt{X}.\lambda\texttt{Y}.\exists\texttt{Z}((\texttt{R X Z}) \wedge (\texttt{closure R Z Y}))$$

A possible query could be: $\leftarrow$ `(closure R a b)` (which intuitively requests for those binary relations such that the pair $(\texttt{a}, \texttt{b})$ belongs to their transitive closure). In a Prolog-like extended syntax, this program would have been written as:

```
closure(R, X, Y) :- R(X, Y).
closure(R, X, Y) :- R(X, Z), closure(R, Z, Y).
```

and the corresponding query as $\leftarrow$ `closure(R, a, b)`.

3.1 SLD Resolution

Definition 4. *The set of* basic expressions of $\mathcal{H}$ *is recursively defined as follows. Every expression of* $\mathcal{H}$ *of type* ι *is a basic expression of type* ι*. Every predicate variable of* $\mathcal{H}$ *of type* π *is a basic expression of type* π*. The propositional constants* false *and* true *are basic expressions of type* o*. A non empty finite union of expressions each one of which has the following form, is a basic expression of type* $\rho_1 \to \cdots \to \rho_n \to o$ *(where* $\mathsf{V}_1 : \rho_1, \ldots, \mathsf{V}_n : \rho_n$*):*

1. $\lambda \mathsf{V}_1 . \cdots \lambda \mathsf{V}_n . \mathsf{false}$
2. $\lambda \mathsf{V}_1 . \cdots \lambda \mathsf{V}_n . (\mathsf{A}_1 \wedge \cdots \wedge \mathsf{A}_n)$*, where each* A_i *is either*
   (a) $(\mathsf{V}_i \approx \mathsf{B}_i)$*, if* $\mathsf{V}_i : \iota$ *and* $\mathsf{B}_i : \iota$ *is a basic expression where* $\mathsf{V}_j \notin fv(\mathsf{B}_i)$ *for all* j*, or*
   (b) *the constant* true *or* V_i*, if* $\mathsf{V}_i : o$*, or*
   (c) *the constant* true *or* $(\mathsf{V}_i\ \mathsf{B}_{11} \cdots \mathsf{B}_{1r}) \wedge \cdots \wedge (\mathsf{V}_i\ \mathsf{B}_{m1} \cdots \mathsf{B}_{mr})$*, where* $m > 0$*, if* $type(\mathsf{V}_i) = \rho'_1 \to \cdots \to \rho'_r \to o$ *and for all* k, l*,* B_{kl} *is a basic expression with* $type(\mathsf{B}_{kl}) = \rho'_l$ *and for all* j*,* $\mathsf{V}_j \notin fv(\mathsf{B}_{kl})$*.*

The B_i *and* B_{kl} *will be called the* basic subexpressions *of* B*.*

The set of *basic templates of* $\mathcal{H}$ is the subset of the set of basic expressions of $\mathcal{H}$ defined as follows: The propositional constants false and true are basic templates; every nonempty finite union of basic expressions in which all the basic subexpressions involved are *distinct* free variables, is a basic template.

A *substitution* θ is a finite set of the form $\{\mathsf{V}_1/\mathsf{E}_1, \ldots, \mathsf{V}_n/\mathsf{E}_n\}$, where the V_i's are different argument variables of $\mathcal{H}$ and each E_i is a positive expression of $\mathcal{H}$ having the same type as V_i. We write $dom(\theta) = \{\mathsf{V}_1, \ldots, \mathsf{V}_n\}$ and $range(\theta) = \{\mathsf{E}_1, \ldots, \mathsf{E}_n\}$. A substitution is called *basic* if all E_i are basic expressions. A substitution is called *zero-order*, if $type(\mathsf{V}_i) = \iota$, for all $i \in \{1, \ldots, n\}$ (notice that every zero-order substitution is also basic). The substitution corresponding to the empty set will be called the *identity substitution* and will be denoted by ϵ. The notions of *unifier* and *most general unifier* is defined in a standard way for the zero-order substitutions. The composition of two substitutions and the application of a substitution to a positive expressions is defined as usual.

Definition 5. *Let* P *be a program and let* $\mathsf{G} = \leftarrow \mathsf{A}$ *and* $\mathsf{G}' = \leftarrow \mathsf{A}'$ *be goal clauses. We say that* A' *is derived in one step from* A *using* θ *(or equivalently that* G' *is derived in one step from* G *using* θ*), and we denote this fact by* $\mathsf{A} \xrightarrow{\theta} \mathsf{A}'$ *(respectively,* $\mathsf{G} \xrightarrow{\theta} \mathsf{G}'$*), if one of the following conditions applies:*

1. $\mathsf{p}\ \mathsf{E}_1 \cdots \mathsf{E}_n \xrightarrow{\epsilon} \mathsf{E}\ \mathsf{E}_1 \cdots \mathsf{E}_n$*, where* $\mathsf{p} \leftarrow_\pi \mathsf{E}$ *is a rule in* P*.*
2. $\mathsf{Q}\ \mathsf{E}_1 \cdots \mathsf{E}_n \xrightarrow{\theta} (\mathsf{Q}\ \mathsf{E}_1 \cdots \mathsf{E}_n)\theta$*, where* $\theta = \{\mathsf{Q}/\mathsf{B}_t\}$ *and* B_t *a basic template.*
3. $(\lambda \mathsf{V}.\mathsf{E})\ \mathsf{E}_1 \cdots \mathsf{E}_n \xrightarrow{\epsilon} (\mathsf{E}\{\mathsf{V}/\mathsf{E}_1\})\ \mathsf{E}_2 \cdots \mathsf{E}_n$*.*
4. $(\mathsf{E}' \bigvee_\pi \mathsf{E}'')\ \mathsf{E}_1 \cdots \mathsf{E}_n \xrightarrow{\epsilon} \mathsf{E}'\ \mathsf{E}_1 \cdots \mathsf{E}_n$*.*
5. $(\mathsf{E}' \bigvee_\pi \mathsf{E}'')\ \mathsf{E}_1 \cdots \mathsf{E}_n \xrightarrow{\epsilon} \mathsf{E}''\ \mathsf{E}_1 \cdots \mathsf{E}_n$*.*
6. $(\mathsf{E}' \bigwedge_\pi \mathsf{E}'')\ \mathsf{E}_1 \cdots \mathsf{E}_n \xrightarrow{\epsilon} (\mathsf{E}'\ \mathsf{E}_1 \cdots \mathsf{E}_n) \wedge (\mathsf{E}''\ \mathsf{E}_1 \cdots \mathsf{E}_n)$*, where* $\pi \neq o$*.*

7. $(\mathsf{E}_1 \wedge \mathsf{E}_2) \xrightarrow{\theta} (\mathsf{E}_1' \wedge (\mathsf{E}_2\theta))$, *if* $\mathsf{E}_1 \xrightarrow{\theta} \mathsf{E}_1'$.
8. $(\mathsf{E}_1 \wedge \mathsf{E}_2) \xrightarrow{\theta} ((\mathsf{E}_1\theta) \wedge \mathsf{E}_2')$, *if* $\mathsf{E}_2 \xrightarrow{\theta} \mathsf{E}_2'$.
9. $(\mathsf{true} \wedge \mathsf{E}) \xrightarrow{\epsilon} \mathsf{E}$
10. $(\mathsf{E} \wedge \mathsf{true}) \xrightarrow{\epsilon} \mathsf{E}$
11. $(\mathsf{E}_1 \approx \mathsf{E}_2) \xrightarrow{\theta} \mathsf{true}$, *where* θ *is an mgu of* E_1 *and* E_2.
12. $(\exists \mathsf{V}\, \mathsf{E}) \xrightarrow{\epsilon} \mathsf{E}$

Let P be a program and G be a goal. Assume that $\mathsf{P} \cup \{\mathsf{G}\}$ has a finite SLD-derivation $\mathsf{G}_0 = \mathsf{G}, \mathsf{G}_1, \ldots, \mathsf{G}_n$ with basic substitutions $\theta_1, \ldots, \theta_n$, such that $\mathsf{G}_n = \Box$. Then, we will say that $\mathsf{P} \cup \{\mathsf{G}\}$ has an *SLD-refutation of length n using basic substitution* $\theta = \theta_1 \cdots \theta_n$. A *computed answer* σ for $\mathsf{P} \cup \{\mathsf{G}\}$ is the basic substitution obtained by restricting θ to the free variables of G.

Example 5. Consider the program of Example 4. A successful SLD-refutation of the goal $\leftarrow$ (`closure Q a b`) is given here (where we have omitted certain simple steps involving lambda abstractions).

`closure Q a b`	$\theta_1 = \epsilon$
`(λR.λX.λY.(R X Y)) Q a b`	$\theta_2 = \epsilon$
`Q a b`	$\theta_3 = \{\mathtt{Q}/(\lambda\mathtt{X}.\lambda\mathtt{Y}.(\mathtt{X} \approx \mathtt{X}_0) \wedge (\mathtt{Y} \approx \mathtt{Y}_0))\}$
`(λX.λY.(X≈X0)∧(Y≈Y0)) a b`	$\theta_4 = \epsilon$
`(a≈X0)∧(b≈Y0)`	$\theta_5 = \{\mathtt{X}_0/\mathtt{a}\}$
`true∧(b≈Y0)`	$\theta_6 = \epsilon$
`(b≈Y0)`	$\theta_7 = \{\mathtt{Y}_0/\mathtt{b}\}$
`true`	

If we restrict the composition $\theta_1 \cdots \theta_7$ to the free variables of the goal, we get the computed answer $\sigma_1 = \{\mathtt{Q}/\lambda\mathtt{X}.\lambda\mathtt{Y}.(\mathtt{X} \approx \mathtt{a}) \wedge (\mathtt{Y} \approx \mathtt{b})\}$. Intuitively, σ_1 assigns to `Q` the relation $\{(\mathtt{a}, \mathtt{b})\}$. Notice that by substituting `Q` with different basic templates, one can get answers that are "similar" to the previous one, such as for example $\{(\mathtt{a}, \mathtt{b}), (\mathtt{Z1}, \mathtt{Z2})\}$, and so on. Answers of this type are in some sense "represented" by the answer $\{(\mathtt{a}, \mathtt{b})\}$.

We say that the basic substitution θ is *a correct answer* for $\mathsf{P} \cup \{\mathsf{G}\}$ if $\mathsf{A}\theta$ is a logical consequence of P.

Theorem 1 (Soundness). *Let* P *be a program and* $\mathsf{G} = \leftarrow \mathsf{A}$ *be a goal. Then, every computed answer for* $\mathsf{P} \cup \{\mathsf{G}\}$ *is a correct answer for* $\mathsf{P} \cup \{\mathsf{G}\}$.

Theorem 2 (Completeness). *Let* P *be a program and* $\mathsf{G} = \leftarrow \mathsf{A}$ *be a goal. For every correct answer* θ *for* $\mathsf{P} \cup \{\mathsf{G}\}$*, there exists an SLD-refutation for* $\mathsf{P} \cup \{\mathsf{G}\}$ *with computed answer* δ *and a substitution* γ *such that* $\mathsf{G}\theta = \mathsf{G}\delta\gamma$.

4 Normal Higher-Order Programs

The basic difference in the types of $\mathcal{H}_{cn}$ from those of $\mathcal{H}$ is the existence of a type μ, which restricts the set of predicate variables that can be existentially

quantified or appear free in goal clauses. The subtypes μ (existential type) and κ (set type) of ρ and π, respectively, are defined as follows.

$$\mu := \iota \mid \kappa$$
$$\kappa := \iota \to o \mid (\iota \to \kappa)$$

Definition 6. *The set of body expressions of the higher-order language $\mathcal{H}$ is recursively defined as follows.*

1. *Every predicate variable (respectively, predicate constant) of type π is a body expression of type π; every individual variable (respectively, individual constant) of type ι is a body expression of type ι; the propositional constants* false *and* true *are body expressions of type o.*
2. *If* f *is an n-ary function symbol and $\mathsf{E}_1, \ldots, \mathsf{E}_n$ are body expressions of type ι, then $(\mathsf{f}\ \mathsf{E}_1 \cdots \mathsf{E}_n)$ is a body expression of type ι.*
3. *If E_1 is a body expression of type $\rho \to \pi$ and E_2 is a body expression of type ρ, then $(\mathsf{E}_1\ \mathsf{E}_2)$ is a body expression of type π.*
4. *If V is an argument variable of type ρ and E is a body expression of type π, then $(\lambda \mathsf{V}.\mathsf{E})$ is a body expression of type $\rho \to \pi$.*
5. *If $\mathsf{E}_1, \mathsf{E}_2$ are body expressions of type π, then $(\mathsf{E}_1 \bigwedge_\pi \mathsf{E}_2)$ and $(\mathsf{E}_1 \bigvee_\pi \mathsf{E}_2)$ are body expressions of type π.*
6. *If $\mathsf{E}_1, \mathsf{E}_2$ are body expressions of type ι, then $(\mathsf{E}_1 \approx \mathsf{E}_2)$ is a body expression of type o.*
7. *If E is a body expression of type o and V is an existential variable of type μ, then $(\exists_\mu \mathsf{V}\, \mathsf{E})$ is a body expression of type o.*
8. *If E is a body expression of type o, then $(\sim\mathsf{E})$ is a body expression of type o.*

We will often write $\hat{\mathsf{A}}$ to denote a (possibly empty) sequence $\langle \mathsf{A}_1, \ldots, \mathsf{A}_n \rangle$ of expressions. For example we will write $(\mathsf{E}\ \hat{\mathsf{A}})$ to denote an application $(\mathsf{E}\ \mathsf{A}_1 \cdots \mathsf{A}_n)$; $(\lambda \hat{\mathsf{X}}.\mathsf{E})$ to denote $(\lambda \mathsf{X}_1. \cdots \lambda \mathsf{X}_n.\mathsf{E})$; $(\exists \hat{\mathsf{V}}\ \mathsf{E})$ to denote $(\exists \mathsf{V}_1 \cdots \exists \mathsf{V}_n\ \mathsf{E})$.

Definition 7. *The set of* clausal expressions *of $\mathcal{H}_{cn}$ is defined as follows:*

1. *If p is a predicate constant of type π and E is a closed body expression of type π, then $\mathsf{p} \leftarrow_\pi \mathsf{E}$ is a clausal expression of $\mathcal{H}_{cn}$, also called a* program clause.
2. *If E is a body expression of type o and each free variable in E is of type μ, then $\mathsf{false} \leftarrow_o \mathsf{E}$ (usually denoted by $\leftarrow_o \mathsf{E}$ or just $\leftarrow \mathsf{E}$) is a clausal expression of $\mathcal{H}_{cn}$, also called a* goal clause.
3. *If p is a predicate constant of type π and E is a closed body expression of type π, then $\mathsf{p} \leftrightarrow_\pi \mathsf{E}$ is a clausal expression of $\mathcal{H}_{cn}$, also called a* completion expression.

All clausal expressions of $\mathcal{H}_{cn}$ have type o. A program of $\mathcal{H}_{cn}$ is a finite set of program clauses of $\mathcal{H}_{cn}$.

Consider a graph represented by a binary relation over the set of its vertices. Then, given a graph of type $\pi = (\iota \to \iota \to o)$, we can express the set of its vertices as a predicate of type $\iota \to o$ of the graph and a path of the graph as a predicate of type $\iota \to \iota \to o$.

$\mathtt{v} \leftarrow (\lambda \mathtt{G}.\lambda \mathtt{X}.\exists \mathtt{Y}(\mathtt{G\ X\ Y})) \bigvee (\lambda \mathtt{G}.\lambda \mathtt{X}.\exists \mathtt{Y}(\mathtt{G\ Y\ X}))$
$\mathtt{p} \leftarrow \lambda \mathtt{G}.\lambda \mathtt{X}.\lambda \mathtt{Y}.(\mathtt{closure\ G\ X\ Y})$

The predicate v succeeds if its second argument is a vertex of its first argument. The predicate p succeeds if there is a path between two vertices of the graph. Note that the path is actually a transitive closure of the graph. We can easily express basic connectivity properties of a graph.

$\mathtt{disconnected} \leftarrow \lambda \mathtt{G}.\exists \mathtt{X} \exists \mathtt{Y}((\mathtt{v\ G\ X}) \wedge (\mathtt{v\ G\ Y}) \wedge \sim(\mathtt{X} \approx \mathtt{Y}) \wedge \sim(\mathtt{p\ G\ X\ Y}))$
$\mathtt{nonclique} \leftarrow \lambda \mathtt{G}.\lambda \mathtt{S}.\exists \mathtt{X} \exists \mathtt{Y}((\mathtt{S\ X}) \wedge (\mathtt{S\ Y}) \wedge \sim(\mathtt{X} \approx \mathtt{Y}) \wedge \sim(\mathtt{G\ X\ Y}))$
$\mathtt{connected} \leftarrow \lambda \mathtt{G}.\sim(\mathtt{disconnected\ G})$
$\mathtt{clique} \leftarrow \lambda \mathtt{G}.\lambda \mathtt{S}.(\mathtt{subset\ S\ (v\ G)}) \wedge \sim(\mathtt{nonclique\ G\ S})$

Let P be a program and let p be a predicate constant of type π. Then, the *completed definition* for p with respect to P is obtained as follows:

- if there exist exactly $k > 0$ program clauses of the form $\mathsf{p} \leftarrow_\pi \mathsf{E}_i$, where $i \in \{1, \ldots, k\}$ for p in P, then the completed definition for p is the expression $\mathsf{p} \leftrightarrow_\pi \mathsf{E}$, where $\mathsf{E} = \mathsf{E}_1 \bigvee_\pi \cdots \bigvee_\pi \mathsf{E}_k$;
- if there are no program clauses for p in P, then the completed definition for p is the expression $\mathsf{p} \leftrightarrow_\pi \mathsf{E}$, where E is of type π and $\mathsf{E} = \lambda \hat{\mathsf{X}}.\mathsf{false}$.

The *program completion* $comp(\mathsf{P})$ of P is the set consisting of all the completed definitions for all predicate constants that appear in P.

4.1 Proof Procedure

An inequality $\sim\exists \hat{\mathsf{V}}(\mathsf{E}_1 \approx \mathsf{E}_2)$ is considered

- *valid* if E_1 and E_2 cannot be unified;
- *unsatisfiable* if there is a substitution θ that unifies E_1 and E_2 and contains only bindings of variables in $\hat{\mathsf{V}}$;
- *satisfiable* if it is not unsatisfiable.

An inequality will be called *primitive* if it is satisfiable, non valid and either E_1 or E_2 is a variable.

Definition 8. *The set of* normal basic expressions *of $\mathcal{H}_{cn}$ of type μ is defined recursively as follows.*

1. *Every expression of $\mathcal{H}_{cn}$ of type ι is a normal basic expression of type ι.*
2. *Every predicate variable of type κ is a normal basic expression of type κ.*
3. *If $\mathsf{E}_1, \mathsf{E}_2$ are normal basic expressions of type κ, then $\mathsf{E}_1 \bigvee_\kappa \mathsf{E}_2$ and $\mathsf{E}_1 \bigwedge_\kappa \mathsf{E}_2$ are normal basic expressions of type κ.*
4. *The expressions of the following form are normal basic expressions of type $\iota^n \rightarrow o$:*
   - $\lambda \hat{\mathsf{X}}.\exists \hat{\mathsf{V}}(\hat{\mathsf{X}} \approx \hat{\mathsf{A}})$
   - $\lambda \hat{\mathsf{X}}.\ \sim\exists \hat{\mathsf{V}}(\hat{\mathsf{X}} \approx \hat{\mathsf{A}})$

where $\hat{\mathsf{X}} = \langle \mathsf{X}_1, \ldots, \mathsf{X}_n \rangle$, $\hat{\mathsf{A}} = \langle \mathsf{A}_1, \ldots, \mathsf{A}_n \rangle$, *each* X_i *is a variable of type* ι, *each* $\hat{\mathsf{A}}_i$ *is a normal basic expressions of type* ι *and* $\hat{\mathsf{V}}$ *is a possibly empty subset of* $fv(\hat{\mathsf{A}})$.

Definition 9. *Let* P *be a program and* E, E' *be body expressions of type* o. *We say that* E *is reduced (wrt. to* P*) to* E' *(denoted as* $\mathsf{E} \rightsquigarrow \mathsf{E}'$*) if one of the following conditions applies:*

1. $\mathsf{p}\ \hat{\mathsf{A}} \rightsquigarrow \mathsf{E}\ \hat{\mathsf{A}}$, *where* E *is the completed expression for* p *with respect to* P
2. $(\lambda \mathsf{X}.\mathsf{E})\ \mathsf{B}\ \hat{\mathsf{A}} \rightsquigarrow \mathsf{E}\{\mathsf{X}/\mathsf{B}\}\ \hat{\mathsf{A}}$
3. $(\mathsf{E}_1 \bigvee_\pi \mathsf{E}_2)\ \hat{\mathsf{A}} \rightsquigarrow (\mathsf{E}_1\ \hat{\mathsf{A}}) \vee (\mathsf{E}_2\ \hat{\mathsf{A}})$
4. $(\mathsf{E}_1 \bigwedge_\pi \mathsf{E}_2)\ \hat{\mathsf{A}} \rightsquigarrow (\mathsf{E}_1\ \hat{\mathsf{A}}) \wedge (\mathsf{E}_2\ \hat{\mathsf{A}})$

Definition 10. *Let* P *a program and let* G_k *and* G_{k+1} *be goal clauses and let* G_k *be a conjunction* $\leftarrow \mathsf{A}_1 \wedge \cdots \wedge \mathsf{A}_n$, *where each* A_i *is a body expression of type* o. *Moreover, let* A_i *one of the* $\mathsf{A}_1, \ldots, \mathsf{A}_n$ *(called* selected expression*) and* $\mathsf{A}' = \mathsf{A}_1 \wedge \cdots \wedge \mathsf{A}_{i-1} \wedge \mathsf{A}_{i+1} \wedge \cdots \wedge \mathsf{A}_n$. *We say that* G_{k+1} *is* derived in one step *from* G_k *using* θ *(denoted as* $\mathsf{G}_k \overset{\theta}{\rightarrow} \mathsf{G}_{k+1}$*) if one of the following conditions applies:*

1. *if* A_i *is* true *and* $n > 1$, *then* $\mathsf{G}_{k+1} = \leftarrow \mathsf{A}'$ *is derived from* G_k *using* $\theta = \epsilon$;
2. *if* A_i *is* $(\mathsf{E}_1 \vee \mathsf{E}_2)$, *then* $\mathsf{G}_{k+1} = \leftarrow \mathsf{A}_1 \wedge \cdots \wedge \mathsf{E}_j \wedge \cdots \wedge \mathsf{A}_n$ *is derived from* G_k *using* $\theta = \epsilon$ *where* $j \in \{1, 2\}$;
3. *if* A_i *is* $(\exists \mathsf{V}\ \mathsf{E})$, *then* $\mathsf{G}_{k+1} = \leftarrow \mathsf{A}_1 \wedge \cdots \wedge \mathsf{E} \wedge \cdots \wedge \mathsf{A}_n$ *is derived from* G_k *using* $\theta = \epsilon$;
4. *if* $\mathsf{A}_i \rightsquigarrow \mathsf{A}'_i$, *then* $\mathsf{G}_{k+1} = \leftarrow \mathsf{A}_1 \wedge \cdots \wedge \mathsf{A}'_i \wedge \cdots \wedge \mathsf{A}_n$ *is derived from* G_k *using* $\theta = \epsilon$;
5. *if* A_i *is* $(\mathsf{E}_1 \approx \mathsf{E}_2)$, *then* $\mathsf{G}_{k+1} = \leftarrow \mathsf{A}'\theta$ *is derived from* G_k *using* $\theta = mgu(\mathsf{E}_1, \mathsf{E}_2)$;
6. *if* A_i *is* $(\mathsf{R}\ \hat{\mathsf{E}})$ *and* $\mathsf{R} : \kappa$ *be a variable, then* $\mathsf{G}_{k+1} = \leftarrow \mathsf{A}'\theta$ *is derived from* G_k *using* $\theta = \{\mathsf{R}/(\lambda \hat{\mathsf{X}}.(\hat{\mathsf{X}} \approx \hat{\mathsf{E}}) \bigvee_\kappa \mathsf{R}')\}$ *where* $\mathsf{R}' : \kappa$ *is a fresh variable;*
7. *if* A_i *is* $\sim\exists \hat{\mathsf{V}}\ \mathsf{E}$ *and* A_i *is negatively reduced to* A'_i, *then* $\mathsf{G}_{k+1} = \leftarrow \mathsf{A}_1 \wedge \cdots \wedge \mathsf{A}'_i \wedge \cdots \wedge \mathsf{A}_n$ *is derived from* G_k *using* $\theta = \epsilon$;
8. *if* A_i *is* $\sim\exists \hat{\mathsf{V}}(\mathsf{R}\ \hat{\mathsf{E}})$, *variable* $\mathsf{R} : \kappa$ *and* $\mathsf{R} \notin \hat{\mathsf{V}}$, *then* $\mathsf{G}_{k+1} = \leftarrow \mathsf{A}'\theta$ *is derived from* G_k *using* $\theta = \{\mathsf{R}/(\lambda \hat{\mathsf{X}}.\ \sim\exists \hat{\mathsf{V}}(\hat{\mathsf{X}} \approx \hat{\mathsf{E}}) \bigwedge_\kappa \mathsf{R}')\}$, *where* $\mathsf{R}' : \kappa$ *is a fresh variable;*
9. *if* A_i *is* $\sim\exists \hat{\mathsf{V}}\ \sim(\mathsf{R}\ \hat{\mathsf{E}})$, *variable* $\mathsf{R} : \kappa$ *and* $\mathsf{R} \notin \hat{\mathsf{V}}$, *then* $\mathsf{G}_{k+1} = \leftarrow \mathsf{A}'\theta$ *is derived from* G_k *using* $\theta = \{\mathsf{R}/(\lambda \hat{\mathsf{X}}.\exists \hat{\mathsf{V}}(\hat{\mathsf{X}} \approx \hat{\mathsf{E}}) \bigvee_\kappa \mathsf{R}')\}$, *where* $\mathsf{R}' : \kappa$ *is a fresh variable.*

Note that the single step derivation will essentially behave similarly with the Definition 5 for definite higher-order programs. The last three cases of the single-step derivation handle the cases of a negative expression.

Definition 11. *Let* P *be a program and let* $\mathsf{B} = \sim\exists \hat{\mathsf{U}}(\mathsf{A}_1 \wedge \cdots \wedge \mathsf{A}_n)$ *be a body expression where* A_i *is a body expression except from conjunction. Let* A_i *be one of* $\mathsf{A}_1, \ldots, \mathsf{A}_n$ *and* $\mathsf{A}' = \mathsf{A}_1 \wedge \cdots \wedge \mathsf{A}_{i-1} \wedge \mathsf{A}_{i+1} \wedge \cdots \wedge \mathsf{A}_n$. *Moreover, let* B' *be a body expression. We say that* B *is* negatively reduced *to* B' *if one of the following conditions applies:*

1. *if* A_i *is* false, *then* $\mathsf{B}' = \mathsf{true}$;
2. *if* A_i *is* true *and* $n = 1$, *then* $\mathsf{B}' = \mathsf{false}$; *otherwise* $\mathsf{B}' = \sim\exists\hat{\mathsf{U}}\ \mathsf{A}'$;
3. *if* A_i *is* $(\mathsf{E}_1 \vee \mathsf{E}_2)$, *then* $\mathsf{B}' = \mathsf{B}'_1 \wedge \mathsf{B}'_2$ *where* $\mathsf{B}'_j = \sim\exists\hat{\mathsf{U}}(\mathsf{A}_1 \wedge \cdots \wedge \mathsf{E}_j \wedge \cdots \wedge \mathsf{A}_n)$ *with* $j \in \{1, 2\}$;
4. *if* A_i *is* $(\exists\mathsf{V}\ \mathsf{E})$, *then* $\mathsf{B}' = \sim\exists\hat{\mathsf{U}}\mathsf{V}(\mathsf{A}_1 \wedge \cdots \wedge \mathsf{E} \wedge \cdots \wedge \mathsf{A}_n)$;
5. *if* $\mathsf{A}_i \rightsquigarrow \mathsf{A}'_i$, *then* $\mathsf{B}' = \sim\exists\hat{\mathsf{U}}(\mathsf{A}_1 \wedge \cdots \wedge \mathsf{A}'_i \wedge \cdots \wedge \mathsf{A}_n)$;
6. *if* A_i *is* $(\mathsf{E}_1 \approx \mathsf{E}_2)$, *then*
   (a) *if* $\sim\exists\hat{\mathsf{U}}(\mathsf{E}_1 \approx \mathsf{E}_2)$ *is valid, then* $\mathsf{B}' = \mathsf{true}$;
   (b) *if* $\sim\exists\hat{\mathsf{U}}(\mathsf{E}_1 \approx \mathsf{E}_2)$ *is not valid and if neither* E_1 *nor* E_2 *is a variable, then* $\hat{\mathsf{X}}$ *is* $dom(\theta)$, $\theta = unify(\mathsf{E}_1, \mathsf{E}_2)$ *and* $\mathsf{B}' = \sim\exists\hat{\mathsf{U}}(\mathsf{A}_1\theta \wedge \cdots \wedge (\hat{\mathsf{X}} \approx \hat{\mathsf{X}}\theta) \wedge \cdots \wedge \mathsf{A}_n\theta)$.
   (c) *if* $\sim\exists\hat{\mathsf{U}}(\mathsf{E}_1 \approx \mathsf{E}_2)$ *is unsatisfiable and either* E_1 *or* E_2 *is a variable in* $\hat{\mathsf{U}}$, *then* $\mathsf{B}' = \sim\exists\hat{\mathsf{U}}(\mathsf{A}'\theta)$, *where* $\theta = \{\mathsf{X}/\mathsf{E}\}$ *and* X *is the one expression that is a variable in* $\hat{\mathsf{U}}$ *and* E *is the other;*
   (d) *if* $\sim\exists\hat{\mathsf{U}}(\mathsf{E}_1 \approx \mathsf{E}_2)$ *is primitive and* $n > 1$, *then* $\mathsf{B}' = \sim\exists\hat{\mathsf{U}}_1\ \mathsf{A}_i \vee \exists\hat{\mathsf{U}}_1(\mathsf{A}_i \wedge \sim \exists\hat{\mathsf{U}}_2\ \mathsf{A}')$, *where* $\hat{\mathsf{U}}_1$ *are in* $\hat{\mathsf{U}}$ *that are free in* A_i *and* $\hat{\mathsf{U}}_2$ *in* $\hat{\mathsf{U}}$ *not in* $\hat{\mathsf{U}}_1$;
7. *if* A_i *is* $(\mathsf{R}\ \hat{\mathsf{E}})$ *and variable* $\mathsf{R} : \kappa$, *then*
   (a) *if* $\mathsf{R} \in \hat{\mathsf{U}}$, *then* $\mathsf{B}' = \sim\exists\hat{\mathsf{U}}'(\mathsf{A}'\theta)$, *where* $\theta = \{\mathsf{R}/(\lambda\mathsf{X}.(\mathsf{X} \approx \mathsf{E}) \bigvee_\kappa \mathsf{R}')\}$, $\mathsf{R}' : \kappa$ *is a fresh variable and* $\hat{\mathsf{U}}' = \hat{\mathsf{U}}\{\mathsf{R}/\mathsf{R}'\}$;
   (b) *if* $\mathsf{R} \notin \hat{\mathsf{U}}$ *and* $n > 1$, *then* $\mathsf{B}' = \sim\exists\hat{\mathsf{U}}_1\ \mathsf{A}_i \vee \exists\hat{\mathsf{U}}_1(\mathsf{A}_i \wedge \sim\exists\hat{\mathsf{U}}_2\ \mathsf{A}') \wedge \mathsf{B}$, *where* $\hat{\mathsf{U}}_1$ *are the variables in* $\hat{\mathsf{U}}$ *that are free in* A_i *and* $\hat{\mathsf{U}}_2$ *in* $\hat{\mathsf{U}}$ *not in* $\hat{\mathsf{U}}_1$;
8. *if* A_i *is* $\sim\exists\hat{\mathsf{V}}\ \mathsf{E}$ *and* A_i *is negatively reduced to* A'_i, *then* $\mathsf{B}' = \sim\exists\hat{\mathsf{U}}(\mathsf{A}_1 \wedge \cdots \wedge \mathsf{A}'_i \wedge \cdots \wedge \mathsf{A}_n)$;
9. *if* A_i *is a primitive inequality* $\sim\exists\hat{\mathsf{V}}(\mathsf{E}_1 \approx \mathsf{E}_2)$, *then*
   (a) *if* $fv(\mathsf{A}_i) \cap \hat{\mathsf{U}} \neq \emptyset$ *and* A' *is conjunction of primitive inequalities, then* $\mathsf{B}' = \sim\exists\hat{\mathsf{U}}\,\mathsf{A}'$;
   (b) *if* $fv(\mathsf{A}_i) \cap \hat{\mathsf{U}}$ *is empty, then* $\mathsf{B}' = \exists\hat{\mathsf{V}}(\mathsf{E}_1 \approx \mathsf{E}_2) \vee \sim\exists\hat{\mathsf{U}}\ \mathsf{A}'$;
10. *if* A_i *is* $\sim\exists\hat{\mathsf{V}}(\mathsf{R}\ \hat{\mathsf{E}})$, *variable* $\mathsf{R} : \kappa$ *and* $\mathsf{R} \notin \hat{\mathsf{V}}$, *then*
   (a) *if* $\mathsf{R} \in \hat{\mathsf{U}}$, *then* $\mathsf{B}' = \sim\exists\hat{\mathsf{U}}'(\mathsf{A}'\theta)$ *where* $\theta = \{\mathsf{R}/(\lambda\mathsf{X}.\ \sim\exists\hat{\mathsf{V}}(\mathsf{X} \approx \mathsf{E}) \bigwedge_\kappa \mathsf{R}')\}$, $\mathsf{R}' : \kappa$ *is a fresh variable and* $\hat{\mathsf{U}}' = \hat{\mathsf{U}}\{\mathsf{R}/\mathsf{R}'\}$;
   (b) *if* $\mathsf{R} \notin \hat{\mathsf{U}}$ *and* $n > 1$, *then* $\mathsf{B}' = \sim\exists\hat{\mathsf{U}}_1\ \mathsf{A}_i \vee \exists\hat{\mathsf{U}}_1(\mathsf{A}_i \wedge \sim\exists\hat{\mathsf{U}}_2\ \mathsf{A}') \wedge \mathsf{B}$, *where* $\hat{\mathsf{U}}_1$ *are the variables in* $\hat{\mathsf{U}}$ *that are free in* A_i *and* $\hat{\mathsf{U}}_2$ *in* $\hat{\mathsf{U}}$ *not in* $\hat{\mathsf{U}}_1$;
   (c) *if* $\mathsf{R} \notin \hat{\mathsf{U}}$, $n = 1$ *and* $\hat{\mathsf{V}} \neq \emptyset$, *then* $\mathsf{B}' = \exists\hat{\mathsf{V}}\ \sim\exists\hat{\mathsf{U}}(\sim(\mathsf{R}\ \hat{\mathsf{E}}) \wedge \sim\exists\hat{\mathsf{V}}'(\mathsf{R}\ \mathsf{E}'))$;
11. *if* A_i *is* $\sim\exists\hat{\mathsf{V}}\ \sim(\mathsf{R}\ \hat{\mathsf{E}})$ *and variable* $\mathsf{R} : \kappa$ *and* $\mathsf{R} \notin \hat{\mathsf{V}}$, *then*
   (a) *if* $\mathsf{R} \in \hat{\mathsf{U}}$, *then* $\mathsf{B}' = \sim\exists\hat{\mathsf{U}}'(\mathsf{A}'\theta)$, *where* $\theta = \{\mathsf{R}/(\lambda\mathsf{X}.\exists\hat{\mathsf{V}}(\mathsf{X} \approx \mathsf{E}) \bigvee_\kappa \mathsf{R}')\}$, $\mathsf{R}' : \kappa$ *is a fresh variable and* $\hat{\mathsf{U}}' = \hat{\mathsf{U}}\{\mathsf{R}/\mathsf{R}'\}$;
   (b) *if* $\mathsf{R} \notin \hat{\mathsf{U}}$ *and* $n > 1$, *then* $\mathsf{B}' = \sim\exists\hat{\mathsf{U}}_1\ \mathsf{A}_i \vee \exists\hat{\mathsf{U}}_1(\mathsf{A}_i \wedge \sim\exists\hat{\mathsf{U}}_2\ \mathsf{A}') \wedge \mathsf{B}$;
   (c) *if* $\mathsf{R} \notin \hat{\mathsf{U}}$, $n = 1$ *and* $\hat{\mathsf{V}} \neq \emptyset$, *then* $\mathsf{B}' = \exists\hat{\mathsf{V}}\ \sim\exists\hat{\mathsf{U}}((\mathsf{R}\ \hat{\mathsf{E}}) \wedge \sim\exists\hat{\mathsf{V}}'\ \sim(\mathsf{R}\ \mathsf{E}'))$.

The computed answer of a successful derivation with primitive goal G' and basic substitution θ is extended as follows: The tuple (σ, G'') is a computed answer for P where σ is the basic substitution obtained by restricting θ to the free variables of G and G'' is the primitive goal G' restricted to the free variables of G and the variables in $fv(range(\sigma))$.

Example 6. Consider the following simple definition for the predicate $\mathtt{q}$.

$$\mathtt{q} \leftarrow \lambda\mathtt{Z}_1.\lambda\mathtt{Z}_2.(\mathtt{Z}_1 \approx \mathtt{a}) \wedge (\mathtt{Z}_2 \approx \mathtt{b})$$

that holds only for the tuple $(\mathtt{a}, \mathtt{b})$. Then, consider the goal: $\leftarrow (\mathtt{R}\ \mathtt{X}) \wedge \sim(\mathtt{q}\ \mathtt{X}\ \mathtt{Y})$ that requests bindings for the variables $\mathtt{R}, \mathtt{X}$ and $\mathtt{Y}$.

0. $\leftarrow \underline{(\mathtt{R}\ \mathtt{X})} \wedge \sim(\mathtt{q}\ \mathtt{X}\ \mathtt{Y})$
1. $\leftarrow \underline{\sim(\mathtt{q}\ \mathtt{X}\ \mathtt{Y})}$ using $\theta_1 = \{\mathtt{R}/\lambda\mathtt{Z}.(\mathtt{Z} \approx \mathtt{X}) \bigvee \mathtt{R}'\}$
2. $\leftarrow \underline{\sim((\lambda\mathtt{Z}_1.\lambda\mathtt{Z}_2.(\mathtt{Z}_1 \approx \mathtt{a}) \wedge (\mathtt{Z}_2 \approx \mathtt{b}))\ \mathtt{X}\ \mathtt{Y})}$
3. $\leftarrow \underline{\sim((\mathtt{X} \approx \mathtt{a}) \wedge (\mathtt{Y} \approx \mathtt{b}))}$
4. $\leftarrow \sim(\mathtt{X} \approx \mathtt{a}) \vee (\mathtt{X} \approx \mathtt{a}) \wedge \sim(\mathtt{Y} \approx \mathtt{a})$
5. $\leftarrow \sim(\mathtt{X} \approx \mathtt{a})$

In step 4 the procedure generates two branches. The first one terminates immediately with a primitive inequality. The computed answer is $\sim(\mathtt{X} \approx \mathtt{a})$ and $\{\mathtt{R}/\lambda\mathtt{Z}.(\mathtt{Z} \approx \mathtt{X}) \bigvee \mathtt{R}'\}$.

Theorem 3 (Soundness). *Let* P *be a program and* G *be a goal. Then, every computed answer for* $\mathsf{P} \cup \{\mathsf{G}\}$ *is a correct answer for* $comp(\mathsf{P}) \cup \{\mathsf{G}\}$.

5 Conclusions

In this dissertation we have extended the study initiated in [7] and derived a complete framework for extensional higher-order logic programming. We have introduced the higher-order language $\mathcal{H}$ that extends the language in [7] and allows uninstantiated predicate variables. We have proposed a sound and complete SLD-resolution with respect to the minimum Herbrand model semantics.

We have also introduced an extension of $\mathcal{H}$ that supports the operator of negation-as-failure. We have proposed a proof procedure that extends the higher-order SLD-resolution in order to handle constructive negation. As a result, the proposed proof procedure, avoids the floundering problem of negation-as-failure. We have also established the soundness of the proof procedure with respect to the completion semantics.

References

1. M. Bezem. An improved extensionality criterion for higher-order logic programs. In *Proceedings of the 15th International Workshop on Computer Science Logic (CSL)*, pages 203–216, London, UK, 2001. Springer-Verlag.
2. D. Chan. Constructive negation based on the completed database. In *ICLP/SLP*, pages 111–125, 1988.
3. A. Charalambidis, K. Handjopoulos, P. Rondogiannis, and W. W. Wadge. Extensional higher-order logic programming. *ACM Trans. Comput. Log.*, 14(3), 2013.
4. W. Chen, M. Kifer, and D. S. Warren. Hilog: A foundation for higher-order logic programming. *Journal of Logic Programming*, 15(3):187–230, 1993.
5. J. W. Lloyd. *Foundations of Logic Programming, 2nd Edition*. Springer, 1987.
6. G. Nadathur. *A Higher-Order Logic as the Basis for Logic Programming*. PhD thesis, University of Pennsylvania, 1987.
7. W. W. Wadge. Higher-order horn logic programming. In *Proceedings of the International Symposium on Logic Programming (ISLP)*, pages 289–303, 1991.
8. D. H. Warren. Higher-order extensions to prolog: are they needed? *Machine Intelligence*, 10:441–454, 1982.

Development of a Conceptual and Methodological Framework for the Identification and Management of Barriers and Opportunities in the Adoption of e-Government services

Haroula N. Delopoulos

Department of Informatics and Telecommunications, University of Athens, Panepistimioupolis Ilissia, Athens, Greece, GR-15784[1]

hadelop@econ.uoa.gr

Abstract. The aim of this thesis is twofold, a conceptual and methodological framework for the management and identification of obstacles and opportunities for the adoption of electronic Government (e-Gov) services was developed. Predicting models of identifying barriers and opportunities of the adoption/use of eGov services were used. A new composition methodology of web usability evaluation of e-Gov services was developed. This method was applied on e-deliberation service of Greece. The synthesis of methodologies consists of the following assessment methods: the Nielsen's Heuristics, the Cognitive walkthrough method, the Inspection method, Expert testing, Policy analysis method, Questionnaire and Scenario as data collective methods. The questionnaire constructed following the HHS web usability guidelines as well as web usability standard ISO9241-151. The synthesis of methodologies can be applied to any e-Gov service. 125 usability points were tested and discovered that 23% of them had major usability problems 14% had small usability problems and 63% had no usability problems. The identified usability problems were categorized according to Nielsen's Heuristics, and specific usability problems were trucked down. In parallel, the adoption model of UTAUT-PBO (Predominant barriers and opportunities) following the Unified Theory of Acceptance and Use of Technology (UTAUT) was proposed. Four models derived (Total e-Gov use, e-Gov-Obtain Info, e-Gov-Download Info, e-Gov-Filling Forms) identifying barriers or opportunities of use of e-Gov services in the EU (European Union). The data are drawn from "Eurostat" and "United Nations" statistical data bases, EU time series indicators for the years 2001-2011. Methodology of linear regression models with step was used. In the independent variables of the models potential obstacles or opportunities were evaluated. The models were evaluated per country and per year and a total of 164 models were estimated. Significant barriers and opportunities in the adoption/use of services trucked down in geographical areas of the EU. The variable "never use internet" which is negatively associated with the dependent variables of the 4 models, was identified as the predominant obstacle to the use/adoption during the period 2000 to 2011 in EU. A methodological gap in the evaluation of e-Gov in the EU exists. Europe focuses on the supply side and after the e-Gov services have been produced. It is suggested EU to follow a framework for adoption of e-Gov services in the design phase of services that takes into account the barriers and opportunities of adoption of e-Gov services.

Subject area: electronic Government or electronic Governance, information systems

Keywords: electronic Government or electronic Governance, information systems of e-Government services, barriers and opportunities of e-Government services, use/adoption of e-Government services, usability evaluation of e-Government services, UTAUT model

1. Introduction

The use of Information Communication Technologies (ICT) in the public administration and services is specified *as Electronic Governance* (e-Gov), which contains organisational changes and new skills for the improvement of public services and democratic processes [1]. The potential of e-Gov exceeds by far the initial achievements of electronic public services. In European Union and in many countries of the planet we notice an intense activation of both leaders and researchers on issues as broadband technology, interoperability, interactive e-Gov services accessible by all, public conventions (electronic supplies, with use of the Internet, public points of access to the internet, e-learning programs, creation of health e-card, web health services, dynamic environment for electronic entrepreneurship, safe infrastructure of information, safe communications between public services, lifting of legislative obstacles [2].

Approximately a decade earlier, the European Union (EU) was discussing European Governance and now the focus is on Internet Governance. In the year 2000, European Union's Lisbon Strategy set out the goal for the EU to become (in 2010) the most competitive knowledge based economy, enjoying full employment. The EU was determined to launch, in early 2000 an initiative to amend European Governance as a strategic objective well in advance of the Nice European Council. Decided to reform governance and how the EU uses the powers given by its citizens. One of the main aims was to open up policy-making and make governance more inclusive and

[1] Dissertation Advisor, Panagiotis Georgiadis, Emeritus Professor

accountable. The promote of new forms of European Governance was a major scope of the European Union at 2000 [3], [4].

EU has set strategic objectives of the information society in relation to the citizens in various sectors of eGov services in everyday life, such as: a) employments in the public sector, Job search, Announcements (e.g. educational possibilities, price changes in transportation means etc), b) Summaries of new laws, Official Government Newspaper, Collection of National Statistical Data, c) Electronic submission of forms, d) Demands on issuing public services documents e) Demands of assistance and guidance by public services, f) Immediate democracy: Contribution of simple people with their opinions, comments and thoughts regarding imminent statutes, bills, protests etc, g) Notifications in the services of the entire public sector: change of address, marital status, number of children etc, h) Moral and political protection of citizens, c) Application of the current national penal legislation on the internet nodes prohibiting: children pornography, drug trafficking etc [2].

For this purpose the EU undertook in Lisbon the Action Plan e-Europe 2002 [5] as a modernization guide of the European economy. Then, in the European Council of Seville, the European Committee proposed an updated action plan [1], called Action Plan e-Europe 2005. Today the EU has set in effect the new action plan eEurope i2010 [6]. The Strategic Framework eEurope i2010 is the new strategic frame of European Committee for information society and the mass media. The Action Plan eEurope i2010 focuses on three priorities: a) the completion of single European information space with the encouragement of an open and competitive internal market for the information society and the mass media, b) the reinforcement of innovation and investments in the research with object the ICT in Education, and c) the achievement of greater productivity through the efficient use of new technologies, which can be achieved with the modification of economic behavior, (making use of new technologies), the adaptation of corporate activities, the web supply of public services and the improvement of skills. It is important to point out that in order for Europe to materialize its strategic plans, not only should European governments produce e-Gov services, but also the citizens should adopt services in their everyday routine, and we must not neglect that a large proportion of the world population does not visit the internet at all [2].

2. The Problem that Triggered this Doctoral Research

The problem focuses on low use of eGov services and located in the geographical area of the EU (see second chapter of this thesis) and existing indicators available from the statistical basis of Europe "Eurostat" were investigated.

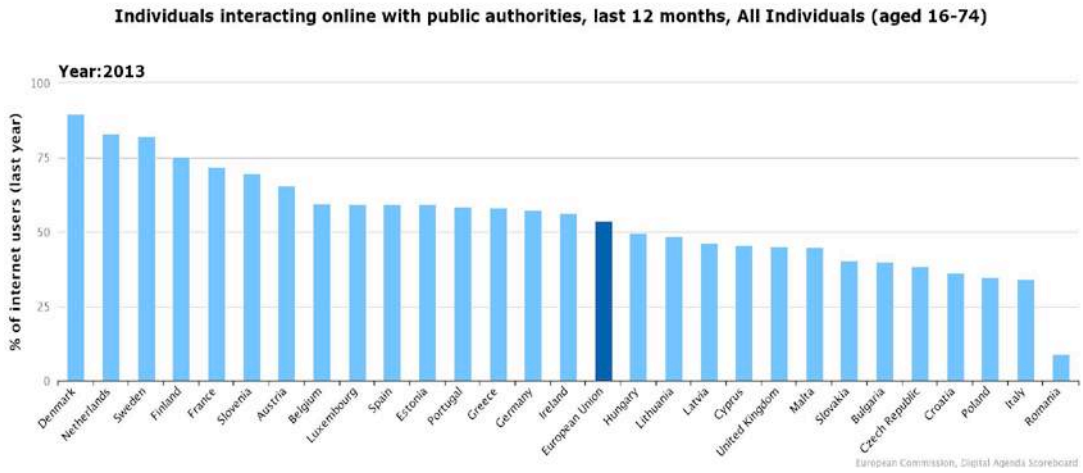


Figure1. Use of e-Gov Services in EU in 2013

According to the last available data (2013) of the Digital Agenda Scoreboard, 53.7% of all Individuals aged between 16 to 74 years old, interact online with public authorities using Internet during last 12 months in EU. They use Internet for obtaining information from public authorities' web sites, downloading official forms and sending filled in forms. Big inequalities also exist among EU member states in this metric; for example, in Romania only 9% of Individuals interact online with public authorities using Internet during last 12 months in EU, while in Denmark the corresponding rate is 89.5%, (see figure 1).

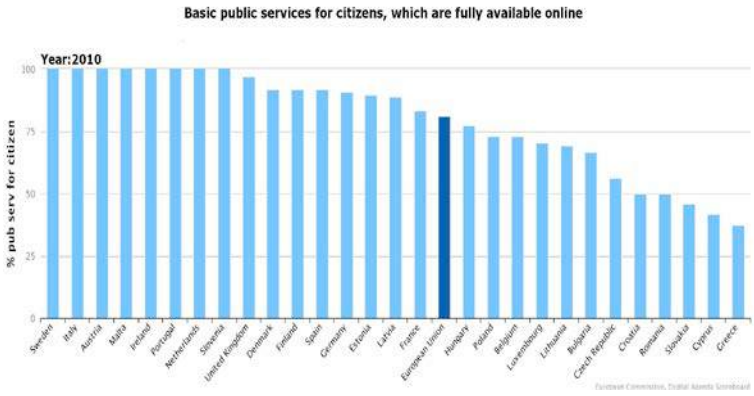


Figure2. Availability of e-Gov Services in EU in 2010

According to the last available data (2010) of the Digital Agenda Scoreboard which assesses progress with respect to the targets set in the Digital Agenda for Europe (DAE), 80.9% of a basket of 12 basic services (income taxes, job search, social security benefits, personal documents, car registration, building permissions, declaration to police, public libraries, certificates, enrolment in higher education, announcement of moving, and health-related services) for which the entire procedure can be completed online, in EU. Big inequalities also exist among EU member states in this metric; for example, in Greece only 37.5% of the services are fully available on line, while in Sweden the corresponding rate is 100%, (European Commission, 2014a) (see figure 2). It is obvious that the use of services is low, and the availability is high (see figure 2), therefore a research effort was started to detect the barriers that prevent people from adopting eGov services or the opportunities that facilitate or encouraging the adoption of eGov services.

3. Obstacles and Opportunities of Adoption of e-Government Services

In the third chapter a literature review of the obstacles and opportunities of adoption/use of e-Gov services was accomplished. The factors that are potential obstacles or opportunities were taxinomised in 3 categories: a) barriers or opportunities from the demand side, b) barriers and opportunities concerning characteristics of e-Gov services, and c) barriers or opportunities affecting demand and related legal, strategic and technological context.

Barriers have been recorded concerning both the side of supply of e-Gov services and the side of demand. If the barriers are pinpointed, it will be possible to take them into consideration while designing e-Gov services which the citizens are likely to use, if the obstacles are raised. The barriers will thus be converted in opportunities that will facilitate the adoption. It is observed worldwide that the governments tend to convert more and more public services in web accessible services. However, no one can guarantee that the web accessible services offered by the state to the citizens will indeed be used [2].

Although governments invest continuously in producing of e-Gov services, citizens face difficulties to adopt these services. Barriers derive and prevent from using them. Barrier is anything preventing the users/citizens from the adoption of e-Gov services. Barriers impede or do not allow the adoption of e-Gov services by the citizens. We see therefore, e-Gov services being continuously produced but no one guarantees that these services are used or will be used by the citizen. Barriers concerning whether e-Gov services will be adopted or not, is interrelated among others, on various factors, among them: to the income, access to the internet, and the saving of time by the citizens, the ease of use, the experience of user, accessibility, and the civic engagement.

Social e-Gov services concerning low economic status cannot be adopted by the citizens because they do not have the required knowledge. In most countries there are many disadvantaged groups, who are much less likely to use e-Gov services. These subgroups of population include elderly individuals, people with special needs, of low socio-economic level, unemployed, low income, low formal education level, national minorities, and immigrants. These disadvantaged groups make very little use of personal computer. So we infer that while the public services become digitalized, they involve the risk that a big part of European and global population might not be able to use them. A large proportion of the world population does not visit the internet at all. Decision makers and policy engravers do not use ICT technologies to take decisions. Another barrier to the use of e-Gov services is education, lack of knowledge, low levels of technology access and concerns about privacy and security by the citizens. Among several barriers is pointed the lack of trust and confidence by the citizens. The civil servants as well as the decision makers may either facilitate or prevent the growth of e-Gov services [2].

For the application of e-Gov however, many obstacles and barriers should be overcome, while extensive investments are required. The change of procedures as far as the organization and mentalities is time-consuming and many years may be required until the combined investments in I Information and Communication Technologies (ICT) funds, organization and skills yield completely their profits. The e-Gov is not only based on technological achievements, but among others, *"is a strategy aimed at offering more effective and more*

functional services" [7], it is, that is to say, a way in which, access by citizens to the volume of information owned by the state can be increased [8]. Also, e-Gov is an innovation of society which is often engaged as flow of information transmitted by the individuals who have the ability to influence the rest of the members of society. However, barriers interfere in this flow [9] [2]. EU must be skeptic and try to identify the barriers of adoption of e-Gov.

4. European e-Government Strategy from eEurope 2002 to Digital Agenda 2020

In the fourth chapter, a extensive Literature review, aims to track down the e-Gov European Strategy, investigating Legal documents of the Lisbon Strategy, the Action Plan eEurope 2002, the Action Plan eEurope 2005, the Strategic Framework i2010 and the Digital Agenda 2020.

The main scope of the Lisbon Strategy was not only to broaden and enrich the public debate on European matters, but also to encourage discussion on European values, issues and decisions using ICT. Europe decided to reform Governance and the concept of European Governance had to follow five political principles: openness, participation, accountability, effectiveness and coherence as well as to promote democracy in Europe. The EU's Action Plan 2002, established policies implemented in the provision of public services for citizens and businesses through Internet. Education and professional occupation where two major areas in which ICT could help citizens be part of the digital era and employees become flexible and specialized in ICT [4].

Furthermore, the EU's Action Plan 2005 promoted its belief that "Europe's public sector is today at crossroads, facing challenging economic and social conditions, institutional change and the profound impact of new technologies. However, there were barriers that had to overcome such as: change mindsets, push through organisational change and sustain investment. Citizens expect authorities to safeguard liberty, justice and security in the Internet as in real life. The main idea was accessibility for all, broadband connections, and interactive public services. For this reason an "Interoperable Framework for Pan European Services, electronic public procurements and public Internet points" should be created, with priority given to e-learning and e-Health. According to i2010 ICT could help make public health and welfare systems more efficient and effective. ICT could have an impact on cultural creativity in a large number of citizens. ICT could be used as a tool for environmental sustainability by using disaster management and by creating low energy efficient production processes. A new Lisbon Governance Cycle was outlined in the Strategic framework i2010, and thus new objectives of European Information Society had to be followed. The main scope was: a) to create a Single European Information Space with the aim of creating an open and competitive internal market, b) to produce better public services that increase quality of life as well as jobs and sustainable development, and c) to increase innovation and invest in ICT in order to increase growth and jobs EU [4].

Last but not least, according to the Digital Agenda 2020, in 2010, taking of course into perspective the economic crisis, the EU realized that ICT via Internet Governance could propose actions for smart, sustainable and economic growth. Priority was given to the Internet but there are seven obstacles that prevent the exploitation of ICT, such as: 1) Fragmented digital markets, 2) Lack of interoperability, 3) Rising of cybercrime and the risk of low trust in networks, 4) Lack of investment in networks, 5) Insufficient research and innovation efforts, 6) Lack of digital literacy and skills, 7) Missed opportunities in addressing societal challenges [4].

It is very important for the EU to create a *"Framework of Adoption of e-Gov Services"* that ensures that e-Gov services will be adopted by the citizens. In this view, the EU should try to ensure that e-Gov services will be used by as many European citizens as possible. This could happen, if services are produced by taking into consideration at the phase of design all the above mentioned barriers of adoption of e-Gov services, it would be more likely to increase e-Gov usage [4].

5. Identification of Barriers of e-Government Implementation

In the fifth chapter, Eurostat Database was searched and a data analysis was accomplished in order to identify barriers of e-Gov Implementation. Strategic objectives are pointed out as well as the barriers of e-Gov implementation such as, low computer use, low internet use, low computer skills, low internet skills, low level of Internet access of households. European citizens that do not have internet access at home estimate that: access costs are too high (telephone, etc.), or there is lack of skills, or they don't need it because content is not useful, or content is not interesting, or the equipment costs are too high, or content is harmful, or there are privacy or security concerns. Concerning barriers there are great differences among countries in EU. Also, it is very important for the EU to create a "Framework of Adoption of e-Gov services" that will enable e-Gov services to be adopted by citizens. In order for the EU to materialize its strategic plans, European governments have to ensure that citizens are using those services in their everyday routine. If e-Gov services are not used by the citizens, then the European relevant strategies will not benefit for the society [4].

Eurostat Database was searched in order to demonstrate some of the e-Gov evaluation metrics that probably affect the use and availability of e-Gov in EU concerning individuals and enterprises. Row data were processed and estimated the annual average and the annual average change of 19 e-Gov indicators for the years 2005-2010. Furthermore, some of the targets of Digital Agenda 2020 were evaluated whether will be accomplished ore

not, at 2015. Therefore, according to existing trends these indicators were estimated how will be, if nothing changes, in 2015. Major differences pointed out for the same indicator among countries member states of EU.

Major conclusions were that some targets of Digital Agenda were very ambitious and might not become reality by 2015, for all European countries member states of EU. Also, major differences appeared in some indicators and Europe should find out the reasons that cause these differences. If nothing changes the same indicators' trends for the years 2005-2010 will probably appear the next 5 years (2011-2015). If this happens, according to our data analysis, some countries will be under the key performance targets of Digital Agenda 2020. Also, for the years 2005-2010 trends of the same indicator are negative for some countries of EU and for others are positive. Future research is needed to investigate the reasons why this is happening [10].

Among these indicators, there are some that might affect the use of e-Gov. Availability of e-Gov in EU for the years 2005-2010 was high. Nevertheless, e-Gov usage by individuals for obtaining information from public authorities, for downloading official forms from public authorities, or for sending filled forms was very low at the same period, 32% at the period 2005-2010. This indicates that is not enough to evaluate mainly the supply side, but Europe should investigate the reasons why e-Gov use is low [10]. In EU a large amount of European citizens do not use a pc or internet, they do not know how to use a pc or internet. The results of the empirical research revealed that e-Gov services are not for "all citizens". The empirical study is a pilot study and a precursor of the investigation carried out in chapter 9. Furthermore this pilot research helped our research should focus on the assessment and management of e-Gov. E-Gov usage by enterprises is high for the years 2005-2010 concerning usage of Internet: for obtaining information from public authorities, or for obtaining forms from public authorities, or for returning filled in forms to public authorities but it is low concerning interaction with public authorities for full electronic case handling. On the contrary, e-Gov usage by enterprises for interaction with public authorities for e-procurement is low. Internet purchases of goods or services, over the Internet, by individuals for private use is very low as well as, online purchases in the last 3 months for the period 2005-2010. Online sales of small medium enterprises (SMEs), without financial sector, 10-249 persons employed, with at least 1% of turnover, are very low for the years 2005-2010 [10].

These metrics showed us that EU, in the evaluation frameworks of e-Gov, did not give a deep focus on some indicators that might affect directly the use of e-Gov or might play the role of "prerequisites" for e-Gov adoption. Therefore a new model was introduced for the adoption/use of e-Gov services in chapter 9, the UTAUT-PBO (predominant barriers or opportunities) of e-Gov Adoption. This model includes variables that are included to the category of "prerequisites of e-Gov adoption". We must not neglect that a major proportion of individuals in EU had never used Internet for the period 2005-2010. Is very important that, before EU implements a new e-Gov strategy, it sees the trends of those indicators that affect directly the adoption/use of e-Gov services by individuals and enterprises .This will help EU to track down a strategy with more realistic targets. Europe should establish a framework that should track down the reasons why e-Gov usage is low in some countries and why is high in other countries. EU may group countries according to its e-Gov readiness or maturity of adoption of e-Gov services. Also in this framework must be defined which indicators affect directly e-Gov use. Is internet use? Is computer use? etc. If these indicators are defined a more realistic European strategy could be outlined for all countries member states of EU [10].

In our days, Europe first sets targets of e-Gov and then evaluates them. In order for the targets of Digital Agenda to be successfully accomplished, EU should produce e-Gov services that will be evaluated at the phase of design. In the design stages, the services will be personalized accordingly to a) citizens' needs of each country, and b) the prerequisites of adoption of e-Gov services. This could be materialized if a "Common Framework of Pre-evaluation" is taking in consideration the prerequisites of adoption of e-Gov services and is followed by all countries at the stage of design of e-Gov services, [10]. Therefore chapter 9 and chapter 10 focus to a "Common Framework of Pre-evaluation" of the "Adoption/Use of e-Gov services".

6. Different Models in the Evaluation and Management of e-Government

In the sixth chapter we explored and recorded the different models in the evaluation and management of e-Gov. In order to explain the adoption, intention to adopt, use or intention to use of e-Gov services several models are used. Some researchers refer to adopt or intention to adopt of e-Gov services, or both. Others deal with intension to use, use of e-Gov services, or both. Some researchers employ the term "use" to imply intention to use e-Gov services. Since "adoption" is more a subjective attitude and does not necessarily imply use, we opted to deal with use and intention to use of e-Gov services directly. These models take into account many of the factors influencing the adoption or intention to adopt, use or intention to use of e-Gov services. Some of the theories predicting the use of information system (IS) can be found in the following works: Theory of Reasoned Action (TRA) [11], Extension of TRA theory is Theory of Attitude toward Behavior ([12], Technology Acceptance Model (TAM) [13], Extension of the Technology Acceptance Model (TAM2) [14], Diffusion of Innovation Model (DOI) [9], [15], Unified Theory of Acceptance and Use of Technology (UTAUT) ([16], Theory of Planned Behavior –TPB model, [17], Technology Acceptance Model- Motivational Model-MM model, Model of PC Utilization, (MPCU model) [18], Social Cognitive Theory (SCT), [19], Unified Theory of Acceptance and Use of Technology-UTAUT, Venkatesh [16].

To explain e-Gov adoption or use, some researchers: (a) suggest models which follow these theories, or (b) propose new models which attempt to predict and explain the behavior using a variety of independent variables, (c) explore new models giving weight to trust, security and usability, or (d) apply data driven methods, i.e. do data analytics without proposing models.

Among others these theoretical frameworks give focus to the following mentioned explanatory variables: Perceived Usefulness, Perceived Ease of Use-PEOU, Previous Positive Experience, Perceived Credibility, Computer self-efficacy, TOG, PU, PEOU, Perceived Risk, Cultural Characteristics which affect Uncertainty avoidance, effective interaction over the net , Risk & Uncertainty Avoidance, PU, PEOU, effective interaction over the net, Trust of the actor providing the service, General Predisposition to trust, Social Demographics (gender, education etc.), Party Affiliation, Cultural factors, Risk perceptions, Time, Environment of Innovation, User Characteristics, Satisfaction, Web-site Design, Perceived Control over the process, Perceived Usefulness, User Expectations, Economic Development, Innovation, Internet Connectivity, National Performance Indicators.

7. E-Government Evaluation Frameworks Used by the EU Action Plan e-Europe-2002 to the Digital Agenda 2020

In the seventh chapter we record the "e-Gov Evaluation Frameworks" used by the EU, from Action Plan e-Europe 2002 [20] [21] to the Digital Agenda 2020. A methodological gap was found in the evaluation of e-Gov services in Europe, while EU evaluates mainly the supply side. European e-Gov evaluation strategy is outlined in Action Plan eEurope 2002, Action Plan e-Europe 2005, the Strategic Framework i2010 and the Digital Agenda 2020. The availability & sophistication of 20 basic public services was evaluated in the evaluation frameworks of Action Plan e-Europe 2002, Action plan e-Europ was adopted. In the last decade, Europe evaluated mainly the supply side of e-Gov by evaluating the availability of 20 basic public services by estimating the indicator "Online availability and interactivity of public services", [10].

An evaluation of Europe's evaluation frameworks of Action Plans from e-Europe 2002 to Digital Agenda 2020 was accomplished and a data analysis in Eurostat Database [22] was conducted. Europe at these evaluation frameworks, until strategic framework12010, focuses mainly on availability and sophistication of e-Gov services.

The assessment of the availability of the 20 basic e-Gov services is included in Action Plans e-Europe 2002, e-Europe 2005, Strategic Framework i2010 [23]. In the last decade Europe gives priority to the supply side of e-Gov services According to Digital Europe measurements e-Gov services treated as advanced services and citizens that are digital divide cannot use them. [24]

Europe follows specific measurement frameworks, by beginning with the "Evaluation Framework of e-Europe 2002" until today following the "Evaluation framework of Digital Agenda 2020". In Action Plan e-Europe 2002 [20, 25-27], in Action Plan e-Europe 2005 [27-30] and in the Strategic Evaluation framework of i2010 [31-32] followed the "4 stage evaluation model" for the assessment of e-Gov indicators. A lot of researchers follow similar model for e-Gov evaluation, [33], [34]. Others focus to the idea of fully or not fully integrated on line e-Gov services [35]. It must be pointed out that the above mentioned methodologies evaluate e-Gov services after they are produced.

8. Usability Evaluation of e-Government Services

In the eighth chapter a new synthesis of methodologies for usability evaluation of e-Gov in the EU was proposed, which can be customized with appropriate adjustments to any service e-Gov. Also this new usability evaluation method could take the place of a "Common Framework of Pre-evaluation" of e-Gov services at the phase of design. The lack of usability is one of the most important obstacles that affect the use of e-Gov services. We apply the proposed methodologies in the e-deliberation service in Greece.

The objective of this chapter is to evaluate the usability of any e-Gov service. The e-deliberation service of Greece was chosen to serve as an object of evaluation. In Greece (a member state of EU), there is one central e-deliberation service called "Opengov", [36] available for all Ministries. In the course of the study the input of the e-deliberation service of Greece and how easily the citizens publicize their comments are examined. The focus is on the e-deliberation service of Greece, and usability problems are detected following a scenario/script: *"The submission of comments by the citizen by using the e-deliberation service of Greek Government",* in order to find out the barriers of usability problems that citizen face when using an e-Gov service. [37]

The combination of methods proposed for the usability evaluation of e-Gov services are as follows: Nielsen's Heuristic, Cognitive Walkthrough, Inspection, a script/scenario, a questionnaire, usability guidelines, usability standards, expert testing and, policy analysis. The questionnaire used was [38] to identify usability problems of e-deliberation service of Greece according to Nielsen's principles, categorized into major and minor problems. The questionnaire followed HHS usability guidelines [39] and ISO 9241-151 usability standard [40] [41]. This could help comparisons between usability evaluations, facilitate exchange of best practices and measure usability in a more structured way [37].

The questionnaire was filled by 4 experts. Usability problems are barriers of adoption/use of e-Gov services. Major and minor usability problems of an e-Gov service were detected according to Nielsen's principles concerning the e-deliberation service of Greece [37] see table 1.

Table 1. Nielsen's Heuristics

1. Visibility system status 2. Match between system and real world 3. Recognition rather than recall 4. Flexibility and efficiency using 5. User control and freedom 6. Consistency and standards 7. Error prevention 8. Aesthetic and minimalist design 9. Help users recognize, diagnose and recover from errors. 10. Help and documentation

The results showed that 23% usability points had major usability problems, 14% had minor usability problems and 63% had no usability problems. Usability problems were grouped in accordance with Nielsen's Heuristics. If designers follow this combination of methods approach at the stage of design of any e-Gov service, the usability will increase [37].

Overall, e-deliberation service of Greece is simply designed and tells the user what the next step is. This is a significant advantage in order to be comprehensible and usable by the average user. Because of its simplicity, it might be possible for citizens to participate in the e-deliberation procedure. From the other hand, the service does not help the unskilled users because they are not guided at all, there are no tools for search. These kinds of functions are essential for deliberation and for real dialogue to be accomplished in practice. Certain usability problems were addressed according to Nielsen's Principles.

9. Adoption Model of e-Government Services "UTAUT-PBO"

In the ninth chapter through empirical research some of the prevailing obstacles or opportunities were identified that affect the use of e-Gov in the EU by applying the "Adoption Model of e-Gov Services UTAUT-PBO" (PBO stands for predominant barriers and opportunities). Four models were evaluated to track do the key barriers and opportunities in Adoption/Use of e-Gov services"

The objective of this chapter is to ascertain the relation of e-Gov use to a variety of explanatory variables in an effort to measure the impact of barriers and opportunities to the ability of the general population in the EU countries to take advantage of the e-Gov services. In this research there is an effort to identify the predominant factors affecting the use of e-Gov services. An empirical study was conducted in 30 countries of the EU. Time series data were conducted for 11 years from 2001 to 2011 and for the 30 countries on 25 variables. Data were extracted from EUROSTAT and the statistical database of the United Nations. This research contributes to existing literature in the scientific area of e-Gov by proposing a model of use of e-Gov services by extending the Unified Theory of Acceptance and Use of Technology (UTAUT) model. Statistical analysis reveals specific barriers and opportunities that affect the use of e-Gov services in EU. The predominant barrier among them is the "never used internet" attribute.

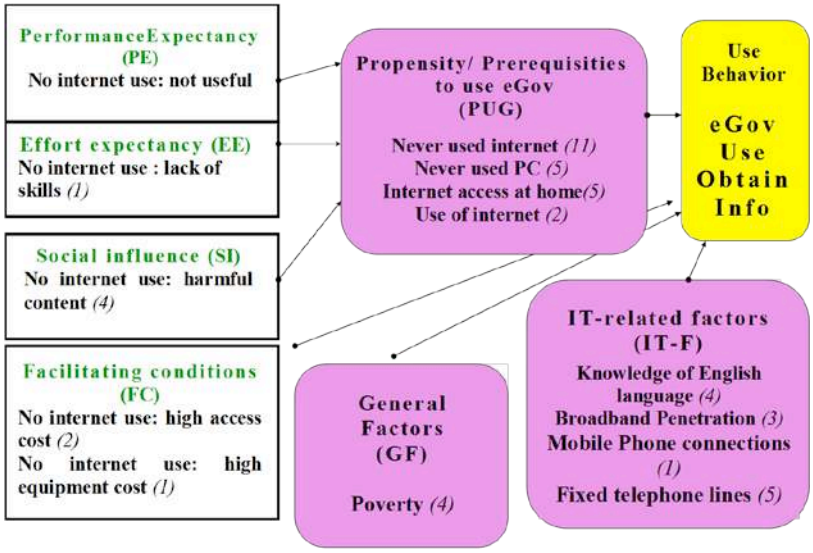


Figure 3. Adoption Model e-Gov Services UTAUT-PBO

Barriers and opportunities were identified influencing e-Gov services use in EU. The most important barrier was the fact that individuals had never used the Internet. Even though this number is decreasing in time, policy makers should explore other channels to offer e-Gov services, if the target is to increase their use in the immediate future.

Further examination of barriers and opportunities such as the ones identified herein (PC use, costs factors etc.) should be carried out before the implementation of e-Gov use enhancement measures, so that the measures themselves could be more effective and realistic.

The empirically validated model could provide a useful framework for e-Gov authorities to develop, implement and promote e-Gov services more likely to be adopted. The findings of this research also provide several important implications concerning each country in separate. The results must be seen in details by each country. Since the variable "never used internet" was identified as the most important variable to use e-Gov services, future research should focus on explaining the reasons behind the lack of internet use (a) is it objective (difficulty in access/absence of access)? (b) is it subjective (citizens do not want to or cannot use the internet)? Depending on the explanation different e-Gov implementation strategies should be adopted. In the former case improvements in access should increase the use of e-Gov services; In the latter other means of e-Gov services access should be exploited, such as mobile/smart phones and information kiosks. Since there exist considerable differences among EU countries in e-Gov, the Union should study the reasons for these differences and develop tailor made solutions which will take under consideration the country environment. To achieve a better explanation of opportunities and barriers of e-Gov use, more factors should be explored. To implement such a project, the European strategy should focus on data collection on an expanded basis.

10. Conclusions

The results of this research are manifold. Obstacles and opportunities of adopting e-Gov services were categorized by the side the demand of e-Gov services as well as to citizen's and service's characteristics. Also some of Europe's strategic objectives of e-Gov will not be achieved in 2015. There is a methodological gap in the evaluation of e-government services and systems as the EU evaluates mainly the supply side.

This thesis contributes to the existing literature in the research field of e-Gov by producing a new synthesis of methodologies for usability evaluation. The produced method can be applied to any e-government service at the phase of the design of e-Gov services and aims to rise the use of e-Gov services through increasing the usability. We must not neglect that, the lack of usability is one of the primary obstacles in the adoption of any IT system, and one of the main factors of not adopting e-Gov services. Usability is a key criterion in the diffusion of ICTs and plays an important role in e-Gov services and electronic participation environments. E-Gov services are used by heterogeneous population groups, so interfaces should have these features that guarantee all citizens may use them. Even though, EU claims that "e-Gov services for all", when e-Gov services are produced EU does not aim to usability. To produce e-Gov services for all citizens must follow the usability principles of design for all (design for all).

Developers, as well designers, could make the appropriate corrections, in order to increase the usability of the particular e-Gov service. The ultimate purpose is to increase usability of any e-Gov service and the overall added value of an e-Gov service. The proposed methodology could help avoid usability problems and increase usability, which is one of the barriers of the adoption/use of e-Gov services. The proposed evaluation methodology can be applied to any e-Gov service in order to identify the barriers of the usability. The advantage of the proposed methodology is that it combines the adoption of standards, the formulation of relevant guidelines and Nielsen's Heuristics, in an integrated whole, leading to the identification of the great majority of usability problems. It also facilitates the classification of problems in a way that the path to their solution is easily understood. ISO 9241-151web usability standard as well as HHS usability web guidelines could be used as a common base of web usability evaluation in all member states countries of EU and evaluators could follow the general guidelines of ISO 9241-151 and produce checklists and questionnaires. This could help comparisons between usability evaluations, facilitate exchange of best practices and measure usability in a more structured way [37].

Expert usability evaluation evaluates e-Gov service at the phase of prototype's design. A user testing could be conducted at the second phase. Many usability professionals first do a usability evaluation and then follow it up with a usability test [42]. Also, an accessibility testing should be designed in the future, because "Usability and Accessibility are looking at User Experience through two Lenses. Usability and accessibility are slightly different lenses to assess user experience. It is possible to be strong in one area and weak in the other. Using either approach alone could result in an inaccurate view of your site's user experience. Evaluating your website with both usability and accessibility in mind gives all users the best possible user experience"[42].

We tried by employing a multi-method approach, to trace the range of usability problems. This multi-method approach and the findings of this study can provide the basis of future explorations of usability problems of e-Gov services in Europe.

This research provides significant findings about factors that influence the adoption/use of eGovernment services by introducing a new adoption model “UTAUT-PBO”. The proposed model “UTAUT-PBO” is an extension of “UTAUT” model. In UTAUT model three new categories of variables were added: the factors which are the prerequisites of use of eGovernment services, general factors and factors related to technology. The main factor influencing the adoption/use of e-Gov services in all EU countries was that "people have not ever use the internet". Other factors affecting the use of e-Gov use in EU are as following “never used internet”, “Internet access at home”, “no basic computer skills”, “never used PC”, “poverty”, “no internet: high access cost”, “no internet: high equipment cost”, “no internet: not useful”, “no internet: lack of skills”, “no internet: harmful content”, “GERD”, “GDP”, “Knowledge of English Language: First Certificate”, “Knowledge of English Language: Proficiency”, “service availability”, “employment rate”, “broadband penetration”, “Use of internet”, “mobile phone connections”, “fixed telephone lines”, “internet use over never used internet residual”, “never used pc over never used internet residual”, “broadband penetration over never used internet residual”, “mobile telephone over never used internet residual”, “internet access home over never used internet residual”

The mediator variables of the UTAUT model were grouped into two structural components: “General Factors” and “IT-related Factors”, involving more dimensions. “General factors” contain factors determining the quality of life within a society as Gross Expenditure on Research and Development (GERD), Gross Domestic Product (GDP), employment rate, poverty. IT-related factors contain the IT framework, infrastructure, the supply channels of e-Gov services such as mobile connections, fixed telephones lines, broadband penetration, mobile phone connections. This is direct result of the fact that in order to use E-Gov services available technologies of ICT must be used. Additionally, PC or internet related technology requires a fair knowledge of the English Language. The empirically validated model could provide a “Useful Framework for e-Gov Authorities” to develop, implement and promote e-Gov services more likely to be adopted. The findings of this research also provide several important implications concerning each country in separate. The results must be seen in details by each country.

References

[1] *eEurope2005 : Information Society for all-An action plan to be presented in view of the Sevilla European Council", (2002),,* Communication from the Commission to the Council, the European Parliament, the European Economic and Social Committee and the Committee of the Regions Vol. COM 263, Final 2002.

[2] H. Delopoulos, "Barriers and Opportunities for the Adoption of eGovernance Services," presented at the Conference of Waset World Academy of Science, Engineering And Technology Paris, France 2010.

[3] *European Governance. A White Paper*, 2001.

[4] H. Delopoulos, "Objectives and Barriers of Implementation of eGovernment European Strategy: From Lisbon Strategy to Digital Agenda 2020," *Journal ePractice -Digital Strategies for Government and Business,* vol. 17, pp. 97-125, 2012.

[5] P. T. Jaeger*, et al.*, "The structures of centralized governmental privacy protection: approaches, models, and analysis," *Government Information Quarterly,* vol. 19, pp. 317-336, 2002.

[6] P. T. Jaeger and K. M. Thompson, "E-government around the world: lessons, challenges, and future directions," *Government Information Quarterly,* vol. 20, pp. 389-394, 2003.

[7] T. R. Davies, "Throw e-gov a lifeline," *Governing* vol. 72, 2002.

[8] P. T. Jaeger and K. M. Thompson, "Social Information Behavior and the Democratic Process: Information poverty, normative behavior and electronic government in the United States," *Library & Information Science Research* vol. 26, pp. 94-107, 2004.

[9] E. Rogers, ""New product adoption and diffusion"," *The Journal of Consumer Research,* vol. No4, pp. 290-301, 1976.

[10] H. Delopoulos, "Evaluation and Metrics of e-GovernmentQ From eEurope 2002 to Digital Agenda 2020," in *Developing e-Government Projects: Frameworks and Methodologies*, Zaigham Mahmood, Ed., ed USA: IGI Global, 2013, pp. 290-322.

[11] Ajzen I. and Fishbein M., *Understanding Attitudes and Predicting Social Behavior*. New Jersey: Prentice- Hall, 1980.

[12] Ajzen I. , "The Theory of Planned Behavior," *Organizational Behavior and Human Decision Processes,* pp. 179-211, 1991.

[13] F. Davis, D.,, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS Quarterly,* vol. 13, pp. 319-340, 1989.

[14] V. Venkatesh and F. Davis, D.,, "A theoretical extension of the technology acceptance model: Four longitudinal field studies," *Management Science,* vol. 46, pp. 186-204, 2000.

[15] E. Rogers, *Diffusion of Innovations*. New York: Free Press, 2003.

[16] V. Venkatesh*, et al.*, "User acceptance of information technology: toward a unified view," *MIS Quarterly,* vol. 27, pp. 425-478, 2003.

[17] I. Ajzen, "The theory of planned behavior," *Organizational Behavior and Human Decision Processes,* vol. 50, pp. 179-211, 1991.

[18] R. Thompson*, et al.*, "Personal computing: toward a conceptual model of utilization," *MIS Quarterly,* vol. 15, pp. 125-143, 1991.

[19] D. Compeau*, et al.*, "Social cognitive theory and individual reactions to computing technology: A longitudinal study," *MIS Quarterly,* vol. 23, pp. 145-158, 1999.

[20] Commission of the European Communities, "eEurope2002 Impact and Priorities,," in *COM(2001) 140 final,*, Communication from the Commission to the Council the European Parliament the European Economic and Social Committee and the Committee of the Regions, Ed., ed. Brussels: Commission of the European Communities,, 2001, p. 20.

[21] *eEurope 2005: Benchmarking Indicators,* Communication from the Commission to the Council the European Parliament, 2002.

[22] *Eurostat - Your key to European Statistics*. Available: http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/themes

[23] European Commission. (2007, 1/10). *The User Challenge Benchmarking The Supply Of Online Public Services, 7th Measurement, Gap Gemini, Ernst & Young repotr conducted for Directorate General for Information Society and Media,* . Available: http://ec.europa.eu/information_society/eeurope/i2010/docs/benchmarking/egov_benchmark_2007.pdf

[24] European Comission, "Europe's Digital Competitiveness Report 2010," Publications Office of the European Union ISBN 978-92-79-15829-2, 2010.

[25] Cap Gemini and Ernst & Young, "Web-based Survey on Electronic Public services, Results of the Third Measurement," 2002.

[26] Cap Gemini and Ernst & Young, "SUMMARY REPORT Web-based Survey on Electronic Public Services (Results of the second measurement: April 2002), Results of the Second Measurement," April 2002 2002.

[27] *eEurope Benchmarking Report eEurope2002,* Communication from the Commission to the Council the European Parliament the European Economic and Social Committee and the Committee of the Regions, 2002.

[28] Cap Gemini and Ernst & Young, "Online Availability of public services: How is Europe Progressing? ," 2004.

[29] *eEurope 2005: Benchmarking Indicators,* Communication from the Commission to the Council the European Parliament, 2002.

[30] Commission of the European Communities, "e Europe 2005 : An Information society for all. An action plan to be presented in view of the Sevilla European Council," in *COM(2002) 263, Final*, Communication from the Commission to the Council the European Parliament the European Economic and Social Committee and the Committee of the Regions, Ed., ed. Brussels: Commission of the European Communities,, 2002, p. 23.

[31] Commission of the European Communities, "i2010 – A European Information Society for growth and employment," in *{SEC(2005) 717}, COM(2005) 229 final*, Communication from the Commission to the Council the European Parliament the European Economic and Social Committee and the Committee of the Regions, Ed., ed. Brussels: COMMISSION OF THE EUROPEAN COMMUNITIES,, 2005, p. 12.

[32] Cap Gemini and Ernst & Young, "The User Challenge Benchmarking The Supply Of Online Public Services, 7th Measurement," conducted for Directorate General for Information Society and Media, European CommissionSeptember 2007.

[33] C. Kaylor*, et al.*, "Gauging e-government: A report on implementing services among American cities," *Government Information Quarterly,* vol. 18, pp. 293-307, 2001.

[34] M. P. Gupta and D. Jana, "E-government evaluation: a framework and case study," *Government Information Quarterly,* vol. 20, pp. 365-387, 2003.

[35] West Darrell M. (2000, 1/10). *Assessing E-Government: The Internet, Democracy, and Service Delivery by State and Federal Governments*. Available: http://www.insidepolitics.org/egovtreport00.html

[36] Ministry of Interior of Greek Government. site of Open Government. *(1/4/2011),* Consultation on the Draft Decision: Establish a maximum capacity of Government Motor Vehicles and other provisions". Available: http://www.opengov.gr/ypes/?p=620

[37] H. Delopoulos (in press), "A usability evaluation of e-Government services: The case of e-deliberation service of Greece," *International Journal Electronic Governance Inderscience,* 2014.

[38] H. Delopoulos. (2012, 1/9/2013). *Questionnaire of Usability Evaluation of an e-Gov service. The case of e-deliberation service of Greece*. Available: http://e-ego.gr/Questionnaire.pdf

[39] U.S.A. Department of Health and Human Services, "Research-Based Web Design & Usability Guidelines,," U.S.A. Department of Health and Human Services, Washington ISBN 0-16-076270-7, 2006.

[40] ISO, "ISO 9241-151:2008 Ergonomics of human-system interaction - Part 151: Guidance on World Wide Web user interfaces," ed, 2008.

[41] H. Delopoulos. (2012, 1/9/2013). *Questionnaire Compliance with HSS Guidelines and ISO 9241-151 standard- Usability Evaluation of an e-Gov service .The case of e-deliberation service of Greece*. Available: http://e-ego.gr/Questionnaire_Compatibility.pdf

[42] 20/1/2013). *Official U.S. Government for usability*. Available: http://usability.gov

High-dimensional polytopes defined by oracles: algorithms, computations and applications

Vissarion Fisikopoulos*

Department of Informatics and Telecommunications
National and Kapodistrian University of Athens 15784, Ilissia, Athens, Greece

April 2014

Abstract

The processing and analysis of high dimensional geometric data plays a fundamental role in disciplines of science and engineering. A systematic framework to study these problems has been developing in the research area of discrete and computational geometry. This Phd thesis studies problems in this area. The fundamental geometric objects of our study are high dimensional convex polytopes defined by an oracle.

The contribution of the thesis is threefold. First, the design and analysis of geometric algorithms for problems concerning high-dimensional convex polytopes, such as convex hull and volume computation and their applications to computational algebraic geometry and optimization. Second, the establishment of combinatorial characterization results for essential polytope families. Third, the implementation and experimental analysis of the proposed algorithms and methods.

Keywords: convex polytopes, volume computation, Newton polytope of sparse resultant, secondary polytope, regular triangulations, geometric predicates, algorithm engineering, experimental analysis

1 Introduction

The processing and analysis of high dimensional geometric data plays a fundamental role in disciplines of science and engineering. In the last decades many successful geometric algorithms have been developed in 2 and 3 dimensions. However, in most cases their performance in

*Dissertation advisor: Ioannis Z. Emiris, Professor

higher dimensions is poor. This behaviour is commonly called *the curse of dimensionality*. A solution framework adopted for the healing of the curse of dimensionality is the exploitation of the special structure of the data, such as sparsity or low intrinsic dimension, and the design of approximation algorithms. This thesis studies problems inside this framework.

The main research area is discrete and computational geometry and its connections to branches of computer science and applied mathematics like polytope theory, algorithm engineering, randomized geometric algorithms, computational algebraic geometry and optimization. The fundamental geometric objects of the study are *polytopes*, with main properties of being *convex* and defined in a *high dimensional* space.

The contribution of this thesis is threefold. First, the design and analysis of geometric algorithms for problems concerning high-dimensional convex polytopes, such as convex hull and volume computation and their applications to computational algebraic geometry and optimization. Second, the establishment of combinatorial characterization results for essential polytope families. Third, the implementation and experimental analysis of the proposed algorithms and methods. The developed software is open-source, publicly available from:

`http://sourceforge.net/users/fisikop`.

It builds on, extends and is competitive with state-of-the-art geometric and algebraic software libraries such as `CGAL` [3] and `polymake` [17].

What follows is a brief presentation of the research topics and results of the thesis, avoiding technical details.

2 Polytopes and oracles

In polytope theory, a (convex) polytope P admits two explicit representations. The first is the set of P vertices, which is called the V-representation or vertex representation. The second is the bounded intersection of a set of linear inequalities or half-spaces, which is called

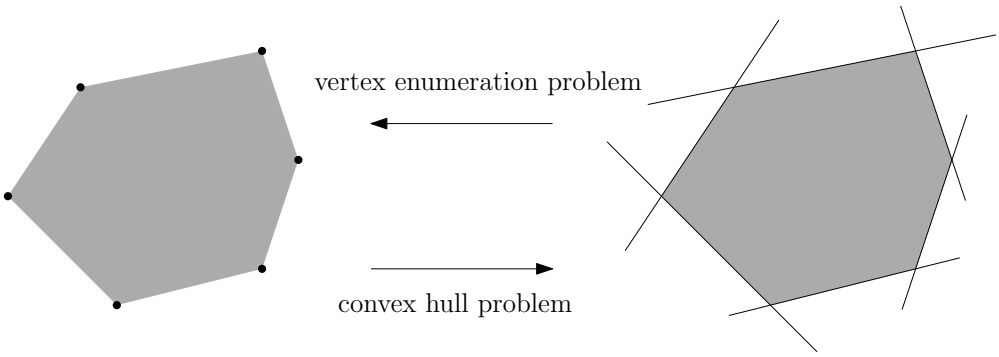


Figure 1: The V- and H-representation of a convex polygon.

H-representation or halfspace representation. Given a polytope in V-representation, computing the H-representation constitutes the *convex hull* problem, while the opposite is the *vertex enumeration* problem. These problems are algorithmically equivalent from a computational complexity point of view by *polytope duality* and establish two of the most important computational problems in discrete geometry. See Figure 1 for an illustration. For a detailed presentation on several aspects related to convex polytopes we refer to [26].

A polytope P can also be given by an implicit representation, called *(polytope) oracle*. An oracle is a black box routine that answers questions regarding P. An *optimization*, or *linear programming (LP)*, or *vertex* oracle given a vector c returns a vertex of P that has the maximum inner product with c among all points in P. Another important implicit representation for P is the *separation* oracle. That is, given a point x the oracle returns yes if $x \in P$ or a hyperplane that separates P from x otherwise. To illustrate the above definitions, let P be given in H-representation. Then an optimization oracle for P given a vector c solves an LP problem on P, while a separation oracle for P given point x evaluates the set of defining inequalities of P with x.

The relations among various oracles have been studied by Grötschel, Lovàsz and Schrijver in [20] by adopting the oracle Turing machine model of computation. To acquire, for example, an optimization oracle for P when P is given by a separation oracle, one has to solve a linear program over P. This can be done by the ellipsoid method [22]. Given an oracle for P, the entire polytope P can be reconstructed and

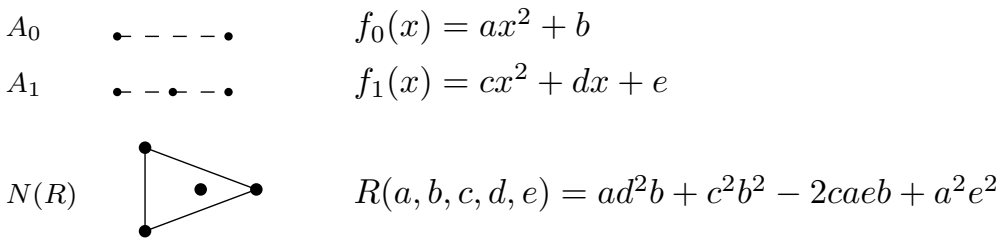


Figure 2: The resultant of a system of two polynomials in one variable.

its explicit representation can be found using an incremental convex hull algorithm such as the Beneath-and-Beyond [4].

3 Algorithms for resultant polytopes

From the algebraic geometry perspective polytopes characterize polynomials better than total degree thus offering the fundamental representation in sparse elimination theory, called *Newton polytopes*. An important class of such polytopes is the Newton polytopes of the *sparse resultant polynomial* or simply the *resultant polytopes*. They have been studied by Gelfand, Kapranov and Zelevinsky in [19] and by Sturmfels in [25]. An example of the resultant of two polynomials f_0, f_1 in one variable x is depicted in Figure 2. It is a polynomial R in the coefficients a, b, c, d, e of the two polynomials which vanishes if the system we get by specializing a, b, c, d, e to numerical values has a solution. Here, the Newton polytope $N(R)$ of the resultant is a triangle.

In [19] the study of resultant polytopes is connected to the study of *secondary polytopes*. The secondary polytope of a pointset A is a fundamental object in geometric combinatorics since it offers a polytope realization of the graph of *regular triangulations* of the pointset. An equivalent realization is the graph of *regular fine mixed subdivisions* of the Minkowski sum of pointsets. Figure 3 depicts an example of secondary and resultant polytopes. In the special case where the points in A are in convex position and two dimensional all triangu-

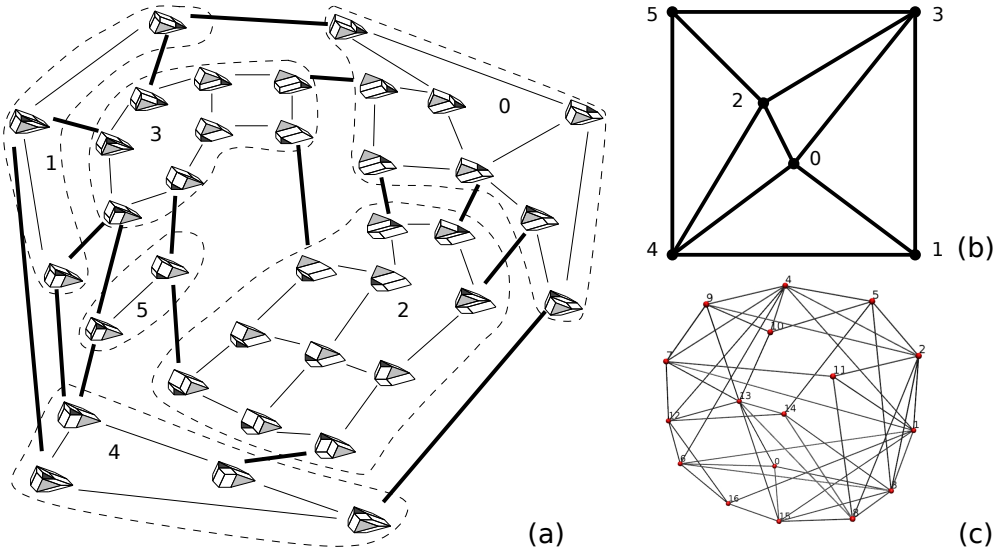


Figure 3: Example of secondary and resultant polytopes: (a) The secondary polytope of two triangles (dark, light grey) and one segment; vertices correspond to mixed subdivisions of the Minkowski sum $A_0 + A_1 + A_2$ and edges to flips between them (b) the resultant polytope, whose vertices correspond to the dashed classes of the secondary polytope. Bold edges of the secondary polytope map to edges of the resultant polytope (c) 4-dimensional resultant polytope of 3 generic trinomials with f-vector $(22, 66, 66, 22)$; figure made with `polymake`.

lations are regular and the secondary polytope is the 3-dimensional associahedron [24].

Chapter 2 of the thesis presents the design and the analysis of the first *output-sensitive* algorithm for computing (projections of) resultant polytopes. The algorithm is output-sensitive as it makes one oracle call per vertex and facet of the polytope. The key ingredients of that algorithm is the compact representation of resultant polytopes by an optimization oracle and the exploitation of their low intrinsic dimension. The oracle constructs regular triangulations in order to compute the optimal vertex in the polytope. Finally, the resultant polytope is reconstructed using an incremental convex hull algorithm that uses this oracle.

The algorithm is implemented in the software package `respol`, which computes 5-, 6- and 7-dimensional polytopes with $35{\cdot}10^3$, $23{\cdot}10^3$ and 500 vertices, respectively, within 2 hours on a standard computer, and the Newton polytopes of many important surface equations encountered in geometric modelling in $<$ 1sec, whereas the enumeration of the vertices of the corresponding secondary polytopes is intractable. `respol` has been used to solve essential problems in CAD [14] as well as to compute discriminant polynomials [15]. We propose and implement a technique called *hashing of determinants*, which avoids duplication of computations by exploiting the nature of determinants computed by the algorithm. In practice, this technique accelerates execution up to 100 times.

The results of this work have been published in [12] and their full version in [13]. An extension of the above method to computing discriminant polytopes is discussed in Section 2.6 and has appeared in [11].

4 Edge-skeleton computation

Motivated by the fact that the above algorithm is impractical in 8 or more dimensions since it relies on an incremental convex hull algorithm, the study extends in finding more efficient, i.e. *total polynomial time*, algorithms for convex hulls. An algorithm runs in total polynomial time if its time complexity is bounded by a polynomial in the input *and* output size. In general dimension finding a total polynomial time algorithm for vertex enumeration is a major open problem in algorithmic geometry. However, total polynomial time algorithms exist for vertex enumeration of special polytope cases, such as simplicial polytopes [1] and 0/1-polytopes [2].

Here we establish another case where total polynomial time algorithms exist. We present the first total polynomial time algorithm for a special case of the vertex enumeration problem where the polytope is given by an optimization oracle and we are also given a superset of its edge directions. In particular the algorithm computes the *edge-skeleton* (or 1-skeleton) of the polytope, which is the graph of polytope vertices and edges. Since the vertices are computed along with

the skeleton, the edge-skeleton computation subsumes vertex enumeration.

We consider two main applications. We obtain total polynomial time algorithms for computing signed Minkowski sums of convex polytopes, where polytopes can be subtracted provided the signed sum is a convex polytope, and for computing secondary, resultant, and discriminant polytopes. Further applications include convex combinatorial optimization and convex integer programming, where we offer an alternative approach, thus removing the exponential dependence on the dimension in the complexity.

The results of this work are presented in Chapter 3 of the thesis. Some preliminary results have been published in [9] and their full version in [10].

5 Approximate volume computation

Vertex enumeration in high dimensions (e.g. one hundred) using the above methods is a futile attempt. Thus, this thesis aims at exploiting the limits of learning fundamental characteristics of a polytope such as its volume. Although volume computation is #-P hard for V- and H-representations of polytopes [7] there exist randomized polynomial time algorithms to approximate the volume of a convex body with high probability and arbitrarily small relative error. Starting with the breakthrough polynomial time algorithm of [6], subsequent results brought down the exponent on the dimension from 27 to 4 [23]. However, the question of an efficient implementation had remained open.

This thesis undertakes this by experimentally studying the fundamental problem of computing the volume of a convex polytope given as an intersection of linear inequalities. We implement and evaluate practical randomized algorithms for accurately approximating the polytope's volume in high dimensions (e.g. one hundred). To carry out this efficiently we experimentally correlate the effect of parameters, such as random walk length and number of sample points, on accuracy and runtime. Moreover, we exploit the problem's geometry by implementing an iterative rounding procedure, computing partial

generations of random points and designing fast polytope boundary oracles. Our publicly available code is significantly faster than exact computation. We provide volume estimations for the Birkhoff polytopes $B_{11}, \ldots, B_{15}$, whereas only the volume of B_{10} has computed exactly.

The results of this work are presented in Chapter 4 of the thesis and published in [8].

6 Combinatorics of resultant polytopes

We study the combinatorics of resultant polytopes. These are known in the case of two polynomials in one variable, also known as the Sylvester case [18] and in the case where the polytope's dimension is up to 3 [25]. We extend this work and at the same time answer an open question raised in [21] by studying the combinatorial characterization of 4-dimensional resultant polytopes, which show a greater diversity and involve computational and combinatorial challenges.

In particular, our experiments, based on `respol`, provide a series of polytopes that establish lower bounds on the maximal number of faces. By studying subdivisions of Minkowski sums, called *mixed subdivisions*, we obtain tight upper bounds on the maximal number of facets and ridges. These yield an upper bound for the number of vertices, which is 28 whereas the general bound yields 6608 [25]. Figure 3(c) shows an instance with f-vector $(22, 66, 66, 22)$ that maximizes the number of facets and ridges.

We establish a result of independent interest, namely that the f-vector is maximized when the input is sufficiently generic, namely full dimensional and without parallel edges. Lastly, we offer a classification result of all possible 4-dimensional resultant polytopes.

The results of this work are presented in Chapter 5 of the thesis and have been published in [5].

7 Geometric predicates

Geometric algorithms involve both combinatorial and algebraic computation. In many cases, such as convex hull computations, the later

boils down to determinant sign computations, also called *geometric predicates*. As the dimension of the computation space grows, a higher percentage of the computation time is consumed by these predicates. Our goal is to study the sequences of determinants that appear in geometric algorithms. We use dynamic determinant algorithms to speed-up the computation of each predicate by using information from previously computed predicates.

We propose two dynamic determinant algorithms with quadratic complexity when employed in convex hull computations, and with linear complexity when used in point location problems. Moreover, we implement them and perform an experimental analysis. Our implementations outperform the state-of-the-art determinant and convex hull implementations in most of the tested scenarios, as well as giving a speed-up of 78 times in point location problems.

The results of this work are presented in Chapter 6 of the thesis and have been published in [16]. The developed software package has been submitted in `CGAL` [3] and is currently under revision.

8 Extensions and open problems

Several intriguing open questions emerge by the study of this thesis. From the geometric combinatorics point of view one question is to understand the symmetry of the maximal f-vector, i.e. vector of polytope's face cardinalities, that appear in the study of the combinatorics of 4-dimensional resultant polytopes.

There are a few questions related to sampling. The first is to study volume approximation algorithms when an optimization oracle is available. The current research focuses on convex bodies, or polytopes, represented by a membership oracle. A special case which is also interesting is to sample random points from polytopes given in V-representation without using membership queries. Other related problems are computing the volume of spectahedra or general semi-algebraic sets, application of the current software to other #P-hard problems like counting linear extensions of partial ordered sets, integration of polynomial functions over convex polytopes, study polytopes that are easy/difficult to sample from under the assumption that

they are rounded, study the quality of sampling or compare point samples, and sampling integer points from polytopes.

Nearest neighbour searching has been considered as one of the most fundamental problems in computer science. Our study in Chapter 4 paves the way for an application of approximate nearest neighbour searching to approximate polytope oracles and polytope volume approximation.

References

[1] D. Avis and K. Fukuda. A pivoting algorithm for convex hulls and vertex enumeration of arrangements and polyhedra. *Discrete & Comput. Geometry*, 8:295–313, 1992.

[2] M. Bussieck and M. Luebbecke. The vertex set of a 0/1-polytope is strongly P-enumerable. *Comput. Geom.: Theory & Appl.*, 11:103–109, 1998.

[3] CGAL: Computational geometry algorithms library. http://www.cgal.org.

[4] B. Chazelle. An optimal convex hull algorithm in any fixed dimension. *Discrete & Computational Geometry*, 10:377–409, 1993.

[5] A. Dickenstein, I.Z. Emiris, and V. Fisikopoulos. Combinatorics of 4-dimensional resultant polytopes. In *ISSAC '13*, pages 173–180, New York, NY, USA, 2013. ACM.

[6] M. Dyer, A. Frieze, and R. Kannan. A random polynomial-time algorithm for approximating the volume of convex bodies. *J. ACM*, 38(1):1–17, 1991.

[7] M.E. Dyer and A.M. Frieze. On the complexity of computing the volume of a polyhedron. *SIAM J. Comput.*, 17(5):967–974, 1988.

[8] I.Z. Emiris and V. Fisikopoulos. Efficient random walk methods for approximating polytope volume. In *Proc. Symp. Comp. Geometry*. ACM, 2014.

[9] I.Z. Emiris, V. Fisikopoulos, and B. Gärtner. Efficient volume and edge-skeleton computation for polytopes defined by oracles. In *Proc. EuroCG 2013*, Braunschweig, Germany, March 2013.

[10] I.Z. Emiris, V. Fisikopoulos, and B. Gärtner. Efficient edge-skeleton computation for polytopes defined by oracles, 2014. Submitted to journal.

[11] I.Z. Emiris, V. Fisikopoulos, and C. Konaxis. A software framework for computing newton polytopes of resultants and (reduced) discriminants. In *MEGA*, Frankfurt, Germany, 2013.

[12] I.Z. Emiris, V. Fisikopoulos, C. Konaxis, and L. Peñaranda. An output-sensitive algorithm for computing projections of resultant polytopes. In *Proc. Symp. on Comp. Geom.*, pages 179–188, 2012.

[13] I.Z. Emiris, V. Fisikopoulos, C. Konaxis, and L. Peñaranda. An oracle-based, output-sensitive algorithm for projections of resultant polytopes. *Intern. J. Comp. Geom. Appl., Special Issue*, 23:397–423, 2013.

[14] I.Z. Emiris, T. Kalinka, C. Konaxis, and T. Luu Ba. Implicitization of curves and (hyper)surfaces using predicted support. *Theor. Comp. Science, Special Issue on Symbolic & Numeric Computing*, 479(0):81–98, 2013.

[15] I.Z. Emiris, T. Kalinka, C. Konaxis, and T. Luu Ba. Sparse implicitization by interpolation: Characterizing non-exactness and an application to computing discriminants. *J. Computer Aided Design*, 45:252–261, 2013.

[16] V. Fisikopoulos and L. Peñaranda. Faster geometric algorithms via dynamic determinant computation. In *ESA 2012*, volume 7501 of *Lecture Notes in Computer Science*, pages 443–454. Springer, 2012.

[17] Ewgenij Gawrilow and Michael Joswig. polymake: a framework for analyzing convex polytopes. In G. Kalai and G.M. Ziegler, editors, *Polytopes — Combinatorics and Computation*, pages 43–74. Birkhäuser, 2000.

[18] I.M. Gelfand, M.M. Kapranov, and A.V. Zelevinsky. Newton polytopes of the classical resultant and discriminant. *Advances in Math.*, 84:237–254, 1990.

[19] I.M. Gelfand, M.M. Kapranov, and A.V. Zelevinsky. *Discriminants, Resultants and Multidimensional Determinants*. Birkhäuser, Boston, 1994.

[20] M. Grötschel, L. Lovász, and A. Schrijver. *Geometric Algorithms and Combinatorial Optimization*. Springer, Berlin, 2nd edition, 1993.

[21] A. Jensen and J. Yu. Computing tropical resultants. *Journal of Algebra*, 387(0):287–319, 2013.

[22] L.G. Khachiyan. A polynomial algorithm in linear programming. *Soviet Math. Doklady*, 20(1):191–194, 1979.

[23] L. Lovász and S. Vempala. Simulated annealing in convex bodies and an $O^*(n^4)$ volume algorithm. *J. Comp. Syst. Sci.*, 72(2):392–417, 2006.

[24] Alexander Postnikov. Permutohedra, associahedra, and beyond. *Int. Math. Res. Not.*, 2009(6):1026–1106, 2009.

[25] B. Sturmfels. On the Newton polytope of the resultant. *J. Algebraic Combin.*, 3:207–236, 1994.

[26] G.M. Ziegler. *Lectures on Polytopes*. Springer, 1995.

Organizational Structure, Operational Strategy, Indexes and Forecasting in the Telecommunication Market

Antonios Kargas*

National and Kapodistrian University of Athens
Department of Informatics and Telecommunications
kargas.antonios@gmail.com

Abstract. The modern business environment is characterized by intense competition, which has led telecommunication companies to a continuous race towards gaining and maintaining a competitive advantage. In order to succeed, telecommunication companies "cultivate" cultures as non – imitable characteristics, capable to ensure long – term corporate viability and growth. This thesis provides an insight into the Greek telecommunication industry by contributing: a) to the creation of national cultural profile in telecommunication industry, which can become a starting point in a wider trial to create a European industry profile, b) to the empirical testing of the correlation between culture and a series of administrative and financial indexes and c) to the examination of the extent to which background factors (such as firms' age and size) should be taken into account during the implementation of a business strategy. Moreover, the thesis scope is to use historical data in order to forecast Greek GDP. By using a range of forecasting models the Greek economy's performance is investigated and the main indicators are revealed. Main findings are the revealing of important economic indexes determining the Greek GDP as retail trade index, industrial production index, unemployment rate and touristic index, and the forecast of high recession for the year 2012.

1 Introduction

During the last two decades of the 20th century, economies' globalization, markets' deregulation and the privatization of governmental monopolies, have significantly changed the way companies and markets, work. The new, complicated competitive environment intensified the need for operational control, but moreover the need for comprehension of the competition.

Historically, the traditional financial tools (profit indexes, return of investment, etc.), which have been the most important source of information for the administrative executives, proved to face limitations compared to the market needs. Scientists realized that the pressure for immediate financial profitability could lead to short-term results, but this could undermine the company's ability to produce future economic

* Dissertation Advisor: Dimitrios Varoutas, Assistant Professor

value, through the sustainable development of elements, such as organizational culture. In the modern, globalized, business environment, where firms compete for a sustainable competitive advantage, profits are still thought to be important, but the way they are obtained, is considered to be first priority.

Telecommunications providers should take into consideration these operational changes, and to adopt, "soft core" factors, such as organizational culture. Managers should be able to give solutions in problems related to the coordination between organizational characteristics and desirable business strategy, as well as to answer the question how this coordination can prevail intensive market competition. Furthermore, understanding leadership's role is of high importance, in order to understand the way managers formulate operational structures and the relationships with "external" associates (customers, suppliers, competitors and state). Moreover, managers and regulators should have the ability to understand how organizational characteristics, such as size and age, affect perceived work conditions, so as to develop financially effective structures and to obtain the best operational result.

Under these circumstances, telecommunication industry has been selected, in order to understand and present its organizational structure, operational strategy and its main business indexes. The specific industry has been chosen as a representative of the whole market because of its substantive organizational and business characteristics that permit international comparisons. The market is characterized by free and intense competition, while no entry barriers exist. The technological, economic and regulatory developments affect directly the market, and phenomena of acquisitions, mergers and strategic alliances redefine its boundaries. In most developed as long as in the developing countries, telecommunication product is a validated entrepreneurial and consumable product, while its usage can determine the countries' overall economic potentials (in terms of Gross Domestic Product), according to several studies. Finally, the telecommunications market is mature, as it has made its way from quantitative financial indexes to a multifactoral system, which consists of both quantitative and qualitative indexes. This fact allows conducting research; in order to conclude about a firm's competiveness and business success, but moreover on how these can be achieved.

Organizational structure and operational strategy, have been determined by measuring fixed and mobile operators' organizational culture, along with their in between interactions. The study's aim is to understand the specifications of: a.) each company, b.) each industry (fixed-mobile telephony) and c.) the market as a whole. Organizational structure, operational strategy and efficiency indexes are determined through the interdependent relationship between organizational culture and:

- market's competition level,
- leadership's character,
- company's size,
- company's age,
- market orientation and
- performance indexes.

All these are examined in order to determine strategies or / and business practices that will enable: researchers to make international comparisons, administrative executives to practice effectively their duties and regulators to define measures for attaining a better level of competition, according to market's characteristics.

1.1 Method

The thesis employed a quantitative research tool, the Organizational Culture Assessment Instrument (OCAI) created by R.E Quinn & K.S Cameron [1] to create the theoretical framework used so as to conduct the study presented hereafter. This instrument recognizes six cultural dimensions: dominant characteristics, leadership style, employees' management, organizational glue, strategy and criteria of success. The model has two dimensions (flexibility and orientation), which create four types of culture (Figure 1). The two dimensions create four distinct quadrants, each one representing a different type of organizational culture. Each type of organizational culture has its own characteristics and its own strengths and weaknesses. Of course, an organization can have elements from more than one type and that is why the reader must be familiar with all types.

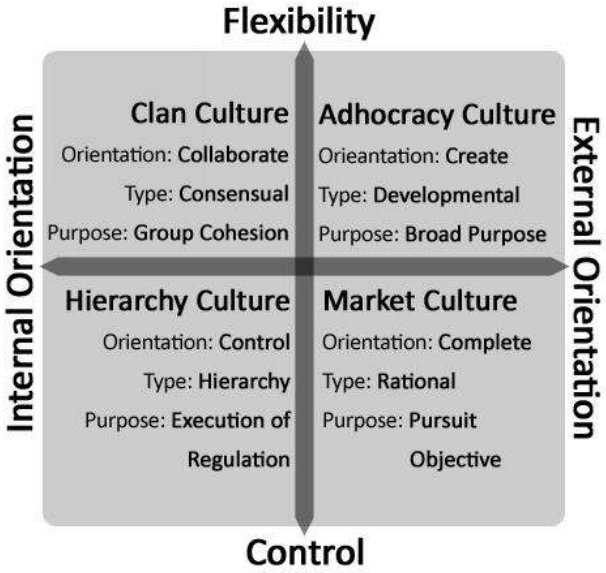


Fig. 1: Quinn & Cameron's model

Author selected a quantitative approach, instead of a qualitative one, taking into account the nature of research and the strengths – limitations both approaches have. The non-experimental, quantitative approach selected gives: a) precision, through quantitative and reliable measurement, b) statistical techniques for sophisticated analyses and c) replicable results. This approach permits to apply conventional standards of reliability and validity, while the results are open to criticism. Moreover, the quantitative approach is conducted in an attempt to answer certain questions and to test hypotheses. It represents an attempt to identify why something happens, what causes some event, or under what conditions an event does occur. To answer such questions researchers have to eliminate the simultaneous influence of many variables to isolate the cause of an effect. Controlled inquiry is absolutely essential to this because without it the cause of an effect could not be isolated. Qualitative approaches

for example appear to be subject of anonymity and confidentiality which have a profound effect in the subjects of study, while the viewpoints of both researcher and participants have to be identified and elucidated because of issues of bias.

Data regarding organizational culture in the Greek telecommunication industry were collected by interviews and mails. No judgmental criteria have been used because of the relatively small number of telecommunication companies operating in the Greek market. The questionnaire comprised of twenty one (21) closed-type multiple choice questions. A trial survey was conducted in the last trimester of 2008 to establish if the answering procedure could be easily understood and complied with, while the research questionnaires were made available a few months later, during 2009. Three hundred and two (302) employees and middle line managers from five fixed operators and three mobile operators participated in the research. Employees were asked to fill out the questionnaires and return them, while firms' managers were interviewed. The purpose of the interviews was to achieve a profound understanding of the framework under which each company operates (in top management level) and to measure its overall organizational culture. Interviews were conducted in order to persuade managers to participate in the research and to save time from exploring and trying to understand the questionnaire. Moreover, interviews provided important qualitative characteristics about the structures and the operational management of each firm.

At the end, 80 questionnaires from managerial representatives were collected by interviews and 222 questionnaires from employees were sent back (out of 374 totally sent to employees – response rate of 59.36%). The survey only includes companies that achieved predetermined limit of 25 questionnaires, in order to gain a statistically significant view for each firm. Although the analysis conducted in the paper was at firm level, the characteristics of respondents were also provided (Table 1) in order to achieve higher comprehension of the participants in the research. Firms' age measured by years since the founding date and data reveal that the majority of companies are relatively new (less than 10 years old). Firms' size measured according to current number of employees and reveals that the vast majority of firms (62.50 per cent) can be characterized as medium or small size enterprises (until 250 employees).

2 Results and discussion

In the second part of the thesis the main results are presented. The results concern: a) the role of organizational culture in the new business environment and b) the methods used in order to forecast the Greek GDP

2.1 Organizational culture and market orientation in Greek telecommunication market

Three research hypotheses were examined in order to reveal the relationship between organizational culture and market orientation in the Greek telecommunication market:

***H1**: The degree of market orientation among the Greek telecommunication providers varies from highest to lowest depending on the dominant culture type. In other words, the greater the prevalence of the market culture, the highest the degree of market orientation.*

***H2**: If H1 holds true then extrovert-type cultures (adhocracy – market) will result in improved performance compared to introvert-type cultures (clan – hierarchy).*

***H3**: The special traits of telecommunication providers (size and age) influence the degree of market orientation and the dominant culture type (introvert or extrovert)*

The internal consistency of the questionnaire variables was controlled with the Cronbach α consistency coefficient and use of the SPSS statistical package. The Cronbach α consistency coefficient constitutes one of the most widely-acknowledged methods for consistency control defining the consistency coefficient for each individual variable but also for the set of variables as a whole. Based on Kurtinaitiene [2] and Nunnaly [3], a tool is considered consistent when the "α" coefficient, ranging between 0 and 1, exceeds 0.7. As illustrated in Table 1 the "α" coefficient for the research tools ranges between 0.691 and 0.703, satisfying the set criterion.

Table 1. Consistency Results

Criteria	Number of questions	Cronbach α
Clan Culture	6	0.700
Adhocracy Culture	6	0.691
Market Culture	6	0.778
Hierarchical Culture	6	0.703

Given the number of companies comprising the sample and the fact that results do not follow normal distribution at all times, the Kendall coefficient was used in order to examine the possible correlation between the different variables in the samples. The findings that emerged are statistically significant and are presented below.

The first research hypothesis examines the link between market orientation among the Greek telecommunication companies and the dominant culture type. The results indicate that extrovert culture types (market and adhocracy) are positively correlated with the degree of market orientation, with the correlation index reaching values of 0.429 and 0.238 respectively, compared to other two types. The Research Hypothesis H1 is therefore confirmed. Market orientation among Greek telecommunication companies is indeed influenced by the dominant culture within the company. In particular, companies with extrovert culture types display the greatest extent of market orientation, compared to those with dominant cultures with introvert characteristics.

Furthermore, comparing culture types in sets of two (clan and hierarchical to adhocracy and market) it is observed that when companies tend towards 'control' as opposed to 'flexibility' they are market orientated to a greater degree. It is therefore derived that the more Greek telecommunication companies tend towards control, the greater their degree of market orientation. In fact, the only culture type negatively correlated to market orientation is the clan culture, which displays a great degree of

flexibility and introvert characteristics. The findings linked to this research hypothesis are presented in Table 2.

Table 2. Link between market orientation and culture type

	Clan Culture	Adhocracy Culture	Market Culture	Hierarchical Culture
Market orientation	**-0.524**	**0.238***	**0.429**	**0.143**

Note: All correlations are significant (α = 0.01) except those marked by * which are only significant at an α = 0.05

The second research hypothesis attempts to shed light on the link between performance and dominant culture type. In order to examine this link, two indicators were used, profitability and financial growth, as observed among the Greek telecommunication companies. The findings indicate positive correlation between the adhocracy and market culture types with the two indicators of financial performance and a negative correlation between the other two culture types and the indicators used. These empirical findings confirm Research Hypothesis H2 and underline the earlier results obtained by Papadimitriou et al. [4] on the Greek telecommunication companies.

It is interesting to note that in regards to performance, the greater tendency is observed among the telecommunication companies with extrovert characteristics and also greater flexibility. Therefore, the companies with an adhocracy culture display higher performance levels in comparison to those with market culture (and decidedly improved are the results for clan culture companies compared to those with hierarchical cultures. The findings linked to the second research hypothesis are presented in Table 3.

Table 3. Link between performance and culture type

	Clan Culture	Adhocracy Culture	Market Culture	Hierarchical Culture
Profitability	**-0,048**	**0.810**	**0.143***	**-0.238***
Growth	**-0.143**	**0.714**	**0.048**	**-0.524**

Note: All correlations are significant (α = 0.01) except those marked by * which are only significant at an α = 0.05

The third research hypothesis focuses on whether the dominant culture type and the degree of market orientation are influenced by company-specific indicators, like size and age. It emerged that the larger the company the more dominant the hierarchical culture (0.619), somewhat expected for large companies where operational control demands bureaucratic structures compared to smaller and more flexible companies [5]. The same findings are observed in the links between culture and company age. Finally, the correlation between market orientation and the two company-specific indicators is positive confirming Research Hypothesis H3. This finding is mainly attributed to the fact that in these companies the application of marketing approaches is a formalized process that is carried out on a frequent basis [6]. The findings linked to the third research hypothesis are presented in Table 4.

Table 4. Link between market orientation, culture type, age and size

	Clan Culture	Adhocracy Culture	Market Culture	Hierarchical Culture	Market orientation
Size	**-0.048***	**-0.238**	**-0.238**	**0.619**	**0.143***
Age	**-0.390**	**-0.390**	**-0.098***	**0.781**	**0.390**

Note: All correlations are significant ($\alpha = 0.01$) except those marked by * which are only significant at an $\alpha = 0.05$

2.2 Greek GDP forecast estimates

There are a lot of methods to forecast economic magnitudes: Time Series Decomposition [7], ARIMA methodology [8], Vector Autoregressive methods (VAR) [9], Dynamic Factor Model (DFM) [10] and regression [11].

Three different methods have been used, in order to forecast the Greek GDP: Time series decomposition, ARIMA methodology and regression of GDP as dependent variable and different economic indexes as independent variables and forecast of the independent variables. All using data are quarterly and originate from Eurostat and Hellenic Statistical Authority. The data are at constant prices of 2005.

All three models forecast recession exceeding 9%, while the most pessimistic forecast (OECD) was referring to a recession of 5,3%. This divergence can become an obstacle during the governmental strategic planning. The Center of Planning and Economic Research, a Greek independent institute, announced on June 2012 that recession reached 6,54% the Q1/2012 and 6,87% the Q2/2012, while there is an estimation for a total recession of 6,7% for the whole 2012 [12]. The forecasting method used by KEPE, ([12], [13]), is the Dynamic Factor Model, with six factors calculated from 144 different variables and maximum variance explained 64%, which is not large.

Table 5. GDP forecast of different organizations

	2009	**2010**	**2011**	**2012**	**2013**
OECD	-3,2	-3,5	-6,9	-5,3	-1,3
IMF		-3,5	-6,9	-4,7	0,0
European Commission		-3,5	-6,9	-4,7	0,0
Greek Budget Program		-3,5	-5,5	-2,8	
Forecast Methodologies					
Time Series Decomposition				-9,4	
ARIMA				-10,1	
Regression				-9,9	
Percentage change, volumes (2005 prices)					

A well defined formula with dependent variable GDP and independent or predictor variables different economic indexes is found. Two equations have been produced with initial value the first quarter of 2004: in the first data are taken into account until

the fourth quarter of year 2010 and in the second until the fourth quarter of year 2011. Respectively, the equations are:

$$GDP = 27.992{,}63 + 62{,}97\ INPR + 0{,}404\ INV + 111{,}31\ RET - 12{,}65\ CON - 229{,}61\ UNR + 51{,}46\ TI + 0{,}10886\ DIFF \quad (1)$$

$$GDP = 29.140{,}36 + 58{,}01\ INPR + 0{,}438\ INV + 109{,}73\ RET - 13{,}33\ CON - 252{,}89\ UNR + 49{,}59\ TI + 0{,}09811\ DIFF \quad (2)$$

Where:

GDP: (Millions)
INPR: Industrial Production Index (Volume).
INV: Public Investment (Millions).
RET: Retail Trade Index (Turnover).
CON: Construction Index (Turnover).
UNR: Unemployment Rate.
TI: Touristic Income Index (Turnover).
DIFF: General Government Expenses minus Total Government Income.

The presented equations reveal a paradox: the negative coefficient of the variable CON. The construction index was due to Olympic Games in very high levels during and before the year 2004 and has been normalized to smaller values after 2 years, during GDP increased. The very high values of negative slope during the years 2004 and 2005 led to the negative coefficient of the variable CON, even if the variable is statistically significant. The contribution of other economic indexes such us "Agriculture, hunting and forestry, fishing index" tends statistically to zero. These equations have the advantage to reveal us which economic variables contribute to the Greek GDP and to understand the reality of the Greek economy.

For the Equation 1 the Durbin-Watson statistic is 1,85022 < 2,071 for 0,05 significant level with the test to be inconclusive and greater than the value 1,847 for 0,01 significant level (n=28, K=7 independent variables). R2=97,1%, R2(adjusted)=96,1% and R2(prediction)=94,27%. The p-values are much smaller 0,05 except for UNR variable with p-value 0,075 and all VIF factors smaller than 4.

For the Equation 2 the Durbin-Watson statistic = 2,05909 > 2,004 for 0,05 significant level and greater than the value 1,788 for 0,01 significant level (n=32, K=7 independent variables). For both cases there is not serial correlation. R2=97,6%, R2(adjusted)=96,9%, and R2(prediction)=95,53%, all VIF factors are smaller than 6 and p-values much smaller than 5%. Using Equation 1 and put to the independent variables the actual values of 2011 it leads to a recession of -6,82% and Equation 2 with the same values leads to a recession of -6,85%, both in very good agreement with the actual recession of -6,91%.

In order to forecast the GDP and the recession for the year 2012 the values of independent variables are needed. There are different models to forecast the economic indexes – independent variables. One of them is using VAR models in different groups of endogenous variables. The absence of suitable data has made the application of such a technique impossible. Instead time series decomposition and

ARIMA models have been used in order to calculate the needed values. ARIMA models are the best univariate forecasting tool for a wide family of functions [14].

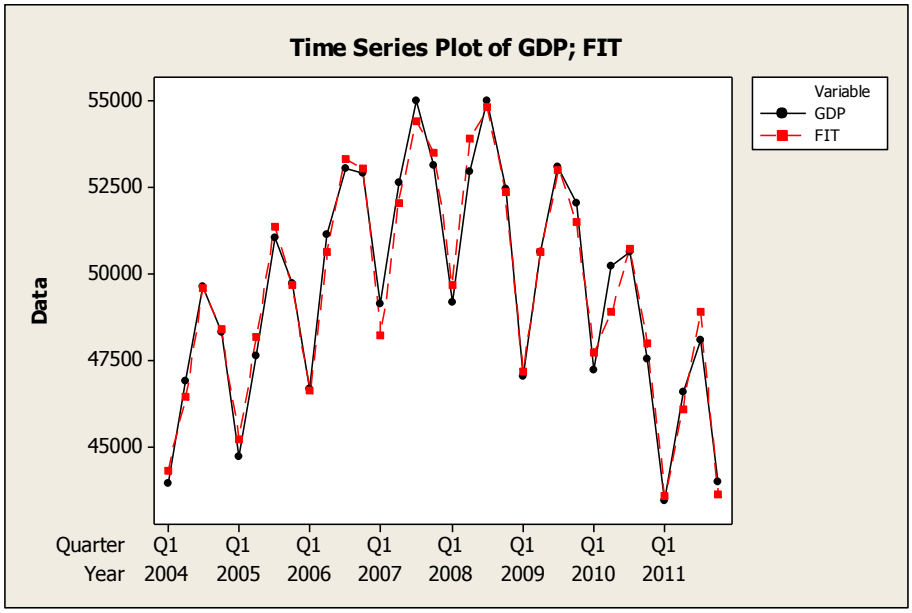


Fig. 1. GDP actual values (black curve) and the FIT (red curve) from Equation 2.

Particularly for every given predictor the following technique is used, while the results are given in Table 4:

INPR: ARIMA model (0,1,1)(1,1,0) with a seasonal period of 4.

RET: For this predictor it was not possible to find a satisfactory ARIMA model. The residuals were in every chosen model strongly autocorrelated and partial autocorrelated, even if the models (coefficients) were statistically well satisfied. Because the time series show a seasonal pattern, time series decomposition has been used. The trend is quadratic and CI is calculated with an ARIMA(0,0,2) model.

CON: ARIMA model (1,1,0)(0,1,0) for Ln(CON) with seasonal period of 4.

UNR: ARIMA model (0,2,1)(0,1,1) with seasonal period of 4.

TI: ARIMA model (2,1,0)(0,1,0) with seasonal period of 4.

INV, DIFF: For these two predictors we have not forecasted any value because they depend on governmental decisions and the memorandum's implementation and not on the dynamic of Greek economy. We use these two predictors as parameters. Particularly for the predictor INV an ARIMA model (0,0,0)(1,0,0) with a seasonal period of 4 is used, in order to complete the missing values for the third and fourth quarter of 2011.

Table 6. Predicted values for the independent variables and the resulting GDP quarterly and annually.

VAR	Q1	Q2	Q3	Q4	SUM 2011
GDP	39.610,38	41.738,39	43.757,14	39.183,15	164.289,06
INPR	72,81	75,30	74,68	66,12	
INV	677	1454	1910,33	2597,21	
RET	87,89	82,98	81,98	98,36	
CON	32,57	36,10	47,34	41,87	
UNR	23,31	24,65	26,89	30,33	
TI	43,22	91,54	148,34	45,44	
DIFF	5.007,00	6.757,00	5.373,00	2.557,00	

The predicted recession is -9,77% taken in to account the values for INV and DIFF from the year 2010. If we halve the values of INV and DIFF the recession reaches a value of -11,10%, also an increase of about 1,33%.

Using Equation 1 with the same values as in Table 6 the recession reaches a value of -9,46% in good agreement with the results of Equation 2. For values outside from the initial intervals the prediction interval becomes wider. In order to examine the influence of each independent variable on GDP, a Tornado Diagram has been constructed. A difference to the values of order of +- 10% has been assumed. As basis we take the values of the second quarter of 2011. RET variable is the most influential, followed by INPR, TI and UNR. The other variables can play a significant role, only if their changes are very large.

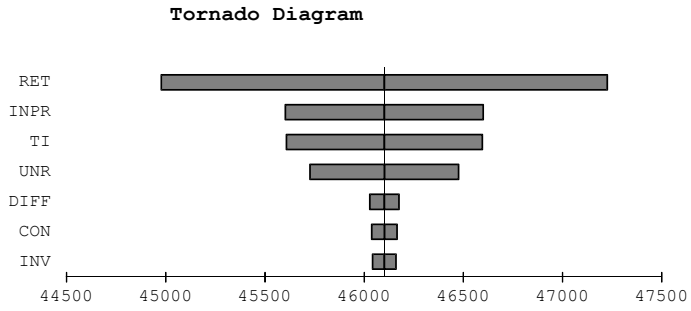


Fig. 2. Sensitivity analysis of the independent variables in a Tornado Diagram

From Fig. 2, it can be understood, why the recession in Greece is very high. All the taken measures in Greece, high taxes on medium to low income and no income citizens - unemployed, have as result the sharp drop of the retail trade. Consequence

of that is the increase of unemployment rate and farther decrease of industrial production and of internal tourism. Taking also into account the difficult international economic conditions and the introvert function of the Greek economy, Greece is led to high recession. When the Greek economy reaches a stability point and which is the stability point, is to find out.

3 Conclusions

Three research hypotheses were examined in order to ascertain the links between culture type and market orientation, culture type and performance, as well as the links between culture type and telecommunication company age and size. Telecommunication companies giving emphasis to their external environment (market and adhocracy oriented cultures) tend to have higher degree of market orientation. In other words, there are indications that clan and hierarchy cultures do not match the demands of their competitive environment.

The results indicate that telecommunication companies are heavily affected by factors such as age and size, while firms' age, the results are consistent with organizational life cycle theories where it is proposed that more hierarchical structures emerge as organizations grow and age, since the growing firms might develop more complex management systems [15]. The Greek telecommunication companies have cultural types which are mainly "control – oriented", which confirms Hofstede's [16] results about a high degree of uncertainty avoidance in this specific national business environment. The type of organizational culture affects the overall performance in terms of profitability and growth and affects companies' viability and future expansion, while cultural type is heavily connected with market orientation [17].

Growing competition, reduced market shares and profitability can create a tough business environment. In such an environment market oriented culture can become a non – imitable characteristic, capable to ensure corporate viability, as it is directly related with profitability and growth [18]. The emerging findings can provide executives with new market opportunities to be considered during their long – term planning determination.

As far as GDP's forecast affects, a series of empirical tests have been conducted in order to evaluate various economic and financial variables before determine which ones play a significant role to GDP's formation. The results indicate a GDP formatted mainly by industrial production, public investment, retail trade, construction, unemployment (as a factor determining the product's demand), tourism and finally, general governmental expenses minus total governmental income [19]. The results indicate the absence of variables such as private investments, exports and consumers' expectations, which are fundamental for an economic renaissance.

The differences between the two proposed equations permit a comparison between 2010 and 2011 in order to understand the Greek economy's depreciation. Equations enrich our understanding about each variable's influence on GDP (positive or negative) and the progress conducted during these two years. Six out of seven variables have either a stronger negative or a smaller positive effect on GDP from

2010 to 2011. Only public investments tend to lead to a greater GDP, but their total effect is rather small as a result of governmental expenses' cutting.

References

[1] R. E. Quinn and K. S. Cameron, *Diagnosing and Changing Organizational Culture*. USA: Addison – Wesley Publishing, 1999.

[2] J. Kurtinaitiene, "Marketing orientation in the European Union mobile telecommunication market," *Marketing Intelligence & Planning*, vol. 23, pp. 104-113, 2005.

[3] J. C. Nunnally, *Psychometric theory* New York: McGraw-Hill, 1978.

[4] A. Papadimitriou and A. Kargas, "The relationship between organizational culture and market orientation in the Greek telecommunication companies," *NETNOMICS*, pp. 1-23, 2012.

[5] C. Carrier, "Intrapreneurship in Large Firms and SMEs: A Comparative Study," *International Small Business Journal International Small Business Journal*, vol. 12, pp. 54-61, 1994.

[6] S. Laforet, "Effects of size, market and strategic orientation on innovation in non-high-tech manufacturing SMEs," *European Journal of Marketing European Journal of Marketing*, vol. 43, pp. 188-212, 2009.

[7] J. E. Hanke and D. Wichern, *Business Forecasting, 9/E*: Prentice Hall, 2009.

[8] G. E. P. Box and G. M. Jenkins, *Time series analysis : forecasting and control*. San Francisco: Holden-Day, 1976.

[9] H. Lótkepohl, *Introduction to multiple time series analysis*. Berlin; New York: Springer-Verlag, 1991.

[10] C. A. Sims, *Macroeconomics and reality*. Evanston, Ill: Econometric Society, 1980.

[11] G. Christou, *An Introduction to Econometrics, 2nd edt*. Athens: Gutenberg, 2004.

[12] KEPE, "Greek Economic Outlook," vol. 15: Centre of Planning and Economic Research, 2012.

[13] KEPE, "Greek Economic Outlook," vol. 18: Centre of Planning and Economic Research, 2012.

[14] J. H. Stock and M. W. Watson, "A Comparison of Linear and Non Linear Univariate models for Forecasting Macroeconomic Time Series," National Bureau of Economic Research, 1998.

[15] A. Kargas and A. Papadimitriou, "The relationship between organizational culture and market orientation in the Greek telecommunication companies," *Netnomics*, vol. DOI: 10.1007/s11066-012-9066-0, 2012.

[16] G. H. Hofstede, *Culture's consequences : international differences in work-related values*. Beverly Hills, Calif.: Sage Publications, 1980.

[17] D. A. Kargas and A. D. Pardalis, "Organizational culture and its relationship with financial statements: the Greek case," *European Journal of Management*, vol. 9, pp. 167-174, 2009.

[18] S. F. Slater and J. C. Narver, "The Positive Effect of a Market Orientation on Business Profitability: A Balanced Replication," *Journal of business research.*, vol. 48, pp. 69-73, 2000.

[19] M. Kiriakidis and A. Kargas, "Greek GDP Forecast Estimates," *Applied Economics Letters*, vol. 20, pp. 767-772, 2013.

Information dissemination and consumption in competitive networking urban environments

E. Kokolaki*

Department of Informatics and Telecommunications,
National & Kapodistrian University of Athens
Ilissia, 157 84 Athens, Greece
evako@di.uoa.gr

Abstract. The focus of this thesis lies on demonstrating, investigating and understanding decision making in human-driven information and communication systems within autonomous networking urban environments and competitive contexts. Indeed, the thesis examines modern networks that integrate mobile communication devices with online social applications and different types of pervasive sensor platforms and hence, foster unprecedented amounts of information. When shared, this information can enrich people's awareness about and enable more efficient management of a broad range of resources, ranging from natural goods such as water and electricity, to human artefacts such as urban space and transportation networks. Especially in environments where users' welfare is better satisfied by the same finite set of resources, it is important to understand how the presence of competition shapes decisions and behaviors regarding the information dissemination and building of collective awareness, on the one hand, and the way collective awareness is exploited under different assumptions about the rationality levels of decision-makers, on the other hand. We investigate these questions by exploiting insights and results from different disciplines ranging from Communication Networks and Decision Theory to Behavioral Economics and Cognitive Science. Our results provide theoretical support for the practical management of limited-capacity resources since they challenge the need for more elaborate information mechanisms. They also reveal useful insights to the dynamics and benefits emerging from human behavior in situations that expose "tragedy of commons" effects.

1 Introduction

The tremendous increase of urbanization necessitates the efficient and environmentally sustainable management of various urban processes and operations. Recent advances in wireless networking and sensing technologies can address this need by enabling efficient monitoring mechanisms for these processes and higher flexibility to control them, thus paving the way for the so-called *Smart Cities*. With the dawn of Smart Cities, the emerging networking environment has dramatically changed the role of end users and resulted in unprecedented rates of information generation and diffusion.

* Dissertation Advisor: Ioannis Stavrakakis, Professor

This information can be intelligently controlled by platforms that collectively enrich people's awareness about their environment, whether this is the natural environment or the physical space they move in while working, driving, or entertaining themselves. In parallel, this knowledge promotes new forms of participatory processes and approaches to managing the resources of their environment, which can range from natural goods such as water and electricity, to human artefacts such as urban space and transportation infrastructure. Besides possibly generating information by themselves via the sensing devices they might be equipped with, the networked entities are also typically involved in disseminating this information widely, contributing to building collective awareness. Furthermore, these same entities may actually exploit this awareness of their environment to meet own needs or achieve certain individual objectives. That is, these entities are involved in the *dissemination* and *consumption* of the information.

If the disseminated information concerns the availability of some limited resources or service, then competition naturally emerges among entities desiring to use such resources. In such environments, it is important to understand how the presence of competition shapes decisions taken by these entities regarding (a) the way these entities participate in disseminating information and creating collective awareness and (b) the way collective awareness is exploited if at all. The first of these very general and fundamental questions amounts to deciding whether a networked entity will deviate from the expected behavior (*misbehave*) by hiding or falsifying resource/service availability information, aiming at reducing the competition to its advantage. The second, amounts to deciding whether a networked entity will *compete or not compete* for some limited resources.

In this thesis, we study scenarios where some finite resource is of interest to a population of distributed users with variable perceptions about the resource supply and demand for it. The high-level question we address is *how efficiently the competition about the resources is resolved under different assumptions about the way the users make their decisions*. We devise analytical and simulation models that describe the decision-making process of users concerning the dissemination and consumption of information, when faced with multiple choices. We instantiate this context in a concrete case that we can study systematically, namely an urban environment in which parking space is the resource of interest to the users-drivers and whose availability is disseminated or becomes accessible to some extent. With this information, drivers can make more informed search for parking, while municipal authorities can address more efficiently the challenge to manage the available parking space and reduce the vehicle volumes that cruise in search of it, in order to alleviate not only traffic congestion but also the related environmental burden.

2 Outline of the thesis

In this section we outline the contents of the thesis in association with the related publications.

In the introductory part, we describe fundamental concepts and principles in networking solutions for the upcoming smart city environments and present socio-tech issues, trends and challenges that arise in various application paradigms that have been developed through these networks and serve as case-studies in our research.

The thesis continues with the study of the effectiveness and side-issues of information within competitive settings. In [4] and [13], we explore how the discovery of service can be facilitated or not by utilizing service location information that is opportunistically disseminated primarily by the consumers of the service themselves. We apply our study to the real-world case of parking service in busy city areas which has attracted the interest of the research community and the private sector in the context of the so-called "Smart City" initiative. As the vehicles drive around the area, they opportunistically collect and share with each other information on the location and status of each parking spot they encounter. The parking space scenario serves as an example of opportunistic networking environments where the user-nodes can collectively gain from the sincere exchange of (parking availability) information (*i.e.*, cooperation), yet each one of them can only gain if certain information is hidden from others (potential competitors); thus, an environment, where the processes of information dissemination (benefiting service discovery) and competition (reducing the service delivery prospects) are coupled and counter-acting. This opportunistically-assisted search is compared against the "blind" non-assisted search and a centralized approach, where the allocation of parking spots is managed by a central server holding global knowledge about the parking space availability. This comparative study concludes with the observation that the availability of information is not always better than the lack of it in competitive environments, as the sharing of information assists nodes by increasing their knowledge about parking space availability but, at the same time, synchronizes nodes' parking choices. This synchronization in turn increases the effective competition and, ultimately, the congestion penalties experienced (*e.g.*, long car cruising when searching for cheap on-street parking spots in busy urban environments).

Being aware of the competition, the nodes are motivated to defer from sharing information or deliberately falsify information to divert others away from a particular area of own interest. In [6] and [12], we implement those facets of misbehaviors in the opportunistically-assisted parking search. We show that as long as the portion of misbehaving nodes is not very high, the overall performance does not deteriorate significantly, nor does the misbehaving node enjoy any notable performance improvement. This observation suggests that the spatial-temporal-interest diversity in large-scale distributed settings and the dynamicity of the environment, which may render falsified data correct or lack of outdated data advantageous, might confer robustness against misbehaviors.

In the sequel, we investigate how the competition awareness affects the decision to compete or not for some limited-capacity resource set. In essence, we are concerned with the comparison of the decision-making under full against bounded rationality conditions. Fully rational users possess all the information they need to reach decisions and, most importantly, are capable of exploiting all information they have at hand. The impact of perfect rationality is investigated in [8] by considering an environment in which the parking space is the resource of interest to the users-drivers and whose availability is disseminated or becomes accessible to some extent. Drivers decide whether to go for the inexpensive but limited on-street public parking spots or the expensive yet over-dimensioned parking lots, incurring an additional cruising cost when they decide for on-street parking spots but fail to actually acquire one. The drivers are viewed as strategic agents who make rational decisions while attempting to minimize the cost of the acquired parking spots. We take a game-theoretic approach and analyze the unco-

ordinated parking space allocation process as *resource selection game* instances. We derive their equilibria and quantify their (in)efficiency with the related *Price of Anarchy* values. In [11] we propose auction-based systems for realizing centralized parking allocation schemes, whereby perfectly informed drivers bid for public parking space and a central authority coordinates the parking assignments and payments to alleviate congestion phenomena. This market-based parking spot allocation is compared against the conventional uncoordinated parking search practice with fixed parking service cost. In line with intuition, the auctioning system increases the revenue of the public parking operator exploiting the drivers' differentiated interest in parking. Less intuitively, the auction-based mechanism does not necessarily induce higher cost for the drivers: by avoiding the uncoordinated search and thus, eliminating the congestion effects, it turns out to be a preferable option for both the operator and the drivers under various combinations of parking demand and pricing policies.

In [5], we relax the assumption of perfect information and study two game variants under incomplete demand information, where the agents either share common probabilistic information about the overall resource demand or are totally uncertain about it. In this case, the game solutions are derived in terms of Bayesian Nash equilibria. Essentially, Game Theory and the Nash equilibrium concept capture the agents' best responses in terms of expected utility maximization. Nevertheless, several experimental data have shown over time the limitations of the Expected Utility Theory framework to consistently explain the way human decisions are made. At the same time, they have revealed cognitive biases in the way people assess the alternatives they are presented with. Thus, we exploit insights from Behavioral Economics and Cognitive Psychology (Prospect Theory, Quantal Response and Rosenthal equilibria, heuristic reasoning) to model agents of bounded rationality who cannot exploit all the available information due to time restrictions and computational limitations [9]. We derive the operational states in which the competing influences are balanced (*i.e.*, equilibria) and compare them against the Nash equilibria that emerge under full rationality and the optimum resource assignment that could be determined by a centralized entity. Although these decision-making models are shown to predict and accommodate people's answers in various experimental data sets, they cannot describe the processes (cognitive, neural, or hormonal) underlying people's decisions. Yet, the efficient and environmentally sustainable management of various urban processes calls for novel solutions that account for behavioral decision-making in a transparent way that reflects the internal reasoning mechanisms. Indeed, transportation engineers need to be able to understand how drivers decide their route to effectively address the plethora of challenges for alleviating the congestion phenomena in city areas. In [7], we model drivers' decision-making with respect to the parking space search, which has been regarded as one of the major causes of traffic congestion. We view the parking search as an instance of *sequential search problems* and present a game-theoretic investigation of the efficiency of heuristic parking search strategies to locate available parking spot at minimum walking and driving overhead. The analytical study concludes by drawing similarities between the parking game and well-known archetypal games that Game Theory examines.

In the last part of the thesis, we seek to experimentally study some fundamental properties of vehicular social applications that have been deployed to assist in the parking search process. In [10], the awareness and incentive mechanisms that are commonly

incorporated in different instances of social parking applications are modelled and simulation scenarios are considered to explore particular aspects of these applications. It is shown that application users experience improved performance due to the increased efficiency they generate in the parking search process, without (substantially) degrading the performance of non-users. This is extremely important since applications managing common (public) goods should not provide benefits to their users by penalizing or almost excluding non-users. The incentive mechanisms are effective in the sense that they do provide preferential treatment to those fully cooperating but they induce rich-club phenomena and difficulties to newcomers. Interestingly, those problems, that may be a concern for all applications managing common (public) goods, seem to be alleviated by free-riding phenomena and dynamic behaviors.

3 The resource selection environment

In this section we define the critical parameters for the resource selection environment, namely, a fairly autonomic networking environment, where each user runs a service resource selection task and seeks to maximize his benefit, driven by self-oriented interests and biases. In this setting, N agents are called to decide between two alternative sets of resources. The first set consists of R low-cost resources while the second one is unlimited but with more expensive items. When the amount of the low-cost resources is large and the interested user population is small, users can readily opt for using it. When, however, the low-cost resources cannot satisfy the demand, an inherent competition emerges that should be factored by users in their decision to opt for accessing these resources or not. Those who manage to use the limited and low-cost resources pay $c_{l,s}$ cost units, whereas those heading directly for the unlimited, but more expensive option pay $c_u = \beta \cdot c_{l,s}, \beta > 1$, cost units. However, agents that first decide to compete for the low-cost resources but fail to acquire one suffering the results of congestion, pay $c_{l,f} = \gamma \cdot c_{l,s}, \gamma > \beta$ cost units. The excess penalty cost $\delta \cdot c_{l,s}$, with $\delta = \gamma - \beta > 0$, captures the impact of congestion phenomena that appear in various ICT sectors when distributed and uncoordinated high volume demand appears for some limited service. Examples include congestion phenomena that emerge on a road that is advertised as the best alternative to a blocked main road due to an accident, the limited on-street public parking space in urban environments or an advertised low-cost wireless access point.

The deployment of advanced (wireless) networking technologies has enabled new services and smart solutions to congestion problems that stem from the blind uncoordinated search for limited resources. However, the efficiency of these systems ultimately depends not just on the quality of the information about resources they can provide to the agents but also on the way the provided information is used by the agents. Therefore, information may be precise and complete or imperfect and limited; whereas the agents may exhibit different levels of rationality in the way they process the provided information and determine their actions.

4 Fully rational decision-making

In the ideal reference model of fully rational decision-making, the decision-maker is a software engine that in the absence of central coordination, acts as rational strategic

agent that explicitly considers the presence of identical counter-actors to make rational, yet selfish decisions aiming at minimizing the cost of the acquired resource. In this case, the main assumption is that users can possess all relevant information, analyze all possible combinations of actions he and the other users can take, assess the cost/gains of each possible outcome, and strategically make the choice that minimizes their own cost. It is notable that provision of sufficient local content for fully rational decision-making is likely not to be cost effective in terms of storage/networking resources and control mechanisms.

4.1 Formulation

The intuitive tendency to head for the low-cost resources, combined with their scarcity in the considered environments, give rise to congestion effects and highlight the game-theoretic dynamics behind the resource selection task [1]. In [5] we have analyzed this task in the context of parking search application. In particular, in center areas of big cities, drivers are often faced with a decision as to whether to compete for the low-cost but scarce on-street public parking space or directly head for the typically over-dimensioned but more expensive parking lots. In the first case, they run the risk of failing to get a spot and having to *a posteriori* take the more expensive alternative, this time suffering the additional *cruising* cost in terms of time, fuel consumption (and stress) of the failed attempt. In general, drivers might make their decisions drawing on information of variable accuracy about the parking demand, capacity and the applied pricing schemes on the parking facilities, that parking assistance systems collect and broadcast. Under the assumption of fully rationality, an assistance service announces information of perfect accuracy about the demand (number of users interested in the parking resources), supply (number of limited, low-cost, on-street parking resources) and pricing policy on the parking resources. The drivers act as rational and strategic selfish agents that try to minimize the cost the actual humans/drivers pay for the acquired parking space. In fact, we consider automatic software agent implementations rather than human decision-makers yet, the actual human/driver undertakes the action with the assumption that he fully complies with the machines' suggestions.

We derive the drivers' behaviors at the equilibrium states of this strategic game and compare the costs paid at the equilibria against those induced by the ideal centralized system that optimally assigns the low-cost resources to minimize the social cost. We quantify the (in)efficiency of the uncoordinated resource selection using the Price of Anarchy (PoA) metric, computed as the ratio of the worst-case equilibrium cost over optimal cost. The analytical investigation shows that PoA deviates from one, implying that, at equilibrium, drivers tend to over-compete for the on-street parking space, giving rise to redundant cruising cost. In particular, for parking demand exceeding the supply ($N > R$), the number of competing drivers in the equilibrium state $N_{l,eq} = \min(N, N_0)$, with $N_0 = \frac{R(\gamma-1)}{\delta}$, exceeds the optimal number R that would compete for and succeed in getting an on-street parking spot in the ideal scenario. These congestion phenomena can be alleviated by properly manipulating the price differentials between the two types of resources. Notably, our results are in line with earlier findings about *congestion pricing* (*i.e.*, imposition of a usage fee on a limited-capacity resource set during times of high demand), in a work with different scope and modelling approach [14]. The results of this study will serve as a benchmark for assessing

the impact of different rationality levels and cognitive biases on the efficiency of the resource selection process.

5 Bayesian and pre-Bayesian models

In the resource selection context, perfectly accurate information about the resource demand is hard to obtain within a dynamic and complex environment. For instance, when the agents do not possess perfect information about the resource availability, one could imagine that resource information will be disseminated in the network following some dynamics resembling epidemics. In the presence of an infrastructure-based information and sensing mechanism, the resource operator may provide the competing agents with different levels of information about the demand for resources; for example, historical statistical data about the utilization of the low-cost resources. Thus, in this case, the information is impaired in accuracy since it contains only some estimates on the parameters of the environment.

5.1 Formulation

This type of bounded rationality where agents have only knowledge constraints, while they satisfy all other criteria of full rationality, *i.e.*, no computational or time constraints deteriorate the quality of their decisions, can be accommodated in Bayesian and pre-Bayesian models that devise prescriptions of the classical Game Theory. In the Bayesian model of the game, the agents determine their actions on the basis of private information, *i.e.*, their types. In the resource selection problem, the type can operate as a binary variable indicating whether an agent is in search of resources (active player). Every agent knows his own type, yet he ignores the real state at a particular moment in time, as expressed by the types of the other players. The agents draw on common prior probabilistic information about the activity of agents (*i.e.*, the probability for an agent to be active, p_{act}, namely, interested in resources) to derive estimates about the expected cost of their actions. Thus, now, the agents try to minimize the expected cost, instead of the pure cost that comes with a strategy, and play/act accordingly. In the resulting Bayesian Nash equilibrium states, the agents perform their best-response actions and no agent can further lower his expected cost by unilaterally changing his strategy.

In the worst-case scenario (strictly incomplete information/full uncertainty), the agents may avail some knowledge about the upper limit of the potential competitors for the resources, yet their actual number is not known, not even probabilistically. In this case, the resulting agents' interactions can be modelled as an instance of pre-Bayesian games and the game dynamics are discussed in terms of safety-level equilibria; namely, operational states whereby every player minimizes over his strategy set the worst-case (maximum) expected cost he may suffer over all possible types and actions of his competitors.

In [5], we extend the game formulation for the full rationality case and analyze Bayesian and pre-Bayesian models that accommodate two expressions of uncertainty, where drivers either share common probabilistic information about the overall parking demand or are totally uncertain about it. Interestingly enough, we show less-is-more phenomena under uncertainty, whereby more information does not necessarily improve the efficiency of service delivery but, even worse, may hamstring users' efforts to minimize the cost incurred by them. In fact, the safety-level mixed-action equilibrium of

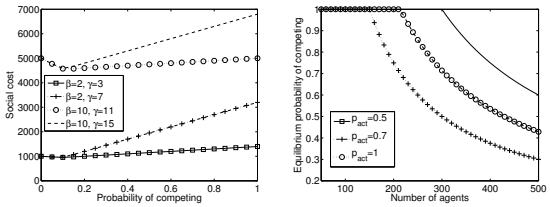


Fig. 1. Social cost for $N = 500$ agents competing for $R = 50$ resources with $c_{l.s} = 1$ (left). Probability of competing in the equilibrium for $R = 50$, $c_{l.s} = 1$, $\beta = 5$, $\delta = 2$ (right).

the pre-Bayesian game corresponds to the mixed-action equilibrium of the strategic game. In the strategic game, the social cost conditionally increases with the equilibrium competing probability, on the one hand, and the equilibrium competing probability decreases with the number of agents, on the other hand (Ref. Fig. 1). Therefore, at the safety-level equilibrium, the agents end up competing with a lower probability than that corresponding to the game they actually play and hence, they may end up paying less than they would if they knew deterministically the competition they face.

6 Behavioral desision theory

Experimental data suggest that human decisions reflect certain limitations and exhibit biases in comparing the expected utilities that come with different alternatives. To accommodate the empirical findings, researchers from economics, engineering, sociology, operations research and cognitive psychology, have tried either to expand/adapt the Expected Utility framework or completely depart from it (and its expressions as embodied in the Nash equilibrium concept) and devise alternative theories as to how decision alternatives are assessed and decisions are eventually taken. The study of the decisions people make is, indeed, the focus of the interdisciplinary behavioral decision theory which has contributed to a re-evaluation of what human decision-making requires.

6.1 Formulation

Cumulative Prospect Theory Tversky and Kahneman in [19] proposed the Cumulative Prospect Theory (CPT) framework to explain, among others, why people buy lottery tickets and insurance policies at the same time, and the fourfold pattern of risk attitude, namely, people's tendency to be risk-averse for alternatives that bring gains and risk-prone for alternatives that cost losses, when these alternatives occur with high probability; and the opposite risk attitudes for alternatives of low probabilities. According to CPT the alternatives are now termed prospects and lead to a number of outcomes that are obtained with a probability. The prospects are valued by an expression of weighted sum of values that resembles the expression of EUT, only now both components of the EUT (*i.e.*, individual outcomes and corresponding probabilities) are modified. However, users are still maximizers, *i.e.*, they try to maximize the expected utilities of their prospects. In [19], the authors propose concrete functions to transform objective probabilities and outcomes with shapes that are consistent with experimental evidence on risk preferences.

In [9], we apply the CPT model to the resource selection problem, where the decisions are made on two alternatives - prospects consisting only of negative outcomes/costs,

and present a comparative study between the per-user costs under the Nash equilibrium, the CPT equilibrium and the optimal resource assignment that could be determined by a centralized entity. When the agents have the opportunity to experience a marginally or significantly lower charging cost by using the low-cost resource set, at low or high risk, respectively, their biased risk-seeking behavior turns to be full rational, and thus, minimizes the expected cost over others' preferences. On the contrary, in the face of a highly risky option reflected in significant extra penalty cost for those who fail in the competition, the risk attitude under the two types of rationality starts to differ; that is, the CPT leads to a more risk-prone behavior when compared to the Nash equilibrium strategy. This is in line with the theory for losses: an agent may decrease the prospect cost by switching his decision from the certain more expensive resource set to the risky low-cost one. The comparison between the Nash and CPT equilibria against the optimal resource allocation shows that both the fully rational and the biased practice are more risk-seeking than they should be, increasing the actual per-user cost (or equivalently, the social cost) over the optimal levels. As a result, being prone to biased risk-seeking behaviors cannot score better than acting fully rationally.

Rosenthal and Quantal Response Equilibria Both casual empiricism as well as experimental work suggested systematic failure of standard Nash equilibrium predictions to track laboratory data, even in some of the simplest two-person games (*e.g.*, generalized matching pennies games). Triggered by this kind of observations, probabilistic choice models have been used to incorporate stochastic elements in the analysis of individual decisions and hence, represent unobserved and omitted elements, estimation/computational errors, individual's mood, perceptual variations or cognitive biases. Rosenthal in [16] and, later, McKelvey and Palfrey in [15], propose alternative solution concepts to the Nash equilibrium in an effort to model games with noisy players. Rosenthal argued that "the difference in probabilities with which two actions are played is proportional to the difference of the corresponding expected gains (costs)". In a similar view of people's rationality, McKelvey and Palfrey explained people's inability to play always the strategy that maximizes (minimizes) the expected utility (cost) by introducing some randomness into the decision-making process. The underlying idea in the proposed Quantal Response equilibrium is that "individuals are more likely to select better choices than worse choices, but do not necessarily succeed in selecting the very best choice". In both equilibrium concepts the rationality of agents is quantified by a degree of freedom which measures the capacity to assess the difference in the utilities between two outcomes. Thus, the models' solutions converge to the Nash equilibria as this rationality parameter goes to infinity.

Let $c(l,p) = \sum_{n=0}^{N-1} g_l(n+1)B(n;N-1,p_l)$, where $g_l(k) = min(1,R/k)c_{l,s} + (1-min(1,R/k))c_{l,f}$ and $B(n;N;p)$ is the Binomial probability distribution, and $c(u,p) = c_u$ denote the expected costs for choosing "low-cost/limited-capacity resource set" and "expensive/unlimited resource set", respectively, when all other agents play the mixed-action $p = (p_l, p_u)$. The Rosenthal equilibrium strategy $p^{RE} = (p_l^{RE}, p_u^{RE})$, $p_u^{RE} = 1 - p_l^{RE}$ and Quantal Response equilibrium strategy $p^{QRE} = (p_l^{QRE}, p_u^{QRE})$, $p_u^{QRE} = 1 - p_l^{QRE}$ are given as fixed-point solutions of equations $p_l^{RE} - p_u^{RE} = -t(c(l,p^{RE}) - c(u,p^{RE}))$ and $p_l^{QRE} = \frac{e^{-tc(l,p^{QRE})}}{e^{-tc(l,p^{QRE})}+e^{-tc(u,p^{QRE})}}$, respectively.

In [9], we compare the fully rational strategies against the two alternative types of equilibrium strategies and the resulting per-user costs in the context of the resource selection task. The implementation of these expressions of bounded rationality increases randomness into agents' choices and hence, draws choice probabilities towards 0.5. Second, the more different the - expected - costs of the two options are, the less the Rosenthal and Quantal Response equilibrium differ from the Nash one, since the identification of the best action becomes easier. Thus, we notice almost no or limited difference when the risk to compete for a very small benefit is high due to the significant extra penalty cost or due to the high demand for the resources. The same reason underlies the differences between the Rosenthal and the Quantal Response equilibrium. Essentially, the three types of equilibrium form a three-level hierarchy with respect to their capacity to identify the less costly resource option, with the Quantal Response equilibrium at the bottom level and the Nash one at the top level. Finally, contrary to the risk attitude as expressed in CPT, the inaccurate but frugal computation of the best action as modelled in these equilibrium concepts decreases the competing probability under low to medium demand and hence, the per-user cost is drawn to near-optimal levels.

Heuristic decision-making In a more radical approach, models that rely on heuristic rules reflect better Simon's early arguments in [17] that humans are satisficers rather than maximizers.

Heuristic decision rule: In an effort to get the satisficing notion in our competitive resource selection setting, we came up with a simple kind of heuristic rule arguing that instead of computing/comparing the expected costs of choices, individuals estimate the probability to get one of the "popular" resources (based on beliefs about the activity of others) and play according to this. In essence, as common sense suggests, one appears overconfident under low demand for the scarce low-cost resources and underconfident otherwise. Interestingly, applying this trivial decision rule in the resource selection problem leads to near-optimal results. Unlike CPT or the alternative equilibrium solutions, it does not take into account the charging costs. Yet, this reasoning mode expresses a pessimistic attitude that takes for granted the failure in a possible competition with competitors that outnumber the resources. As a result, it implicitly seeks to avoid the tragedy of common effects and eventually, yields a socially beneficial solution.

Cognitive heuristics: Cognitive science suggests that people draw inferences (*i.e.*, predict probabilities of an uncertain event, assess the relevance or value of incoming information *etc.*), exploiting heuristic principles. The cognitive heuristics could be defined as fast, frugal, adaptive strategies that allow humans (organisms, in general) to reduce complex decision tasks of predicting, assessing, computing to simpler reasoning processes. In the salient of heuristic-based decision theory, notions such as recognition, priority, availability, fluency, familiarity, accessibility, representativeness and adjustment - and - anchoring stand out.

The various analytical models of bounded rationality that are presented in previous paragraphs, depart from the norms of classical rationality as expressed in the Expected Utility Theory framework. However, people do not seem to perform the calculations that these models require, at least not under all conditions and especially in situations where there is pressure to be "rational" (*e.g.*, route and parking spot selection). In other words, a criticism against these models is that they no longer aim at describing the

processes (cognitive, neural, or hormonal) underlying a decision but just at predicting people's final choices for a large chunk of choice problems. Furthermore, they give no insight as to how should the corresponding models be parametrized each time.

Models that rely on cognitive heuristics originate from the cognitive psychology domain and specify the underlying cognitive processes while they make quantitative predictions. In connection to this, in [7] we analytically investigate drivers' decision-making concerning parking spot selection in city environments drawing on results from experiments with driving emulators [3]. In particular, we address the parking search problem within the framework of sequential search/optimal stopping problems (*e.g.*, mate choice, secretary problem), whereby people devise simple heuristic strategies (rules of thumb) to overcome the complexity of finding the optimal decision. Interestingly, albeit the human cognitive limitations, time constraints and lack of full information in those reasoning contexts, simple rules of thumb can frequently perform as well as more sophisticated search approaches by exploiting the structure of the information in the environment (Ref. *ecological rationality* in [2]). In this investigation, we envisage that drivers use a decision rule based on their distance from the destination, namely the *fixed-distance heuristic*, which ignores all places until the driver reaches a specific distance from the destination and then takes the first vacant one [18]. This instance of heuristics incorporates two fundamental practices in behavioral decision theory: one-at-a-time processing of pieces of information and the use of thresholds. Through a game-theoretic investigation, we show that when the drivers are risk-averse (namely, they prefer walking than driving), the simple fixed-distance heuristic strategy leads to optimal parking spot allocation and hence, minimum social cost.

7 Conclusions

In this thesis, we study networking environments where some finite resource is of interest to a population of distributed users with variable perceptions about the resource supply and demand for it. In such competitive environments, the easier acquisition of environmental information has its negative side, since it synchronizes the perception of different users about the state of resources and, at a second and most important level, their decisions. Consequently, the competition awareness should be factored in the decision (a) to distribute the availability information as expected or misbehave and (b) to go for some limited resources (compete) or not (not compete). The first question has been investigated by considering an urban environment in which the parking space is the resource of interest to the users-drivers and whose availability is disseminated through an opportunistic assistance service. The investigation of the vulnerability of this service to misbehaving nodes that either defer from sharing information or deliberately falsify information, reveals a remarkable resilience as long as the portion of misbehaving nodes is not high and a persistent fate-sharing effect between what misbehaving and cooperative nodes achieve. The second question has been investigated by considering various levels of users' rationality as expressed in the amount of available knowledge and users' computational capacity. We draw on bayesian models to capture the impact of imperfect information and exploit analytical insights from Behavioral Decision Theory to model users with processing limitations. Interestingly, counterintuitive less-is-more effects emerge where more information does not necessarily improve the efficiency of service delivery but, even worse, may hamstring users' efforts to maximize their benefit.

Likewise, very simple heuristic reasoning approaches that are devised to override the complexity of computing the optimal strategy, are shown to yield near-optimal results with respect to the social cost incurred by the user population.

References

1. Ashlagi, I., Monderer, D., Tennenholtz, M.: Resource selection games with unknown number of players. In: Proc. AAMAS. Japan (2006)
2. Goldstein, D.G., Gigerenzer, G.: Models of ecological rationality: The recognition heuristic. Psychological Review 109(1) (2002)
3. Katsikopoulos, K.V.: Advanced guide signs and behavioral decision theory. D. L. Fisher, M. Rizzo, J. K. Caird, and J. D. Lee. Boca Raton, FL: CRC Press (2011)
4. Kokolaki, E., Karaliopoulos, M., Stavrakakis, I.: Value of information exposed: Wireless networking solutions to the parking search problem. In: Eighth International Conference on Wireless On-Demand Network Systems and Services (WONS). Italy (2011)
5. Kokolaki, E., Karaliopoulos, M., Stavrakakis, I.: Leveraging information in parking assistance systems. IEEE Transactions on Vehicular Technology 62(9) (2013)
6. Kokolaki, E., Kollias, G., Papadaki, M., Karaliopoulos, M., Stavrakakis, I.: Opportunistically-assisted parking search: a story of free riders, selfish liars and bona fide mules. In: Proceedings of the 10th International Conference on Wireless On-demand Network Systems and Services (IFIP/IEEE WONS). Banff, Canada (2013)
7. Kokolaki, E., Stavrakakis, I.: Equilibrium analysis in the parking search game with heuristic strategies. In: Second IEEE VTC Workshop on Vehicular Traffic Management for Smart Cities. Seoul, Korea (May 2014)
8. Kokolaki, E., Karaliopoulos, M., Stavrakakis, I.: On the efficiency of information-assisted search for parking space: A game-theoretic approach. In: 7th International Workshop on Self-Organizing Systems (IFIP IWSOS'13). Palma de Mallorca (2013)
9. Kokolaki, E., Karaliopoulos, M., Stavrakakis, I.: On the human-driven decision-making process in competitive environments. In: First Internet Science Conference. Brussels, Belgium (2013)
10. Kokolaki, E., Karaliopoulos, M., Stavrakakis, I.: Parking assisting applications: effectiveness and side-issues in managing public goods. In: 3rd IEEE SASO workshop on Challenges for Achieving Self-Awareness in Autonomic Systems. Philadelphia, USA (2013)
11. Kokolaki, E., Karaliopoulos, M., Stavrakakis, I.: Trading public parking space. In: First IEEE WoWMoM workshop on Smart Vehicles. Sydney, Australia (2014)
12. Kokolaki, E., Karaliopoulos, M., Kollias, G., Papadaki, M., Stavrakakis, I.: Vulnerability of opportunistic parking assistance systems to vehicular node selfishness. Computer Communications 48, 159 – 170 (2014)
13. Kokolaki, E., Karaliopoulos, M., Stavrakakis, I.: Opportunistically assisted parking service discovery: Now it helps, now it does not. Pervasive and Mobile Computing 8(2) (2012)
14. Larson, R.C., Sasanuma, K.: Congestion pricing: A parking queue model. Journal of Industrial and Systems Engineering 4 (2010)
15. McKelvey, R., Palfrey, T.: Quantal response equilibria for normal form games. Games and Economic Behavior pp. 6–38 (1995)
16. Rosenthal, R.: A bounded-rationality approach to the study of noncooperative games. Int. J. Game Theory pp. 273–292 (1989)
17. Simon, H.A.: Rational choice and the structure of the environment. Psychological Review 63(2), 129–138 (1956)
18. Todd., P.M., Gigerenzer, G., the ABC Research Group: Ecological rationality: intelligence in the world. Oxford Univ. Press, New York (2012)
19. Tversky, A., Kahneman, D.: Advances in prospect theory: cumulative representation of uncertainty. Journal of Risk and Uncertainty 5 (1992)

Rate-Optimum Beamforming Transmission in MIMO Rician Fading Channels

Dimitrios E. Kontaxis*

National and Kapodistrian University of Athens
Department of Informatics and telecommunications

Abstract

In this doctoral thesis, the focus is on the capability of MIMO systems to increase channel capacity. The capacity achieved by MIMO systems is closely related to the "channel knowledge" model which is assumed at both ends of the MIMO link. Considering the case of MIMO complex Gaussian ergodic channels, with perfect Channel State Information at the receiver and Channel Distribution Information at the transmitter, the "ergodic capacity" is the maximum average mutual information between transmitter and receiver and is achieved by a unique optimum spatial precoding transmission. For the case of beamforming transmission, the maximum average mutual information is achieved by the "optimum beamformer" and is referred to as "ergodic beamforming capacity". Considering spatially correlated MIMO Rician flat fading channels, there is no closed-form expression for the optimum beamformer. In this case, its calculation is performed numerically and is very complex for real time applications. In this work, it is proven that the aforementioned complex, multi-dimensional, convex constrained optimization problem can be transformed to an 1-D optimization problem, which can be solved very fast using standard 1-D algorithms. This proof was based on geometrical properties, basis transformations and the Karush-Kuhn-Tucker (KKT) conditions. Simulations demonstrate that the proposed 1-D method has significantly lower computational complexity compared to multi-dimensional algorithms and that in some operational environments the ergodic beamforming capacity is very close or equal to the ergodic capacity. Additionally, the 3GPP MIMO channel model is employed in order to study further (via simulations) the performance of the optimum beamformer in practical operational scenarios (urban micro/macro-cellular with/without LOS component and suburban macro-cellular environments). Simulations demonstrate that the optimum beamformer shows high performance in all cases, a fact that

*Doctoral Thesis Advisor: Serafim Karaboyas, Assistant Professor

justifies the significance of the proposed solutions and the contribution of this work.

1 Introduction

1.1 Ergodic capacity

Multiple-Input Multiple-Output (MIMO) systems employ multiple transmit and receive antennas and exploit the fluctuations in the (received) signal level due to multipath propagation (multipath fading) in order to increase spectral efficiency, improve the Quality of Service and coverage and mitigate interference. These benefits are achieved without the expense of additional bandwidth and make MIMO a very attractive and promising option for future mobile communication systems, especially when combined with the benefits of orthogonal frequency-division multiplexing (OFDM). The most important techniques employed by MIMO systems in order to achieve the aforementioned benefits are beamforming, diversity and spatial multiplexing. In this doctoral thesis, the focus is on the capability of MIMO systems to increase spectral efficiency. A MIMO system can achieve much higher channel capacity than a conventional Single-Input Single-Output (SISO) system, and it can be proven that the achieved capacity increases linearly with the number of transmit or receive antenna elements [1]. However, the capacity achieved by MIMO systems is closely related to the channel knowledge model which is assumed at both ends of the link. Assuming perfect channel knowledge, referred to as perfect Channel State Information (CSI), at both ends of the link (transmitter-receiver), the spatial pre-coding transmission scheme that achieves capacity was presented in [1]-[2], and includes transmission along the right singular vectors of the channel matrix combined with "water-filling" for optimum power allocation between the transmit directions. However, perfect CSI at the transmitter is practically unrealistic, mainly due to the inevitable delay in the control channel which is used to feed back the CSI from the receiver or due to the delay in the channel estimation algorithm employed at the transmitter. Instead, it is more realistic and practical to assume that the transmitter has knowledge of the parameters of the MIMO channel distribution, since the channel statistics usually remain invariant in a large time window, (tens to hundreds of times larger than the coherence time). This channel knowledge model is referred to as Transmitter Channel Distribution Information (CDIT) [3]. In a CDIT model the rate-optimum transmission maximizes the average mutual information between transmitter and receiver and the maximum rate achieved in this case is referred to as "ergodic capacity". Considering MIMO complex Gaussian ergodic channels with perfect CSI at the receiver and CDIT, the optimum spatial pre-coding transmission has been addressed in the literature (i.e. methods for the calculation of the optimum transmit covariance matrix have been proposed) for the following channel models [4]-[5]:

a. MIMO Rayleigh flat fading channels. This CDIT model is referred to as Channel Covariance Information (CCI).

b. Spatially uncorrelated MIMO Rician flat fading channels with a unit covariance matrix. This CDIT model is referred to as Channel Mean Information (CMI).

c. Spatially correlated or uncorrelated with a non-unit covariance matrix MIMO Rician flat fading channels. This CDIT model is referred to as combined CMI-CCI model.

1.2 Ergodic beamforming capacity

In MIMO systems, when the transmit covariance matrix is constrained to be rank-1, then all the available power is transmitted along a unique direction with the help of a beamforming vector, as it is shown in Figure 1.

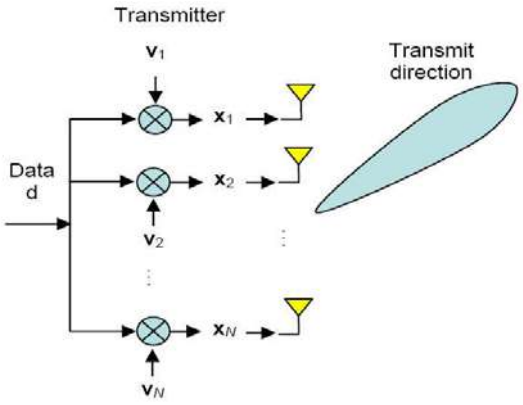


Figure 1: Beamforming transmission. $\mathbf{v} = [\mathbf{v}_1, \mathbf{v}_2, \dots \mathbf{v}_N]$ is the beamforming vector and $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \dots \mathbf{x}_N] = d\mathbf{v}$ is the transmit signal vector.

The beamforming vector that maximizes the average mutual information for this constrained scenario is referred to as "optimum beamformer" and the achieved average mutual information as "ergodic beamforming capacity". There are several reasons why it is important to consider the optimum beamforming transmission in MIMO systems:

a. The complexity of the system and as a consequence the overall cost are significantly reduced.

b. There are operational scenarios (i.e. channels) where the ergodic beamforming capacity is very close to the ergodic capacity, which is achieved by higher rank transmission schemes.

c. The ergodic beamforming capacity does not coincide with the ergodic capacity of the channel, however, this is possible when a specific necessary and sufficient condition is satisfied by the channel distribution. This condition is expressed by a mathematical inequality and is referred to in the literature as the "optimality of beamforming condition" [6].

The solution of the optimum beamforming problem has been addressed extensively in the literature for the CCI and CMI models. For these two cases, closed-form solutions have been derived: the optimum beamformer coincides with the dominant eigenvector of the channel correlation matrix. However, the corresponding solution for the combined CMI-CCI model has received less attention. For this CDIT model, there is no closed-form expression for the optimum beamformer and hence, the solution of the related optimization problem remains very complex for real time applications.

In this work, it is proven that the aforementioned complex, multi-dimensional, convex constrained optimization problem for the combined CMI-CCI model can be transformed to a simple and equivalent 1-D optimization problem, which can be solved very fast using standard 1-D algorithms (gradient-based or direct search methods). This proof was based on geometrical properties of complex vector spaces, basis transformations and the Karush-Kuhn-Tucker (KKT) conditions. Moreover, a special (simpler) solution is proven for MIMO $2 \times M$ and special cases of MIMO $N \times M$ systems (e.g. MIMO systems with rank deficient transmit covariance matrix), where N/M is the number of transmit/receive antenna elements, respectively. The proof of this special case was based on a geometric approach, where the definition of the external product between vectors in high-dimensional complex vector spaces was exploited.

2 Rate-Optimum Beamforming Transmission in MIMO Rician flat fading channels

2.1 MISO systems

We consider a MISO $N \times 1$ flat fading channel with a complex Gaussian distribution $\mathbf{h} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{R})$, with mean $\boldsymbol{\mu} \neq \mathbf{0}$ and covariance $\mathbf{R} \neq \mathbf{I}_N$ (i.e. a spatially correlated or uncorrelated with a non-unit covariance matrix MISO flat fading channel is assumed, with a Rician distribution for the amplitude of the elements of the channel vector $\mathbf{h}$). Assuming perfect CSI at the receiver and CDIT, the rate-optimum beamforming transmission for the channel model under consideration is the solution of an 1-D optimization problem, which is expressed by the following theorem [7]-[8]:

Theorem 1 The optimum beamformer $\mathbf{v}_{\text{opt}}$, for a MISO Rician flat fading channel with N transmit antenna elements ($N \geqslant 2$), mean value $\boldsymbol{\mu}$ ($\boldsymbol{\mu} \in \mathbb{C}^{1\times N}$, $\boldsymbol{\mu} \neq \mathbf{0}$) and transmit covariance matrix $\mathbf{R}$ ($\mathbf{R} \in \mathbb{H}_+^N$, $\mathbf{R} \neq \mathbf{I}_N$), can be calculated from the following 1-D optimization problem:

$$\mathbf{v}_{\text{opt}} = arg \max_{\mathbf{v}\in\mathbf{S}_o} \mathcal{I}_{\text{bf,avg}}(\text{SNR}, \mathbf{v}) \tag{1}$$

$$\mathbf{S}_o = \{\mathbf{v}_\theta \,; \theta \in [0, \phi]\} \tag{2}$$

where:

a. The average mutual information in (1) is expressed as:

$$\mathcal{I}_{\text{bf,avg}}\left(\text{SNR}, \mathbf{v}\right) = \mathcal{E}_{\mathbf{h}}\left[\log_2 \det\left(\mathbf{I}_M + \text{SNR}\mathbf{h}\mathbf{v}^{\dagger}\mathbf{v}\mathbf{h}^{\dagger}\right)\right] \tag{3}$$

b. ϕ is the angle between $\boldsymbol{\mu}$ and the complex conjugate transpose of the dominant eigenvector of the channel transmit covariance matrix $\mathbf{R}$, denoted as $\mathbf{U}_{\bullet 1}^{\dagger}$, i.e.

$$\phi = \cos^{-1}\left(|\mathbf{m}\mathbf{U}_{\bullet 1}|\right) \tag{4}$$

with $\mathbf{m} = \boldsymbol{\mu}/\|\boldsymbol{\mu}\|_2$

c. $\mathbf{v}_\theta$ in (2) is expressed as:

$$\mathbf{v}_\theta = \cos\theta\left[1\ \mathbf{Z}(r_\theta \mathbf{I}_{N-1} - \mathbf{G})^{-1}\right]\mathbf{W}^T\mathbf{U}^{\dagger} \tag{5}$$

where:

i. $\mathbf{U}$ is the eigenvector matrix of $\mathbf{R}$ and $\mathbf{W}$ is a complex $N \times N$ orthonormal matrix with its first column defined as $\mathbf{W}_{\bullet 1} = \mathbf{U}^T\mathbf{m}^T$, whereas the rest of its columns ($\mathbf{W}_{\bullet i}$, $i = 2, ..., N$) are arbitrarily chosen, with the restriction that $\mathbf{W}^{\dagger}\mathbf{W} = \mathbf{I}_N$. Moreover, $\mathbf{G}$ and $\mathbf{Z}$ are defined as:

$$\mathbf{G} = \begin{pmatrix} \mathbf{K}_{22} & \cdots & \mathbf{K}_{2N} \\ \vdots & \ddots & \vdots \\ \mathbf{K}_{N2} & \cdots & \mathbf{K}_{NN} \end{pmatrix} \tag{6}$$

$$\mathbf{Z} = [\mathbf{K}_{12}\ \mathbf{K}_{13}\ \cdots\ \mathbf{K}_{1N}] \tag{7}$$

where $\mathbf{K}_{lm}$ is the l^{th} row and m^{th} column element of matrix $\mathbf{K}$, defined as:

$$\mathbf{K} = \sum_{i=1}^{N} \lambda_i(\mathbf{R})\mathbf{W}_{i\bullet}^T\mathbf{W}_{i\bullet}^* \tag{8}$$

with $\lambda_i(\mathbf{R})$ the i^{th} eigenvalue of $\mathbf{R}$.

ii. r_θ is the maximum real root of the $2(N-1)$-degree polynomial:

$$P(x;\theta) = \cos^2\theta \sum_{i=1}^{N-1} |\mathbf{Z}\mathbf{g}_i|^2 \left[\prod_{\substack{j=1 \\ j\neq i}}^{N-1} (x - \lambda_i(\mathbf{G}))^2\right] - \sin^2\theta \prod_{i=1}^{N-1}(x - \lambda_i(\mathbf{G}))^2 \tag{9}$$

where $\mathbf{g}_i \in \mathbb{C}^{(N-1)\times 1}$ is the i^{th} eigenvector of matrix $\mathbf{G}$.

Theorem 1 implies that the optimum beamformer belongs to a continuous trajectory (the continuity of the trajectory can be mathematically proven) that is defined by the vectors of $\mathbf{S}_o$ (see (2)), which lies on the surface of the unit-radius Euclidean ball, starts from $\mathbf{m}$ (for $\theta = 0$) and ends to $\mathbf{U}_{\bullet 1}^{\dagger}$ (for $\theta = \phi$). This is visualized in Figure 2.

Theorem 2 below ([7]-[8]) provides an alternative geometrically-based approach, especially for MISO systems with $N = 2$ transmit antenna elements.

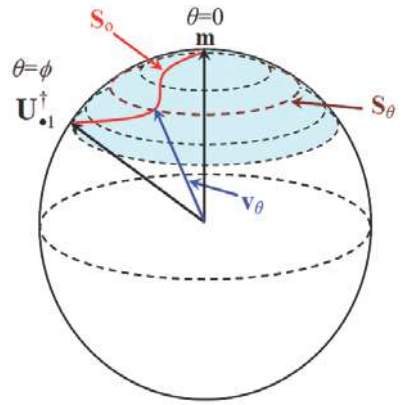


Figure 2: Geometric interpretation of 1-D method (Theorem 1).

Moreover, this theorem is also mathematically valid for the following special cases, with $N > 2$:

a. When $\boldsymbol{\mu}$ is a point in the hyperplane defined by $\mathbf{U}^\dagger_{\bullet 1}$ and $\mathbf{U}^\dagger_{\bullet 2}$ (i.e. the two dominant eigenvectors of $\mathbf{R}$).

b. When the channel covariance matrix has two eigenvalues, $\lambda_1(\mathbf{R})$ and $\lambda_2(\mathbf{R})$ ($\lambda_1(\mathbf{R}) \geqslant \lambda_2(\mathbf{R})$) with algebraic multiplicity one and $N-1$, respectively, or it is rank deficient, with $rank\{\mathbf{R}\} \leqslant 2$.

Theorem 2 For MISO systems with $N = 2$, $\mathbf{v}_\theta$ can be expressed by the following (closed-form) equation:

$$\mathbf{v}_\theta = \cos\theta \frac{\mathbf{U}^\dagger_{\bullet 1}\mathbf{m}^\dagger\mathbf{m}}{\|\mathbf{U}^\dagger_{\bullet 1}\mathbf{m}^\dagger\mathbf{m}\|_2} + \sin\theta \frac{\mathbf{m}^*\left(\mathbf{m}^T\mathbf{U}^\dagger_{\bullet 1} - \mathbf{U}^*_{\bullet 1}\mathbf{m}\right)}{\|\mathbf{m}^*\left(\mathbf{m}^T\mathbf{U}^\dagger_{\bullet 1} - \mathbf{U}^*_{\bullet 1}\mathbf{m}\right)\|_2} \tag{10}$$

Moreover, in the context of this work, it is also proven that in MISO systems the average mutual information for the beamforming scenario can be calculated by an infinite-series, which converges fast (only a few tens of terms are required) to the corresponding value calculated by using Monte Carlo integration with thousands of channel samples [7]-[8]:

$$\mathcal{I}_{\text{bf,avg}}(\theta) = \mathcal{I}_{\text{bf,avg}}(\text{SNR}, \mathbf{v}_\theta) = (ln2)^{-1}\, exp\left(\frac{1 - \text{SNR}\,|\boldsymbol{\mu}\mathbf{v}^\dagger|^2}{\text{SNR}\,\mathbf{v}\mathbf{R}\mathbf{v}^\dagger}\right) \times$$
$$\sum_{n=0}^{\infty}\left[\frac{1}{n!}\left(\frac{|\boldsymbol{\mu}\mathbf{v}^\dagger|^2}{\mathbf{v}\mathbf{R}\mathbf{v}^\dagger}\right)^n \sum_{k=0}^{n}\left(\frac{1}{\text{SNR}\,\mathbf{v}\mathbf{R}\mathbf{v}^\dagger}\right)^k \Gamma\left(-k, \frac{1}{\text{SNR}\,\mathbf{v}\mathbf{R}\mathbf{v}^\dagger}\right)\right]\Bigg|_{\mathbf{v}=\mathbf{v}_\theta} \tag{11}$$

Using the infinite-series (11) in Theorem 1 and 2 (see equations (1)-(2)) the 1-D method for the calculation of the optimum beamformer in MISO systems

can be further simplified and hence, the relative computational complexity is further reduced.

2.2 MIMO systems

We consider a MIMO $N \times M$ flat fading channel with a complex Gaussian distribution $\mathbf{H} \sim \mathcal{N}(\mathrm{vec}(\mathbf{H}_m), \mathbf{R})$, with a *rank*-1 channel mean $\mathbf{H}_m \neq \mathbf{0}$ ($\mathbf{H}_m$ represents the LOS component) and covariance $\mathbf{R} = \mathbf{R}_t^T \otimes \mathbf{R}_r \neq \mathbf{I}_{MN}$, with $\mathbf{R}_t/\mathbf{R}_r$ the channel transmit/receive covariance matrices respectively. Assuming perfect CSI at the receiver and CDIT, it can be proven [9] that the rate-optimum beamforming transmission for the channel model under consideration (spatially correlated or uncorrelated with a non-unit covariance matrix MIMO Rician flat fading channel) is the solution of an 1-D optimization problem, which is expressed using Theorem 1 with the following modification: the normalized channel mean vector in the MISO case, which was denoted in Theorem 1 as $\mathbf{m}$, is replaced by the complex conjugate transpose of the right singular vector of $\mathbf{H}_m$, denoted as $\mathbf{q} \in \mathbb{C}^{1\times N}$, in the MIMO case[1].

In the same manner, for MIMO $N \times M$ flat fading channels with $N = 2$ or $N \geqslant 3$ and $rank\{\mathbf{R}_t\} \leqslant 2$, $\mathbf{v}_\theta$ is expressed using equation (10) (Theorem 2) and replacing $\mathbf{m}$ with $\mathbf{q}$ [9].

2.3 Results for the computational complexity of the 1-D method

The computational complexity of the proposed 1-D method - expressed as the runtime (in seconds) per iteration - is presented via simulations with respect to:

a. The number of channel samples which where used for the calculation of the ergodic beamforming capacity with the Monte Carlo method.

b. The number of transmit antenna elements (N).

The aforementioned complexity is compared with the corresponding complexity of the following multi-dimensional algorithms, which can also be employed in order to calculate the optimum beamformer:

a. An interior-point algorithm with logarithmic barrier function, for MISO and MIMO systems [10].

b. An iterative asymptotic (and hence, sub-optimum) approach, for MISO systems. This algorithm was recently developed in [11] and calculates the optimum beamformer (only) when the optimality of beamforming condition is satisfied.

The simulations were for Uniform Linear Arrays (ULAs) under the two-path delay spread correlation model, which was studied by Winters in [12].

Results are presented in Figure 3 for various scenarios[2] and demonstrate that the proposed 1-D method has significantly lower computational complexity, as follows:

[1] Since $rank(\mathbf{H}_m) = 1$, it is $\mathbf{H}_m = \mu\mathbf{p}\mathbf{q}$, with μ the unique eigenvalue of $\mathbf{H}_m$ and $\mathbf{p} \in \mathbb{C}^{M\times 1}, \mathbf{q}^\dagger \in \mathbb{C}^{N\times 1}$ its left and right singular vectors, respectively.

[2] The exact parameters of these scenarios can be found in [8]-[9].

a. For the simulated scenarios related to MISO systems, the runtime of the 1-D method is on average 5 to 7 times faster than the interior-point method and 2 to 10 times faster than the asymptotic approach.

b. For the simulated scenarios related to MIMO systems, the runtime of the 1-D algorithm is approximately 8.5 times faster than the interior-point method.

3 Results for the optimality of beamforming condition

As referred to in paragraph 1.2, the optimum beamformer achieves ergodic capacity when a necessary and sufficient optimality of beamforming condition is satisfied. This condition was proven in [6] an is expressed by the following inequality:

$$\lambda_{max}\Big((\mathbf{I}_N - \mathbf{v}_{\text{opt}}^{\dagger}\mathbf{v}_{\text{opt}})\mathbf{K}(\mathbf{I}_N - \mathbf{v}_{\text{opt}}^{\dagger}\mathbf{v}_{\text{opt}})^{\dagger}\Big) \leqslant \mathbf{v}_{\text{opt}}\mathbf{K}\mathbf{v}_{\text{opt}}^{\dagger} \tag{12}$$

where $\lambda_{max}(\cdot)$ stands for the maximum eigenvalue and $\mathbf{K} \in \mathbb{H}_{+}^{N}$ is expressed as:

$$\mathbf{K} = \mathcal{E}_{\mathbf{H}}\Big[\mathbf{H}^{\dagger}(\mathbf{I}_M + \text{SNR}\mathbf{H}\mathbf{v}_{\text{opt}}^{\dagger}\mathbf{v}_{\text{opt}}\mathbf{H}^{\dagger})^{-1}\mathbf{H}\Big] \tag{13}$$

Condition (12) is studied in this work for correlated MIMO Rician flat fading channels (assuming perfect CSI at the receiver and CDIT), i.e. the combined CMI-CCI model [13]. The parameters that affect condition (12) - and hence, the *optimality region*, which is defined as the set of channel distribution parameters that satisfy condition (12)- are studied via simulations with the help of a probabilistic approach, leading to important observations:

Observation 1. Beamforming becomes the rate-optimum strategy as the SNR decreases.

Observation 2. Beamforming becomes the rate-optimum strategy as the singular value of $\mathbf{H}_m$, μ, increases.

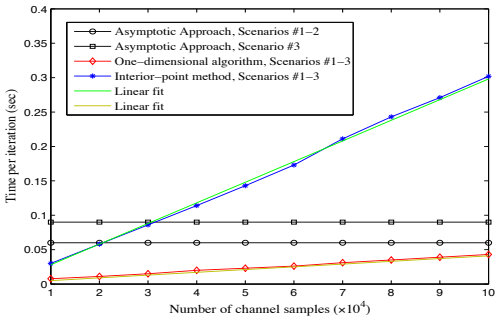
Observation 3. Beamforming becomes the rate-optimum strategy as the channel variance β decreases.

Observation 4. Beamforming becomes the rate-optimum strategy as ϕ (see (4)) decreases.

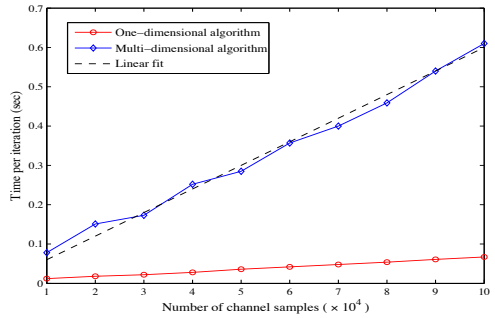
Observation 5. Relatively low β values lead to abrupt increase of the optimality region for relatively high values of the transmit antenna correlation coefficient ρ_t. Moreover, in the low-ρ_t regime, the optimality region seems to be less "sensitive" (i.e. is less affected) to an increase of the SNR, β and ϕ, compared to the high-ρ_t regime.

Observation 6. Beamforming becomes the rate-optimum strategy as the number of receive antenna elements (M) decreases.

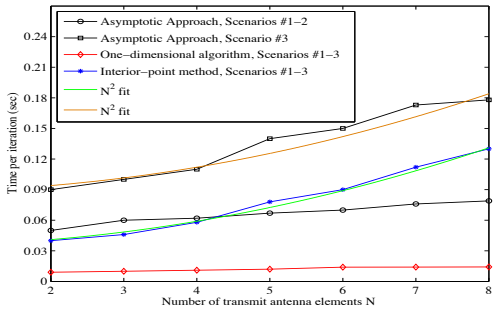
The results show that the CDIT model under consideration incorporates the basic characteristics of the uncorrelated MIMO Rician model (addressed with observations 1-3 and 6), however, the model also reveals new characteristics presented for the first time in this work (addressed with observations 4 and 5).



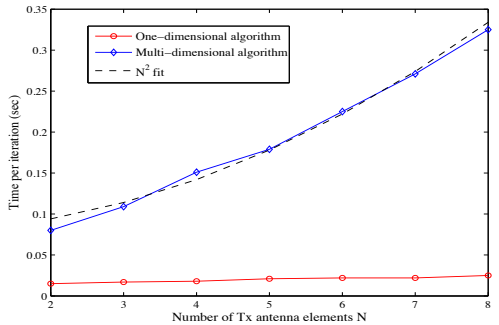
(a)



(b)



(c)



(d)

Figure 3: Runtime vs. the number of channel samples for a MISO 4×1/MIMO 4×4 system (a)/(b), and N with 2×10^4 channel samples and $M = 1/M = 4$ (c)/(d).

Observations 1-5 are visualized in Figure 4. This figure shows a set of curves on the $\mu - \rho_t$ plane, for different β values and $\phi = 35^o/65^o$, SNR $= 0/3$dB and Rx Angular Spread $\Delta_r = 68^o$. Each curve represents a bound: any $\{\mu, \rho_t\}$ point above this bound - i.e. in the region indicated with $Pr_{bf} = 1$ - corresponds to an operational scenario where the optimum beamformer achieves ergodic capacity, i.e. (12) is statistically always satisfied. The $\mu - \rho_t$ region where $Pr_{bf} = 1$ is referred to as the "optimality region".

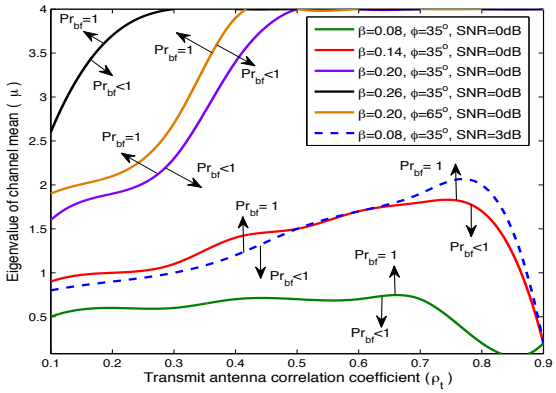


Figure 4: Optimality region $\mu - \rho_t$, for a MIMO 4×4 system and $\{$SNR $= 0/3$dB, $\phi = 35^o/65^o$, $\Delta_r = 68^o\}$.

4 Results using the 3GPP MIMO channel model

In the last part of this work the 3GPP MIMO channel model [14] is employed in order to study the performance of the optimum beamformer with respect to condition (12), using a probabilistic analysis and assuming perfect CSI at the receiver and CDIT. Results are derived for the following cases:

a. Urban micro-cellular environment with LOS component. The long term statistics of this environment simulates best a correlated MIMO Rician flat fading channel and hence, the long term combined CMI-CCI model can be employed [15]. In this case, the optimum beamformer achieves ergodic capacity with probability 0.9 for a wide SNR range.

b. Suburban and urban macro/micro-cellular environments [16]. The long term statistics of these environments simulate a MIMO Rayleigh flat fading channel, where both the CMI/CCI model can be employed, as a short/long term model, respectively. The analysis showed that in both CDIT models the optimum beamformer achieves ergodic capacity with a probability $\geqslant$0.45, in all operational environments and for a wide SNR range.

5 Conclusions

In this doctoral thesis the multi-dimensional and computationally complex optimization problem for the calculation of the rate-optimum beamforming transmission in correlated MISO/MIMO Rician flat fading channels (combined CMI-CCI model) is transformed into a simple 1-D optimization problem, which can be subsequently solved using standard 1-D search algorithms, reducing system's complexity. The reduced complexity can be exploited to either reduce cost by using devices with lower processing power or in order to: (a) operate in environments with smaller coherence time, proportional to the relative processing gain, and hence, support operational scenarios with higher mobility, (i.e. higher speeds, proportional to the relative processing gain), (b) increase the available processing power required by the system for other supplementary techniques. Moreover, the optimality of beamforming condition is studied via simulations for the combined CMI-CCI using a probabilistic analysis, and the optimality region is plotted for different values of the channel distribution parameters and the SNR. Generally, the knowledge of the optimality region can be valuable during the system design and deployment phases: if information for the targeted operational scenarios/channels is available, it can be used to produce the optimality regions and hence, decide if optimum beamforming can be employed as the main transmission strategy, which ultimately leads to reducing the system's complexity and cost. Results demonstrate that there is a wide range of operational scenarios and SNR values where the optimum beamformer achieves ergodic capacity or shows a relatively high (or near-optimum) performance, a fact that justifies the significance of the proposed solutions and the contribution of this work.

References

[1] E. Telatar, "Capacity of multi-antenna Gaussian channels", *European Trans. Telecommun.*, vol. 10, no. 6, pp. 585-596, Nov. 1999.

[2] G. J. Foschini, and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas", *Wireless Pers. Commun.*, vol. 6, no. 3, pp. 311-335, Mar. 1998.

[3] A. Goldsmith, S. A. Jafar, N. Jindal, and S. Vishwanath, "Capacity limits of MIMO channels", *IEEE J. Select. Areas Commun.*, vol. 21, no. 5, pp. 687-702, Jun. 2003.

[4] E. Visotsky, and U. Madhow, "Space-time transmit precoding with imperfect feedback", *IEEE Trans. Inf. Theory*, vol. 47, no. 6, pp. 2632-2639, Sep. 2001.

[5] S. A. Jafar, and A. Goldsmith, "Transmitter optimization and optimality of beamforming for multiple antenna systems with imperfect feedback", *IEEE Trans. Wireless Commun.*, vol. 3, no. 4, pp. 1165-1175, Jul. 2004.

[6] S. Srinivasa, and S. A. Jafar, "The optimality of transmit beamforming: a unified view", *IEEE Trans. Inf. Theory*, vol. 53, no. 4, pp. 1558-1567, Apr. 2007.

[7] D. E. Kontaxis, G. V. Tsoulos, and S. Karaboyas, "Optimum beamforming for correlated Rician MISO channels", *in Proceedings of Vehicular Technology Conference (VTC), Spring 2011.*

[8] D. E. Kontaxis, G. V. Tsoulos, and S. Karaboyas, "Ergodic capacity optimization for single-stream beamforming transmission in MISO Rician fading channels", *IEEE Trans. Veh. Technol.*, vol. 62, no. 2, pp. 628-641, Feb. 2013.

[9] D. E. Kontaxis, G. V. Tsoulos, and S. Karaboyas,"Beamforming capacity optimization for Rician MIMO wireless channels", *IEEE Wireless Commun. Letters*, vol. 1, no. 3, pp. 257-260, June 2012.

[10] M. Vu, and A. Paulraj, "Capacity optimization for Rician correlated MIMO wireless channels", *in Proc. 2005 Asilomar Conference*, pp. 133-138.

[11] J. Dumont, W. Hachem, S. Lasaulce, Ph. Loubaton, and J. Najim, "On the capacity achieving covariance matrix for Rician MIMO Channels: an asymptotic approach", *IEEE Trans. Inf. Theory*, vol. 56, no. 3, pp. 1048-1069, Mar. 2010.

[12] J. Salz, and J. Winters, "Effect of fading correlation on adaptive arrays in digital mobile radio", IEEE Transactions on Vehicular Technology, vol.43, no.4, pp.1049-1056, November 1994.

[13] D. E. Kontaxis, G. V. Tsoulos, G. E. Athanasiadou, and S. Karaboyas, "Optimality of transmit beamforming for spatially correlated MIMO Rician fading channels", *submitted.*

[14] 3GPP TR 25.996 V6.1.1 (2003-09).

[15] D. E. Kontaxis, G. V. Tsoulos, and S. Karaboyas, "Optimum ergodic beamforming capacity in urban microcellular operational environments", *in Proceedings of International Wireless Communications and Mobile Computing (IWCMC), 2012.*

[16] D. Kontaxis, G. Tsoulos, and S. Karaboyas, "Performance of multiple antenna systems in different operational environments", *in Proceedings of Personal Indoor Mobile Radio Communications (PIMRC), 2007.*

Algorithms for the Analysis and Processing of Autostereoscopic Images

Efthymios T. Koufogiannis*

Department of Informatics and Telecommunications
National and Kapodistrian University of Athens
efthimis@di.uoa.gr

Abstract. Nowadays, acquisition and display of three-dimensional (3D) images requires the use of special tracking devices or glasses. However specialized techniques provide the ability of 3D content delivery to end users without such limitations. These methods are called autostereoscopic and the resulting images are called autostereoscopic images.
A promising type of autostereoscopic imaging is called Integral Imaging (InI). InI provides the ability of capturing Integral Images (InIms) that contain embedded 3D information and are additionally able to display it to the end user without the need for specialized equipment. But the existence of even slight misalignments between the optical components in the acquisition device results in geometrical aberrations in the structure of the acquired InIm. These result in total loss of the displayed 3D content as well as failure of all InI analysis and processing algorithms that depend on predetermined geometric dimensions of the acquired InIm. In this doctoral dissertation robust image processing frameworks were developed in order to successfully correct these geometrical aberrations.
In detail, using computer vision methodologies, the problems of geometrical aberrations in arrays of square, hexagonal-triangular as well as circular lenses were extensively studied, a process that resulted in the development of robust InIm processing and rectification algorithms.

Keywords: Three dimensional image, Autostereoscopic image, Integral photography, Computer vision, Projective distortion

1 Introduction

Current technological developments in sensor and display technologies as well as in the manufacturing of optical components have made possible the creation of novel three-dimensional (3D) capturing as well as display devices. These devices are currently penetrating consumer market applications, therefore there is a need for improving them and increasing their robustness. This doctoral dissertation focuses on images acquired from Integral Imaging (InI), systems. The InI principle which was initially formulated by the Nobel laureate G. Lippman

* Dissertation Advisor: Manolis Sangriotis, Associate Professor.

[1] back in 1908, is currently considered as one of the most promising techniques for delivering 3D content, featuring full color, adequate detail and depth levels as well as support for multiple simultaneous viewers [2].

The basic InI acquisition setup shown in Fig. 1(a) consists of a charged coupled device (CCD) and a lens array (LA) that projects a real world scene on the CCD, forming a number of Elemental Images (EIs) depicting different parts of the acquired scene. The corresponding display setup shown in Fig. 1(b)

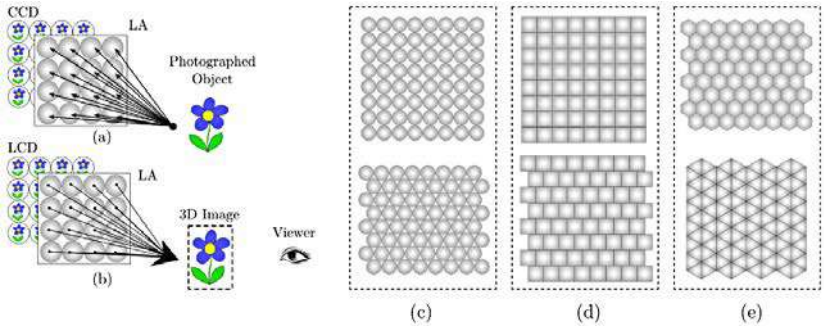


Fig. 1. (a) Integral Imaging acquisition, (b) InI display setup, (c), (d), (e) arrays of circular, square and hexagonal-triangular lenses.

reverses the previous procedure by using a liquid crystal display (LCD) between the image formed on the CCD and the LA. This results in a 3D image being projected between the LA and the viewer. It should be noted that LAs come in different configurations containing circular, round, and hexagonal-triangular lenses as shown in Figs. 1(c), 1(d), 1(e). Additional details of the InI display and acquisition processes can be found in [3], [4].

Since misalignments almost always occur during the InIm acquisition stage, perspective distortion is introduced which alters the expected shape of the resulting EIs. This results in degrading the resulting EI grid since the acquired EIs do not have constant geometric properties and are not properly aligned and sized. The result is total loss of the 3D information on the display setup. Furthermore all subsequent InI processing tasks such as compression [5] and 3D object reconstruction [6] fail since they depend on the accurate location as well as dimension of the acquired EIs.

Previous work on this field initially [7] focused on slight rotational and translational misalignments between the CCD and arrays of square lenses during the InI acquisition phase. In [8] an initial approach to the problem of perspective distortion on square lens setups was proposed where a single rectangle detected using the Hough Transform inside the area of a distorted InIm was used to extract the necessary rectification parameters.

2 Dissertation Summary

In this doctoral dissertation efficient and robust image analysis and processing algorithms were proposed in order to rectify geometric distortions not only for InIms acquired using arrays of square lenses, but for all commonly used lens array configurations.

Using computer vision methodologies geometric distortions corresponding to InIms acquired using arrays of square [9], hexagonal-triangular [10] and circular [11], [12] lenses were corrected by utilizing statistical image wide features. Furthermore the rectification process for square, hexagonal and triangular lens arrays was further streamlined by using a least squares approach that led to a simple and efficient implementation that bypassed all intermediate matrix computations.

It should be noted that in this thesis and for the first time in current literature, an effective approach was suggested in order to alleviate geometric distortions occurring from arrays of circular lenses. Circular lenses are still widely used in existing setups [13] and in this work effective algorithms where implemented both for rotational [11] as well as perspective distortion [12] in arrays of circular lenses regardless of lens packing configuration.

3 Perspective Rectification

According to [14] the transformation corresponding to perspective distortion is mathematically represented by the 3×3 real value matrix H^{-1}. The inverse procedure results in the perspective rectification matrix H that maps a point $\mathbf{x}$ from the distorted image to its corresponding point $\mathbf{x}'$ on the rectified image using

$$\mathbf{x}' = H\mathbf{x} \tag{1}$$

where $\mathbf{x}$ and $\mathbf{x}'$ are three-vectors in homogeneous coordinates representing the relevant points on the two-dimensional (2D) projective space. In a similarly way distorted lines on the 2D projective space are rectified using H^{-T} [14]:

$$\mathbf{l}' = H^{-T}\mathbf{l} \tag{2}$$

Furthermore and according to Liebowitz [15] the perspective transformation H can be stratified as a breakdown structure that consists of the three matrices H_p, H_a, H_s:

$$H = H_s H_a H_p \tag{3}$$

where

$$H_p = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ l_1 & l_2 & l_3 \end{bmatrix} \tag{4}$$

$$H_a = \begin{bmatrix} \frac{1}{\beta} & \frac{-\alpha}{\beta} & 0 \\ 0 & 1 & 0 \\ l_1 & l_2 & l_3 \end{bmatrix} \tag{5}$$

$$H_s = \begin{bmatrix} sR & \mathbf{t} \\ \mathbf{0}^T & 1 \end{bmatrix} \tag{6}$$

This stratified rectification approach and its results from the subsequent application of H_p, H_a, H_s is demonstrated in Fig. 2 for a perspectively distorted checkerboard image.

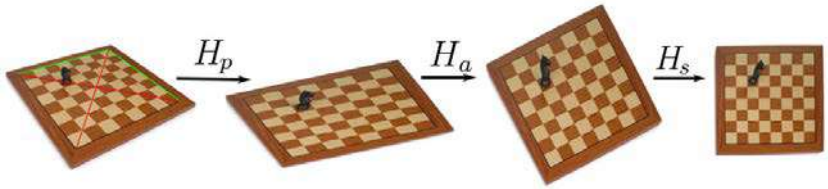


Fig. 2. Subsequent application of the matrices H_p, H_a and H_s results in a perspectively rectified image.

More analytically, matrix H_p consists of the parameters l_1, l_2, l_3. These parameters define the vanishing line (horizon line) $l_1x+l_2y+l_3 = 0$ of the distorted image plane. As seen in the second image of Fig. 2 applying H_p removes the projective component of the distortion resulting in an image where line parallelism has been restored but length and angle ratios are wrong.

Matrix H_a consists of the parameters α, β that restore the metric properties on the affine image. These parameters are calculated by measuring foreknown geometric properties such as length and line ratios on the affine image and subsequently applying the metric constraints described in [15]. Application of H_a results in an arbitrarily rotated and scaled image with correct metric properties as seen in the third image of Fig. 2.

Finally applying H_s results in correct rotation and scaling as seen in the rightmost image of Fig. 2. It should be noted that R is a rotational matrix, s is a scaling parameter while $\mathbf{t}$ is a translation vector.

3.1 Least Squares based Rectification

For our specific InIm rectification purposes translation and scaling in the final stage of the rectification procedure are not needed since they can be calculated during the InI display stage. Therefore the perpective rectification matrix H will have six degrees of freedom and the form of

$$H = \begin{pmatrix} h_1 & h_2 & 0 \\ h_3 & h_4 & 0 \\ h_5 & h_6 & 1 \end{pmatrix} \tag{7}$$

According to Eq. 2 the corresponding line transformation matrix has the form H^{-T} and is given by

$$H_L = (H^{-1})^T = \begin{pmatrix} g_1 & g_3 & g_5 \\ g_2 & g_4 & g_6 \\ 0 & 0 & 1 \end{pmatrix} \tag{8}$$

In the case of a hexagonal-triangular EI grid and by using the foreknown ideal line angles of 90°, −30°, 30° the line rectification matrix can be estimated by forming a Homogeneous Overdetermined Linear System (HOLS) using only the distorted line parameters $\{\mathbf{d}, \mathbf{b}\}$ as seen in Eq. 9.

$$\begin{pmatrix} \mathbf{0}^T & \mathbf{d}_{v1}^T \\ \mathbf{0}^T & \mathbf{d}_{v2}^T \\ \vdots & \vdots \\ \mathbf{0}^T & \mathbf{d}_{vN_v}^T \\ \mathbf{d}_{a1}^T & \lambda\mathbf{d}_{a1}^T \\ \mathbf{d}_{a2}^T & \lambda\mathbf{d}_{a2}^T \\ \vdots & \vdots \\ \mathbf{d}_{aN_a}^T & \lambda\mathbf{d}_{aN_a}^T \\ \mathbf{d}_{b1}^T & -\lambda\mathbf{d}_{b1}^T \\ \mathbf{d}_{b2}^T & -\lambda\mathbf{b}_{a2}^T \\ \vdots & \vdots \\ \mathbf{d}_{bN_b}^T & -\lambda\mathbf{b}_{bN_b}^T \end{pmatrix} \begin{pmatrix} g_1 \\ g_3 \\ g_5 \\ g_2 \\ g_4 \\ g_6 \end{pmatrix} = \mathbf{0} \tag{9}$$

On the previous system, $\lambda = \sqrt{3}/3$, $\mathbf{0}^T = [0, 0, 0]$ and $\mathbf{0}$ is column vector containing as many zeros as the total number of distorted lines detected. The least squares solution of the previous system results in the estimation of the parameters $g_1, g_2, \ldots, g_6$.

3.2 Rectification of InIms acquired from Arrays of Circular Lenses

An initial approach to geometric distortion of InIms acquired from arrays of circular lenses was proposed in [11] where the problem of InIm rotation was resolved using the Gradient augmented Circular Hough Transform [16] and a subsequent Delaunay triangulation on the centers of the resulting centers in order to identify the rotation angle.

But in the case of perspective distortion the imaged EIs are being mapped to elliptical shapes. Therefore all available line or circle detection methods are of no use since the only available structures found inside the distorted InIms are elliptical contours. In this case a different mathematical approach is applied to calculate the necessary rectification matrices H_p, H_a, H_s.

This approach is based on the theory of the Circular Points [14] that are the two complex conjugate three-vectors $(1, \pm i, 0)^T$ located at the infinity on the undistorted InIm plane. All coplanar circles verify their coordinates [17], [18].

The corresponding Images of the Circular Points are the conjugate three-vectors $(\alpha l_3 \mp i l_3 \beta, l_3, -l_2 - l_1 \alpha \pm i l_1 \beta)^T$ that are located on the vanishing line of the distorted InIm plane after the perspective distortion H^{-1} has occured. All coplanar ellipses on the distorted plane verify their coordinates.

Obtaining the coordinates $(x_c, y_c, 1)^T$ and $(\overline{x_c}, \overline{y_c}, 1)^T$ for the two conjugate ICPs results in the extraction of all necessary rectification parameters in the matrices H_p, H_a, H_s as shown in Fig. 3.

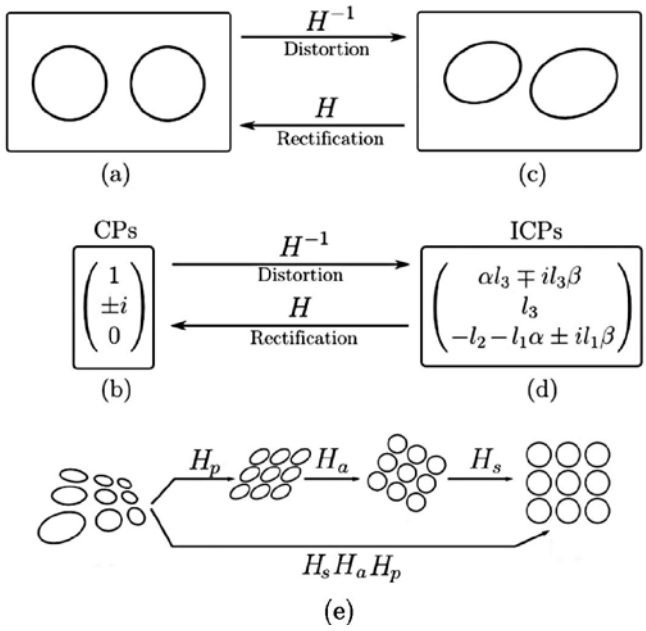


Fig. 3. (a) Any pair of coplanar circles verifies the CPs shown in (b). Under the perspective distortion H^{-1}, the pair of ellipses in (c) corresponding to the circles in (a) verifies the ICPs shown in (d). (e) Subsequent application of the matrices H_p, H_a and H_s results in a perspectively rectified image.

These parameters are given by

$$(l_1, l_2, l_3)^T = (x_c, y_c, 1)^T \times (\overline{x_c}, \overline{y_c}, 1)^T \tag{10}$$

$$\begin{cases} \alpha = \text{Real}\left(\frac{-l_2 x_c}{l_3 + l_1 x_c}\right) = \text{Real}\left(\frac{l_3 + l_2 y_c}{-l_1 y_c}\right) \\ \beta = \left|\text{Imag}\left(\frac{-l_2 x_c}{l_3 + l_1 x_c}\right)\right| = \left|\text{Imag}\left(\frac{l_3 + l_2 y_c}{-l_1 y_c}\right)\right| \end{cases} \tag{11}$$

A fully detailed methodology containing all the necessary steps followed to extract the ICPs as well as the corresponding rectification parameters $l_1, l_2, l_3, \alpha, \beta$ in a real case distorted InIm scenario is seen in [12].

4 Line Segment Detection and Clustering

In order to efficiently detect line segments corresponding to EI border edges for arrays of square and hexagonal lenses we used the LSD algorithm [19] which is robust against noise, minimizes the number of false segment detections and has linear execution time.

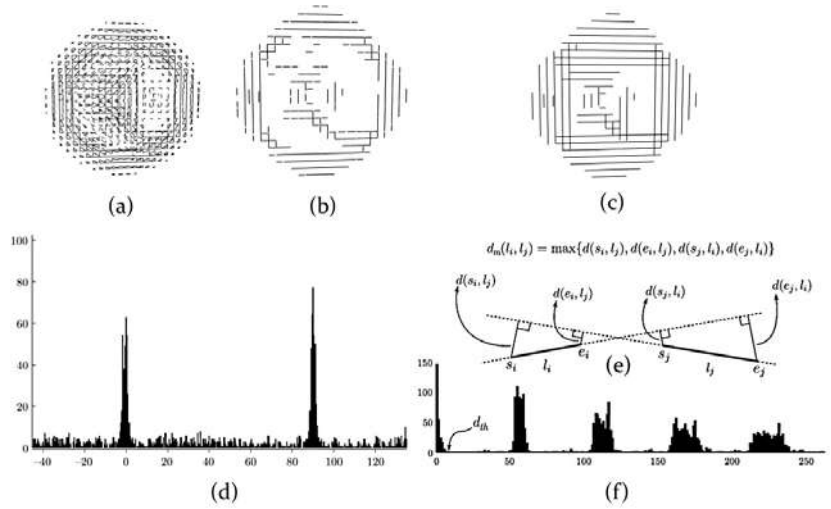


Fig. 4. (a) Acquired InIm after applying the LSD algorithm. (b) Segments resulting after isolating the histogram lobes shown in (d). (c) Clustering of line segments from (b) after using the clustering function of (e) and using its histogram in (f).

Application of the LSD on an acquired InIm results in the EI segments shown in Fig 4(a). Further calculation of the segment angles and isolation of the segments corresponding to the main lobes of the angles histogram results in a significant reduction of the noisy segments as shown in Fig 4(b).

Finally a hierarchical clustering [20] approach using a custom metric function operating between segments and a subsequent least squares fitting of the line segments in the resulting clusters results in very accurate lines corresponding to EI borders as shown in Fig 4(c). For a more concrete analysis of this line isolation framework the reader can refer to [9].

5 Elliptical Contour Extraction

To efficiently estimate analytical equations of ellipses corresponding to distorted EI borders an edge linking [21] approach was applied followed by elliptical fitting using a least squares approach. The best 20% of the fitting results (according

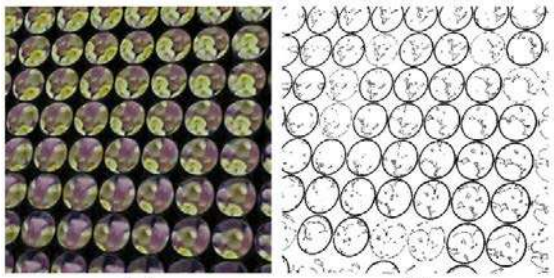


Fig. 5. (a) Perspectively distorted InIm acquired using an array of circular lenses. (b) Elliptical contours extracted after applying an edge linking and ellipse fitting approach.

to their Mean Squared Error) were retained in order to obtain accurate elliptical borders. This approach is demonstrated in Fig 5 and for all the relevant implementation details the reader may refer to [12].

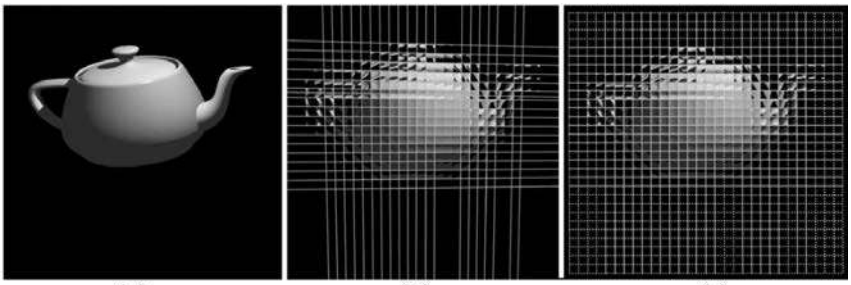
6 Experimental Results and Discussion

A large number of different artificially generated InIms featuring different types of LAs, varying levels of texture details and object complexity were generated using the methodology proposed in [22]. In Fig. 6 an artificial "Teapot" has been rendered using an array of square lenses while in Fig. 8 a "3D Objects" scene has been rendered using an array of circular lenses.

Furthermore to effectively assess the robustness of our InIm analysis and rectifications algorithms, the computer generated images were contaminated with Gaussian noise resulting in image qualities of 30, 25 and 20 dB. Following this approach provided full control over the introduced perspective distortion as well as prior knowledge of the ground truth values for the rectification parameters.

In addition to the computer generated images a large number of optically acquired InIms was acquired featuring different scene complexities, utilizing arrays of square, hexagonal as well as circular lenses. Figure 7 features an optically acquired "Dice" using an array of square lenses, while Fig. 9 features an optically acquired "Toy" using an array of circular lenses.

We used two geometric consistency metrics that statistically characterize the consistency of the rectified InIm grid after applying the rectification methodologies proposed in this dissertation. For square lens packing configurations we used the fact that the reconstructed grid separating the lenses is characterized by equally spaced intersecting lines forming angles of 90°. To evaluate the deviations from an ideal grid we calculated the angles $\{\omega\}$ as well as the segment lengths $\{\lambda\}$ formed between the intersecting lines of the grid. We used a similar approach for arrays of hexagonal lenses [10] by considering line angle values of $-30°$, $30°$ and $90°$.

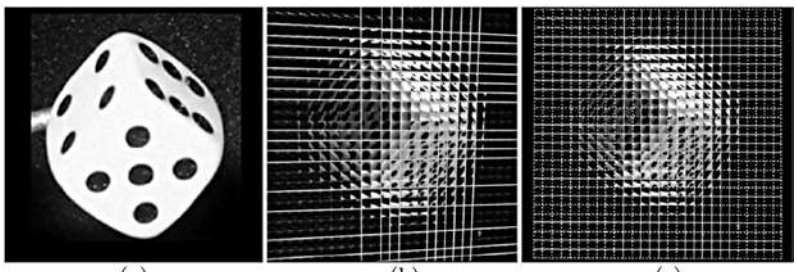


(a) (b) (c)

	$\bar{\omega}_i \pm \sigma_{\omega_i}$	$\bar{\lambda}_i \pm \sigma_{\lambda_i}$
noiseless	$90.000° \pm 0.060°$	0.996 ± 0.011
30 dB	$90.00° \pm 0.10°$	1.009 ± 0.013
25 dB	$90.00° \pm 0.12°$	1.009 ± 0.017
20 dB	$90.00° \pm 0.16°$	0.999 ± 0.024

(d)

Fig. 6. Rectification results for a representative artificially generated InIm. (a) The original 2D image of a 3D "Teapot", (b) the corresponding perspectively distorted InIm, (c) the rectified InIm along with the registered grid lines superimposed, (d) the corresponding geometric consistency parameter values for various noise levels.

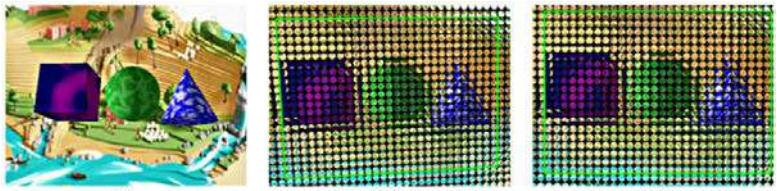


(a) (b) (c)

$\bar{\omega}_i \pm \sigma_{\omega_i}$	$\bar{\lambda}_i \pm \sigma_{\lambda_i}$
$90.000° \pm 0.055°$	1.006 ± 0.010

(d)

Fig. 7. Rectification results for an optically acquired InIm. (a) The original 2D acquired image of a real "Dice", (b) the corresponding perspectively distorted InIm, (c) the rectified InIm along with the registered grid lines superimposed, (d) the corresponding geometric consistency parameter values.



(a) (b) (c)

	$\overline{\omega_i} + \sigma_{\omega_i}$	σ_{λ_i}
Noiseless (dB)	$90.00 \pm 0.15°$	0.014
30	$90.00 \pm 0.17°$	0.022
25	$90.00 \pm 0.52°$	0.037
20	$90.00 \pm 0.87°$	0.041

(d)

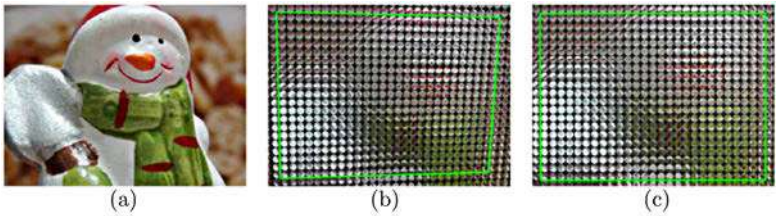
Fig. 8. Rectification results for a representative artificially generated InIm acquired using an array of circular lenses. (a) The original 2D image of the "3D Objects", (b) the corresponding perspectively distorted InIm, (c) the rectified InIm, (d) the corresponding geometric consistency parameter values for various noise levels. The green border in (b) and (c) has been drawn for illustrative purposes.

In the tables shown in Fig. 6(d) and Fig. 7(d) the results for the rectification of two representative InIms using arrays of square lenses are shown. For the artificial "Teapot" of Fig. 6 we can observe that the standard deviation of $\{\omega\}$ does not exceed 0.16° while the standard deviation of $\{\lambda\}$ does not exceed the value of 0.024 even for high noise contamination levels. These are typical results for the whole artificial InIm set, while we can further observe that the corresponding results for the optically acquired "Dice" of Fig. 7 are in line with the artificial simulation data.

Similarly in Fig. 8 and in Fig. 9 the rectification results for two characteristic InIms using arrays of circular lenses are shown. For the artificial "3D Objects" scene of Fig. 8 we can observe that the standard deviation of $\{\omega\}$ does not exceed the value of 0.87° while the standard deviation of $\{\lambda\}$ does not exceed the value of 0.041 even within high noise contamination levels. These results characterize the whole artificial InIm set, while it is further verified in Fig 9 and the corresponding table that the results for the optically acquired "Toy" are in line with the artificial InIm rectification data.

7 Conclusions

In this dissertation automated solutions were proposed to alleviate geometric distortions affecting the InIm acquisition process. The occurrence of this issue was studied over the whole range of available InIm configurations and effective solutions where proposed for its removal.



(a) (b) (c)

$\overline{\omega_i} + \sigma_{\omega_i}$	σ_{λ_i}
$90.00 \pm 0.46°$	0.039

(d)

Fig. 9. Rectification results for an optically acquired "Toy" InIm acquired using an array of circular lenses. (a) The original 2D image acquired for the "Toy", (b) the corresponding perspectively distorted InIm, (c) the rectified InIm, (d) the corresponding geometric consistency parameter values. The green border in (b) and (c) has been drawn for illustrative purposes.

The rectification of InIms acquired using square as well as hexagonal-triangular lenses was further optimized using a least squares approach while for the first time in current literature a novel and effective rectification methodology was proposed for arrays of circular lenses.

During the evaluation process, large sets of optically acquired as well as raytraced InIms were utilized to examine the parameters that may affect the robustness of proposed rectification frameworks. It should be noted that the usage of computationally generated InIm sets allowed joint control of both perspective distortion and noise levels.

The geometric consistency for the rectified InIms that are presented in this dissertation synopsis and the relevant publications [9–12] is fully retained as verified by the corresponding data. Since noise greatly affects the rectification procedure at the early processing stages, we conclude that the proposed methodologies are robust against high noise contamination levels and effective to to be used as an integrated and useful InIm analysis and rectification framework.

References

1. Lippmann, G.: La photographie integràle. Comptes-Rendus Academie des Sciences **146** (1908) 446–451
2. Son, J.Y., Javidi, B.: Three-dimensional imaging methods based on multiview images. J. Display Technol. **1**(1) (Sep 2005) 125
3. Park, J.H., Kim, Y., Kim, J., Min, S.W., Lee, B.: Three-dimensional display scheme based on integral imaging with three-dimensional information processing. Opt. Express **12**(24) (Nov 2004) 6020–6032
4. Jang, J.S., Javidi, B.: Formation of orthoscopic three-dimensional real images in direct pickup one-step integral imaging. Optical Engineering **42**(7) (2003) 1869–1870

5. Sgouros, N., Kontaxakis, I., Sangriotis, M.: Effect of different traversal schemes in integral image coding. Appl. Opt. **47**(19) (Jul 2008) D28–D37
6. Passalis, G., Sgouros, N., Athineos, S., Theoharis, T.: Enhanced reconstruction of three-dimensional shape and texture from integral photography images. Appl. Opt. **46**(22) (Aug 2007) 5311–5320
7. Sgouros, N.P., Athineos, S.S., Sangriotis, M.S., Papageorgas, P.G., Theofanous, N.G.: Accurate lattice extraction in integral images. Opt. Express **14**(22) (Oct 2006) 10403–10409
8. Hong, K., Hong, J., Jung, J.H., Park, J.H., Lee, B.: Rectification of elemental image set and extraction of lens lattice by projective image transformation in integral imaging. Opt. Express **18**(11) (May 2010) 12002–12016
9. Koufogiannis, E.T., Sgouros, N.P., Sangriotis, M.S.: Robust integral image rectification framework using perspective transformation supported by statistical line segment clustering. Appl. Opt. **50**(34) (Dec 2011) H265–H277
10. Koufogiannis, E.T., Sgouros, N.P., Sangriotis, M.S.: Perspective rectification of integral images produced using hexagonal lens arrays. In: Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), 2012 Eighth International Conference on. (july 2012) 75–78
11. Koufogiannis, E., Sgouros, N., Ntasi, M., Sangriotis, M.: Grid reconstruction and skew angle estimation in integral images produced using circular microlenses. In: Digital Signal Processing (DSP), 2013 18th International Conference on, IEEE (2013) 1–7
12. Koufogiannis, E.T., Sgouros, N.P., Sangriotis, M.S.: Perspective rectification of integral images produced using arrays of circular lenses. Appl. Opt. **52**(20) (Jul 2013) 4959–4968
13. Lim, Y.T., Park, J.H., Kwon, K.C., Kim, N.: Resolution-enhanced integral imaging microscopy that uses lens array shifting. Opt. Express **17**(21) (Oct 2009) 19253–19263
14. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision. Second edn. Cambridge University Press (2004)
15. Liebowitz, D., Zisserman, A.: Metric rectification for perspective images of planes. In: Proceedings. 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No.98CB36231), IEEE Comput. Soc (1998) 482–488
16. Kimme, C., Ballard, D., Sklansky, J.: Finding circles by an array of accumulators. Commun. ACM **18**(2) (February 1975) 120–122
17. Ip, H., Chen, Y.: Planar rectification by solving the intersection of two circles under 2d homography. Pattern Recognition **38**(7) (2005) 1117–1120
18. Lourakis, M.: Plane metric rectification from a single view of multiple coplanar circles. In: Proceedings - International Conference on Image Processing, ICIP. (2009) 509–512
19. von Gioi, R., Jakubowicz, J., Morel, J.M., Randall, G.: Lsd: A fast line segment detector with a false detection control. Pattern Analysis and Machine Intelligence, IEEE Transactions on **32**(4) (april 2010) 722–732
20. Theodoridis, S., Koutroumbas, K.: Pattern Recognition, Third Edition. 3 edn. Academic Press (2006)
21. Kovesi, P.: Matlab and octave functions for computer vision and image processing
22. Athineos, S.S., Sgouros, N.P., Papageorgas, P.G., Maroulis, D.E., Sangriotis, M.S., Theofanous, N.G.: Photorealistic integral photography using a ray-traced model of capturing optics. Journal of Electronic Imaging **15**(4) (2006) 043007

Testing and Performance Calibration Techniques for Integrated RF Circuits

John G. Liaperdos*

Department of Informatics and Telecommunications
National and Kapodistrian University of Athens
gliaperd@teikal.gr

Abstract. In this dissertation, a unified approach is proposed for the testing and calibration procedures of integrated RF circuits, that exploits a common set of optimally selected observables. In order to address the problem of accessibility to these observables, a built-in measurement technique is presented, while a method to minimize the uncertainty introduced in the measurement system itself is described, as well. By the application of selection algorithms, the number of observables is reduced through an optimization procedure, leading to test cost savings due to the reduction of test conduction complexity and time.

Keywords: Defect Detection, Integrated Circuits, Mixers, Performance Calibration, Wireless Transceivers

1 Introduction

Specifications of analog integrated circuits (ICs), especially radio frequency (RF) circuits, have become increasingly strict as their applications tend to be more complex and demanding. To meet these specifications, an often painstaking and time-consuming series of repetitive design cycles has to be undertaken which, however, cannot guarantee that all fabricated circuit instances are acceptable in terms of their expected performance.

In order to assure reliability, each fabricated IC should be subject to a testing procedure aiming to ascertain that the circuit is functional and, furthermore, compliant to its specifications. In the conventional RF IC testing approach, automatic test equipment (ATE) is used to sequentially measure the performance characteristics of a circuit under test (CUT). Although these measurements are simple, they require a variety of test resources which, together with the long test application times, increase the total manufacturing cost. In many cases testing cost turns to be unacceptable, tending to be comparable to the rest manufacturing cost [1]. To overcome the inabilities of conventional testing, alternative low-cost techniques have been proposed, the most notable being defect-oriented testing (DOT) and 'alternate' test.

* Dissertation Advisor: Angela Arapoyanni, Professor

Defect-oriented testing, or structural testing, follows the assumption that most or all defect mechanisms that commonly lead to malfunctioning circuits manifest themselves in more fundamental observables than the specifications, thus simplifying the test procedure and also reducing cost. DOT efficiency is primarily determined by the defect detection capability obtained by the selected observables and the cost for their stimulation and measurement. On the other hand, the objective of the alternate test methodology [2] is to find a suitable test stimulus and to accurately predict the circuit's performance from the corresponding alternate test response. Appropriate selection of the set of observables which compose the test response crucially determines prediction accuracy; however, test simplicity is often compromised leading to increased cost [3].

As an outcome of the test procedure, a portion of the tested circuits is, unavoidably, rejected: Catastrophic failures due to physical limitations, together with the variability of a large number of parameters affecting the IC production processes, constitute the problem of yield loss. Although process variations do not necessarily lead to a defective circuit in terms of functionality, a large amount of manufactured circuits might, however, fail to comply with their performance specifications, leading to an increased yield loss which turns to be significant in nanometer technologies [4, 5].

Several calibration methods have been proposed, which address the issue of increased parametric yield loss by compensating for parametric variations using on-chip resources [6, 7]. A critical issue to be addressed in the calibration procedure is the appropriate selection of the specific circuit's state at which performance is restored to acceptable levels.

A key problem in both RF IC testing and calibration is that it is not always possible for the ATE to have direct access to all or even part of the internal signals of an IC, especially in system-on-chip (SoC) or system-in-package (SiP) designs. Although some internal signals can be made available to the external tester, frequency limitations due to the lower speed of the I/O interface may not permit their direct observation.

A unified low-cost approach for the testing and calibration procedures of integrated RF circuits is proposed in this dissertation, which exploits a set of optimally selected observables. The exploitation of these observables enables both defect detection and prediction of the circuit's performance, allowing the examination of compliance with the specifications and performance calibration, as well. A combination of the DOT and alternate test methodologies is used to maximize fault coverage, while both alternate test accuracy and calibration efficiency are increased by the exploitation of the adjustable features of the circuits under consideration. In order to address the problem of accessibility to test observables, a built-in technique is proposed, while a method to minimize the uncertainty introduced in the measurement system itself is also described. The application of selection algorithms is explored, aiming to reduce the number of test observables through an optimization procedure that leads to test cost savings due to the reduction of the test conduction complexity and time.

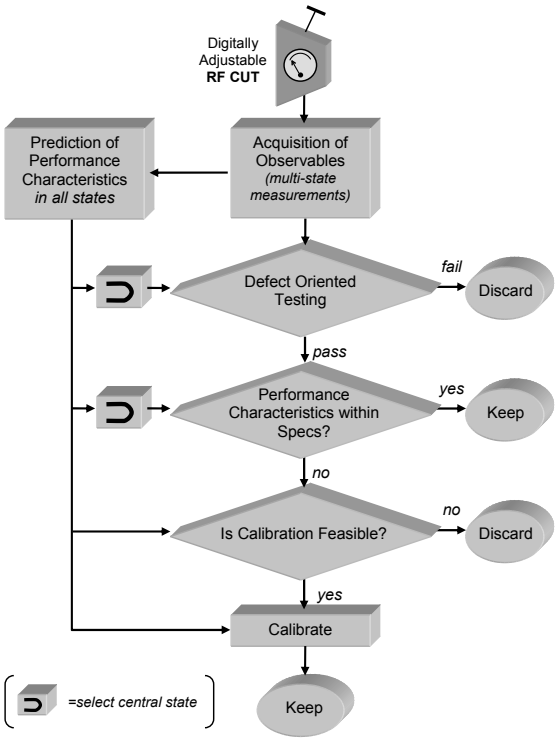


Fig. 1. Testing and calibration flow

The efficiency of the proposed techniques is validated by their application to a typical RF mixer designed in a 0.18m CMOS technology. Simulation results are obtained and assessed, while comparison with similar conventional methodologies is also provided.

2 A Test and Calibration Strategy for Adjustable RF Circuits

2.1 Methodology

We consider adjustable RF circuits, that are designed such as to be able to operate in several discrete states, thus providing the capability to vary their performance characteristics (PCs) around their post-fabrication values. This functionality is obtained by the use of an adjustable element, the value of which is related to the circuit's performance characteristics under consideration.

The proposed test and calibration methodology is illustrated in Fig. 1 and synopsized as follows. First, the adjustable RF CUT is measured in various states to obtain a specific set of test observables according to the alternate

test approach. Then, its performance characteristics are predicted for all states of operation using pre-developed regression models [8]. The set of measured observables, together with the predicted performance characteristics, are used for defect detection, while the predicted performance characteristics in a single state (the central state) of operation are sufficient for defect detection, as shown in [9]. Defect-free circuits are examined to determine if their predicted performance characteristics in the central state comply with the specifications. The predicted performance characteristics in the remaining states are used to explore the ability to calibrate each circuit found to be non-compliant. Finally, circuits for which their predicted performance characteristics in at least one state are compliant with the specifications are calibrated exploiting the existing adjustable element.

Testing Procedure The first step towards defect detection is the derivation of the expected range of values, due to process variations and device mismatches, for each individual observable as well as for each predicted performance characteristic. The derivation of these ranges, for which we adopt the term "variation bands", assumes defect-free circuits.

To maximize defect detection efficiency, we may consider to extend the initial set of observables ($\boldsymbol{O}$) to a superset of extended observables ($\boldsymbol{E}$), whose elements represent either an initial observable or a simple linear combination of these observables. The latter can be determined by specifying correlations between elements of $\boldsymbol{O}$, either through empirical observation or via principal component analysis (PCA). The derivation of the corresponding variation bands is performed by statistical analysis on a sample of either actual or Monte-Carlo simulated instances, as explained in detail in [9]. Figure 2(a) summarizes the overall procedure followed for the derivation of the variation bands.

Defect detection is carried out after the extended observables are calculated, as illustrated in Fig. 2(b), while the predicted performance characteristics for the circuit's central state of operation are also considered. Defect detection is accomplished according to the following rule: If at least one of the extended observables (in a DOT sense) or at least one of the PCs (in an alternate test sense) fails to fall within its corresponding variation band, the CUT is classified as defective and discarded, otherwise it is considered to be free of defects. If the latter is true, a calibration procedure is initiated in the case where predicted PCs in the central state do not comply with the specifications, in order to reduce parametric yield loss.

Calibration Procedure The calibration procedure determines the circuit's state of operation for which all predicted performance characteristics comply with their specifications. This is possible by the exploitation of the regression models that have been constructed for all PCs and all states of the circuit's operation, according to the principle described in Fig. 3. In this figure, an example of a correctable circuit is shown, since for the state *S2* all predicted PCs meet the specifications simultaneously. On the contrary, for the uncorrectable circuit

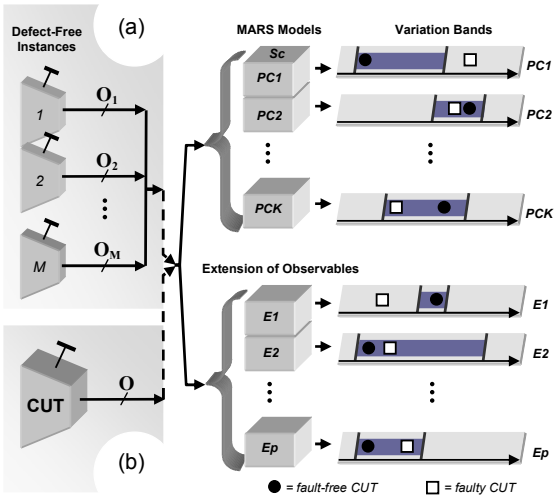


Fig. 2. Procedure for (a) the derivation of the variation bands (shaded areas) and (b) defect detection

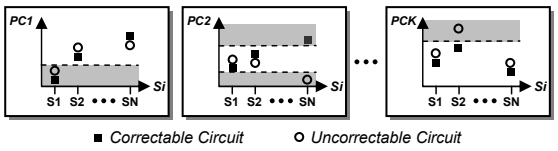


Fig. 3. Calibration principle. (Shaded areas indicate non-compliant performance characteristic ranges, while S_i (i=1,2, ... ,N) correspond to the circuit's states of operation)

instance shown in the same figure, no state exists for which all predicted PCs fall inside their acceptable ranges.

2.2 Evaluation

The effectiveness of the proposed methodology has been evaluated by simulations on a typical RF mixer. The mixer under consideration, presented in Fig. 4, is designed in the 0.18μm Mixed-Signal/RF CMOS technology of UMC (Vdd=3.3V) with an intermediate frequency (IF) of 150MHz. A digitally controllable resistor R_{var} in the mixer's bias circuitry has been used as the adjustable element, by which the mixer's current is controlled and states of the circuit's operation are provided, as summarized in Table 1 where the values of each PC of interest – namely, gain (G), 1dB compression point (1dB CP) and input referred 3rd order intercept point (IP3) – are also presented for each state.

In this case study, we adopt the use of the local oscillator's (LO) signal as the test stimulus at the RF inputs of the mixer [3, 7, 10]. The self-mixing of the LO signal forces the mixer to operate in homodyne (zero IF) mode, generating

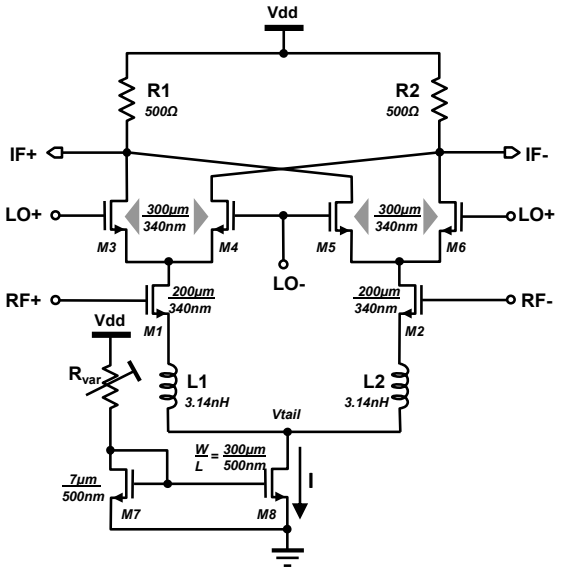


Fig. 4. The adjustable RF mixer under consideration

Table 1. RF mixer states of operation

State ID *(Si)*	G *(dB)*	1dB CP *(dBm)*	IP3 *(dBm)*	I *(mA)*
S1	3.68	-1.77	7.52	3.63
S2	4.03	-0.92	8.60	4.08
*S3**	*4.39*	*-0.19*	*9.73*	*4.67*
S4	4.72	0.62	10.88	5.46
S5	4.85	0.85	10.72	6.58

* central state (S_c=*S3*)

DC voltage levels at its "IF" outputs. The aforementioned DC levels (IF_+, IF_-) are used as the main observables, together with the DC voltage component of the mixer's tail voltage (denoted as V_{tail} in Fig. 4).

It has been proven that prediction accuracy improves significantly if the voltage observables are obtained from more than one of the mixer's states. Specifically, only two states are enough to provide very high prediction accuracy [11], namely the central state (*Sc=S3*) and the maximum tail current state (*S5*). Furthermore, observables are extended such as to include the differential mixer output voltage in test mode, since this inclusion increases DOT effectiveness [9,10].

Defect Detection All possible defects (38 opens, 43 shorts, 13 bridgings) have been simulated in the presence of process variations and device mismatches, set-

Table 2. Defect detection probability results

Type of Defect	Defect Detection Probability (%)		
	Defect-Oriented ***(E)***	*Alternate* ***(PC)***	*Combined* ***(E, PC)***
Shorts	78.58	99.14	100
Opens	100	89.26	100
Bridgings	100	70.15	100
Overall	*90.20*	*91.14*	***100***

ting the mixer in both selected states. Defect detection probabilities have been calculated and the results are summarized in Table 2, where columns labeled "Defect-Oriented" and "Alternate" correspond to the probabilities obtained by the extended observables (***E***) and the predicted performance characteristics (***PC***), respectively, while "Combined" indicates the result obtained by using both the DOT and the alternate test approaches. According to these results, all defects can be detected successfully since a detection probability of 100% is provided. This ensures that all mixer instances entering the succeeding calibration phase are free of defects and, hence, candidate for calibration.

Calibration It is assumed that specifications for the mixer under consideration require: 4dB≤G≤5dB, 1dB CP≥–0.5dBm and IP3≥9dBm. It has been observed that 48.57% of the instances involved in the calibration procedure are found to comply with the specifications before calibration. After applying the proposed calibration technique, the amount of compliant mixer instances corresponds to 88.57%, which indicates a +82.35% relative yield improvement. Similar improvement has also been reported for different specification requirements [11].

3 A Built-In Voltage Measurement Technique for the Calibration of RF Mixers

The proposed built-in technique addresses the problem of accessibility to the alternate test response signals that are necessary for the conduction of an RF mixer calibration procedure. The procedure described in the previous section is adopted.

3.1 Design and Implementation

By utilizing a ring-type voltage-controlled oscillator (VCO) and a counter, a low-cost time-based analog to digital converter (ADC) is constructed which is used as a voltage acquisition circuit (VAC, shown in Fig 5) that provides digital readings for the alternate test voltage observables.

A setup that allows the application of the test stimulus and the connection of the VAC to the appropriate mixer node for DC voltage acquisition follows the

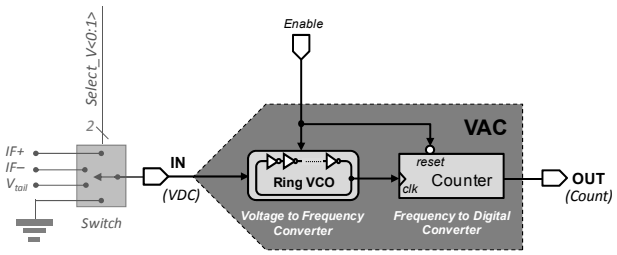


Fig. 5. Voltage acquisition circuit (VAC)

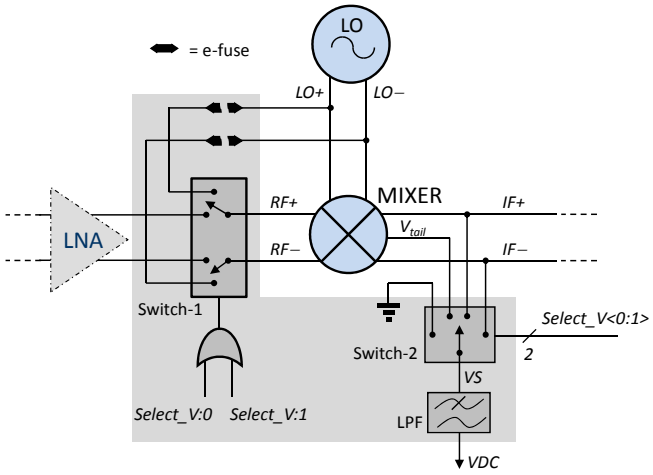


Fig. 6. RF mixer design modifications (shaded area)

scheme presented in Fig. 6, which illustrates the case of a differential RF mixer in a receiver. However, the proposed setup can be easily extended to cover both mixers in transceiver circuits, following the shared resource approach presented in [12]. In order to provide a built-in solution, an analog switch (Switch-1) disconnects the mixer's differential input from the low noise amplifier (LNA) and connects it to the LO. A second analog switch (Switch-2), as presented in Fig. 6, is used to select a voltage observable among IF+, IF− and V_{tail}, one at a time and also provides the ground level required for the correction of the VAC readings, as it is explained in [13]. A common RC low-pass filter (LPF) is connected to the output of the second switch to reject any remaining high-order frequency components and to provide a DC voltage signal (VDC).

Aiming to avoid the influence of the LO signal on the RF signal path in the normal mode of operation, through the first switch, low cost electrical fuses (e-fuses) or laser-cut fuses can be optionally exploited to eliminate the LO-RF test path after the completion of the measurements procedure.

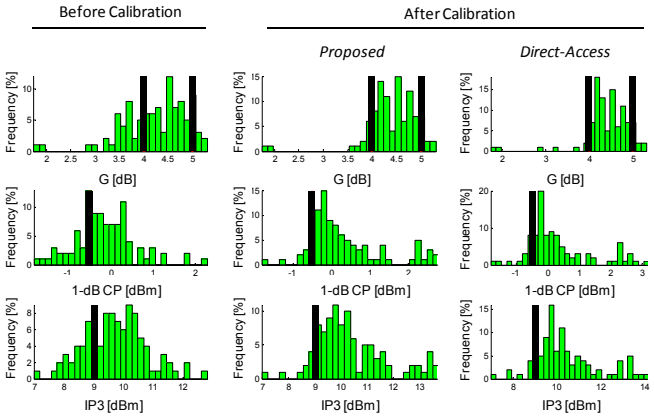


Fig. 7. Distributions of performance characteristics, before and after calibration

3.2 Evaluation

To assess the efficiency of the calibration procedure, as conducted using the proposed measurement technique, an evaluation set consisting of extremely perturbed defect-free mixer instances was generated using Monte Carlo simulations. Specification requirements for the mixer under consideration were defined as follows: 4dB≤G≤5dB, 1 dB CP≥0.5dBm and IP3≥9dBm.

The reported efficiency of the calibration procedure is illustrated in Fig. 7 where the distributions of performance characteristics before and after calibration are presented both for the proposed technique and the direct-access case, for the sake of comparison. Bold vertical lines indicate the margins of acceptable performance as set by the specifications. For the mixer's specifications under consideration, 42% of the instances involved in the calibration procedure are found to comply with the specifications before calibration. After the application of the proposed calibration procedure, the amount of compliant mixer instances corresponds to 75%, which indicates a +78.6% yield improvement, while for the direct-access approach a slightly higher yield (77%) is reported, which corresponds to a difference of only 2.5%.

4 Adjustable RF Mixers' Alternate Test Efficiency Optimization by the Reduction of Test Observables

Alternate tests for adjustable RF mixers are considered, where the alternate test response (ATR) consists of DC voltage levels while the mixer operates in homodyne mode, as presented in the previous sections. Selection techniques are proposed to determine the set of optimum test response observables. This is a subset of all available observables, obtained from all states, which is selected through specific optimization procedures in order to minimize a certain cost criterion that incorporates both test accuracy and complexity.

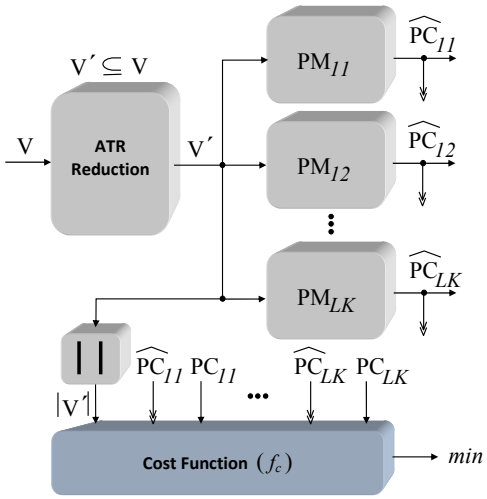


Fig. 8. Principle of ATR reduction (global approach)

4.1 Methodology

The proposed methodology aims to reduce the number of observables that are used as inputs to the predictive models, without a significant compromise of the corresponding alternate test accuracy. It has been found [14] that a 'global' approach is more efficient compared to its 'local' counterpart. Instead of the selection of an optimum subset of observables per individual model, the global ATR reduction approach attempts to minimize a cost function using a single common subset V' of the full set of potential model inputs (V) for all predictive models (PM_{lj}), as shown in Fig. 8. Actual performance characteristic values (PC_{lj}) and their predicted counterparts ($\widehat{PC_{lj}}$) corresponding to all predictive models are used as inputs to the cost function in order to provide a global measure of the corresponding accuracy, while the cardinality $|V'|$ of the common reduced subset of observables is used as a measure of test complexity.

Since an exhaustive examination of all subsets V'_{lj} of V in order to find the optimum subset which leads to a cost function minimum would require a rather large number of predictive model construction and evaluation trials, low complexity selection algorithms are adopted, namely sequential forward selection (SFS) and sequential backward selection (SBS). Furthermore, the inherent input selection capabilities of the regression algorithm (i.e., the MARS algorithm [8]) are explored, as well.

4.2 Evaluation

By the application of the proposed methodology, conduction of alternate tests on the adjustable RF mixer presented in the previous sections has led to the results

Table 3. Prediction error vs. test complexity for responses obtained by different global observable selection methods

SFS

error (e)	<1%	<1.2%	1.6%
reduction of observables	5/15 (33%)	7/15 (47%)	10/15 (67%)
accuracy degradation *(with respect to minimum achievable error)*	4.5%	28.1%	79.8%

SBS

error (e)	<1%	<1.1%	1.3%
reduction of observables	6/15 (40%)	7/15 (47%)	9/15 (60%)
accuracy degradation *(with respect to minimum achievable error)*	5.2%	19.7%	45.5%

MARS

error (e)	<1%	<1.3%	<1.6%
reduction of observables	6/15 (40%)	9/15 (60%)	11/15 (73%)
accuracy degradation *(with respect to minimum achievable error)*	6.7%	38.2%	76.4%

shown in Table 3. These results indicate that several cases exist for which a significant reduction of observables is associated to only a small accuracy degradation with respect to the minimum achievable error. For example, a 33% reduction of observables is followed by an accuracy degradation of only 4.5% in the SFS case presented in Table 3. However, even when a significant relative accuracy degradation is reported, a relatively low corresponding absolute variation is observed (i.e. an increase of 0.6% in the absolute error corresponds to a 76.4% relative accuracy degradation).

5 Conclusions

In this dissertation we have shown that testing and calibration procedures for integrated RF circuits can be viewed in a common framework, leading to reliable low-cost solutions. High defect coverage and significant reduction in parametric yield loss are reported, while the prediction of the performance characteristics is significantly improved by exploiting the adjustable features of the RF circuit under test.

It has also been shown that it is feasible to conduct highly efficient calibration procedures on integrated RF circuits, even when the measured alternate test response consists of voltage components which appear at internal nodes, overcoming the accessibility limitations met in embedded systems. A significant reduction in parametric yield loss is reported for the proposed built-in techniques,

which is very close to that achieved by direct measurements of the alternate test response.

Finally, it has been proven that alternate test complexity or, equivalently, cost can be further reduced by the selection of the optimal test response, with a negligible degradation of accuracy.

References

1. SIA - The International Technology Roadmap for Semiconductors. [Online]. Available: `http://public.itrs.net`
2. P. Variyam, S. Cherubal, and A. Chatterjee, "Prediction of analog performance parameters using fast transient testing," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 21, no. 3, pp. 349–361, Mar. 2002.
3. E. Garcia-Moreno, K. Suenaga, R. Picos, S. Bota, M. Roca, and E. Isern, "Predictive test strategy for CMOS RF mixers," *Integration, the VLSI Journal*, vol. 42, pp. 95–102, Jan. 2009.
4. S. R. Nassif, "Design for variability in DSM technologies," in *Proc. IEEE 1st Int. Symp. Quality Electron. Des. (ISQED)*, San Jose, CA, USA, Mar. 2000, pp. 451–454.
5. R. Goering and R. Wilson. (2003, Mar.) Yield, packages hang up design below 100 nm, EE Times. [Online]. Available: `http://www.eetimes.com`
6. T. Das, A. Gopalan, C. Washburn, and P. Mukund, "Self-calibration of input-match in RF front-end circuitry," *IEEE Trans. Circuits Syst. II*, vol. 52, no. 12, pp. 821–825, Dec. 2005.
7. A. Goyal, M. Swaminathan, and A. Chatterjee, "Self-calibrating embedded RF down-conversion mixers," in *Proc. IEEE Asian Test Symp. (ATS)*, Taichung, Taiwan, 2009, pp. 249–254.
8. J. H. Friedman, "Multivariate adaptive regression splines," *Ann. Stat.*, vol. 19, no. 1, pp. 1–141, 1991.
9. I. Liaperdos, L. Dermentzoglou, A. Arapoyanni, and Y. Tsiatouhas, "Fault detection in RF mixers combining defect-oriented and alternate test strategies," in *Conf. Design of Circuits and Integrated Systems (DCIS)*, 2011, pp. 315–320.
10. I. Liaperdos, L. Dermentzoglou, A. Arapoyanni, and Y. Tsiatouhas, "A test technique and a BIST circuit to detect catastrophic faults in RF mixers," in *Conf. Design and Technology of Integrated Systems in the Nanoscale Era (DTIS)*, 2011, paper st1a.
11. J. Liaperdos, A. Arapoyanni, and Y. Tsiatouhas, "A test and calibration strategy for adjustable RF circuits," *Analog Integrated Circuits and Signal Processing*, vol. 74, no 1, pp. 175–192, Jan. 2013.
12. L. Dermetzoglou, J. Liaperdos, A. Arapoyanni, and Y. Tsiatouhas, "Testing wireless transceivers' RF front-ends utilizing defect-oriented BIST techniques," in *Proc. 19th IEEE Int. Conf. Electronics, Circuits and Systems (ICECS)*, Seville, Dec. 2012, pp. 961–964.
13. J. Liaperdos, A. Arapoyanni, and Y. Tsiatouhas, "A built-in voltage measurement technique for the calibration of RF mixers," *IEEE Trans. Instr. Meas.*, vol. 62, no 4, pp. 732–742, Apr. 2013.
14. J. Liaperdos, A. Arapoyanni, and Y. Tsiatouhas, "Adjustable RF mixers alternate test efficiency optimization by the reduction of test observables," *IEEE Trans. Comp.-Aided Des. Integ. Circ. and Syst.*, vol. 32, no 9, pp. 1383–1394, Sep. 2013.

Interconnection between a Satellite Interactive Network and wireless broadband networks

Anastasia Lygizou*

Dept. of Informatics and Telecommunications
National and Kapodistrian University of Athens, Greece

Abstract. This doctoral thesis deals with the problem of interconnection between a satellite interactive network and broadband networks. An interconnection mechanism is proposed, which consists of three parts: 1) an entity at the satellite terminal responsible for capacity requests, 2) resource allocation to the satellite terminals and 3) sharing the capacity of a satellite terminal among the subscribers of the broadband network. The main contribution of the proposed mechanism is the introduction of a prediction mechanism in the first part for bandwidth requests. Simulation results show that the proposed mechanism can provide a differentiated treatment to services of different type, leading to improved performance, especially in terms of throughput and packet delay, compared to a simpler one that does not use prediction.

We propose solutions for all three parts of the interconnection mechanism aiming to improve the overall performance of the system, especially for real time traffic that can tolerate less delay. The NLMS (Normalized Least Mean Square) algorithm is chosen to be used in the first part of the proposed mechanism. In addition, we extend the second part of the mechanism for performing the slot allocation in MF-TDMA. Finally, we improve the scheduler of the third part of the proposed mechanism. The target of this improvement is to schedule traffic of real time connections of the broadband network based on Quality of Experience (QoE) metrics. We propose a new quality of experience metric and a new algorithm, based on this metric, for resource allocation. This use of QoE metrics for scheduling is rather novel, since the main use of such metrics so far has been for the assessment of video quality. Simulation results show that the proposed algorithm attains a sizeable reduction of the mean delay and a considerable improvement of the quality of experience for video connections.

Keywords: traffic scheduling, quality of service, quality of experience, video prediction

1 Introduction

IEEE 802.16 1 is a standard that aims at filling the gap between local and wide area networks, by introducing an advanced system for metropolitan environments. In such a system, also known as WiMAX, both point-to-multipoint (cellular) and mesh mode configurations can be supported, while node mobility is also covered by the recent amendment 802.16e [2]. One of the main advantages of the standard is the large de-

*Dissertation Advisor: Lazaros Merakos, Professor

gree of flexibility it provides by supporting a wide range of traffic classes with different characteristics and quality of service (QoS) requirements. This is attained through a large set of parameters that allow users to describe in detail their traffic profiles and service needs. On the other hand, the standard of Digital Video Broadcasting – Return Channel Satellite (DVB-RCS) ([3],[4]) describes the uplink direction of a satellite network, providing advanced QoS capabilities for requesting and acquiring capacity for demanding services.

The advantage of combining the two technologies is that a satellite network can be used for interconnecting WiMAX islands with the Internet and avoiding layout of expensive backbone infrastructures. This can be advantageous, especially in rural areas or locations affected by environmental factors, e.g. islands, mountains, etc. However, a satellite network experiences large round trip delays that can deteriorate quality especially for real-time applications. In this doctoral thesis, we investigate how the two networks can co-operate, especially in terms of QoS, in order to reduce end-to-end delays and packet losses due to expiration.

In the literature there is a number of proposals for interworking between satellite and WiMAX networks, however these focus mainly on the architectural aspects. For example, in [5] a resource reservation scheme for a hybrid wireless scenario, between satellite, WiMAX, and WiFi networks is proposed, based on the use of intelligent mobile agents, while [6] presents a new hybrid network solution relying on synergy between IEEE 802.16-based and DVB-RCS networks.

A set of proposals focus on the way a satellite terminal makes its capacity requests. These proposals follow mainly a bandwidth-on-demand approach. In [7] and [8], a dynamic bandwidth allocation in satellite networks is addressed, using adaptive predictive control methods. In [9], a new connection admission control algorithm is proposed with the aim of efficiently managing only real-time multimedia video sources both with constant and high variable data rate transmissions. According to [10], whenever a capacity assignment is performed, the NCC predicts the length of the queues which will be experienced at the RCST. This prediction is achieved through a sliding window based mechanism tailored to the satellite traffic. In [11], an efficient but complex method for optimal timeslot scheduling in an interactive satellite multimedia network is developed, so that the system's (weighted) throughput is maximized. The timeslot assignment problem is formulated as a binary integer programming problem, with a vast number of decision variables.

Finally, in [12] a dynamic traffic management strategy for the return channel of a DVB-RCS satellite system has been presented, consisting of two parts: the Resource Manager scheduler and the Terminal scheduler.

2 Dissertation Summary

In [14], we propose an interconnection of a satellite and a WiMAX network, assuming that one or more of the RCSTs are also WiMAX BSs serving a number of SSs as shown in Fig. 1. This integrated scheduling provision mechanism consists of three main parts:

PartA is an entity at the RCST/BS that makes the capacity requests following a prediction-based approach, PartB is an entity at the NCC that allocates resources and creates the TBTP, while PartC is an entity at the RCST/BS that shares the given capacity among its WiMAX subscribers. PartB accepts the capacity requests made from all PartAs, processes them and creates the TBTP in order to allocate the capacity of a superframe among the different RCSTs. PartC, located at the RCST/BS, contains the scheduling algorithm that is responsible to share the allocated capacity, to the uplink traffic arriving from the WiMAX network. In more detail, PartC classifies uplink traffic arriving from the SSs into five queues: UGS_queue, rtPS_queue, ertPS_queue, nrtPS_queue, BE_queue based on each packet's QoS service type. It then interprets TBTP (knows exactly which slots has been assigned to it) and selects which packets will be transmitted. This selection is made based on a priority scheme: it first selects packets from the UGS_queue, then from the rtPS_queue, then from the ertPS_queue, then from the nrtPS_queue and finally from BE_queue. Finally, it is also responsible to discard packets that are expired based on the deadlines set for their transmission to the satellite network and keep statistics on the packets transmitted and discarded.

The main contribution of the proposed mechanism is that it takes into account the different QoS characteristics of the WiMAX traffic and proposes a prediction mechanism used in PartA for bandwidth requests. Simulation results show that the proposed mechanism can provide a differentiated treatment to services of different type, leading to improved performance, especially in terms of throughput and packet delay, compared to a simpler one that does not use prediction.

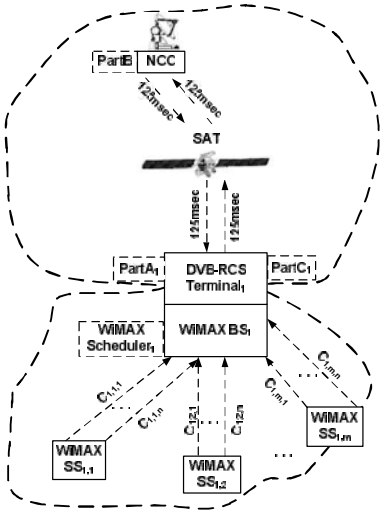


Fig. 1. Network architecture with three BS/RCST

An accurate traffic predictor for a satellite terminal is crucial in order to enhance channel utilization and guarantee the QoS requirements of real-time connections. In [15] and [16], we extend [14] towards improving the video prediction mechanism.

The traffic prediction algorithm used in PartA in [14] is based on the mean data rate of the connections, named BMDR (Based on Mean Data Rate).

After bibliographic search, we can conclude that there are three main categories of methods for traffic prediction. The first one uses the characteristics of traffic of real data, like self-similarity and long range dependency and tries to model the traffic in order to make the prediction ([17], [18], and [19]). The main drawback of this category is that the values of parameters of the different models must be predetermined, in order to achieve the optimal performance, particularly for real-time videos in which the traffic characteristics are unknown in advance. This is the main reason we do not consider it in our selection.

The second category uses neural networks [20], which are powerful tools for traffic prediction, but their implementation can be quite complicated resulting in large computational overheads. Besides, the training procedure of a neural network may suffer slow convergence and can be time consuming. These disadvantages make neural networks unattractive for use in applications with limited computational capability like satellite networks.

The third category make stochastic prediction of data that may arrive in the queue in the time interval between a request is made and the time this request is granted ([21]- [23] and [24]). We consider the third category as the most appropriate for our goal. Among them, [21] and [24] show better results.

For the prediction of rtPS traffic, the NLMS algorithm is used with three different alternative mechanisms. The first one proposes the implementation of the NLMS algorithm in the WiMAX BS, the second one proposes the implementation of the NLMS algorithm in the satellite terminal, while the third one proposes the implementation of the NLMS algorithm in both the WiMAX BS and the satellite terminal.The simulation results and the complexity analysis lead us to choose the second alternative, named VPNLMSb, as the most suitable for our system, which is presented and evaluated.

In [25], we start with a bibliographic search for slot allocation methods in static MF-TDMA satellite systems, followed by a multi-frequency extension of the algorithm used in PartB. Simulation results show improved performance of slot allocation in MF-TDMA, especially in terms of throughput and delay.

[26] improves the scheduler of rtPS connections in Part C based on Quality of Experience (QoE) metrics. After a bibliographic search on QoE metrics, the FC-MDI (Frame Classification-Media Delivery Index) metric is chosen to be used in the proposed algorithm named FC_MDI_S, for the scheduling of real time connections. Two versions of the algorithm are proposed and evaluated. Simulation results show that the proposed algorithm considerably improves the QoE and the mean delay of the real time connections.

3 Scheduler based on QoE metrics

The target of this part of the doctoral thesis is to improve a previously proposed mechanism, in order to make the scheduling of rtPS connections based on the use of

QoE metrics. QoE metrics are usually used for the assessment of the transmission of video on different network conditions, and rarely used in scheduling solutions, while they have never been used till now for scheduling in satellite networks. Subjective metrics are the most accurate for QoE measurements, as they are evaluated by real-human. Their main shortcoming is that they are time-consuming and high-cost in man power. Thus, they cannot be easily repeated several times nor used in real-time (being a part of an automatic process). As we need the proposed improvement to be part of an automatic procedure, subjective and hybrid QoE metrics are excluded in our case. From the already proposed solutions in other kind of networks, the solutions proposed in [27-29] have the drawback of using the PSQA metric for scheduling and QoE management. On the other hand, the solution proposed in [30-31] is considered complex, as it calculates the QoE produced by every possible packet dropping. Our proposal aims to be simpler in order to be used in satellite networks, which have the drawback of delays. For all these reasons, the FC-MDI metric was chosen to be used in the existing mechanism [32], as it is an objective metric that gives a different weight to the loss of different categories of voice and video frames. The FC-MDI (Media Delivery Index based on Frame Classification) metric is an extension of the MDI (Media Delivery Index) metric [33], an objective metric that contains two numbers separated by colon: the delay factor (DF) and the media loss rate (MLR). DF is time value indicating how many milliseconds the buffer must be able to contain to eliminate jitter, while MLR is computed difference between number of media packets received during an interval and number of media packets expected during an interval, everything scaling in the value of one second. Because the MLR is a rate, some important information is lost, such as whether the IP packets lost are consecutive or not. It does not consider the quality degradation that suffered some propagated loss from previous temporally related frames, so [33] proposes FC-MDI which takes frame classification into account to improve the performance of the MDI measurement. It distinguishes the packet loss based on the frame classification, and gives the different frame a different weight. In all types of frames, I-frame plays the most important role, as the rest frame of the whole group of picture (GOP) cannot decode normally if the I-frame is lost. Compared with B-frame, P-frame relies less on its previous I-frames and P-frames. FC-MLR (Media Loss Rate based on Frame Classification) improves the definition of the MLR and takes frame classification into account as follows:

$$FC - MLR = \frac{\alpha * I_{PLoss} + \beta * P_{PLoss} + \gamma * B_{PLoss}}{interval},$$

where α, β, γ are weights with ($3 \geq \alpha > \beta > \gamma \geq 0$, $\alpha + \beta + \gamma = 3$) and I_{PLoss}, P_{PLoss}, and B_{PLoss} are respectively the number of lost I, P and B frames. The results of experiments demonstrate that when two videos of different qualities have a same number of total dropped-packet, the traditional MDI measurement cannot tell the difference between them, as MDI does not take into account the quality degradation that suffers some propagation loss from previous temporally related frames, while FC-MDI possesses a distinguishing feature.

The FC-MDI takes frame classification into account by giving different weights to the number of I-frames lost, P-frames lost and B-frames lost. However, it does not take into account if the frames lost from a specific category are consecutive or not,

which makes a difference. In [34], the LA-MDI is proposed (which is an improvement of FC-MDI), in order to give a greater importance to the consecutive lost frames of a specific category. In the LA-MDI, the definition of the DF is the same with its definition in the simple MDI, where the LA-MLR improves the definition of the FC-MLR in order to take into account the consecutive lost frames as follows:

$$LA-MLR = \frac{\alpha*\frac{I_{PLoss}}{ngI}+\beta*\frac{P_{PLoss}}{ngP}+\gamma*\frac{B_{PLoss}}{ngB}}{interval},$$

where α, β, γ are weights with ($3 \geq \alpha > \beta > \gamma \geq 0$, $\alpha + \beta + \gamma = 3$), I_{PLoss}, P_{PLoss}, and B_{PLoss} are respectively the number of lost I, P, B frames, and ngI, ngP, and ngB are respectively the number of group of lost I, P, B frames. The greater the number of group of lost frames, the more dispersed the lost frames are, and so the QoE is better. Generally, as the FC-MLR and the LA-MLR grows, the QoE becomes worse as the number of lost frames increases.

We further improve the scheduling algorithm of PartC, in order to make the scheduling of rtPS connections based on the use of the proposed LA-MDI metric.

In the beginning of every superframe, the proposed algorithm, referred to as LAQoE, drops the packets that are expired due to delay factor. Then, it sorts the video connections based on their mean LA-MLR. The mean LA-MLR of a connection in superframe t is defined as $\frac{\sum_{i=t-T}^{t-1} LA-MLR_i}{T}$, where T is a small number of superframes (time window), in order to reflect the quality of the connection in the recent past. Two alternatives are studied for sorting the connections according to the mean LA-MLR. The first alternative is named LAQoEG and has a greedy logic. In order to preserve the connections that have good quality, the connections are sorted based on mean LA-MLR in ascending way, from the best quality to the worst. This may lead to the maintenance of the quality of some connections and the starvation of some others. The second alternative is named LAQoEF and has a fair logic. In order to be fair and maintain all connections (even in worse quality), the connections are sorted in the opposite way than the previous algorithm based on the mean LA-MLR of the connections from the worst quality to the best.

In the beginning of every superframe, the PartC has accepted the TBTP generated from the NCC, so it has the knowledge of the available capacity for transmission. For every connection with the order of the previous sorting, the PartC creates a binary tree named QoE Tree (QoET) based on the available capacity for this connection.

PartC knows from PartA the sequence of packets that have arrived during the previous superframe. For every rtPS connection, PartC constructs a QoET that represents the possible combinations of packet transmission in this superframe. If, for example,

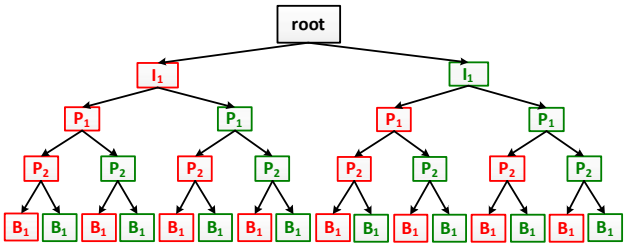


Fig. 2. Example of a QoE tree

PartC wants to transmit the sequence of I1P1P2B1 packets, then the QoET that is constructed is shown in Fig. 2.

Every path of the tree represents a combination of packet transmission, where a red node shows that a packet is not transmitted and a green node that a packet is transmitted. Knowing the TBTP, PartC can compute if a packet will expire due to delay before it's time for transmission. If the packet expires, then naturally it is not transmitted. In addition, the construction of a path stops, if its capacity comes to the available capacity that this connection has for transmission.

The leaves of the tree also contain the information of the LA-MLR metric for the specific path, which is easy to compute as we know the sequence of lost frames from every different category, as well as the total amount of bytes to be transmitted.

The LAQoE selects the path (sequence of packets) from the QoET of this connection with the best LA-MLR value. The available capacity for the next connection is reduced by the size of transmitted bytes of the selected path.

During the superframe, the PartC transmits, whenever it has available capacity based on the TBTP, the packets from the path selected of a connection based on the order of the sorted connections. If the packets of the selected paths of all connections are transmitted and PartC has still available capacity, then it transmits packets that have arrived in this superframe, using the logic of the FC_MDI_S algorithm. The transmission of these packets as well as the dropping of the packets is admeasured to the computing of the LA-MLR of the connections to the next superframe.

The LAQoERA algorithm is an improvement of the LAQoE algorithm that makes rate adaptation. PartC has the possibility of transmitting video in three rates: high, medium and low. Low quality is corresponded to rate 1, medium quality to rate 2 and high quality to rate 3. The greater the LA-MLR metric becomes, the worse it is. In the LAQoERA algorithm, the corresponding path of the QoET is not able to be transmitted upon a LA-MLR threshold. Instead, the connection transmits to a lower quality. If it is already in the lowest quality, then the connection transmits the best path that it is able to.

The LAQoERA algorithm differentiates the sorting of video connections, the creation of the QoET and the selection of the transmitting path so as to take into account the rate of video connections. The sorting of the video connections is based on the mean LA-MLR of the connections and the mean rate (mR) of the connections. The mR of a connection in superframe t is defined as $\frac{\sum_{i=t-T}^{t-1} Rate_i}{T}$, where T is a small number of superframes (time window), in order to reflect the rate of the connection in the recent past. The connections are sorted according to the mean LA-MLR and mR under two versions. The first version is named LAQoERAG and has a greedy logic. In order to preserve the connections that have good quality, the connections are sorted based on mR in descending way, from the best rate to the worst, and then based on mean LA-MLR in ascending way, from the best quality to the worst. The second version is named LAQoERAF and has a fair logic. In order to be fair and maintain all connections (even in worse quality), the connections are sorted in the opposite way than the previous algorithm based on mR in ascending way, from the worst rate to the

best, and then based on mean LA-MLR in descending way, from the worst quality to the best.

The difference in the creation of the QoET from the previous algorithm is that there are flags in every path showing if this path is able to be transmitted in rate 3, rate 2 and rate 1. The flag of one rate becomes false only when the capacity of a path overcomes the available capacity of this connection. If the flags of the three rates are false, then the path stops. In addition there are flags in the whole tree showing the existence of a path in rate 3, rate 2 or rate1.

Finally, the LAQoERA algorithm selects the path (sequence of packets) with the best LA-MLR value in the best rate that this connection has the ability to transmit. This is shown from the flags of the QoET. If the flag of the whole tree in rate 3 is true, then the path with the best LA-MLR metric will be selected (from these paths that have the respective flag in rate 3 set to true). If the LA-MLR metric of the selected path is over a threshrate (threshrate3), then PartC prefers to transmit in lower grade but in better quality. The same procedure is repeated for rate 2. If the path selected in rate 2 has the LA-MLR metric over a threshrate (threshrate2), then PartC will select the path with the best LA-MLR metric in rate 1.

The available capacity for the next connection is reduced by the size of transmitted bytes of the path selected in the respective rate.

In order to measure the performance of the proposed algorithms, we accommodated the simulation program presented in [14]. The program is constructed in C++ and simulates the full operation of WiMAX network, as well as the DVB-RCS for the return link of a satellite network. We use the simulation scenario presented in [14] with three DVB-RCS terminals each one interconnecting a WiMAX network, all with the same number of subscribers. In order to present the difference of the proposed mechanisms regarding the QoE of the video connections, in the present simulation scenario every SS has only one video connection. The same video trace is used for every SS, in order to present the difference between the greedy and fair versions. The source of this video trace is the "Alladin" film from "http://trace.eas.asu.edu/TRACE/ltvt.html" in high quality ("Verbose_Alladin.dat" file). Especially, for the LAQoERAG and LAQoERAF algorithms, we also use the same video trace in medium ("Verbose_Alladin_10.dat" file) and low quality ("Verbose_Alladin_10_14_18.dat" file).

Fig. 3 presents that the LAQoE, and LAQoERA algorithms reduce the mean delay of the video connections. This is due to the philosophy of the algorithms that take into account the TBTP to the construction of QoET and the selection of packets for transmission with the best QoE metric. This is a substantially improvement, as we prefer video connections to have reduced delay.

The two lastly proposed algorithms use the LA-MLR metric for their QoE evaluation. Fig. 4 shows that the FC_MDI_S and the LAQoE algorithms have the same mean FC-MLR value, while Fig. 5 shows that the LAQoE algorithm improves the LA-MLR value regarding to the FC_MDI_S one, as it takes account the number of group of lost frames of different categories. This is a proof of the differentiation and improvement of the LA-MLR metric. Fig. 5 presents the mean LA-MLR value for all connections of a SS. This figure shows that the two lastly proposed algorithms sub-

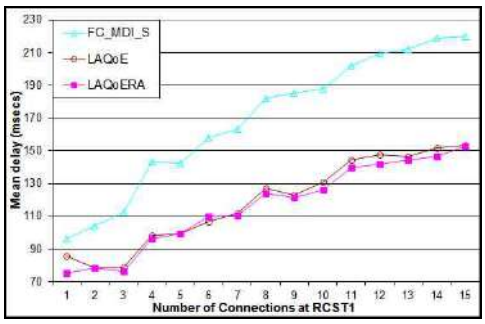


Fig. 3. Mean delay per proposed algorithm

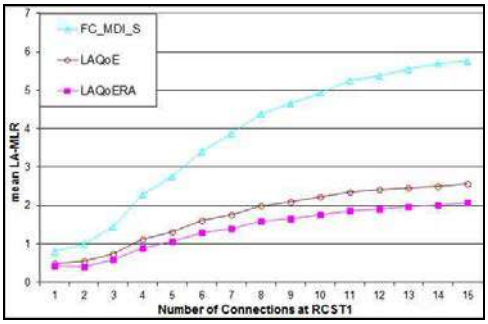


Fig. 5. Mean LA-MLR per proposed algorithm

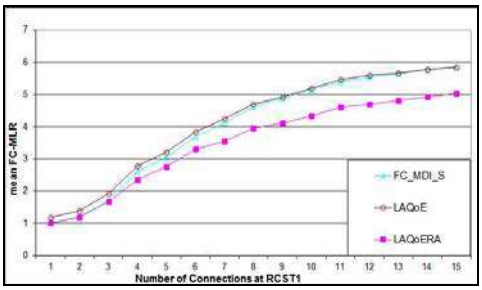


Fig. 4. Mean FC-MLR per proposed algorithm

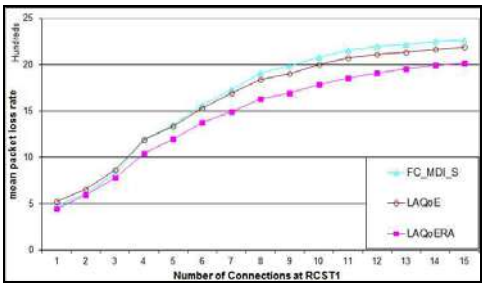


Fig. 6. Loss rate per proposed algorithm

stantially improve the QoE performance of the video connections. Especially, the LAQoERA algorithm has the best QoE performance. This is due to the rate adaptation of this algorithm, which loses the least of the transmitted information. It may transmit in lower quality but it transmits more information. This is better presented in Fig. 6, which presents the percentage of lost bandwidth.

From the presented results, we conclude that the LAQoE algorithm further reduces the mean delay of the connections, and improves the QoE performance of the video connections relatively to the FC_MDI_S algorithm. This is due to the philosophy of this algorithm which serves the sequence of packets with the best QoE metric. Finally, the LAQoERA algorithm has the best mean delay and QoE performance for video connections, as it loses less of the transmitted information due to the rate adaptation that it makes.

4 Conclusions

In this doctoral thesis, a scheduling mechanism for interconnecting satellite and WiMAX networks is presented. The proposed mechanism consists of three parts: PartA, located at the RCST, is responsible for making capacity requests, PartB, located at the NCC, is responsible for assigning bandwidth per RCST and creating the TBTP, while PartC, located at the RCST, is responsible for sharing the given capacity among its WiMAX subscribers. The main contribution of the proposed mechanism is that it takes into account the different QoS characteristics of the WiMAX traffic and

proposes a prediction mechanism used in PartA for bandwidth requests. Simulation results show that the proposed mechanism can provide a differentiated treatment to services of different type, leading to improved performance, especially in terms of throughput and packet delay, compared to a simpler one that does not use prediction.

Following a bibliographic search for video prediction in WiMAX and satellite networks, we select NLMS as the most suitable algorithm in order to improve the prediction of rtPS traffic in joint WiMAX/Satellite networks. Although the NLMS algorithm has been proposed for prediction of traffic in satellite networks, it is novel the research of the most effective way using it for prediction of video in an integrated satellite and WiMAX network. Three alternatives for the extension of an existing scheduling mechanism were investigated. Both the simulation results and the complexity analysis lead us to choose the second alternative, named VPNLMSb, as the most suitable for our system, which is presented and evaluated.

A bibliographic search for slot allocation in static MF-TDMA satellite systems is followed by multi-frequency extension of the scheduling algorithm lays in PartB proving the multi-frequency capability. Simulation results show improved performance, especially in terms of throughput and packet delay, compared to single frequency slot allocation.

We further improve the proposed scheduling algorithm used in PartC named RTFS. This algorithm is responsible to share the allocated capacity to the uplink traffic arriving from the WiMAX network in an integrated satellite/WiMAX network. After a bibliographic search for QoE metrics in WiMAX and satellite networks, the FC_MDI QoE metric is selected to be used in the proposed algorithm named FC_MDI_S. This is considered novel, as QoE metrics are mainly used for the assessment of video quality and not for scheduling. Especially in satellite networks, QoE metrics have never been used in management tools. We proposed and evaluated two versions for FC_MDI_S, and simulation results show that it considerably improves the QoE of video connections and reduces their mean delay.

Finally, we propose an improvement of the FC_MDI metric named LA_MDI. We propose and evaluate two alternative algorithms based on this new metric named LAQoE and LAQoERA. The second algorithm is an improvement of the first one that also makes rate adaptation. Simulation results show that the proposed algorithms, and especially the second one, considerably improve the QoE of video connections and reduce their mean delay.

5 References

1. D IEEE Std 802.16-2004, “IEEE Standard for Local and Metropolitan Area Networks – Part 16: Air Interface for Fixed Broadband Access Systems”, October 2004.
2. IEEE Std 802.16e-2005, “Amendment to IEEE Standard for Local and Metropolitan Area Networks - Part 16: Air Interface for Fixed Broadband Wireless Access Systems- Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands”, February 2006.
3. ETSI EN 301 790 V1.4.1, “Interaction channel for satellite distribution systems”, September 2005.

4. ETSI TR 101 790 V1.3.1, “Interaction channel for Satellite Distribution Systems; Guidelines for the use of EN 301 790”, September 2006.
5. E. Puşchiţă, T. Palade, and A. Căruntu, “Design of a resource reservation model for a hybrid wireless network architecture based on the use of intelligent mobile agents”, Eighteenth International Conference on Systems Engineering (ICSE2006), September 2006.
6. A. Centonza and S. McCann, “Architectural and Protocol Structure for Composite DVB-RCS/IEEE 802.16 Platforms”, IET Seminar on Digital Video Broadcasting over Satellite: Present and Future, London, UK, November 2006.
7. L. Chisci, R. Fantacci and T. Pecorella, “Dynamic Bandwidth Allocation via Distributed Predictive Control in Satellite Networks”, First International Symposium on Control, Communications and Signal Processing, Hammamet, Tunisia, March 2004.
8. L. Chisci, R. Fantacci and T. Pecorella, “Predictive Bandwidth Control for GEO Satellite Networks”, IEEE International Conference on Communications, Paris, France, June 2004.
9. P. Pace and G. Aloi, “Effective Admission Policy for Multimedia Traffic Connections over Satellite DVB-RCS Network”, ETRI Journal – Electronic and Telecommunications Research Institute, 2006, Vol. 28, No. 5, pp. 593-606.
10. F. Priscoli, D. Pompili and G. Santoro, “A QoS-aware Bandwidth on Demand Assignment Mechanism in a GEO Satellite System”, EU Information Society Technology Mobile and Wireless Communications Summit (IST Summit), Lyon, France, June 2004.
11. K.-D. Lee, H.-J. Lee, Y.-H. Cho and D. Oh, “Throughput-Maximizing Timeslot Scheduling for Interactive Satellite Multiclass Services”, IEEE Communications Letters, 2003, Vol. 7, No. 6, pp. 263-265.
12. M.Costabile, C. Follino, A. Iera, and A. Molinaro, “QoS Differentiation in DVB-RCS multimedia platforms”, 15th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications”, Barcelona, Spain, September 2004.
13. A. Lygizou, S. Xergias, N. Passas, L. Merakos, “A prediction-based scheduling mechanism for interconnection between WiMAX and satellite networks”, International Wireless Communications and Mobile Computing Conference 2008 (IWCMC 2008), August 2008.
14. A. Lygizou, S. Xergias, N. Passas, L. Merakos, “A prediction-based scheduling mechanism for interconnection between WiMAX and satellite networks”, International Journal of Autonomous and Adaptive Communications Systems, 2009, Vol. 2, No.2, pp. 107-127.
15. A. Lygizou, S. Xergias and N. Passas, “Video traffic prediction for improved Scheduling in Joint WiMAX / Satellite Networks”, 8th International Wireless Communications and Mobile Computing Conference 2012 (IWCMC 2012), Cyprus, August 2012.
16. A. Lygizou, N. Passas, S. Xergias and L. Merakos, “Effective Prediction for Video traffic in Joint WiMAX / Satellite Networks”, in Springer Wireless Personal Communications, 2012, pp.21-39.
17. B.N. Bhandari, R.V. Kumar and S.L. Maskara, “Performance of IEEE 802.16 MAC layer protocol under conditions of self-similar traffic”, TENCON 2008- 2008 IEEE Region 10 Conference, Hyderabad, India, November 2008.
18. G. Boudour, “MPEG-4 Traffic Prediction Using Density Estimation for Dynamic Bandwidth Allocation in IEEE 802.16 Networks”, Global Telecommunications Conference (GLOBECOM 2011), December 2011.
19. F. Chiti, R. Fantacci and F. Marangoni, “Advanced Dynamic Resource Allocation Schemes for Satellite Systems”, IEEE International Conference on Communications, Seoul, May 2005.
20. G. Moayeripour, A. Aghakhani, M. N. Moghadam and H. Taheri, “Reducing Bandwidth Allocation Delay in a DVB-RCS Network Using Bayesian Neural Network”, 11th International Conference on Advanced Communication Technology, February 2009.

21. R. Mukul, P. Singh, D. Jayaram, D. Das, N. Sreenivasulu, K. Vinay, and A. Ramamoorthy, “An Adaptive Bandwidth Request Mechanism for QoS Enhancement in WiMAX Real Time Communication”, IFIP International Conference on Wireless and Optical Communications Networks, Singapure, July 2007.
22. Z.Peng, Z. Guangxi, L. Hongzhi and S. Haibin, “Adaptive Scheduling Strategy for WiMAX Real-time Communication”, International Symposium on Intelligent Signal Processing and Communication Systems, Xiamen, China, Nov 2007.
23. W.-K. Kuo, “Efficient Traffic Scheduling for Real Time VBR MPEG Video Transmission Over DOCSIS-Based HFC Networks”, Journal of Lightwave Technology, 2009, Vol. 27, No. 6, pp. 639-654.
24. P. Pace and G. Aloi, “Effective Prediction Scheme for Bandwidth Allocation in Interactive Satellite Terminals”, Wireless Communication Systems. 2008. ISWCS '08. IEEE International Symposium on, October 2008.
25. A. Lygizou, and N. Passas, "MF-TDMA slot allocation in Joint WiMAX / Satellite Networks", 6th ACM International Wireless Communications and Mobile Computing Conference (IWCMC 2010), Caen, France, June 2010.
26. A. Lygizou, S. Xergias and N. Passas, “rtPS Scheduling with QoE metrics in Joint WiMAX / Satellite Networks”, in 4th International Conference on Personal Satellite Services (PSATS), March 2012.
27. K. Piamrat, K.D. Singh, A. Ksentini, C. Viho and J.M. Bonnin, “QoE-aware scheduling for video-streaming in High Speed Downlink Packet Access”, Wireless Communications and Networking Conference (WCNC), 2010 IEEE, April 2010.
28. K. Piamrat, A. Ksentini, J.-M. Bonnin, and C. Viho, “Rate Adaptation Mechanism for Multimedia Multicasting in Wireless Networks”, Broadband Communications, Networks, and Systems, 2009. BROADNETS 2009. Sixth International Conference on, September 2009.
29. K. Piamrat, A. Ksentini, J.-M. Bonnin, and C. Viho, “Q-DRAM: QoE-based Dynamic Rate Adaptation Mechanism for Multicast in Wireless Networks” Global Telecommunications Conference, GLOBECOM 2009. IEEE, November 2009.
30. A.B. Reis, J. Chakareski, A. Kassler and S. Sargento, “Quality of experience optimized scheduling in multi-service wireless mesh networks”, Image Processing (ICIP), 2010 17th IEEE International Conference on, Sept. 2010.
31. A.B.Reis, J. Chakareski, A. Kassler and S. Sargento, “Distortion Optimized Multi-Service Scheduling for Next-Generation Wireless Mesh Networks”, INFOCOM IEEE Conference on Computer Communications Workshops, March 2010.
32. J Krejci, “MDI measurement in the IPTV”, Systems, Signals and Image Processing, 2008. IWSSIP 2008. 15th International Conference on, June 2008.
33. F. Shaofen and H. Liang; “A Refined MDI Approach Based on Frame Classification for IPTV Video Quality Evaluation”, Education Technology and Computer Science (ETCS), 2010 Second International Workshop on, March 201.
34. A. Lygizou, S. Xergias and N. Passas, “Improved rtPS Scheduling with QoE metrics in Joint WiMAX / Satellite Networks”, in the International Journal of P2P Network Trends and Technology, 2012, Vol. 2, No. 1, pp 4-12.

Distributed Signal Processing and Data Fusion Methods for Large Scale Wireless Sensor Network Applications

Dimitris V. Manatakis *

National and Kapodistrian University of Athens,
Department of Informatics and Telecommunications
dmanatak@di.uoa.gr

Abstract. In this PhD dissertation we study the problem of continuous object tracking using large scale WSNs. We propose a novel practical WSN-based scheme that is able to track and predict the evolution behavior of a continuous object's boundary under realistic assumptions. The proposed scheme consists of two main components: a) A collaborative in-network WSN algorithm that estimates the local evolution parameters (orientation, direction and speed) of an evolving continuous object, and b) a novel algorithm which combines the produced local estimates, as they become available to a fusion center, to reconstruct the overall continuous object's boundary. Extensive computer simulations demonstrate the ability of the proposed collaborative algorithm to estimate accurately the evolution characteristics of complex continuous objects (e.g. with time-varying evolution rates and/or irregular boundary shapes) using reasonably dense WSNs. Moreover, it shown that the algorithm is robust to sensor node failures and communication link failures which are expected in harsh environments. Finally, we show that the proposed boundary reconstruction algorithm is able to track with accuracy the evolution of different types of continuous objects, using a small number of local front estimates that may be distorted with error.

Keywords: machine learning, distributed estimation, Bayesian estimation, continuous object tracking, environmental hazards, wireless sensor networks.

1 Introduction

Wireless Sensor Networks (WSNs) is a rapidly maturing technology with a wide range of applications (e.g. target tracking, surveillance, environmental monitoring, patient monitoring to name a few). A WSN typically consists of a large number inexpensive autonomous electronic devices (sensor nodes) which are deployed over a geographical region and monitor physical or environmental parameters. Apart from "sensing" the environment, WSN nodes are also able to process

* Dissertation Advisor: Elias S. Manolakos, Associate Professor.

data and exchange information. Recent advances in microelectronics and wireless communication have made WSN technology an ideal candidate for large-scale decision and information-processing tasks.

Tracking objects (i.e. determining their location over time) has been a well studied problem with numerous civilian and military applications. Apart from finding the trajectory of the objects, it is also important to estimate their motion characteristics (i.e. direction and speed) in real time, since this information can be used to predict their future locations and understand their evolution behavior.

Wireless sensor network technology has been extensively used for single and multiple target tracking applications. Due to the rapidly dropping cost of the sensor nodes, WSNs are also gaining popularity in environmental monitoring applications. Recently, sensor network-based methods have been proposed for detecting the boundaries of diffusive hazardous phenomena [1–3], modeled as "continuous objects" (such as expanding wildfires, oil spills, diffusing bio-chemical materials etc.). Continuous objects are usually spread in large regions and their size and shape is dynamically changing with time. The ability to track and predict, with reasonable accuracy, the location of a diffusive hazard's boundary is of paramount importance since it helps the authorities to organize efficiently their responses (hazard suppression, possible evacuation etc).

The key idea behind the reported WSN-based continuous object tracking methods has been an attempt to identify over time the sensor nodes located closest to the evolving object's front line (boundary nodes). Although these methods can estimate implicitly the boundaries of a continuous object (evolving hazard) using the locations of the boundary nodes, they have important limitations that renders them impractical for the development of real world application systems for hazard tracking.

The main limitations that appear in almost all reported WSN-based continuous object tracking schemes are:

L1: They require unrealistic sensor nodes densities (thousands sensors per km^2) to determine with reasonable accuracy the boundary of an evolving continuous object. Although, the cost of the sensor nodes has been significantly reduced, it still remains prohibitive to cover large geographical regions (many km^2) with high density WSNs.

L2: They do not consider node or communication failures. However, these failures are certainly expected in large scale WSNs applications, and especially in the harsh environments created by the evolving hazardous phenomena (e.g. wildfires).

L3: They require synchronization between the sensor nodes, a capability that is difficult to achieve even in small scale WSNs.

L4: They assume an idealized sensing mechanism (i.e. fixed sensor nodes detection distance, do not consider sensing functionality disruptions etc) that renders them impractical for hazard tracking.

L5: They are incapable to provide information about the spatiotemporal evolution characteristics (e.g. direction and speed) of the continuous object's

boundary. This limitation makes them incapable to be used for predictive modeling as part of decision support systems.

L6: They are incapable to assess the processing, memory and energy requirements before a real field deployment.

L7: They propose naive techniques to reconstruct the continuous object's boundary or are incapable to reconstruct it without using the human ability to identify the boundary's shape from the boundary nodes locations.

The main contribution of this dissertation is the conception, design and development of a WSN-based continuous object tracking scheme that addresses all the aforementioned limitations. We have to mention that most parts of the doctoral dissertation have been published in peer reviewed scientific journals and high quality referred Conferences Proceedings at the time of its writing.

2 Modeling Detection Distance Uncertainty

It is usually assumed that a sensor node can detect an event inside a disk area of radius R_d. Although this may not always hold in real applications, it is frequently adopted since it simplifies the analysis. Many disk based sensing models have been proposed in the literature e.g. the binary, staircase, probabilistic, etc. Among the most popular ones is the probabilistic sensing model given below,

$$p(x) = \begin{cases} 1 & x \leq R_s \\ e^{\lambda(x-R_s)^\gamma} & R_s < x < R_d \\ 0 & x \geq R_d \end{cases} \quad (1)$$

where the probability for a sensor node to detect an event is exponentially decreasing with distance x in the range $[R_s, R_d]$ and it is assumed that the sensor will detect an event with probability 1 (perfect sensor) if it occurs within the inner circle of radius R_s (see Figure 1a). The value of R_s (in the range $[0, R_d]$) is application dependent. The parameters γ and λ in equation (1) control the rate of probability decrease and can be determined considering the physical properties of the sensor, the noise in sensor measurements, the characteristics of the sensed physical quantity etc.

We introduce a novel variation of the probabilistic sensing model which, in addition to describing the detection distance uncertainty, it also accounts for the real possibility of a sensor node malfunctioning in a harsh environment as the hazard's front gets closer. This sensing model variation was inspired by the analysis of real WSN data collected from two outdoor experimental burns that took place at Gestosa's experimental field site in Portugal [4]. The data analysis has shown that in many cases the sensors were unable to detect the approaching fire front since abrupt increases in temperature (usually due to sudden flame fluctuations) destroyed the sensors before they detected the phenomenon (their measurements overcome a predetermined threshold).

The sensing range of a node S_i is assumed to be a circular region of radius R_d (see dotted circle in Figure 1b) centered at the sensor's location (L_i), as for

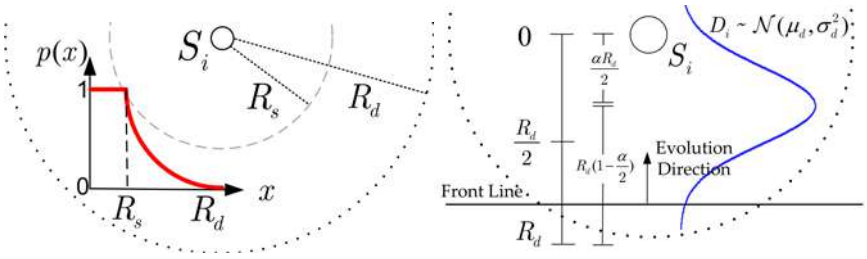


Fig. 1: Sensing Modeling: (a) Probabilistic exponential sensing model. (b) The proposed shifted Gaussian sensing model.

the probabilistic model. The value of R_d is hazard specific and depends on: (i) The sensor's technical specifications (e.g. its sensitivity), (ii) how the monitored phenomenon affects the physical quantity measured by the sensor. Using this information we can estimate the expected distance at which the evolving front is detected by the sensor [5]. We set this distance equal to $\frac{\alpha R_d}{2}$, where $0 \leq \alpha \leq 1$ (see Figure 1b). However, due to the stochastic nature of a hazard's detection this distance may actually deviate from its expected value. To account for this stochasticity we treat the detection distance as a normally distributed random variable, $D_i \sim \mathcal{N}(\mu_d, \sigma_d^2)$, with parameters:

$$\mu_d = \frac{\alpha R_d}{2}, \quad 3\sigma_d = R_d(1 - \frac{\alpha}{2}) \Rightarrow \sigma_d = \frac{R_d}{3}(1 - \frac{\alpha}{2}). \tag{2}$$

In setting the standard deviation as in (2) above we assumed that the probability for a sensor to detect the approaching diffusive phenomenon at a distance larger than R_d is negligible.

As observed in Figure 1b the probability of detection increases monotonically as the distance of the local front from the sensor decreases in the range $\left[\frac{\alpha R_d}{2}, R_d\right]$. However, in close range $\left[0, \frac{\alpha R_d}{2}\right]$ the probability of detection decreases. This modeling decision is justified considering that the inability of a sensor to detect the approaching front at the expected detection range $\left[\frac{\alpha R_d}{2}, R_d\right]$ is an indication of a potential hazard-induced malfunction reducing the probability of detecting the hazard as it gets closer to the sensor node. This simple and realistic, sensing model in the presence of propagating hazards allows us to capture both the inherent stochasticity associated with the detection distance as well as the sensor node's increasing probability to malfunction as the hazard gets in close range. Importantly, it does not harm at all the generality since by setting the parameter $\alpha = 0$ in equation (2) (i.e. $\mu_d = 0$) we can relax the assumption that a node may malfunction and revert back to a monotonic probabilistic sensing model centered at the sensor node's location. The proposed "shifted" Gaussian model is therefore very flexible since it can cover both scenarios: diffusive hazards which may, or may not, affect the functionality of deployed sensor nodes. This is in contrast to

the classical monotonic probabilistic model which ignores the real possibility of sensing mechanism failures as the hazard propagates in close range.

3 Collaborative WSN algorithm for estimating the spatiotemporal evolution characteristics of a continuous object

In this section we present the collaborative WSN algorithm for estimating and tracking the local evolution characteristics of continuous objects [6–8].

The key idea of the proposed in-network collaborative algorithm is the following: As soon as the deployed sensor nodes detect the evolving front line of a propagating hazard they are dynamically organized into ad-hoc local clusters (see Figure 2a) of 3 nodes *(triplets)*. Each triplet consists of a Master sensor (S_i^M) who initiates cluster formation and two Helper sensors $\{S_j^H, S_k^H\}$ that the Master selects among the nodes in its neighborhood and uses (without them knowing it!) to update its current (prior) local front evolution belief model. The parameters of the updated (posterior) model (speed, orientation and evolution direction) are then propagated forward to other sensor nodes residing in the area where the evolving phenomenon is moving into.

3.1 Model Updating

In our example we assume *w.l.o.g.* that S_i^M has notified by its Helper neighbors $N_i^H = \{S_j^H, S_k^H\}$ (see Figure 2c) that they have detected the evolving front at time instances t_{ij} and t_{ik} respectively, where *w.l.o.g.* $t_{ij} < t_{ik}$. When the Master S_i^M receives the notifications from the pair of Helpers, initiates the model updating procedure described below.

The updating starts with the calculation of the "new" (posterior) local front speed model parameters ($U_i^* \sim \mathcal{N}(u_i^*, s_i^{*2})$). Master node S_i^M uses the expressions in (3) and calculates the parameters of the Normal speed models $U_{ih} \sim \mathcal{N}(u_{ih}, s_{ih}^2)$ of the two Helper projection points p_{ih} where $h = \{j, k\}$.

$$u_{ih} = \frac{\mu_{ih}}{t_{ih}} = \frac{2d_{ih} - \alpha R_d}{2t_{ih}}, \qquad s_{ih} = \frac{\sigma_{ih}}{t_{ih}} = \frac{R_d(1 - \frac{\alpha}{2})}{3t_{ih}} \tag{3}$$

By substituting these parameter values in (4), S_i^M calculates the Gaussian mixture weights w_{ij} and w_{ik}.

$$w_{ij} = \frac{1}{1+C}, \quad w_{ik} = \frac{C}{1+C}, \quad C = \frac{s_{ij}|u_i - u_{ij}|}{s_{ik}|u_i - u_{ik}|}. \tag{4}$$

Then, by applying the resulting mixture weight values into (5) and (6) the Master calculates the parameters ($\hat{u}_i$ and $\hat{s}_i$) of the Normal distribution that best approximates the Gaussian mixture (see equation (7)).

$$\hat{u}_i = w_{ij}u_{ij} + w_{ik}u_{ik} \tag{5}$$

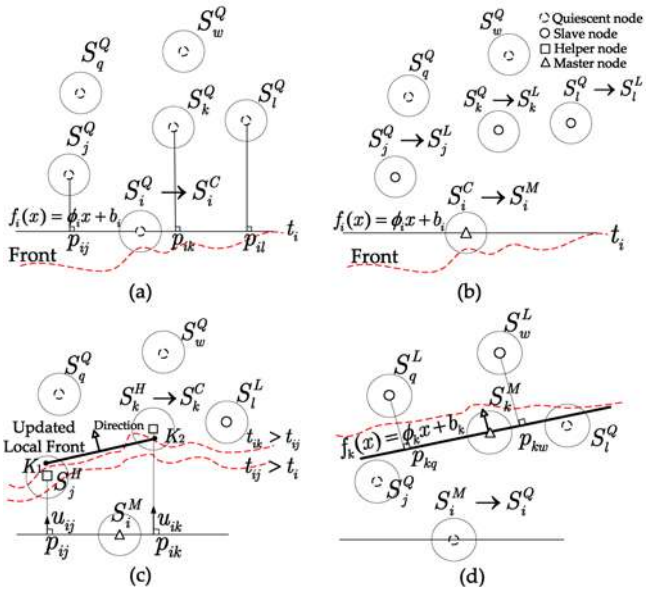


Fig. 2: Local front model updating procedure: (a) Node S_i becomes Master candidate and checks if it satisfies the conditions to become a Master, (b) node S_i becomes a Master and "enslaves" its neighbors S_j, S_k and S_l, (c) Master S_i^M uses the information received from its two Helpers (S_j^H and S_k^H) and updates the local front's line parameters, (d) node S_k becomes the new Master and S_i releases its slaves.

$$\hat{s}_i^2 = w_{ij}s_{ij}^2 + w_{ik}s_{ik}^2 + w_{ij}w_{ik}(u_{ij} - u_{ik})^2 \tag{6}$$

$$p(u) = \sum_{h\in\{j,k\}} w_{ih}\mathcal{N}(u|u_{ih}, s_{ih}^2), \tag{7}$$

Having available these parameters ($\hat{u}_i$ and $\hat{s}_i$), along with the prior model parameters (u_i, and s_i), S_i^M applies them to equation (8) to obtain parameters (u_i^*, s_i^{*2}) of the posterior speed model.

$$u_i^* = \frac{u_i\hat{s}_i^2 + \hat{u}_i s_i^2}{\hat{s}_i^2 + s_i^2}, \quad s_i^{*2} = \frac{\hat{s}_i^2 s_i^2}{\hat{s}_i^2 + s_i^2} \tag{8}$$

Next, Master S_i^M estimates the local front's orientation, ϕ_i^*. To update this parameter the Master finds the coordinates of two points, $K_1 = (x_1, y_1)$ and $K_2 = (x_2, y_2)$, which are expected to lie on the "new" local front line (see Figure 2c) and applies them directly to equation (9).

$$\phi_i^* = \frac{y_2 - y_1}{x_2 - x_1}. \tag{9}$$

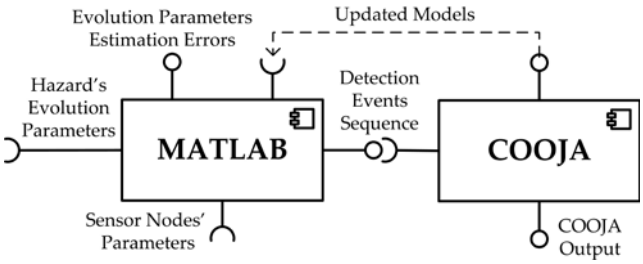


Fig. 3: UML component diagram of Matlab-COOJA based simulation workflow.

To update the direction of evolution parameter δ_i^*, node S_i^M derives the equation of the line $f_i^*(x)$ that is determined by points $K_1(x_1, y_1)$ and $K_2(x_2, y_2)$ (see Figure 2).

$$f_i^*(x) = \phi_i^* x + b_i^* \tag{10}$$

where $b_i^* = y_1 - \phi_i^* x_1$

Subsequently, node S_i^M substitutes its abscissa (x_i) in (10) and checks the $sgn(f_i^*(x_i))$. If $sgn(f_i^*(x_i)) > 0$ ($sgn(f_i^*(x_i) < 0$) then Master node S_i^M infers that the new local front line evolves into the *negative (positive)* half plane and it updates the direction parameter $\delta_i^* = -1(1)$ accordingly.

All model parameters are updated using closed form expressions that can be realized easily by embedded microprocessors commonly used in WSN node architectures.

3.2 Evaluation of the Algorithm

For the evaluation we have developed a flexible simulation workflow which allows us to generate and execute realistic WSN simulation scenarios with different sensor node densities, deployment strategies, sensor node failure probabilities, communication (Rx and Tx) failure probabilities, and propagating hazard front properties (shape, speed and acceleration).

Simulation Workflow
The WSN simulation workflow includes two main components: i) The flexible WSN simulator COOJA (COntiki Os JAva) for the Contiki sensor node operating system, and ii) a Matlab-based component which prepares the COOJA input file and evaluates the estimation accuracy of the proposed in-network algorithm.

As shown in the UML component diagram of Figure 3, the Matlab component takes as input information about: a) the deployed sensor nodes (location, prior model parameters, etc.), and b) the propagating hazard's front properties, and determines the sequence in which the deployed sensor nodes detect the evolving hazard. After that, it generates a file (*Detection Events Sequence*) which contains for each sensor node the following information: {*ID, location, time of detection, prior model parameters*}. This file is passed as input to COOJA used to simulate

the behavior of the proposed distributed algorithm, as if it was implemented by a WSN consisting of Atmel's AVR RAVEN nodes. Using COOJA we simulate the IEEE 802.15.4 MAC protocol's byte stream and we can evaluate the proposed algorithm's behavior under different Rx/Tx failure probabilities.

At the end of a simulation, a *COOJA Output* file is produced which contains: a) The updated model parameters, b) the number of Rx and Tx messages/Bytes exchanged in the WSN, and c) the energy consumed for communication (Rx and Tx). To evaluate the estimation accuracy of the proposed algorithm, the updated models information is passed back as input to the Matlab component which compares the corresponding models' orientation and speed with the ground truth values.

Results and Discussion
In the conducted experiments the diffusive phenomenon (continuous object) was simulated using either a Matlab program, that simulates multi-source diffusive hazards, or FLogA a wildfires behavior simulator [9] that allow us to simulate complex diffusive hazards with irregular shapes.

Extensive computer simulation results show that the proposed algorithm is able to estimate with accuracy the evolution parameters (speed, orientation and evolution direction) of the diffusive hazardous phenomena. Its accuracy seems be insensitive to changes in sensor nodes density, node failure probability, and Rx/Tx failure probability. This was also confirmed by comparing pairwise the means of the error densities using Student's t-test. For all cases the difference of the means was found to be insignificant at the 0.05 significance level. Moreover, the results indicate that the accuracy of the proposed algorithm slightly decreases when the sensing radius R_d increases. This can be explained if we consider that an increase of the sensing radius implies increasing the uncertainty associated with the front line's location at the time of the hazard's detection (see Section 2).

4 Assesing Requirements for Large Scale implementation using Simulation-Driven WSN Emulation

For all WSN schemes, computer simulations can be used to assess the expected WSN behavior as a function of its density. However simulations fail to provide: a) accurate energy consumption estimates and how they scale with the size of the network, and b) information about the processing and memory requirements of the distributed algorithm's implementation. Since having such estimates is very important before attempting to deploy a large-scale WSN for environmental monitoring application the real question becomes, how can we meet this requirement without having to deploy a large-scale WSN?

To address this question we introduce a simulation-driven WSN emulation workflow (see Figure 4) which allows us to emulate the operation of a large-scale WSN deployment for environmental applications by reutilizing only a small number of real sensor nodes. The key idea of the proposed method is to re-allocate

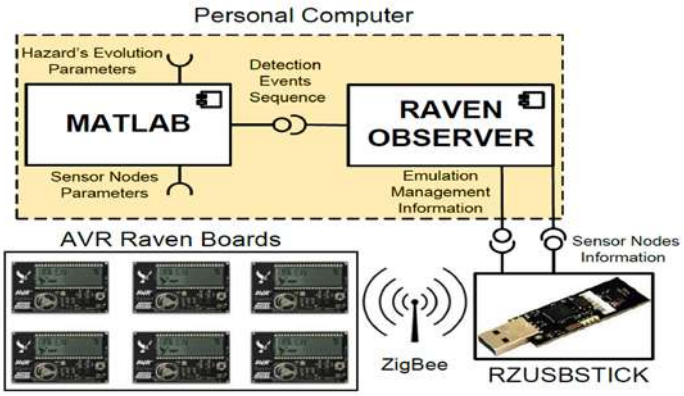


Fig. 4: Simulation driven emulation workflow.

(virtually reposition) the available sensor nodes so that they implement WSN nodes located close to the hazards front line as it evolves [10]. We demonstrate its capabilities using the distributed algorithm we introduced in Section 3 for estimating the spatiotemporal evolution parameters of diffusing environmental hazards. The WSN implementation of the proposed collaborative algorithm was based on the affordable Atmel Raven evaluation kit.The distributed WSN algorithm was coded in C on the IPv6 ready RTOS Contiki, an open source operating system for networked, memory-constrained systems with a particular focus on low-power wireless Internet of Things devices.

Emulation results clearly indicate that our algorithm is suitable for a large-scale WSN deployment, since it respects WSNs' communication, processing, memory and energy constraints. The proposed emulation approach can be followed to assess the practicality of large-scale WSN deployment of other in-network algorithms of similar nature for environmental monitoring applications.

5 Continuous Object Boundary Reconstruction Algorithm

In this Section we present a novel algorithm which reconstructs with accuracy the boundary of an evolving continuous object using a small number of local front estimates [11]. Each local front estimate describes locally the evolution characteristics (orientation angle, direction and speed) of the continuous object's boundary. When a sufficient number of local front estimates becomes available at a fusion center the algorithm combines their information and determines a "smooth" curve that approximates the object's boundary.

The key idea of the proposed boundary reconstruction algorithm is as follows: Let's assume that a monitoring system (e.g. based on WSN technology -

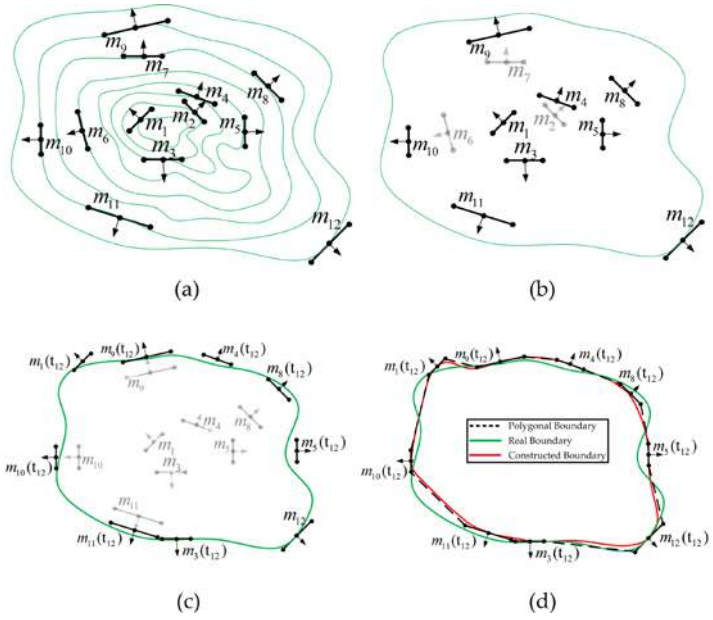


Fig. 5: a) Each green curve shows different instances of an evolving continuous object's boundary; each boundary corresponds to the time instance where a local front estimate (black segment) appears. to the diffusive continuous object's boundary at the times instance where the corresponding local front estimate (black segments) becomes available. b) The dark segments correspond to the subset local front estimates that will be used to determine the boundary at time $t = t_{12}$. c) The predicted locations of the selected local front segments at time $t = t_{12}$. d) The polygon (black dashed) and the smooth curve (red curve) that approximates the continuous object's boundary (green curve).

see Section 4) is able to estimate the evolution characteristics (orientation angle, direction and speed) of a continuous object at different locations and/or time instances (see black segment in Figure 5a). As soon as a sufficient number (application dependent) of local front estimates becomes available, the proposed algorithm combines their information and determines the set of local front estimates (black segments in Figure 5b) that will be used to reconstruct the boundary of the continuous object. In sequence, using their evolution characteristics it determines their locations at the time instance that we wish to reconstruct the continuous object's boundary (time t_{12} see Figure 5c). Using the "new" location coordinates and the evolution direction parameters of the local fronts, the proposed algorithm determines a polygon that approximates the continuous object's boundary (see black dashed polygon in Figure 5d). Subsequently, using uniform cubic B-splines the algorithm determines a "smooth" curve which is

the reconstruction of the continuous object's boundary (see red curve in Figure 5d). Finally, based on the estimation uncertainties of local fronts' parameters, the algorithm computes a probability field, that indicates for each point, the probability to be reached by the continuous object at a given time.

Extensive simulation results demonstrate that the proposed algorithm is able to track accurately the boundary of different types of continuous objects (e.g. time-varying evolution rates and/or irregular boundary shapes), while using a small number of local fronts estimates which may be distorted with error.

6 Conclusions

We proposed a flexible probabilistic sensing modeling approach which in contrast with the existing works that assume a perfect sensing mechanism (see L4 in Section 1), can capture the detection distance uncertainty and the possibility for a sensor node to malfunction in a harsh environment created by an approaching hazard. This simple, yet realistic, sensing model allows us to formulate a local front models' parameters estimation problem in a Bayesian manner. We analytically solved this Bayesian problem and derived closed-form algebraic expression that can be easily implemented by microprocessors of the commodity sensor nodes.

To address limitations L1, L3, L4 and L5 (presented in Section 1) of the state of the art schemes, we developed an *asynchronous* collaborative algorithm that is able, using WSNs of *realistic* density, to estimate with accuracy the spatiotemporal evolution parameters (orientation, direction and speed) of a continuous object's boundary. The proposed parameters estimation procedure implemented in a collaborative fashion by dynamically formed clusters (triplets) of sensor nodes. The algorithm updates the local front model parameters and propagates them to sensor nodes situated in the direction of the hazard's propagation in a fully decentralized manner.

To realistically asses the requirements and behavior of the proposed algorithm (see L6 in Section 1), we developed a simulation-driven WSN emulation workflow which allows us to estimate, before attempting to deploy a large scale WSN, the energy, processing and memory requirements of collaborative algorithms as the WSN's size increases.

The state of the art are incapable to delineate automatically the boundary of an evolving continuous object (see L7 in Section 1). To address this limitation we developed an algorithm which combines dynamically the information of a small number of estimated local front models, as they become available to a fusion center, and determines a smooth curve that approximates the boundary of the continuous object at a specific time instance. By exploiting the estimation uncertainty of the local fronts evolution parameters, the proposed algorithm generates a probability field that indicate for each point of the considered area, the probability to be affected by the continuous object.

Acknowledgment

This research has been co-financed by the European Union (European Social Fund ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Heracleitus II. Investing in knowledge society through the European Social Fund.

References

1. J.Kim, K. Kim, S.Chauhdary, W. Yang, M. Park, "DEMOCO: Energy-Efficient Detection and Monitoring for Continuous Objects in WSN", IEICE Trans.on Communications, vol.E91-B, no. 11, pp.3648-3656, Nov. 2008.
2. B. Park, S. Park, E. Lee, and S. H. Kim, "Detection and Tracking of Continuous Objects for Flexibility and Reliability in Sensor Networks," in Proc. IEEE Inter. Conf. International Conference on Communications, Cape Town pp.1-6, 2010.
3. H. Hong, S. Oh, J. Lee, S. Kim, "A Chaining Selective Wakeup Strategy for a Robust Continuous Object Tracking in Practical Wireless Sensor Networks," Advanced Information Networking and Applications (AINA), 2013 IEEE 27th International Conference on , vol., no., pp.333,339, 25-28 March 2013
4. D. V. Manatakis, E. S. Manolakos, A. Roussos, G. Xanthopoulos, D. X. Viegas,"In-silico estimation of the temperature field induced by moving fire. Predictive modelling and validation using prescribed burn data". Coimbra, Portugal. In Proc. of VI International Conference on Forest Fire Research, November 2010.
5. E. S. Manolakos, D. V. Manatakis, G. Xanthopoulos, "Temperature Field Modeling and Simulation of Wireless Sensor Network Behavior During a Spreading Wildfire",In Proc. 16th European Signal Processing Conference (EUSIPCO), August, 2008.
6. D. V. Manatakis, E. S. Manolakos, "Collaborative Sensor Network algorithm for predicting the spatiotemporal evolution of hazardous phenomena," Int. Conf. on Systems Man and Cybernetics (SMC 2011), October 2011, Anchorage-Alaska, pp. 3439-3445.
7. D. V. Manatakis, E. S. Manolakos, "Predictive modeling of the spatiotemporal evolution of an environmental hazard and its sensor network implementation" In Proc. IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP 2011), May 2011, Prague-Czech pp. 2056-2059
8. D. V. Manatakis, E. S. Manolakos, "Estimating the Spatiotemporal Evolution Characteristics of Diffusive Hazards using Wireless Sensor Networks," Accepted for publication, Parallel and Distributed Systems, IEEE Transactions on, doi: 10.1109/TPDS.2014.2357033, 2014.
9. N. Bogdos, E. S. Manolakos, "A tool for simulation and geo-animation of wildfires with fuel editing and hotspot monitoring capabilities," Elsevier Journal of Environmental Modelling and Software, Vol.46, August 2013, pp. 182-195.
10. D.V. Manatakis, M.G. Nennes, I.G. Bakas, E.S. Manolakos, "Simulation-driven emulation of collaborative algorithms to assess their requirements for a large-scale WSN implementation," Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on , vol., no., pp.8360,8364, 4-9 May 2014
11. D. V. Manatakis, E. S. Manolakos,: Boundary Predictive Tracking of Continuous Objects Using Sparse Local Estimates, IEEE Transactions on Systems Man and Cybernetics, March 2015 (in preperation).

Advance BPEL execution adaptation using QoS parameters and collaborative filtering techniques

Dionisios D. Margaris*

National and Kapodistrian University of Athens
Department of Informatics and Telecommunications
margaris@di.uoa.gr

Abstract. In this thesis, frameworks for providing runtime adaptation for BPEL scenarios are proposed. The adaptation is based on (a) quality of service parameters of available web services (b) quality of service policies specified by users and (c) collaborative filtering techniques, allowing clients to further refine the adaptation process by considering service selections made by other clients.

1 Introduction

Web Services are considered a dominant standard for distributed application communication over the internet. Consumer applications can locate and invoke complex functionality, through widespread XML-based protocols, without any concern about technological decisions or implementation details on the side of the service provider. Web Services Business Process Execution Language (WS-BPEL) [1] allows designers to orchestrate individual services so as to construct higher level business processes; the orchestration specification is expressed in an XML-based language, and it is deployed in a BPEL execution engine, made thus available for invocation by consumers. -BPEL has been designed to model business processes that are fairly stable, and thus involve the invocation of web services that are known beforehand.

In this dissertation an adaptation algorithm uses both QoS specifications and semantic-based collaborative filtering personalization techniques to decide which offered services best fit the client's profile is presented. To achieve this goal, the metasearch algorithm paradigm [4] is followed, using two different candidate adaptation ranking algorithms, the first examining the QoS aspects only and the second being based on collaborative filtering techniques. The adaptation rankings produced by these two algorithms are combined to generate the overall ranking, which then drives the adaptation. The combination of the results is performed using a weighted metasearch score combination algorithm ([4][5]), however varying weights are used to address issues associated with collaborative filtering, such as cold start (i.e. few entries recorded in the rating database, thus no good matches can be obtained) and

*Dissertation Advisor: Panayiotis Georgiadis, Emeritus Professor

gray sheep (i.e. unusual users, which cannot be matched with other users even after the database has been adequately populated).

The adaptation is based on (a) quality of service parameters of available web services (b) quality of service policies specified by users and (c) collaborative filtering techniques, allowing clients to further refine the adaptation process by considering service selections made by other clients.

The combined proposed BPEL execution framework includes provisions for

(a) specifying QoS requirements for invocations of web services within a WS-BPEL scenario
(b) specifying specific bindings for selecting services and designating which services are subject to adaptation,
(c) adapting the WS-BPEL scenario execution according to the results of the service selection algorithm,
(d) monitoring the behavior of the invoked services regarding their QoS aspects,
(e) collecting user satisfaction feedback about the invoked services and taking these data into account when formulating recommendations and
(f) caters maintaining the transactional semantics that invocations to multiple services offered by the same provider may bear.

This approach follows the horizontal adaptation paradigm since, as noted in [2], horizontal adaptation preserves the execution flow which has been crafted by the designer to reflect particularities of the business process, while it also allows the exploitation of specialized exception handlers.

2 Related Work

As stated above, existing adaptation approaches follow either the horizontal or the vertical adaptation approach. [4] performs horizontal QoS-based adaptation, taking into account the sequential and parallel execution structures within the BPEL scenario. Note that none of above approaches incorporates CF techniques to enhance the quality of the adaptation. [5] presents CF techniques to drive the adaptation, and an associated execution framework, it however uses very limited QoS-based criteria (only a lower and an upper bound for each QoS attribute), hence it runs the risk of formulating solutions whose QoS is much inferior to the optimal composition QoS that can be attained, especially in cases that CF has known issues (e.g. cold start and gray sheep). In order to perform QoS/CF-based adaptation or exception resolution, all approaches employ means to formally specify the services' functionality; techniques involving QoS characteristics need additionally to have access the services' QoS attribute values. In this work, we adopt the subsumption relationship approach [18] due to its expressiveness and flexibility. An important aspect of QoS attributes is that their values may vary, according to server load, network conditions or other relevant factors. To this end, work in [19] is adopted to allow a more accurate estimation of QoS attribute values; this increased accuracy can be used in adaptation systems to improve the quality of the adaptation.

3 QoS AND CF UNDERPINNINGS

In the following subsections we summarize the underpinnings from the areas of QoS and CF, which are used in our work.

3.1 QoS concepts

For conciseness purposes, in this paper we will consider only the attributes responseTime (rt), cost (c) and availability (av), adopting their definitions from [9]. This does not lead to loss of generality, since the algorithms can be straightforwardly extended to accommodate more attributes. The QoS specifications for a service within the BPEL scenario may include an upper bound and a lower bound for each QoS attribute, i.e. for each service sj included in a BPEL scenario, the designer formulates two vectors MINj=(minrt,j, minc,j, minav,j) and MAXj=(maxrt,j, maxc,j, maxav,j). Additionally the designer formulates a weight vector W=(rtw, cw, avw), indicating how important each QoS attribute is considered by the designer in the context of the particular operation invocation. The values of the QoS attributes are assumed to be expressed in a "larger values are better" setup, e.g. a service having cost = 6 means that that it is cheaper than a service having cost = 4. In order to compute the QoS attribute values of a service S composed from constituent services s1, …, sn having QoS attributes equal to (rt1, c1, av1), …, (rtn, cn, avn), respectively, the formulas given in table 1 [10] can be used. Note that these formulas do not take into account the possibility that some service is executed multiple times within a loop or conditionally with a probability of p; these aspects will be considered in our future work.

	QoS attribute		
	responseTime	cost	availability
Sequential composition	$\sum_{i=1}^{n} rt_i$	$\sum_{i=1}^{n} c_i$	$\prod_{i=1}^{n} av_i$
Parallel composition	$\max_i(rt_i)$	$\sum_{i=1}^{n} c_i$	$\prod_{i=1}^{n} av_i$

Table 1. QoS of composite services

Most works dealing with QoS-based BPEL scenario execution adaptation, consider given QoS attribute values for each service, which can be for instance declared by the service provider within an SLA. However, in the real world, QoS metrics such as response time and availability may vary, due to server or network conditions (failures, overloads, bottlenecks etc). To tackle this issue, in this paper, we employ prediction

models for QoS attribute values, in order to use in the recommentation process values that are closer to the actual ones, improving thus the accuracy of the adaptation. In particular, we adopt [7] and [8] for predicting the service response time and service availability, respectively. Both these algorithms predict future performance of services by examining past measurements; the platform proposed in this work collects these measurements when invoking services in the context of BPEL scenario executions and makes them available to the modules predicting the future QoS values.

3.2 Subsumption relationship representation

In order to adapt the BPEL scenario execution, the adaptation engine needs to be able to find which services offer the same functionality, and are thus candidate for invocation when this particular functionality is needed. In this work, we represent this information using subsumption relationships [6] which, for any pair of services S1 and S2 defined as follows: (i) S1 exact S2, iff S1 provides the same functionality with S2 (ii) S1 plugin S2, iff S1 provides more specific functionality than S2; in this case S1 could be used whenever the functionality of S2 is needed, since it delivers (a specialization of) the functionality delivered by S2 (iii) S1 subsume S2, iff S2 provides more generic functionality than S2. In this case S1 cannot unconditionally be used whenever the functionality of S2 is needed and (iv) S1 fail S2, in all other cases; in this case, S2 cannot be substituted for S2. Under these definitions, a service A can be unconditionally substituted by a service B if (A exact B or A plugin B); this setup provides more flexibility as compared to strict service equivalence (A exact B) regarding the formulation of the adapted execution plan, and is hence adopted in this paper. Effectively, subsumption relationships organize services in a tree, where generic services are located towards the root and more specific services towards the leaves [6]. Tree nodes, besides service identity, can accommodate QoS values for the services they represent; this information can be stored in repositories such as OPUCE [11]. Fig. 1 shows an excerpt of a subsumption relationships tree.

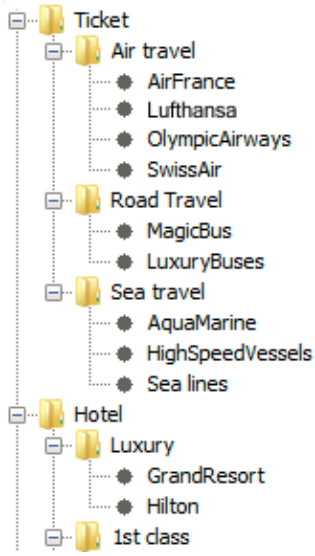


Fig. 1. Example subsumption relationships tree

3.3 Designations on specific service bindings and functionality omissions

As noted above, users may wish to designate exact services to be invoked for realizing specific functionalities, while asking for recommendations on other ones. For instance, in a travel planning scenario the consumer may request that s/he travels by "Sea Lines". Further, the consumer may also specify that some functionality optionally included in the BPEL scenario is not executed; for example, a tourist may not want to rent a car, while such a provision is present in the scenario. Typically, the BPEL code will examine input parameters and decide using a conditional execution construct (<switch>) whether to invoke the functionality or not. Finally, functionalities that are neither explicitly bound to specific services, nor are designated as "not to be executed" are subject to adaptation. We consider that specific bindings and designations for functionality omissions are explicitly expressed in the request for scenario invocations.

3.4 Usage patterns repository

In order to perform CF-based adaptation, a repository with user ratings for services is required. In this paper, we adopt the representation used in [5], where the ratings repository is modeled as a table having a number of columns equal to the functionalities present in the BPEL scenario, and one row for each BPEL scenario execution. Cell i,j is filled with value S if during the ith execution of the BPEL scenario, service S was used to implement functionality j; cell (i, j) may be also blank, if during the ith execution of the BPEL scenario functionality j was omitted. In order to accommodate user ratings, we extend this repository by adding one column per functionality. This col-

umn stores an integer value from the domain [1, 10], corresponding to the rating given by the user that executed the particular scenario instance. For the cases that the user has not provided a rating, a null value is stored and the CF-based algorithm uses a default value, as explained in section 4. The BPEL scenario adaptation module inserts new records to the usage patterns repository (UPR), when the concrete services that will be invoked in the context of a particular BPEL scenario execution are decided, while the user evaluation collection module arranges for storing user rankings in the relevant columns. Table 2 presents an example UPR.

# exec	Travel	Hotel	Event
1	OlympicAirways	YouthHostel	ChampionsLeague
2	SwissAir	Hilton	GrandConcert
3	HighSpeedVessels	YouthHostel	
4	LuxuryBuses		EuroleagueFinals
5	Lufthansa	YouthHostel	GrandConcert
6	AirFrance	Hilton	
7	SwissAir	YouthHostel	ChampionsLeague

Table 2. Example usage patterns repository

4 THE SERVICE RECOMMENDATION ALGORITHM

As stated in section 1, our approach follows the horizontal adaptation paradigm, leaving the composition logic intact and adapting the execution by selecting which concrete service implementation will be used in each specific invocation. To perform this task, the algorithm takes into account the following criteria:

- The consumer's QoS specifications (bounds and weights).
- Designations on which exact services should be invoked, if such bindings are requested by the consumer (e.g. a user wanting to travel using Air France).
- Designations on which functionalities should not be invoked

(e.g. a user wanting to book a trip without scheduling any event attendance).

- The QoS characteristics of the available service implementations, including monitored values of the QoS attributes of the services.
- The service subsumption relationships.
- The UPR, which includes user ratings.

The approach proposed in this paper incorporates two different candidate service ranking algorithms, the first examining the QoS aspects only ([4]) and the second being based on CF techniques ([5]). The algorithms run in parallel to formulate their suggestions regarding the services that should be used in the adapted execution, and subsequently their suggestions are combined, through a metasearch score combination algorithm with varying weights.

4.1 The QoS-based adaptation algorithm

The QoS-based adaptation algorithm initially identifies the services which are candidate to be used for delivering functionalities in the context of the current BPEL scenario, respecting the QoS-bounds set by the user, and subsequently computes the k-best service assignments to the functionalities requested for the particular scenario execution. In more detail, the algorithm proceeds as follows:

- For each functionality f_i for which adaptation has been requested, the algorithm retrieves from the semantic service repository the concrete services that (a) deliver this functionality and (b) respect the QoS bounds set by the users. These are the candidates for implementing functionality f_i. Formally, this is expressed as

$$CF(i) = \{s \in Repository: (funct_i\ exact\ s\ \vee funct_i\ plugin\ s)\ \wedge\ [(min_{rt,i} \leq rt_s \leq max_{rt,i})\ \wedge\ (min_{c,i} \leq c_s \leq max_{c,i}) \wedge\ (min_{rel,i} \leq rel_s \leq max_{rel,i})]\}$$

Note that in all steps of this algorithm, the QoS values for response time and availability considered for each service are those returned by predictor methods [7] and [8], respectively.

- Subsequently, the algorithm formulates an integer programming problem to compute the k-best solutions regarding the assignment of concrete services $s_{i,j}$ to each functionality f_i. To express the integer programming optimization problem in this work we adopt the concrete service utility function used in [13], which is b $U(s_{j,i}) = \sum_{k=1}^{3} \frac{Q_{max}(j,k) - q_k(s_{j,i})}{Q_{max'}(k) - Q_{min'}(k)} * w_k$ where $q_k(si,j)$ is the value of the kth QoS attribute of concrete service $s_{i,j}$ (the first QoS attribute being response time, the second cost and the third one availability), w_k being the weight assigned to the kth QoS attribute, [i.e. the maximum value of QoS attribute k among possible concrete service assignments for functionality fi], and $Q_{max'}$ [resp. $Q_{min'}$] being the overall maximum (resp. minimum) value of QoS attribute k within the service repository. Using the utility function, the computation of the best solution is expressed as the following integer programming problem: maximize the overall utility value given by $OUV_{QoS} = \sum_{i=1}^{F} \sum_{j=1}^{T(i)} U(s_{j,i}) * x_{j,i}$, where F is the number of functionalities fi requiring adaptation, and each $x_{j,i}$ is a binary variable taking the value 1 if $i_{j,j}$ is selected for delivering functionality f_i, and 0, otherwise. Since each functionality f_i is delivered in the final execution plan by exactly one concrete service, the maximization of the utility value is subject to the constraint b

$$\sum_{i=1}^{T(j)} x_{j,i} = 1, 1 \leq j \leq Fb$$

This problem is then solved and the k-best solutions are obtained. Note that this formulation employs the sum function to rate the availability of the composite service taking into account the availability values of the constituent services, rather than the product function, as denoted in table 1.

Note that although integer programming is NP-hard, in practice, solving techniques employ a number of speed up factors namely cutting planes, presolve, branching rules, heuristics, node presolve and probing on dives [17], with each speed up factor providing a speed up ranging from 53.7% (cutting planes) to 1.1% (probing on dives);

therefore the time taken by solvers to compute the solution is much lower than the worst-case (NP-hard) complexity. The solutions are saved, together with their overall utility score, for perusal in the combination step. In order to solve the integer programming problem computing the k-best solutions, the IBM ILOG CPLEX (www.ibm.com/software/commerce/optimization/cplexoptimizer/) optimizer was used. In our implementation, we have set k=20.

4.2 The CF-based algorithm

The CF-based algorithm employed in our proposal is an adaptation of the standard GroupLens algorithm [14], modified to take into account the semantic distance of the services realizing the same functionality. For instance, rows 2 and 5 of table 2 are considered "semantically close", since they both list air transport for travel, a first class hotel for accommodation and classical music events; on the other hand rows 2 and 7 of the same table are considered "semantically distant", since all three services correspond to diverse real world counterparts (air travel vs. bus, 1st class hotel vs. 3rd class, concert vs. sports). Taking this into account, when a request arrives asking for travel via AirFrance and accommodation in GrandResort and requesting a recommendation for event attendance, the ratings in rows 2 and 7 must be taken more strongly into account than those in row 7, since the former two rows are "closer" to the one under adaptation. To accommodate this adaptation, we extend the formula of cosine similarity between two rows $\vec{X}$, $\vec{Y}$ of the UPR as follows: *b*

$$r(\vec{X},\vec{Y}) = \frac{\sum_{k=1}^{n}(\vec{X}[k] * \vec{Y}[k] * d(\vec{X}[k],\vec{Y}[k]))}{\|\vec{X}\| * \|\vec{Y}\|}$$

We can observe in equation (2) that the standard cosine similarity metric has been extended to accommodate the semantic distance between the services that realize the same functionality in rows $\vec{X}$ and $\vec{Y}$; this is accomplished by multiplying each term of the sum in the nominator by a metric of the semantic distance between the two services, which is denoted as $d(s_1, s_2)$ and is computed using the formula introduced in [15]: $d(s_1,s_2) = C - lw*PathLength - NumberOfDownDirection$, where *C* is a constant set to 8 [15], *lw* is the level weight for each path in subsumption tree (cf. Fig. 1), *PathLength* is the number of edges counted from functionality s_1 to functionality s_2 and *Number-OfDownDirection* is the number of edges counted in the directed path between functionality s_1 and s_2 and whose direction is towards a lower level in the subsumption tree. For more details in the computation of the semantic distance, the interested reader is referred to [15]. We further normalize this similarity metric in the range [0, 1] by dividing the result computed in the above formula by 8; this way, the multiplication by the normalized similarity metric in equation (2) reduces the correlation coefficient between the two rows by a factor proportional to the semantic distance of the services employed in these rows to realize the same functionality. For items not explicitly rated, we follow the rationale of [5] according to which usage of a service is an indication of preference, and we choose a rating equal to the 80% of the maximum rating. This is inline with the findings of [16], which asserts that dissatisfied users will provide negative feedback with a very high probability (≥89%). Rows that have not been rated at all (and therefore have a default value for all ratings) are the reason behind choosing the cosine similarity against the Pearson similarity, since

the latter disregards rows whose ratings have no variance (i.e. are all equal). Using the modified cosine similarity, the CF-based algorithm operates as follows:
1. It retrieves from the UPR all rows that contain a service implementing the functionality on which a recommendation is requested. For example, if a recommendation on event attendance is requested, only rows 1, 2, 4, 5 and 7 of table 2 will be retrieved.
2. The rows retrieved from step 1 are filtered to retain only those that fulfill the QoS criteria requested by the user.
3. The similarities between the request and each row are computed using the modified cosine similarity metric. The request is represented here as a vector vector $\vec{R}$, having a rating equal to 10 for each functionality included in the scenario and a rating equal to 0 for each functionality designated as not to be executed.
4. For each distinct service implementing the requested functionality that is included in the remaining rows, we compute its rating prediction using the standard rating prediction formula *b*

$$p(\vec{R}[k]) = \underset{m}{mean}(\vec{R}[m]) + \frac{\sum_{\vec{N} \in raters(\vec{R}[k])} (\vec{N}[k]) * r(\vec{R}, \vec{N})}{\sum_{\vec{N} \in raters(\vec{R}[k])} r(\vec{R}, \vec{N})}$$

[14] (we again do not subtract the mean $\overrightarrow{N}$ from $\overrightarrow{N}[k]$, so as not to render useless the rows having only default values).
5. Finally, we retain the 20-best services, for each functionality requiring adaptation, for perusal in the combination step.
After the lists of candidates for each individual service that is subject to adaptation have been computed, the algorithm selects the top-20 execution plans with respect to their CF-score. Given an execution plan containing services (s1,i, …, sN,k) with the similarity scores of the services computed in step 5 being (CFS(s1,i), …, CFS(sN,,k)), then the CF-score of the execution plan is equal to CFS(s1,i)+…+CFS(sN,k). Computing the top-20 execution plans is modeled as an integer programming optimization problem, formulated in a similar fashion to the one described in section 4.1. Full details on the formulation of the integer programming optimization problem are given in [12]. The CF module has been implemented using Apache Mahout (https://mahout.apache.org/), by subclassing the UncenteredCosineSimilarity class and reimplementing in the subclass the UserSimilarity method, to accommodate the semantic similarity metric described above.

4.3 The combination step

The combination step synthesizes the results given by individual algorithms to produce a single result. Recall from the previous two subsections that each algorithm produces a set of candidate execution plans, with each execution plan being tagged with the relevant normalized score (QoS-score or CF-score). In order to combine the scores, we use the CombMNZ metasearch algorithm, since it has been found to have the best performance [3] [the CombMNZ rating of a solution is computed by multiplying the sum of the individual scores by the number of non-zero scores, i.e. where mi is the number of algorithms giving non-zero rating to item *i* and $r_j(i)$ is the rating given by algorithm *j* to item *i*]. After computing the CombMNZ metasearch for all

candidate execution plans, the combination step selects the execution plan with the highest score, which will be used to drive the adaptation process.

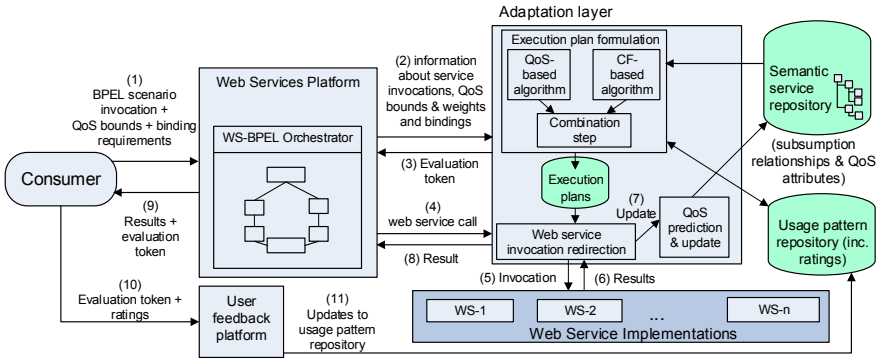


Fig. 2. The execution adaptation architecture

4.4 The execution adaptation architecture

The execution adaptation architecture, illustrated in Fig. 2, follows the middleware-based approach, with an adaptation layer intercepting web service invocations and appropriately directing them to the services chosen by the adaptation algorithm. As shown in Fig. 2, the BPEL scenario execution initially passes to the adaptation layer the information regarding service invocations that will be performed, QoS bounds and weights as well as specific service bindings. When the adaptation layer receives this information, it applies the adaptation algorithm to formulate the execution plan for the particular scenario execution (i.e. decide the actual services that will be invoked to deliver each functionality) and stores the execution plan for later perusal. Subsequently, when a web service invocation is intercepted by the adaptation layer, the respective execution plan is retrieved from the execution plan storage, the web service decided to deliver the specific functionality is extracted and the invocation is routed to that service. Note that steps (4)-(8) depicted in Fig. 2 are repeated multiple times within each BPEL scenario execution, once per web service invocation performed. When the invocation to a service implementation has concluded, the data regarding the service's response time and availability are passed to the QoS prediction and update module, which computes the predicted values for the respective QoS parameters and updates the corresponding elements in the semantic service repository. Additionally, the BPEL scenario returns at the end of its execution, along with the result, an evaluation token, which the consumer may use to enter the ratings for the services s/he has used in the context of the BPEL scenario execution. The evaluation token is returned in the response headers, to retain the response payload schema intact. To accommodate this additional functionality (passing the necessary information to the adaptation layer and returning the evaluation token), the BPEL scenario is preprocessed as described in [5] before being deployed to the web services platform, with the preprocessing step injecting the necessary invocations to the adaptation layer into the scenario, and the result of the preprocessing step is then deployed and made available for invocations.

5 CONCLUSIONS AND FUTURE WORK

In this thesis we have presented a framework for adapting the execution of BPEL scenarios, taking into account data from the monitoring of the QoS offered by the services, as well as user ratings. To perform the adaptation, we follow the metasearch paradigm, by combining two candidate execution plan ranking algorithms. The first one examines the execution plan QoS aspects only, while the second is based on CF techniques. The framework provides means for monitoring the QoS parameters of the services and adjusting accordingly the values of the services' QoS attributes, as well as accepting user ratings for the services they have used, which are taken into account by the CF-based algorithm. The proposed framework is complemented with an execution architecture for enacting the adaptation, which adopts the middleware approach, with an adaptation layer intervening between the BPEL execution platform and the web services and arranging for redirecting service invocations to the services selected by the adaptation algorithm. The proposed framework has been experimentally validated regarding (i) its performance, (ii) the quality of execution plans generated and (iii) the effectiveness of the QoS monitoring and estimation mechanisms. The proposed approach has been also found to be scalable, exhibiting a linear increase in the imposed overhead.

Our future work will focus gathering statistical information from prior scenario executions and using it as input to the adaptation process. This information will quantify aspects regarding the behavior of control constructs in the scenario. We also plan to examine how the algorithm can be extended to consider different adaptation strategies.

6 REFERENCES

[1] OASIS WSBPEL TC. WS-BPEL 2.0. http://docs.oasisopen.org/wsbpel/2.0/OS/wsbpel-v2.0-OS.html

[2] Kareliotis, C., Vassilakis, C., Rouvas, S., Georgiadis, P. QoS-Driven Adaptation of BPEL Scenario Execution. In: *Proceedings of ICWS 2009*, 271-278, 2009.

[3] Montague, M., Aslam, J.A. Relevance score normalization for metasearch. In: *Proceedings of CIKM 2001*, 427-433, 2001.

[4] Margaris, D., Vassilakis, C., Georgiadis, P. An integrated framework for QoS based adaptation and exception resolution in WS-BPEL scenarios. In *Proceedings of the 28th ACM SAC*, Coimbra, Portugal, 1900-1906, 2013.

[5] Margaris, D., Vassilakis, C., Georgiadis, P.: Adapting WSBPEL scenario execution using collaborative filtering techniques. In: *Proceedings of the IEEE 7th RCIS Conference*, Paris, France, 2013.

[6] Paolucci, M., Kawamura, T., Payne, T., Sycara, T. Semantic Matching of Web Services Capabilities. In: *Proceedings of the 2002 International Semantic Web Conference*, 333-347, 2002.

[7] Shao, L., Guo, Y., Chen, X., He, Y. Pattern-Discovery-Based Response Time Prediction. In: *Advances in Automation and Robotics*, vol. 2, LNEE, vol. 123, 355-362, 2012.
[8] Duan, Y., Huang, Y. Research on availability prediction model of web service. In: *Proceedings of the 2011 International Conference on Computer Science and Service System*, 1590–1594, 2011.
[9] O'Sullivan, J., Edmond, D., Ter Hofstede, A. What is a Service?: Towards Accurate Description of Non-Functional Properties. *Distributed and Parallel Databases*, 12, 2002.
[10] Canfora, G., Di Penta, M., Esposito, R., Villani, M.L. An
Approach for QoS-aware Service Composition based on Genetic Algorithms. In: *Proceedings of the 2005 Conference on genetic and evolutionary computation*, 1069-1075, 2005.
[11] Yu, J., Sheng, Q., Han, J., Wu, Y., Liu, C. A semantically enhanced service repository for user-centric service discovery and management. *Data & Knowledge Engineering*, 72, 202- 218, Feb. 2012.
[12] Margaris, D., Vassilakis, C., Georgiadis, P. Combining Quality of Service-based and Collaborative filtering-based techniques for BPEL scenario execution adaptation. University of Peloponnese SDBS Technical report TR-14002, 2014, available at http://sdbs.dit.uop.gr/?q=TR-14002
[13] Alrifai, M., Risse, T. Combining Global Optimization with Local Selection for Efficient QoS-aware Service Composition. In: Proceedings of the 18th international conference on World Wide Web, 881-890, 2009.
[14] Saric, A., Hadzikadic, M., Wilson, D Alternative Formulas for Rating Prediction Using Collaborative Filtering. In: Proceedings of the 18th International Symposium on Foundations of Intelligent Systems, 301-310, 2009.
[15] Bramantoro, A., Krishnaswamy, S., Indrawan, M. A semantic distance measure for matching web services. In: Proceedings of the 2005 International Conference on Web Information Systems Engineering,. 217-226, 2005.
[16] Chelminski, P., Coulter, R. An examination of consumer advocacy and complaining behavior in the context of service failure. Journal of services marketing, 25, 5, 361–370, 2011.
[17] Bixby R.E., Fenelon M., Gu Z., Rothberg E., Wunderling R. Mixed integer programming: A progress report. Chapter in Martin Grötschel (ed.), The sharpest cut: The impact of Manfred Padberg and his work, MPS-SIAM Series on Optimization, Vol. 4, 2004
[18] D. Margaris, C. Vassilakis, P. Georgiadis, "An integrated framework for adapting WS-BPEL scenario execution using QoS and collaborative filtering techniques", Science of Computer Programming, Volume 98, Part 4, 1 February 2015, Pages 707–34, http://www.sciencedirect.com/science/article/pii/S0167642314004778
[19] D. Margaris, C. Vassilakis, P. Georgiadis, "A hybrid framework for WS-BPEL scenario execution adaptation, using monitoring and feedback data", to appear in the 30th ACM Symposium on Applied Computing, Salamanca, Spain, 2015.

Image Analysis and Processing with Applications in Proteomics and Medicine

Eleftheria A. Mylona*

National and Kapodistrian University of Athens
Department of Informatics and Telecommunications
emylona@di.uoa.gr

Abstract. This thesis introduces unsupervised image analysis algorithms for the segmentation of several types of images, with an emphasis on proteomics and medical images. The presented algorithms are tailored upon the principles of deformable models, with an emphasis on region-based active contours. Two different objectives are pursued. The first is the core issue of unsupervised parameterization in image segmentation, whereas the second is the formulation of a complete model for the segmentation of proteomics images, which is the first to exploit the appealing attributes of active contours. The first major contribution is a novel framework for the automated parameterization of region-based active contours. The presented framework aims to endow segmentation results with objectivity and robustness as well as to set domain users free from the cumbersome and time-consuming process of empirical adjustment. It is applicable on various medical imaging modalities and remains insensitive on alterations in the settings of the acquisition devices. The experimental results demonstrate that the presented framework maintains a segmentation quality which is comparable to the one obtained with empirical parameterization. The second major contribution is an unsupervised active contour-based model for the segmentation of proteomics images. The presented model copes with crucial issues in *2D-GE* image analysis including streaks, artifacts, faint and overlapping spots. In addition, it provides an alternate to the laborious, error-prone process of manual editing, which is required in state-of-the-art *2D-GE* image analysis software packages. The experimental results demonstrate that the presented model outperforms *2D-GE* image analysis software packages in terms of detection and segmentation quantity metrics.

Keywords: Segmentation, Active Contours, Proteomics Images, Medical Images.

1 Introduction

Segmentation is a challenging task in image analysis with essential applications in biomedical engineering, remote sensing, robotics and automation. Typically, the

*Dissertation Advisor: Dimitris Maroulis, Professor

target region is separated from the rest of image regions utilizing defining features including intensity, texture, color or motion cues. Moreover, the separation of the target regions is impeded by several daunting factors such as: background clutter, the presence of noise and artifacts as well as occlusions on multiple target regions. This thesis focuses on image segmentation using deformable models and specifically region-based Active Contours (*ACs*) [1] because of their strong mathematical foundation and their appealing properties.

ACs are formulated according to an energy functional defined so as to be minimized when approximating target boundaries. The argument of the energy functional is typically a curve or surface, which evolves and defines the partitioning of the image based on external forces that hinge on image features such as intensity and/or texture. Additionally, internal constraints generate tension and stiffness, which preserve the smoothness and continuity of the model by preventing the formation of sharp corners. The corresponding Euler-Lagrange equation constitutes a Partial Differential Equation (*PDE*), i.e. an iterative gradient descent algorithm, which guides the evolution towards the minimum. The numerical implementation of the evolution is performed by the level set method, which endows the model with topological adaptability, i.e. splitting or merging, appearing or disappearing during the surface evolution.

In this thesis, a novel framework for automated region-based *AC* parameterization is developed, aiming to endow segmentation results with objectivity and robustness as well as to set domain users free from the cumbersome and time-consuming process of empirical parameter adjustment. In addition, an unsupervised *AC*-based model for the segmentation of proteomics images is developed to provide an alternate to the laborious, error-prone process of manual editing by gel analysts.

All ideas presented in this thesis have been published [2],[3], accepted with revisions [4] and submitted [5] in four (4) international peer-reviewed journals, eleven (11) international peer-reviewed conferences [6]-[16], one (1) book chapter [17] and one (1) Hellenic conference [18].

2 Proposed Framework for Automated Parameterization of Region-Based *ACs*

ACs are a rather mature image segmentation paradigm, with several variations proposed in literature. However, their parameterization remains a challenging, open issue, with strong implications on the quality, objectivity and robustness of the segmentation results. Very often, parameters are empirically adjusted on a trial and error basis, a process which is laborious and time-consuming, based on subjective as well as heuristic considerations. On one hand, non-expert users such as Medical Doctors (*MDs)* and biologists require technical support since they are not familiar with the algorithmic inner mechanisms. On the other hand, parameter configurations empirically determined by image analysis experts are usually suboptimal and applicable to specific datasets. A novel framework is proposed for automated adjustment of region-based *AC* regularization and data fidelity parameters based on local image geometry information. Starting from the observation that these parameters

and the eigenvalues of structure tensors are associated with the same orthogonal directions, local image geometry is encoded by the orientation coherence in edge regions. The latter can be mined by means of Orientation Entropy (*OE*), a measure which is an increasing function of the variability in edge orientation, obtaining low values in structured regions containing edges of similar orientations and high values in unstructured regions containing edges of multiple orientations. *OE* is calculated on directional subbands in each scale of the Contourlet Transform (*CTr*) [19], which apart from intensity also represents textural information. As a result, those forces that guide contour away from randomly oriented, high-entropy edge regions are amplified and iterations dedicated to misleading local minima are avoided, speeding up contour convergence. On the other hand, forces imposed within the proximity of structured edges, naturally related to target edge regions, are reduced, enhancing segmentation accuracy.

In the context of the proposed framework, each $q \times q$ image block is fed into the *CTr* filter-bank through an iterative procedure and is decomposed into one pyramidal level, which is then transformed into four directional subbands: 0°, 45°, 90° and 135°. The band-pass directional subbands represent the local image structure. *OE* is calculated on each directional subband image I_{jk} as follows:

$$OE_{jk} = -\sum_{n=1}^{N_{jk}} \sum_{m=1}^{M_{jk}} p_{jk}(m,n) \cdot \log p_{jk}(m,n) \tag{1}$$

$$p_{jk}(m,n) = \frac{|\, I_{jk}(m,n)\,|^2}{\sqrt{\sum_{n=1}^{N_{jk}} \sum_{m=1}^{M_{jk}} [\, I_{jk}(m,n)]^2}} \tag{2}$$

where OE_{jk} is the *OE* of the subband image I_{jk} in the k^{th} direction and the j^{th} level of decomposition, M_{jk} is the row size and N_{jk} the column size of the subband image. Among the *OE* values calculated for each subband image, the maximum value OE_{jk} of the most informative direction *k* is calculated and assigned to all pixels of the corresponding block. The result is considered as an *OE* 'heatmap' reflecting local image structure.

Regularization and data fidelity parameters are matrices of the same dimensions as the original image, and are calculated according to the following equations:

$$w_{reg}^{auto} = a \cdot \left(\frac{1}{w_{df}^{auto}}\right), \quad w_{df}^{auto} = \arg_{I_{jk}} \max(OE_{jk}(I_{jk})) \tag{3}$$

where *a* depends on the dimensions of the image block. The core idea is to guide the active contour towards structured, target edge regions in the early stages of evolution by appropriately amplifying data fidelity forces in randomly oriented, high-entropy regions. As a result the contour will be repelled and iterations dedicated to misleading local minima will be bypassed, speeding up contour convergence towards target

edges. The pipeline of the presented framework is portrayed in the block diagram of Fig. 1.

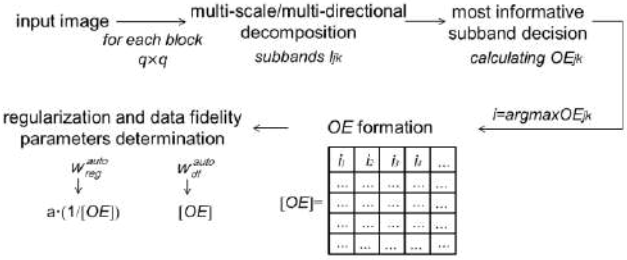
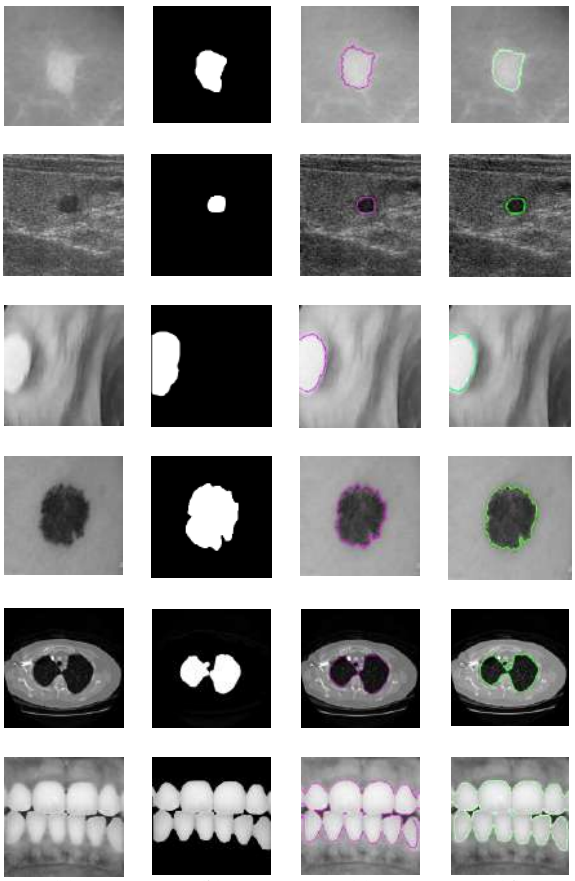


Fig. 1. Block diagram of the pipeline of the presented framework.

The presented framework has been integrated into two region-based [1], [20] and one hybrid [21] *AC* model, in order to evaluate the segmentation performance of the automated versus empirical parameterization. Experiments are conducted on databases of natural and textured images as well as on various medical imaging modalities (mammograms, thyroid ultrasound images, endoscopy images, dermoscopy images, *CT*-scans of lung parenchyma, labial teeth and gingiva photographic images) so as to confirm the framework's generality with respect to image content. The shape of all abnormalities on medical images as well as the irregularity of their margins are malignancy risk factors which are highly considered by *MDs* before proceeding to fine needle aspiration biopsy. Fig. 2 illustrates segmentation results obtained by the automated version on samples of the utilized databases, as well as by the empirically fine-tuned version. The segmentation results depicted in Fig. 2 demonstrate that the presented framework achieves *comparable segmentation quality* to the one obtained by the empirically fine-tuned version *in an automated fashion*. The experimental results are quantitatively evaluated by means of two metrics: the Tanimoto Coefficient (*TC)* [22] and the Hausdorff distance *H* [23] defined as:

$$TC = \frac{N(A \cap B)}{N(A \cup B)}, \quad H(A,B) = \max_{a \in A} \min_{b \in B} \| a - b \| \tag{4}$$

where *A* is the ground truth set, *B* the set under evaluation, *N()* indicates the number of pixels of the enclosed region and *a, b* the points defined in sets *A, B*, respectively. Table 1 presents *TC* and *H* values, obtained by both versions, for each utilized database. The automated version achieves an average *TC* and *H* value of 83.1±1.4% and 40.9±1.8 mm, respectively *with regards to all images tested*, which is comparable to the *TC* and *H* value of 82.0±1.5% and 42.3±3.8 mm respectively obtained by the empirically fine-tuned version. This comparable segmentation accuracy verifies the value of the presented framework for automated parameter adjustment, without the need for laborious fine-tuning from *MDs*.



(a) (b) (c) (d)

Fig. 2. (a) Sample images of the utilized databases, (b) corresponding ground truth images, (c) segmentation results of the empirically fine-tuned version, (d) segmentation results of the automated version.

Table 1: *TC* and *H* values for each utilized database

Database	***TC (%)***		***H (mm)***	
	Empirical	**Automated**	**Empirical**	**Automated**
Mini-MIAS	82.3±1.8	83.4±1.2	42.3±2.5	41.2±1.7
Thyroid US	82.8±1.2	83.7±0.8	43.7±3.3	41.7±2.1
Endoscopy	81.4±1.5	82.3±1.4	41.4±3.8	40.8±1.3
Dermoscopy Atlas	81.7±0.9	82.8±1.6	41.2±4.2	40.1±1.5
CT scans	80.2±1.5	81.8±1.7	40.7±2.6	39.3±2.2
LTG-IDB	82.9±1.6	84.2±1.8	44.8±5.7	42.4±2.5

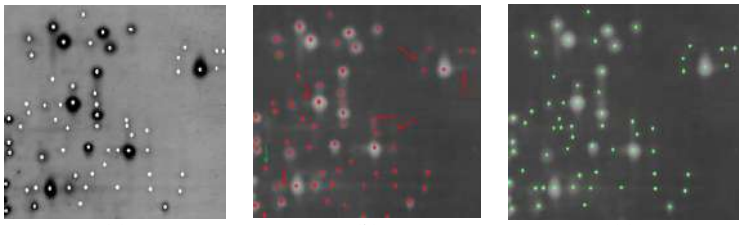
3 Unsupervised *AC*-Based Model for the Detection and Segmentation of Proteomics Images

In this thesis, a novel analysis method is also presented for the detection and segmentation of protein spots in *2D-GE* images. This is the first complete analysis model exploiting the appealing properties of the *AC* formulation in order to cope with crucial issues in *2D-GE* image analysis, including the presence of noise, streaks, multiplets and faint spots. In addition, it is unsupervised, providing an alternative to the laborious, error-prone process of manual editing, which is still required in state-of-the-art *2D-GE* image analysis software packages.

The detection technique utilizes the dilation image operator, which embeds a disk-shaped Structuring Element (*SE*) [24], adjusted to the dominant roundish shape of protein spots. The disk-shaped *SE* limits the falsely detected streaks. *SE* size is set considering that a certain radius value minimizes the detection of false negatives, whereas it allows the detection of local maxima associated with small spots, even in cases where they overlap with larger spots in complex regions.

The accompanying segmentation scheme comprises four main processes, namely: (a) a detection process capable of identifying boundaries of spot overlap in regions occupied by multiplets, based on the observation that such boundaries are associated with local intensity minima, (b) histogram adaptation and morphological reconstruction so as to avoid unwanted amplifications of noise and streaks, as well as to facilitate the identification of faint spots, (c) a contour initialization process aiming to form a level set surface initializing the subsequent level set evolution, based on the observation that protein spots are associated with regional intensity maxima and (d) a level set evolution process guided by region-based energy terms determined by image intensity as well as by information derived from the previous processes.

The presented technique has been experimentally evaluated on a dataset of 13 real *2D-GE* images, containing approximately 26.000 protein spots. This dataset of images was provided by the Biomedical Research Foundation of the Academy of Athens. Melanie 7 [25] software package is used for comparisons. Fig. 3 illustrates: (a) the ground truth image, (b) detection results obtained by Melanie 7 software package and (c) detection results obtained by the presented detection technique. It can be observed that, much more actual protein spots are missed (red arrows), whereas more artifacts are falsely detected as spots (green arrow) by Melanie 7 software package.



(a) (b) (c)

Fig. 3. (a) Ground truth image, (b) detection results obtained by Melanie 7, (c) detection results obtained by the presented detection technique.

The detection results are quantified by means of the Predictive Value (*PV*), Specificity (*SP*) and Detection Sensitivity (*DS*), which are defined as:

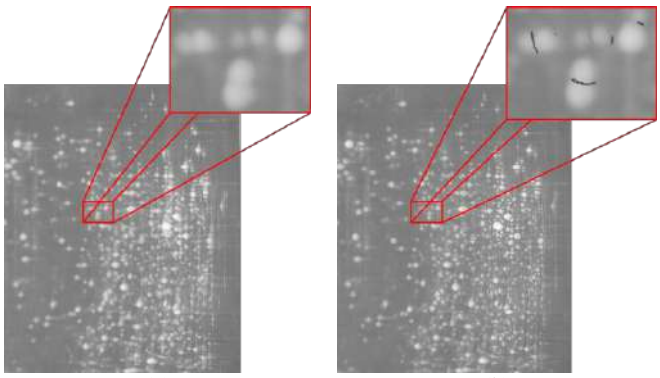
$$PV = \frac{TP}{TP + FP}, \quad SP = \frac{TN}{TN + FP}, \quad DS = \frac{TP}{FN + TP} \tag{5}$$

where *TP, TN, FN* are defined as true positive, true negative and false negative spots. Table 2 presents the *PV*, *SP* and *DS* obtained by the presented detection technique and Melanie 7, in a total of approximately 26.000 protein spots appearing in the dataset of 13 *2D-GE* images. Considering the experimental evaluation it can be concluded that the presented detection technique achieves a *PV,* a *SP* and a *DS* which exceed 80%, outperforms Melanie 7, distinguishes multiple overlapping spots, locates spots within streaks and ignores artifacts.

Table 2: Overall detection results obtained by Melanie 7 and the presented detection technique

	Melanie 7	**Presented Detection Technique**
PV(%)	73.6±17.4	88.2±4.2
SP(%)	33.2±13.5	81.6±5.3
DS(%)	77.4±12.6	87.3±6.2

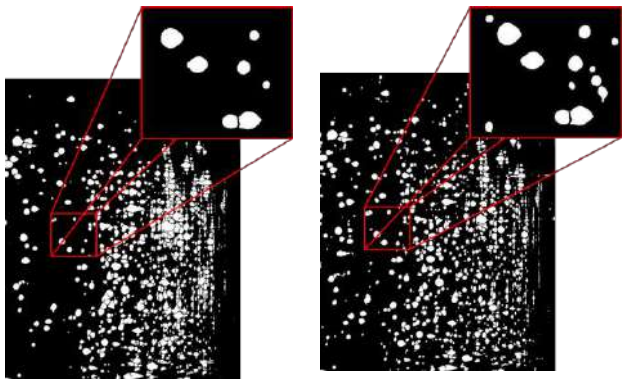
In the context of protein spot segmentation, the presented segmentation scheme is based on the Chan-Vese model [1] and comprises four main processes: (a) separation of multiplets, (b) histogram adaptation and morphological reconstruction, (c) level set initialization and (d) contour evolution. The original *2D-GE* image is scanned with parallel straight-line segments of variable lengths and multiple directions, so as to facilitate the detection of local intensity minima, associated with each particular direction. Local intensity minima are identified for each parallel straight-line segment. Fig. 4 illustrates: a) a real *2D-GE* image, b) the detection results obtained by the local intensity minima process, with each minimum marked as black. It is evident that, the detection process actually identifies boundaries of spot overlap. Therefore, alterations in the pre-processing techniques as well as further manual editing are not required.



(a) (b)

Fig. 4. (a) Real *2D-GE* image, (b) detection results obtained by the local intensity minima process.

A popular histogram equalization variant called Contrast-Limited Adaptive Histogram Equalization (*CLAHE*) [26] is utilized to enhance the segmentation performance of the presented scheme with respect to the presence of faint spots in *2D-GE* images. The enhanced image is binarized according to a threshold value and the flood-fill morphological operation is applied so as to eliminate holes as a result of intensity inhomogeneity. Fig. 5 illustrates the results obtained by the flood-fill morphological operation on: (a) the *2D-GE* image illustrated in Fig. 4(b), (b) the enhanced image which is generated by the application of *CLAHE*. It is evident that, the utilization of *CLAHE* is essential, since most faint spots are missed when *CLAHE* is omitted.



(a) (b)

Fig. 5. Results obtained by the flood-fill morphological operation on: (a) the image illustrated in Fig. 4(b), (b) on the enhanced image which is generated by the application of *CLAHE*.

The level set function is initialized so that the associated zero levels approximate the actual protein spots. Starting from the observation that regional intensity maxima of a *2D-GE* image are associated with protein spots, the presented initialization process constructs a level set surface of multiple cones centered at maxima positions. This surface can serve as a spot-targeted initialization of the level set function. Aiming to enhance segmentation performance, contour evolution is initialized by the spot-targeted level set surface generated by the previous initialization process. The *AC* converges according to the following equation:

$$\begin{aligned}\frac{\partial \varphi}{\partial t} &= w_{reg}^{fixed} \cdot \delta(\varphi(x,y)) \cdot div\left(\frac{\nabla \varphi}{|\nabla \varphi|}\right) \\ &- w_{df_1}^{fixed} \cdot (I_1(x,y) - c_1)^2 + w_{df_1}^{fixed} \cdot (I_1(x,y) - c_2)^2 - \\ &- w_{df_2}^{fixed} \cdot (I_2(x,y) - c_3)^2 + w_{df_2}^{fixed} \cdot (I_2(x,y) - c_4)^2\end{aligned} \tag{6}$$

where I_1, I_2 are the original image and the binarized image which is the output of morphological processing, respectively, c_1, c_2 and c_3, c_4 the average intensities inside and outside of the contour of I_1 and I_2, respectively.

The experimental evaluation of the presented segmentation scheme has been conducted on the dataset of 13 real digital grayscale *2D-GE* images provided by the Biomedical Research Foundation of the Academy of Athens, as well as on a dataset of 30 synthetic *2D-GE* images, so as to facilitate qualitative and quantitative comparisons with state-of-the-art *2D-GE* image analysis software packages. Fig. 6 illustrates segmentation results obtained by the application of Melanie 7 [25], Delta2D, PDQuest 8.0.1 [27] and the presented segmentation scheme on a real *2D-GE* image. It is evident that, the presented segmentation scheme results in more plausible spot boundaries than all three image analysis software packages. PDQuest 8.0.1 results in elliptical boundaries, which do not correspond to the irregular shape of the actual spot boundaries, whereas such elliptical boundaries tend to include background regions. In the cases of Melanie 7 and Delta2D, the obtained segmentation results suffer from over-segmentation and are subject to laborious, error-prone and time-consuming correction process by the expert biologists.

In order to quantitatively evaluate the presented segmentation scheme, experiments were performed on the set of synthetic images generated by the synthetic *2D-GE* image generation software, developed by the Real-time Systems & Image Analysis Lab of our Department. The segmentation performances are quantitatively evaluated in terms of Volumetric Overlap (*VO*) and Volumetric Error (*VE*), which are defined as follows:

$$VO = \frac{ASV}{ASV + FBV}, \quad VE = \frac{FSV}{ASV + FBV} \tag{7}$$

based on the spot volume defined as: $V = \sum_{x,y \in \mathrm{Re}gion} I(x,y)$.

The spot volumes which are calculated according to Eq. (7) correspond to the "Actual Spot Volume" (*ASV*), "False Spot Volume" (*FSV*) and "False Background Volume" (*FBV*), respectively. Table 3 presents the results obtained by Melanie 7, Delta2D, PDQuest 8.0.1 and the presented segmentation scheme. It is evident that, the presented scheme outperforms all three software packages in terms of *VO* and *VE*. Moreover, the presented scheme demonstrates a remarkably lower variance in both performance measures, as a result of its robustness over streaks, multiplets and faint spots.

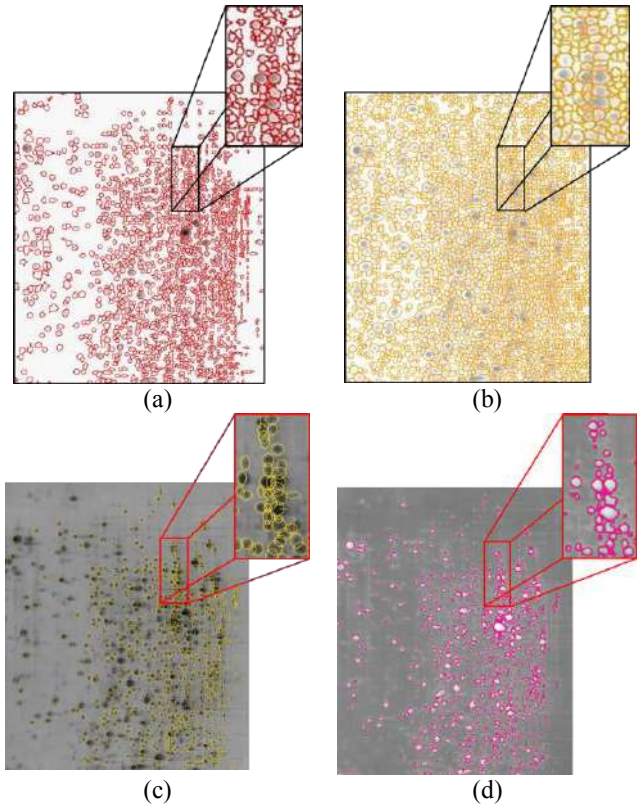


Fig. 6. Segmentation results obtained by the application of: (a) Melanie 7, (b) Delta2D, (c) PDQuest 8.0.1 and (d) the presented segmentation scheme.

Table 3: Segmentation results

	Melanie 7	**Delta2D**	**PDQuest 8.0.1**	**Presented Scheme**
VO	86.5±3.2%	82.4±3.6%	80.2±4.6%	92.0±1.2%
VE	55.0±6.7%	64.3±7.6%	83.1±8.9%	20.0±3.2%

4 Conclusions and Future Work

In this thesis, unsupervised image analysis algorithms have been presented for the detection and segmentation of various types of images focusing on proteomics and medical images. The presented framework for automated adjustment of region-based *AC* parameters was compared to the empirical fine-tuned version and achieved to: a) maintain a high segmentation quality comparable to the one stemmed from each empirically fine-tuned approach, b) speed up contour convergence by selectively amplifying data fidelity forces, c) enrich segmentation results with objectivity and reproducibility and d) relieve domain users from the tedious and time-consuming process of empirical adjustment. Additionally, the presented model for the detection and segmentation of proteomics images achieved to: a) endow detection and segmentation results with objectivity and reproducibility by automatically initializing the level set function based on regional intensity maxima associated with actual spots, b) generate more plausible spot boundaries than commercial image analysis software packages, c) outperform image analysis software packages in terms of *VO* and *VE* segmentation quality measures and d) provide an alternate to the laborious, error-prone and time-consuming process of manual editing, which is required by gel analysis experts in state-of-the-art *2D-GE* software packages.
Future work includes the investigation of active surfaces for 3D segmentation and machine learning algorithms for automated parameter adjustment.

References

[1] T.F. Chan, L.A. Vese, Active contours without edges, IEEE Trans. Im. Proc. 10 (2) (2001) 266-277.

[2] M.A. Savelonas, E.A. Mylona, D. Maroulis, Unsupervised 2D gel electrophoresis image segmentation based on active contours, Pattern Recognition 45 (2) (2012) 720-731.

[3] E.A. Mylona, M.A. Savelonas, D. Maroulis, S. Kossida, A computer-based technique for automated spot detection in proteomics images, IEEE Trans. Inf. Tech. Biomed. 15 (4) (2011) 661-667.

[4] E.A. Mylona, M.A. Savelonas, D. Maroulis, Automated adjustment of region-based active contour parameters using local image geometry, accepted with revisions to IEEE Trans. Cyber. (26/11/2013).

[5] E.A. Mylona, M.A. Savelonas, D. Maroulis, Self-parameterized active contours based on regional edge structure for medical image segmentation," submitted to Med. & Biol. Eng. & Comp. (26/8/2013).

[6] E.A. Mylona, M.A. Savelonas, D. Maroulis, Automated parameterization of active contours: a brief survey, in Proc. IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Athens, Greece, 2013.

[7] E.A. Mylona, M.A. Savelonas, D. Maroulis, Self-adjusted active contours using multi-directional texture cues, in Proc. IEEE International Conference on Image Processing (ICIP), Melbourne, Australia, 2013.

[8] E.A. Mylona, M.A. Savelonas, E. Zacharia, D. Maroulis, C. Pattichis, Unsupervised level set parameterization using multi-scale filtering, in Proc. IEEE International Conference on Digital Signal Processing (DSP), Santorini, Greece, 2013.

[9] E.A. Mylona, M.A. Savelonas, D. Maroulis, A.N. Skodras, Autopilot spatially-adaptive active contour parameterization for medical image segmentation, in Proc. IEEE International Symposium on Computer-Based Medical Systems (CBMS), Porto, Portugal, 2013.
[10] E.A. Mylona, M.A. Savelonas, D. Maroulis, Entropy-based spatially-varying adjustment of active contour parameters, in Proc. IEEE International Conference on Image Processing (ICIP), Orlando, FL, USA, 2012.
[11] M. Savelonas, E. Mylona, D. Maroulis, An automatically initialized level-set approach for the segmentation of proteomics images, in Proc. IEEE International Workshop on Biomedical Engineering, Kos, Greece, 2011.
[12] E.A. Mylona, M.A. Savelonas, D. Maroulis, M. Aivaliotis, 2D-GE image segmentation based on level-sets, in Proc. IEEE International Conference on Image Processing (ICIP), Brussels, Belgium, 2011.
[13] E. Mylona, M. Savelonas, D. Maroulis, A two-stage active contour-based scheme for spot detection in proteomics images, in Proc. IEEE International Conference on Information Technology Applications in Biomedicine (ITAB), Corfu, Greece, 2010.
[14] E. Mylona, M. Savelonas, D. Maroulis, Protein spot detection in 2D-GE images using morphological operators, in Proc. IEEE International Symposium on Computer-Based Medical Systems (CBMS), Perth, Australia, 2010.
[15] M. Savelonas, E. Mylona, D. Maroulis, A level set approach for proteomics image analysis, in Proc. European Signal Processing Conference (EUSIPCO), Aalborg, Denmark, 2010.
[16] M.A. Savelonas, D. Maroulis, E. Mylona, Segmentation of two-dimensional gel electrophoresis images containing overlapping spots, in Proc. IEEE International Conference on Information Technology Applications in Biomedicine (ITAB), Larnaca, Cyprus, 2009.
[17] E.A. Mylona, M.A. Savelonas, D. Maroulis, Towards self-parameterized active contours for medical image segmentation with emphasis on abdomen," A.S. El-Baz et al. (eds.), Abdomen and Thoracic Imaging: An Engineering & Clinical Perspective, Springer Science + Business Media, New York, 2014.
[18] E. Mylona, M. Savelonas, D. Maroulis, Computer-based methodology for protein spot detection and its contribution to health services, 14th Panhellenic Conference of the Greek Physicists Union, Kamena Vourla, Greece, 2012.
[19] M.N. Do, M. Vetterli, The Contourlet transform: an efficient directional multiresolution image representation, IEEE Trans. Im. Proc. 14 (12) (2005) 2091-2106.
[20] X. Bresson, S. Esedoglu, P. Vandergheynst, J. Thiran, S. Osher, Fast global minimization of the active contour/snake model, J. Math. Im. Vis. 28 (2) (2007) 151-167.
[21] C. Li, C. Xu, C. Gui, M.D. Fox, Distance regularized level set evolution and its application to image segmentation, IEEE Trans. Im. Proc. 19 (12) (2010) 3243-3254.
[22] W.R. Crum, O. Camara, D.L.G. Hill, Generalized overlap measures for evaluation and validation in medical image analysis, IEEE Trans. Med. Im. 25 (11) (2006) 1451-1461.
[23] D. Huttenlocher, G. Klanderman, W. Rucklidge, Comparing images using the Hausdorff distance, IEEE Trans. Patt. Anal. Mach. Intell. 15 (1993) 850-63.
[24] P. Soille. Morphological Image Analysis-Principles and Applications. Springer, Berlin, 1999.
[25] R.D. Appel, J.R. Vargas, P.M. Palagi, D. Walther, Melanie II – a third-generation software package for analysis of two-dimensional electrophoresis images: II. Algorithms, Electrophoresis 18 (1997) 2735-2748.
[26] S.M. Pizer, E.P Amburn, J.D. Austin, Adaptive histogram equalization and its variations, Comp. Vis. Graph. Im. Proc. 39 (1987) 355-368.
[27] J.I. Garrels, The QUEST system for quantitative analysis of two-dimensional gels, J. Biol. Chem. 264 (1989) 5269-5282.

Internet Content Management using Complex Network Analysis techniques

Panagiotis Pantazopoulos*

Department of Informatics and Telecommunications,
National & Kapodistrian University of Athens
Ilissia, 157 84 Athens, Greece
ppantaz@di.uoa.gr

Abstract. With the explosion of (user-generated) content and the heterogeneity of users/devices, today's Internet has evolved into a system of extreme complexity hindering the design of effective network protocols. The availability of global information, either topological or related to the users' profiles is non-realistic if not impossible to obtain. As such, a significant open challenge that the thesis seeks to address involves the management of Internet content in a distributed manner with the utilization of local-scope information. Towards this end, the socio-technical insights provided by the interdisciplinary framework of Complex Network Analysis (CNA) have attracted the interest of the research community, yet remain under-explored. The thesis addresses instances of content management problems over mobile opportunistic networks as well as wired ISP networks at the router-level. Among the mobile opportunistic nodes, we seek to identify the appropriate nodes to store and efficiently provide content to the rest of the network. Over the ISP topologies, we study the distributed Internet service placement and the content search driven by local information. The vulnerability of these networks to node attacks is also explored from both the connectivity and content-related standpoint. The novelty dimension of the thesis lies in the effort to introduce and promote centrality metrics to important parameters for the design of Internet content operations. Our results obtained by analysis and simulations over real-world data, reveal useful insights and provide guidelines as to how (local) centrality information can efficiently drive the management of Internet content.

1 Introduction

Content management is a summary term for a broad range of operations that become relevant from the moment some content is produced somewhere in the communication network infrastructure till the time this content is consumed by end-users, including content search, and discovery, content placement and dissemination. With the role of content continuously growing and the latest advances in Web 2.0 technologies fueling user-generated content, end-users increasingly participate in these operations, departing from the monolithic role of content consumer.

The new role of end-users occasionally acting as content providers as well as the interactions between them and the network devices, traditionally, has not been systematically taken into account in the design of network protocols. The emerging structures

* Dissertation Advisor: Ioannis Stavrakakis, Professor

and properties that are due to these interactions present network environments with "social dimensions" that have recently gained momentum in the context of more user-oriented approaches to data networking. Indeed, there has been recent evidence, which motivates the proposed research, that taking explicit account of these social attributes of end-users does benefit content management operations substantially [2] [3]. In light of the content steadily gaining importance in communication networks, the need to make its management more efficient motivates the introduction of novel, socio-aware approaches. The primary concern is to devise distributed and scalable Internet solutions that utilize locally available information.

The highly interdisciplinary framework of Complex Network Analysis (CNA) seeks to describe the research for modeling the behavior of networked interacting elements. It typically utilizes a set of graph-theoretic metrics for describing relations between nodes or groups of nodes. Expectations within the networking community are that the relevant socio-technical insights could assist simplifying the high networking complexity and benefit the design of efficient protocols. The thesis aims to pursue in-depth this research thread by analyzing pervasive CNA-driven content management mechanisms over the current/future Internet. Content-related operations are nowadays carried out over wired ISP networks as well as over dynamically changing (often time-evolving) topologies such as mobile ad-hoc and opportunistic networks. Those networking settings tend to become highly user-centric network paradigms, realized by the end-user participation. Devising topology- and resource-control mechanisms that accommodate the content management operations is a challenging yet promising task. In particular, the proposed research seeks to exploit the CNA-based knowledge and relevant insights for the optimization of content-centric operations in the context of current and future communication network paradigms. At the same time, studying CNA-concepts over complex real-world Internet topologies, the thesis makes solid contributions to the emerging and much promising scientific discipline of Network Science.

2 State-of-the art and outline of the thesis

In this section we briefly present the related publications and outline the corresponding contents of the thesis.

2.1 Background in opportunistic content placement and data forwarding

The opportunistic communications paradigm Opportunistic networks [15] are self-organizing wireless mobile networks formed by user devices such as PDAs or Smartphones, without requiring any pre-existing network infrastructure. Communications in this environment have been typically supported through Mobile Ad hoc Networks (MANETs). However, MANET solutions work only if a rather stable topology can be established among the nodes, which is often not the case in the presence of users' mobility and low node density. Opportunistic networks support communication among nodes even when no stable multi-hop paths between communication endpoints can be established. In an opportunistic network a node carrying a content addressed to a given destination evaluates if any other node it comes in direct contact is "better suited" than

itself to bring the content to the destination. In other words, each contact is opportunistically exploited to bring the content closer and closer to the destination.

Content placement in opportunistic networks Content placement in opportunistic networks can be seen as an instance of the broad data dissemination problem. The latter amounts to the storage of files or, more generally, information objects, in specific network nodes, so that they can be provided to requesting nodes at smaller access costs. Those problems have recently attracted the attention of the networking community. ContentPlace, which is a social-oriented framework for data dissemination constitutes a state-of-the-art relevant solution [1]. ContentPlace assumes that nodes can be aware of the social communities they belong to. Using a general utility-based optimization framework, ContentPlace defines distributed algorithms for nodes to select which content to locally replicate, out of what is available on encountered nodes. These algorithms consider the estimated distribution of content in the network, and the interests of the users with respect to content. Similar approaches have shown that mobility and cooperative content replication strategies can help bridge social groups [11].

In the second Section of the thesis briefly discussed in 3.1, we address the content placement problem in an opportunistic network by incorporating a social dimension to its solution. The proposed algorithm can provide for easy adaptation to dynamic environments due to its local-information-requiring approach and can therefore, cope with complex multimedia content increasingly generated by the network users.

Opportunistic Data Forwarding The opportunistic forwarding problem amounts to the decision that a node has to take on whether some encountered node is more appropriate (or not) to physically carry the data to a given destination. Initial approaches to the problem have essentially been variants of controlled flooding across the network. These schemes reduce the cost of pure epidemic dissemination by setting upper bounds on the message replication(*e.g.*, [18]). More informed decisions are made by forwarding schemes that try to assess the relaying significance (*utility*) of encountered nodes. They may account for the frequency of encounters with the destination node or more general social context such as preferably visited places, common to the candidate relay and the destination node. Recently, social information has been introduced into the node relaying utility functions through direct reuse of Complex Network Analysis concepts. Examples of this approach are the SimBetTS and BubbleRap protocols. In both cases, the CNA metrics are computed over contact graphs that effectively aggregate the sequence of node encounters over certain time windows. In SimBetTS [3] the nodes' utilities are sums of their centrality, similarity, and tie strength values, the latter reflecting the frequency, duration, and recency of the contacts with other nodes. Whereas, BubbleRap [4] explicitly assumes the existence of nodes' communities and manipulates the centrality of encountered nodes, both within their communities and globally across the whole network, to route messages within and across these communities, respectively. Both protocols have reported enhanced performance over the controlled flooding and identified centrality as the metric with the dominant impact on routing even when it is combined with other social metrics.

What we seek to assess in the third Section of the thesis, summarized here in 3.2, is the inherent weaknesses of centrality-based routing that pose hard limits to the performance of socioaware opportunistic routing. Relevant questions amount to but are not limited to: How close-to-optimal can routing decisions based on centrality metrics be? How do different alternatives for *computing* node centrality affect routing performance?

2.2 Background in content/service placement over physical topologies

Efficient content/service placement can dramatically reduce access speeds and related costs, whether viewed in technical or economical terms, and, therefore, improve the quality of the provided service. The optimal placement of content (or a service facility) within a network structure has been typically tackled as an instance of the facility location problem [7]. Input to the problem is the topology of the network nodes that may store the content or (host the service), their costs of installation and the distribution of service demand across the network users. An appropriate formulation of the problem is needed to cope with the current networking trends.

Nowadays, the Web2.0 technologies have enabled a paradigm shift towards more user-centric approaches to content generation and provision. This shift is strongly evidenced in the abundance of *User-Generated Content*(UGC) in social networking sites or video distribution sites (*e.g.*, YouTube). The generalization of the UGC concept towards services has become a major trend in user-oriented networking. The user-oriented service creation concept engages end-users in the generation and distribution of service components. In parallel with the proliferation of the so-called *User-Generated Service* (UGS) paradigm, significant research is being carried out on the design and deployment of energy-efficient data storage architectures. Numerous peer devices such *e.g.*, home-gateways are instrumented using virtualization to create a distributed Internet service platform that leverages end-user proximity. The "in-network storage" argument has also been a key-concept of the emerging Information-centric networking paradigm [2]. In the intersection of these trends, we anticipate a rich ecosystem of service instances that will be generated in almost every network location and will have access to storage resources in various network locations. The technical challenge then is how to place them in a way that minimizes their access cost.

To identify the optimal location a large optimization problem needs to be solved. Thus, approximation algorithms have been so far proposed, the majority of which address the problem employing global knowledge. The Greedy algorithm [16] is a relevant heuristic solution that sequentially places one replica at a time; the current one is placed at the lowest-cost location exhaustively determined under the assumptions that a) the so far placed replicas remain fixed b) a node's requests are directed to the closest replica. The approach achieves placements within a factor of 1.1-1.5 of the optimal for synthetic and real-world network topologies under real-world demand patterns. However, in distributed settings it is infeasible to acquire global input information. The problem is further amplified by the massive *UGC*/*UGS* trend that especially calls for scalable solutions. A few distributed approaches have been so far proposed; most of them are heuristics that employ locally-determined information and migrate the service towards prominent locations. The authors in [17] propose the R-ball heuristic. They iteratively solve multiple small-scale optimization problems within an area of R-hops from the

current location of each service facility and move each service towards (near-)optimal locations. A slightly different instance exploits the shortest-path tree structures induced on the network graph by the routing protocol operation to estimate upper bounds for the aggregate cost when the service migrates to its immediate neighbors.

In the forth Section of the thesis we present a scalable heuristic approach to deal with the complexity and limitations of the distributed service placement. Node centrality insights help us iteratively migrate service facilities towards near-optimal locations achieving very good accuracy and fast convergence. A brief description of the algorithm and the corresponding protocol implementations can be found in this document, in 4.1.

2.3 Background in local centrality metrics

The analytical (graph-theoretic) tools that the Complex Network Analysis (CNA) provides are expected to benefit the design of efficient network protocols. Indeed, there has been evidence that CNA insights can improve network functions such as content-caching strategies in wired networks [2] and routing/forwarding in opportunistic networks [3]. Common denominator to these efforts is the use of CNA-driven metrics for assessing the relative centrality (*i.e.*, importance) of individual network nodes, whether humans or servers. The computation of these metrics, however, typically demands *global* information about all network nodes and their interconnections. The distribution and maintenance of this information is problematic, if not infeasible, in large-scale networks such as the Internet ISP topologies. A more realistic alternative for assessing node centrality draws on its ego network, *i.e.*, the subgraph involving itself, its 1-hop neighbors, and their interconnections. *Egocentric* measurements, carried out within their immediate locality, let nodes derive *local approximations* of their centrality. Lending to simpler computations, egocentric metrics have, in fact, found their way into protocol implementations [3], [2]. Nevertheless, the capacity of these local approximations to substitute the globally computed sociocentric metrics over the Internet is almost always taken for granted rather than evaluated.

In Section 5 of the thesis, summarized in 4.2, we employ ISP network topologies and question how well do sociocentric node centrality metrics, computed under global topological information, correlate with their egocentric variants, as computed locally over the nodes' ego networks. More importantly, we proceed to study what the measured correlation coefficients can reveal regarding the capacity of rank-preserving local centrality metrics to substitute the original global metrics in some elementary network operations *i.e.*, instances of content-related Internet protocols.

3 Content Management over Opportunistic Networks

In this section we briefly describe the solutions proposed in the thesis regarding the placement of content as well as the message forwarding over opportunistic networks.

3.1 Content Placement

CNA insights are being exploited here to determine the optimal physical location of the node to host some content by solving iteratively a spatially restricted, low-cost

1-median problem (instead of a costly, global one) [11]. Consequently, the proposed algorithm can provide for easy adaptation to dynamic environments due to its local-information-requiring approach. As approximations are involved, the achieved solution may not always lead to the identification of the optimal location; nevertheless, the results presented in the thesis show that the divergence error is not substantial.

First, the thesis argues that contact patterns among users naturally hint to a graph representation of the network, where a link exists among nodes if they are "enough frequently" in touch with each other. With this representation and the assumption of full topology knowledge over a limited region around some node currently hosting the content, the thesis introduces a CNA-based criterion, for selecting a number of fairly neighboring nodes (forming a subgraph) to take part in a small-scale, local solution. This, may be derived by employing any well-studied centralized approach. The selected nodes (a percentage of total number) are the ones, having the top values of an innovative metric, inspired by typical CNA centrality metrics, that plays a twofold key role. First, it captures a node's significance, regarding its capability to transport content efficiently. Second, it captures the contribution (to the *1-median* solution) of incoming demand from the rest of the network nodes, not included in the above subgraph. In mathematical terms the introduced metric called Conditional Betweenness Centrality (CBC), captures the topological centrality of a network node with respect to a specific node t. If σ_{st} denotes the number of shortest paths between any two nodes s and t in a connected graph G=(V, E) and $\sigma_{st}(n)$ is the number of shortest paths passing through node $n \in V$, then CBC is:

$$CBC(u;t) = \sum_{s \in V, u \neq t} \frac{\sigma_{st}(u)}{\sigma_{st}} \tag{1}$$

The thesis shows that the demand, following a uniform model, can be deduced from this metric which is a measure of the role of every node in the information flow, towards the one having the content. After solving the small-scale $1 - median$ problem, the node that minimizes the content provision cost (if hosting the content) is identified. Then, the content is assumed to be placed in this node. The new subgraph is determined around this node and the new small-scale optimization problem is solved. This iterative procedure repeats until no further movement of the content is possible. The content therefore, moves according to the outcome of the optimization, on a cost-decreasing path, trying to reach the optimal location. In the thesis we analyze the performance of the proposed heuristic studying its degree of approximation of the optimal centralized solution on two types of graphs representing different user contact patterns, i.e. Erdős-Rényi and the Barabási-Albert graph model. Our simulation shows that the heuristic achieves satisfying convergence to the optimal solution. Interestingly, it performs much better under the B-A model, which is known to capture user social interactions and contact patterns.

3.2 Message Forwarding

In the thesis we present our study of centrality-based opportunistic routing employing real traces of pairwise node encounters. We have used five well-known experimental traces, gathered in the context of the Haggle Project [8]. Rather than consid-

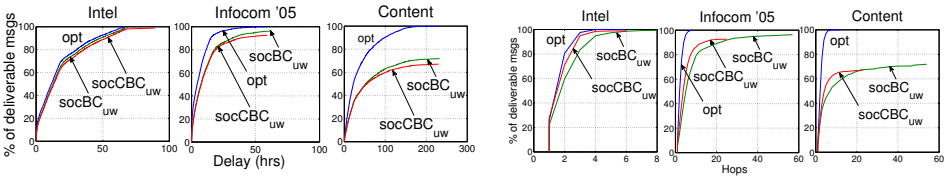


Fig. 1. a) Message delivery delay distribution for destination-aware ($socCBC_{uw}$) and -unaware ($socBC_{uw}$) centrality-based forwarding. b) Message hopcount distribution for destination-aware ($socCBC_{uw}$) and -unaware ($socBC_{uw}$) centrality-based forwarding.

ering a particular protocol, we have assessed the basic routing primitive of centrality-based schemes. We have studied three fundamental alternatives to the sociocentric (*i.e.*, global-info-requiring) computation of the betweenness centrality over contact graphs. Methodologically, our approach amounts to the following steps: First we transform the encounters' history to a graph, over which the node centrality values are computed. We produce two different static graph representations for each trace: one unweighted and one weighted, where the edge weight equals the inverse of their encounters' count. We then generate message triplets $msg(s, d, t)$, where the message source s, destination d, and generation time t are randomly chosen, and emulate their paths over the traces [5]. As the trace is replayed (sequentially read), network nodes compute online their BC values and make forwarding decisions for each message. The generated $msg(s, d, t)$ messages are routed to their destination through successive 'greedy' forwarding decisions. More specifically, if the message is with node u, then u will forward the message to another node, say k, upon its next encounter with it, as long as its betweenness centrality value $BC(k)$ is higher than its own BC value, $BC(u)$ We consider five possible ways to compute BC values assuming that nodes avail perfect information about the history of encounters in the network. Hence, any routing performance penalty is due to the (lack of) informative power of the metric rather than information unavailability. The various BC computations are marked by the abbreviation terms $soc(C)BC_{uw}$, $egoBC_{uw}$, $socBC_w$, $egoBC_w$, for routing based on global (*i.e.*, sociocentric) and local (*i.e.*, egocentric) (C)BC values estimated on unweighted or weighted graphs, respectively. We measure the message delivery delay and the number of message forwarding hops.

A representative result in Fig. 1 compares the performance of optimal, BC- and CBC-based routing schemes over unweighted contact graphs. As expected, the centrality-based approaches perform worse than the optimal method both in terms of message delay and hops. Up to 30% of messages are trapped and do not reach their destination. Note that the performance lag of centrality-based schemes varies from trace-to-trace and heavily depends on the extent that the mobility patterns of users mix with each other. Less intuitively, the CBC-based forwarding does not outperform the BC-based forwarding with respect to delays. Although the use of CBC implies message routing through more appropriate relays towards the destination, there are numerous graph in-

stances where its interpretation turns out to be problematic. When the aggregation of contacts yields non-connected clusters of nodes, the CBC values of nodes outside the destination's cluster are by default zero. Messages stay long with the source node or are trapped quickly at some intermediate node. On the contrary, with use of the BC metric, nodes take on easier non-zero values and the resulting variance of the metric across nodes lets the message hop from one node to the other. On the other hand, this extra message agility comes at higher cost; under BC-based forwarding, messages end up traversing up to 50% longer paths than under the use of CBC (Fig 1.b), *i.e.*, messages end up traveling far more in the network.

4 Content management over ISP network topologies

In this section we describe the solutions proposed in the thesis regarding the distributed service placement and the utility of local centrality metrics when used to drive basic network operations.

4.1 The Distributed Service Placement Problem

Given the network topology and demand dynamics, the k-$median$ problem prescribes the locations for instantiating a fixed number of service facilities so as to minimize the aggregate cost of accessing them over all network users [7]. We focus on the single facility scenario that matches better the UGS paradigm, *i.e.*, various service generated in the network raising small-scale interest so that replication of their facilities be less attractive [14]. The network is represented by an undirected connected graph $G(V, E)$ of $|V|$ nodes and $|E|$ edges. If $w(n)$ denotes the aggregate service demand generated by node n and $d(k, n)$ is the minimum cost path between nodes k and n, then the 1-median problem formulation seeks to minimize the access cost of a service located at node $k \in V$: $Cost(k) = \sum_{n \in \mathcal{V}} w(n) \cdot d(k, n)$.

The proposed heuristic (called cDSMA) relies on a traffic-aware variant [12] of the earlier introduced CBC metric (eq.1) that we call *weighted*-CBC ($wCBC$). It helps each service instance to migrate towards its final location through a finite number of steps:

Step 1*: Initialization*. The algorithm execution starts at the node s that initially generates the service facility. The service placement cost at node s is assigned an infinite value to secure the first iteration

Step 2*: Metric computation and 1-median subgraph derivation*. Next, the computation of $wCBC(u; s)$ metric takes place for every node u in the network graph $G(V, E)$. Nodes featuring the top $\alpha\%$ $wCBC$ values together with the current service $Host$ form the subgraph G^i_{Host} (i^{th} iteration), over which the 1-median problem will be solved.

Step 3*: Mapping the demand of the remaining nodes on the subgraph*. By restricting the solution domain to the G^i_{Host} subgraph, the contribution of the "outside world" to the service provisioning cost would be totally neglected. To allow for its inclusion, the demand for service from the $G \setminus G^i_{Host}$ nodes is mapped on the G^i_{Host} ones.

Step 4*: 1-median problem solution and service migration to the new host node*. Any centralized technique may be used to solve this small-scale optimization and determine

Table 1. cDSMA accuracy assessment for various datasets under different demand distributions

ISP	AS number	Nodes	Diameter	Mean Degree	Minimum $\lvert G^i_{Host}\rvert$ nodes to yield cost$\leq$1.025opt Uniform demand	Zipf demand (s=1)
Global Crossing	3549	100	9	3.78	7	6
NTTC-Gin	2914	180	11	3.53	18	10
Sprint	1239	184	13	3.06	8	7
DFN-IPX-Win	680	253	14	2.62	7	6
Level-3	3356	378	25	4.49	4	4

the optimal location of the new $Host$ among the G^i_{Host} nodes. If the value of the emerging cost is smaller than the current one, the service is moved to the new $Host$ and the algorithm iterates steering the service to (near-)optimal locations.

Practical protocol implementations for different routing strategies The $wCBC(u;t)$ metric practically represents the service demand that node u routes towards node t, including its own demand and the transit demand flowing from other network nodes through u towards t. Therefore, individual nodes may, in principle, *estimate* their own metric values $wCBC$ through passive measurements of the service demand they route towards the current service host node. What is actually computed theoretically demanding global information about the network topology and service demand, can be locally approximated by u providing the basis for the practical implementation of a distributed solution. The approximation lies in the fact that what is measured, even with perfect accuracy, is not always equal to the nominal $wCBC$ value. In the thesis we show that what matters is the estimate of the actual demand routed through the node.

Using this locally-obtained estimate values of each node's $wCBC$ the thesis proposes a real-world *distributed* implementation for cDSMA, catering for all challenges related to distributed operation: how the node each time hosting the service collects topological and demand information and how it uses it to reconstruct the inputs needed by the algorithm. Each network node communicates the $wCBC$ estimates via dedicated messages to the current service host. This information can then be processed by the service host to extract partial topological information about the 1-median subgraph and determine the next service host on its migration path.

Results on ISP topologies We present a portion of our experimental results regarding the performance of the theoretical solution *i.e.*, when the involved metrics are globally determined. We employ real-world ISP network topologies [10] that feature adequate variance in size, diameter, and connectivity degree statistics. Table 1 reports the minimum number of nodes $|G^i_{Host}|$ required to achieve a solution that lies within 2.5% of the optimal for different levels of service demand asymmetry. The $|G^i_{Host}|$ values show remarkable insensitivity to both topological structure and service demand dynamics. Employing 4.5% of the total number of nodes or 6% for the least favorable case suffices to obtain very good accuracy across all ISP topologies. Likewise, the required 1-median subgraph size remains in almost all experiments practically invariable with the demand distribution skewness.

More importantly, the thesis shows that the cDSMA heuristic lends to realistic protocol implementations that realize the distributed service placement. Our relevant re-

sults suggest that the proposed practical implementations preserve similar advantages with the theoretical cDSMA algorithm regardless of the employed routing policies. Our experiments also show that the cDSMA implementations exhibit welcome scalability properties with respect to the number of the involved nodes and message overhead. Finally, the concentration of Internet services (over highly central nodes) is almost negligible when the cDSMA implementations operate under realistic service demand scenarios.

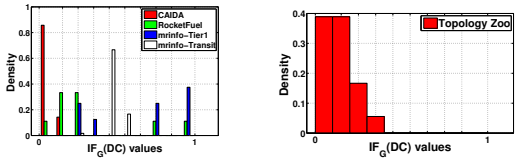
4.2 The utility of local centrality metrics

A realistic alternative for assessing node centrality yet suitable for distributed Internet environments draws on local structures such a node's ego network [6]; the latter consists of the subgraph involving the ego-node, its 1-hop neighbors, and their inter-connections. Egocentric measurements, carried out within their immediate locality, let nodes derive local approximations of their centrality.

In view of the fact that most centrality-driven protocol implementations rely on the (BC-determined) ranking of nodes rather than the absolute values we first investigated how well do BC metrics computed under global topological information, correlate with their locally-determined variants over real-world Internet topologies [13]. Our results presented in Chapter 5 of the thesis suggest that the original BC and its destination-aware variant, exhibit high rank-correlation (in the order of 0.8-0.9) with their local counterparts across almost all 20 ISP snapshots studied. On the other hand, the match between the two variants is much worse when we compare the top-k nodes selected by each of them. Then, we have tried to assess what the algebraic values of the correlation coefficients reveal regarding the performance of network functions, when the original metrics are substituted by their local approximations. Both a simple navigation and a content search scheme employing local centrality metrics produce significantly different navigation patterns and lower hitrates, respectively, than their counterparts with the original global metrics. These results suggest that, despite the positive correlations, local variants can hardly offer effective approximations to the original metrics.

5 Vulnerability of ISP networks to intelligent node attacks

In the sixth Section of the thesis we have studied a problem that relates implicitly to the Internet content management; namely, we have experimentally assessed how vulnerable are the router-level Internet topologies to attacks (*i.e.*, removals) targeting their top central nodes [9]. Content-related operations will benefit if these topologies can maintain high levels of network connectivity and volumes of accommodated traffic, in face of the attacks. To identify the top central nodes of each topology we have used the seven most popular centrality metrics and first studied how different are the node rankings (measured by correlation and top-k overlap), they induce. Included in the set of these centralities, is the only local metric *i.e.*, the degree centrality(DC) which is of apparent interest since it requires almost no computation to be obtained. Then, with respect to those rankings, highly central nodes are removed and the impact on both the connectivity and the traffic-carrying capacity of each network are assessed.



a. b.

Fig. 2. Empirical probability mass function of the normalized distance measure $IF_G(DC)$ computed *w.r.t* the size of the giant component (a) and the maximum accommodated flow (b) for different ISP network datasets.

In terms of the matching of node rankings, we have found (inline with intuition) that it is the top-5% overlap rather than the full rank correlation that predicts more accurately when the node removals that are determined by two different indices have similar impact on the network. Regarding the centrality-driven attacks, the use of different centralities for the choice of the nodes to-be-attacked varies significantly their impact on the network. Focusing on the only local metric (*i.e.*, DC) we have introduced a normalized distance metric called IF to quantify how close to the most catastrophic attack is the impact of the DC-driven removals. Our results suggest that the locally computed DC approximates closely the most catastrophic global centrality metrics with respect to the traffic-carrying capacity (Fig. 2.b). On the contrary, such an approximation is more topology-dependent in terms of connectivity (Fig. 2.a).

6 Conclusions

The thesis proposes algorithms and protocols to enhance conventional content (or service) management functions over the Internet. The task of content management may include the generation, transfer and provision of some piece of various-typed informative content over underlying networks of diverse characteristics. The relevant research is triggered by both the high complexity of emerging distributed Internet environments as well as the preliminary yet promising results of the upcoming Complex Network Analysis (CNA) framework. Employing centrality metrics, the thesis seeks to devise content management solutions that can cope with today's content explosion as well as the need for distributed approaches utilizing local information. The focus lies on two instances of the current/emerging Internet environment, the mobile opportunistic networks and the wired ISP router-level topologies. In the former case, we seek to a) identify the nodes to store a content item that will be provided to the rest of the network nodes at minimum access cost and b) experimentally assess the impact of centrality computations on the efficacy of socio-aware opportunistic forwarding. In the latter case, the question investigated pertains to the distributed placement of Internet services and the utility of local centrality metrics to drive basic network operations. Finally, we study CNA metrics in the Internet vulnerability context. The thesis introduces and promotes centrality metrics to important parameters for the design of Internet content-related operations. Our results obtained by analysis and simulations over real-world data, essentially provide yet another evidence that the incorporation of complex network structures information is a

promising path to the development of efficient content-centric networking protocols for future Internet users.

References

1. Boldrini, C., Conti, M., Passarella, A.: Design and performance evaluation of contentplace, a social-aware data dissemination system for opportunistic networks. Computer Networks 54(4), 589–604 (2010)
2. Chai, W.K., He, D., Psaras, I., Pavlou, G.: Cache "less for more" in information-centric networks. In: Proc. of the 11th IFIP Networking. Prague, Czech Republic (May 2012)
3. Daly, E.M., Haahr, M.: Social network analysis for information flow in disconnected delay-tolerant manets. IEEE Trans. Mob. Comput. 8(5), 606–621 (2009)
4. Hui, P., Crowcroft, J., Yoneki, E.: Bubble rap: Social-based forwarding in delay-tolerant networks. IEEE Trans. Mob. Comput. 10(11), 1576 –1589 (nov 2011)
5. Karaliopoulos, M., Pantazopoulos, P., Jaho, E., Stavrakaki, I.: Trace-driven Analysis of Data Forwarding in Opportunistic Networks. In: Proc. of the 2nd Conference on the Analysis of Mobile Phone Datasets and Networks (NetMob'11). Cambridge, MA, USA (2011)
6. Marsden, P.: Egocentric and sociocentric measures of network centrality. Social Networks 24(4), 407–422 (October 2002)
7. Mirchandani, P., R.Francis: Discrete location theory. John Wiley and Sons (1990)
8. Nikolopoulos, P., Papadimitriou, T., Pantazopoulos, P., Karaliopoulos, M., Stavrakakis, I.: How much off-center are centrality metrics for routing in opportunistic networks. In: Proc. of the 6th ACM Workshop on Challenged Networks. CHANTS '11, Las Vegas, USA (2011)
9. Nomikos, G., Pantazopoulos, P., Karaliopoulos, M., Stavrakakis, I.: Comparative assessment of centrality indices and implications on the vulnerability of ISP networks. In: 26th International Teletraffic Congress (ITC 2014). Karlskrona, Sweden (Sep 2014)
10. Pansiot, J.J., et al.: Extracting intra-domain topology from mrinfo probing. In: Proc. PAM. Zurich, Switzerland (April 2010)
11. Pantazopoulos, P., Stavrakakis, I., Passarella, A., Conti, M.: Efficient social-aware content placement for opportunistic networks. In: IEEE WONS. Kranjska Gora, Slovenia (2010)
12. Pantazopoulos, P., Karaliopoulos, M., Stavrakakis, I.: Centrality-driven scalable service migration. In: The 23rd International Teletraffic Congress (ITC). San Francisco, USA (2011)
13. Pantazopoulos, P., Karaliopoulos, M., Stavrakakis, I.: On the local approximations of node centrality in internet router-level topologies. In: IFIP IWSOS. Mallorca, Spain (Mar 2013)
14. Pantazopoulos, P., Karaliopoulos, M., Stavrakakis, I.: Distributed placement of autonomic internet services. IEEE Trans. on Parallel and Distributed Systems 25(7), 1702–1712 (2014)
15. Pelusi, L., Passarella, A., Conti, M.: Opportunistic networking: data forwarding in disconnected mobile ad hoc networks. IEEE Communications Magazine 44(11), 134 (2006)
16. Qiu, L., Padmanabhan, V.N., Voelker, G.M.: On the placement of web server replicas. In: Proceedings of the IEEE INFOCOM'01. vol. 3, pp. 1587–1596 (2001)
17. Smaragdakis, G., et al.: Distributed Server Migration for Scalable Internet Service Deployment. IEEE/ACM Transactions on Networking 22(3) (2014)
18. Spyropoulos, T., et al.: Efficient routing in intermittently connected mobile networks: The Multiple-Copy case. IEEE/ACM Trans. Netw. 16(1), 77–90 (Feb 2008)

Advanced modulation schemes and signal processing techniques for transmission in highly multimode fibers

Evangelos Pikasis*

National and Kapodistrian University of Athens
Department of Informatics and Telecommunications
vag_pik@di.uoa.gr

Abstract. In this dissertation, advanced modulation schemes and digital signal processing techniques are proposed and investigated through both numerical simulations as well as experiments, in order to overcome the limitations of multimode step index plastic optical fibers (SI-POF) to support data rates in the order of Gbps. In particular, novel multi-carrier modulation techniques with the inherent property of symbol spreading (spreading multicarrier modulation schemes) are proposed and applied. These schemes are Discrete Fourier Transform Spread Discrete Multitone (DFT-Spread DMT) and Code Division Multiple Access Discrete Multitone (CDMA-DMT), which, in this thesis, are suitably adapted for short-range IM/DD transmission via SI-POF and compared against conventional Discrete Multitone (DMT) in terms of achieved transmission rate given a target bit error rate. Evaluation is performed for links of 50m and 100m by exploring various cases affecting the overall performance, including rate adaptation and artificial PAPR reduction. In all cases it is found that the DFT-spread DMT perfoms better than all other schemes, followed by CDMA-DMT hence, paving the way for more elaborated research for optimizing its transmission properties.

Keywords: Step Index Optical Fiber, SI-POF, Plastic Optical Fiber, Discrete Multitone, Discrete Fourier Transform, Code Division Multiple Access, Multicarrier Spreading Modulation Schemes, OFDM.

1 Motivation

As the need for very high bit rate transmission become more imperative, conventional transmission media such as copper and air, which are extensively used in wired and wireless transmission respectively, seem to either reaching their available capacity or their usage is not advantageous. The use of optical fiber as a medium for short range networks is now becoming a promising solution to achieve increased transmission rate while at the same time supporting different (broadband) services, such as, for example, the simultaneous transmission of data, video, voice (e.g., VO.IP, IPTV, HDTV). In particular, plastic optical fibers of large core diameter i.e., 1mm SI-POF (Step Index-Plastic Optical Fiber), exhibit a number of advantages regarding not only low manufacturing cost and maintenance but also immunity to mechanical stress, avoidance of electromagnetic interference and enhanced bandwidth. On the other hand, compared to silica fibers, limitations on their usage

*Dissertation Advisor: Dimitris Syvridis, Professor

include large optical power loss (e.g. 160-180dB/Km) and limited transmission bandwidth i.e., typically about 100MHz/50m. Therefore, in order to efficiently support transmission rates in the order of Gbps, it is necessary to develop and employ complex modulation schemes together with digital signal processing techniques.

2 Plastic Optical Fiber

High speed short range communications employing large core Plastic Optical Fiber (POFs) have attracted a significant research interest over the past few years and have been assessed for multi gigabit transmission. The plastic optical fibers are multimode waveguides which are made from plastic materials such as polymethyl mathacrylate (PMMA) and fluorinated polymers (PF).

There are two types of PMMA plastic optical fibers: the Step Index Plastic Optical Fibers (SI-POFs) and Graded Index Plastic Optical Fibers (GI-POFs). The fibers of the first type have core with large diameter value (980um) and Numerical Aperture (N.A.) 0.5. Because of the large number of guided modes, these fibers exhibit large modal dispersion values. As a consequence, the available bandwidth of these fibers is reduced to several MHz (50MHz/100m). Another disadvantage of these fibers is the large attenuation value with a typical value 180dB/Km for the case of wavelength of 650nm. The fibers of the second category (GI-POF) have similar geometric features as the SI-POF but they have smaller numerical aperture and reduced modal dispersion.

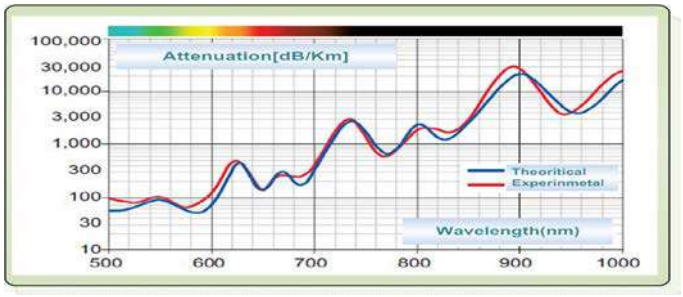


Fig. 1: Attenuation spectra of 1-mm diameter of step index PMMA plastic fiber

In Fig.2, shows the block diagram of a typical communication system based on plastic optical fiber with Intensity Modulation and Direct Detection (IM/DD).

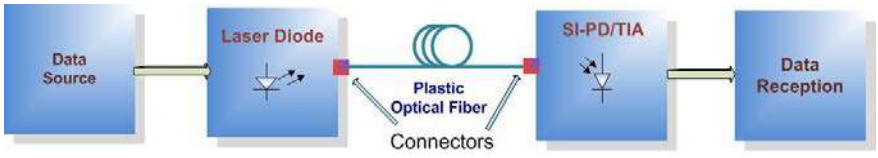


Fig. 2: Block diagram of a typical communication system based on plastic optical fiber

It is obvious that the large values of attenuation constraints the length of an optical link with plastic optical fiber. Also the effect of the large modal dispersion values do not allow the transmission data rates in the order of Gbit/s. As a result, the transmission rate using NRZ-OOK modulation for fiber length in the cases of 25, 50 and 100m is limited in 200, 100 and 50Mbps respectively.

In order, to overcome the constraining factors that plastic optical fibers introduce, it is necessary to develop and employ complex modulation schemes together with digital signal processing techniques. The investigated schemes belong to multicarrier modulation schemes such as DMT with emphasis on inherent spreading properties such as DFT-Spread DMT and CDMA-DMT. In order to maximize the transmission rate of spreading multicarrier modulation schemes over SI-POF, rate adaptive bit loading techniques are also explored.

2 Investigated Modulation Schemes

2.1 Conventional DMT Modulation scheme

The DMT is derived from the well-known Orthogonal Frequency Division Multiplexing (OFDM) and constitute its baseband counterpart. DMT as a multicarrier scheme, provide efficient bandwidth utilization and robustness against dispersive channels. Also, it has tolerance in frequency selective channels in contrast with single carrier schemes, since it divides the available channel bandwidth in smaller narrower sub-channels. In addition, it can be combined with Frequency Domain Equalization (FDE) and can be efficiently implemented using Fast Fourier Transform at the transmitter and receiver.

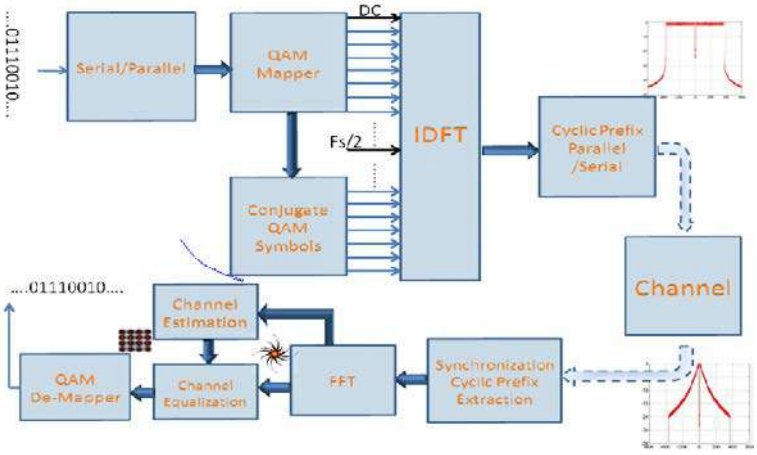


Fig. 3: Block diagram of the DMT transceiver

The principle of DMT is shown in Fig.3. A high speed binary data stream is divided into N parallel data streams. Every M number of bits are given as input to a QAM mapper and produce the QAM complex symbols. These symbols are driven

with their complex conjugate symbols to IFFT and the real valued time signal is produced. In order to combat the multipath delay spread and to allow for FDE the last samples of every DMT time signal are copied in the start of DMT signal. These additional samples are the Cyclic Prefix or Guard Interval (CP or GI). The length of CP is chosen to be longer than the channel's maximum delay spread.

2.2 DFT Spread DMT modulation scheme

In this dissertation, the Discrete Fourier Transform Spread Discrete Multitone modulation (DFT-Spread DMT) scheme is proposed for the first time in short range optical IM/DD transmission (over 50 and 100m) with with 1-mm SI-POF link. This scheme combines the advantages of Single Carrier (SC) and Multi Carrier (MC) transmission. The DFT-Spread DMT, likewise DMT, derives from it's wireless counterpart, namely Single Carrier Frequency Division Multiple Access (SC-FDMA). This scheme also has in essence a SC nature hence the inherent advantage of a lower Peak to Average Power Ratio (PAPR) and also allows for FDE. These properties naturally motivate for its exploration in the context of high speed short-range optical transmission over SI-POFs.

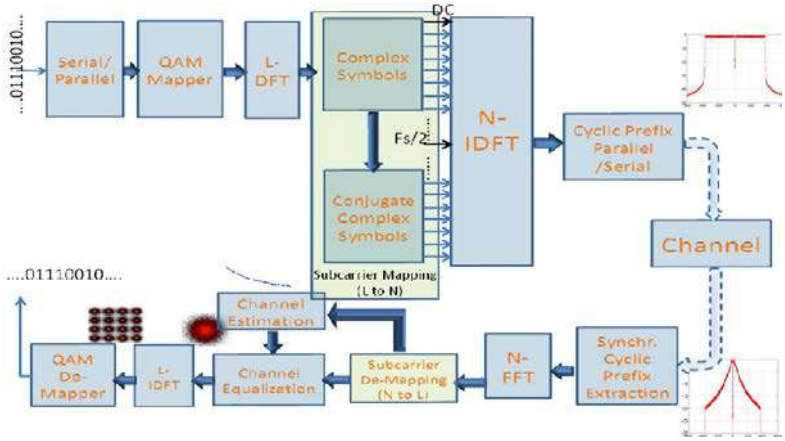


Fig. 4: Block diagram of the DFT Spread DMT transceiver

In Fig.4, the block diagram describing the DFT Spread DMT transceiver structure is shown. Compared to DMT transceiver structure, the DFT Spread DMT has an extra L-point DFT stage at the transmitter combined with a Subcarrier Mapping stage. Also, at the receiver structure there is an extra L-point IDFT stage together with a Subcarrier De-Mapping module [1-2,4].

2.3 CDMA-DMT modulation scheme

In this dissertation, another symbol spreading technique, the Code Division Multiple Access Discrete Multitone (CDMA-DMT) modulation scheme is investigated for transmission over IM/DD link of 50 and 100m of 1mm PMAA SI-POF. CDMA-DMT is a baseband modulation scheme of CDMA-OFDM (also

referred as MC-CDMA) which combines the advantages of OFDM (i.e., MC transmission) and CDMA (i.e., spreading of data symbols via orthogonal codes). It is not only a multiple access scheme but a spread spectrum technique aiming to boost and offer enhanced immunity in the presence of high transmission loss and increased noise. This fact motivates for its exploration in short range high speed large core SI-POF transmission.

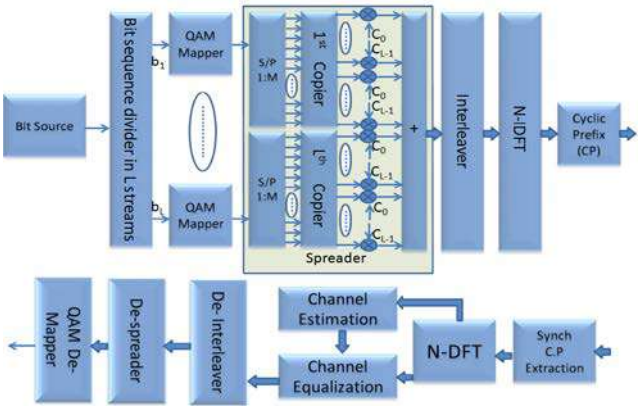


Fig. 5: Block diagram of CDMA-DMT transmitter and receiver

The block diagram describing the CDMA-DMT transceiver structure is shown in Fig. 5. The transmitter consists of the followings parts: parallel divider of initial bit stream in L bit streams, QAM mapper, spreader, interleaver and DMT modulator [3].

2.4 Comparison of DFT Spread and DMT - Experimental results

In order to evaluate the performance of DFT-Spread DMT as well as to compare it against DMT, an offline experiment using conventional components was performed, as is shown in Fig.6.

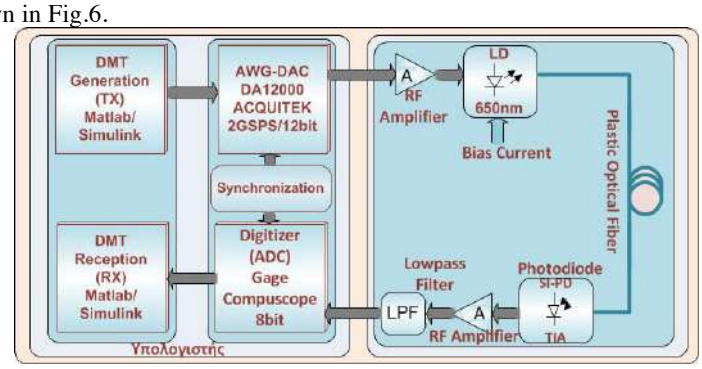


Fig. 6: Block diagram of experimental setup

We have measured the achieved Bit Error Rate (BER) against a range of transmission rates by varying the number of the L data subcarriers as well as using

different modulation formats 32, and 64-QAM and SI-POF lengths 50 and 100m. The results are shown in Figs 7 and 8.

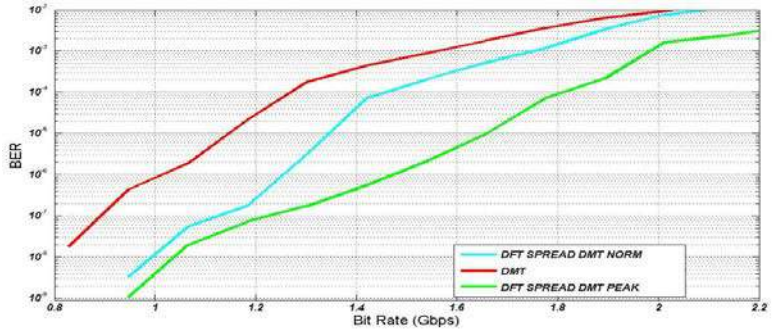


Fig. 7: Comparative results of BER versus net transmission rate for DFT-spread DMT and DMT for the case of 32-QAM and 50m SI-POF

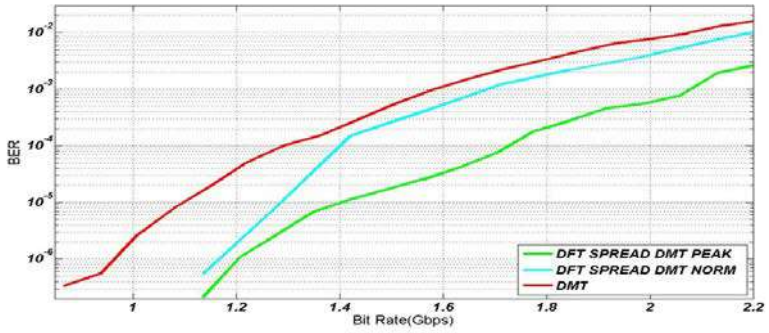


Fig. 8: Comparative results of BER versus net transmission rate for DFT-spread DMT and DMT for the case of 64-QAM and 50m SI-POF

A general observation can be made for all cases evaluated in this study related to the performance enhancement of the proposed technique relative to the pure DMT approach. Spreading of the symbols in the frequency domain via the DFT Spread procedure seems to offer a performance enhancement relative to conventional DMT. For the DFT Spread DMT, the case where it is generated at the transmitter with an average electrical power greater than that of conventional DMT (PEAK case) outperforms DMT in terms of achieved transmission rate for a given threshold BER. A reason for this performance enhancement can be attributed to the fact that DFT-spread DMT has a lower PAPR compared to DMT, as depicted in Fig.7 and Fig.8. Enhanced performance is observed for DFT Spread DMT in the standard case, where both schemes have equal peak power (NORM case). Table 1 summarizes the achievable data bit rates for all investigated schemes. The same performance is observed for the case of transmission over 100m SI-POF (Figs 9 and 10).

Length of SI-POF	50m		100m	
Modulation Scheme	32-QAM Rate (Gbps)	64-QAM Rate (Gbps)	32-QAM Rate (Gbps)	64-QAM Rate (Gbps)
DMT	1.56	1.58	0.81	0.89
DFT SPREAD (PEAK)	1.97	2.05	1.01	1.05
DFT SPREAD (NORM)	1.75	1.68	0.85	0.91

Table 1: Transmission rates for a target ber of 1E-3

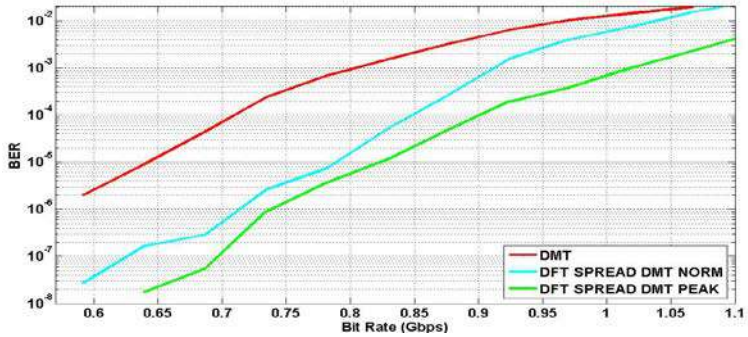


Fig. 9: Comparative results of BER versus net transmission rate for DFT-spread DMT and DMT for the case of 32-QAM and 100m SI-POF

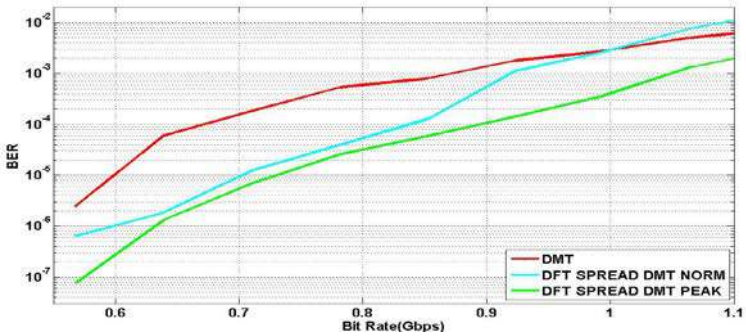


Fig. 10: Comparative results of BER versus net transmission rate for DFT-spread DMT and DMT for the case of 64-QAM and 100m SI-POF

2.5 Comparison of CDMA-DMT and DMT - Experimental results

In order to evaluate the performance of CDMA-DMT as well as to compare it against DMT, we have used the same experimental setup, as is shown in Fig.6. The comparative results are depicted in Figs 11 and 12 for 32- and 64-QAM modulation per subcarrier.

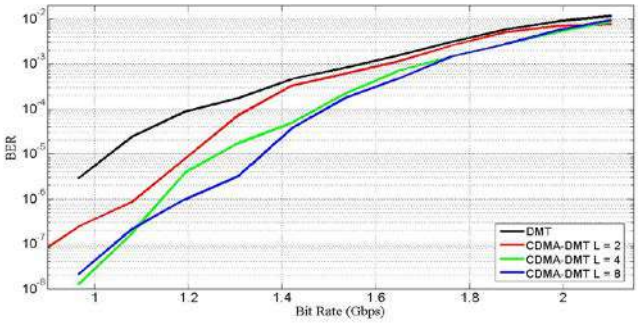


Fig. 11: Comparative results of BER versus net transmission rate for CDMA-DMT and DMT for the case of 32-QAM and 50m SI-POF

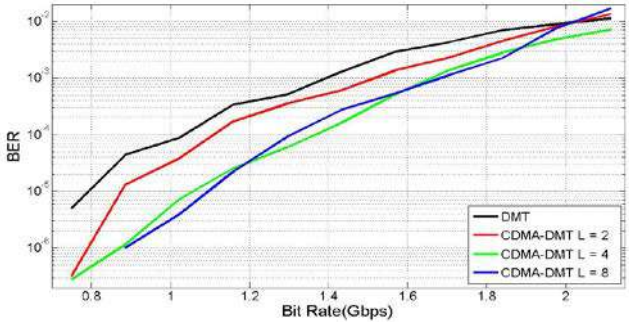


Fig. 12: Comparative results of BER versus net transmission rate for CDMA-DMT and DMT for the case of 64-QAM and 50m SI-POF

A general observation can be made for all cases evaluated in this study related to the performance enhancement of the proposed technique relative to the pure DMT approach. Spreading of the symbols in the frequency domain via the CDMA procedure seems to offer a performance enhancement relative to conventional DMT. This can be attributed to the inherent spreading property of CDMA-DMT where parts of the same QAM symbol are transmitted in different subcarriers due to the chipping process, combined with the fact that the receiver employs the energy of all the received symbols which are scattered in the frequency domain. Finally, as the spreading factor L increases, the *CDMA-DMT* performs better than DMT, especially for the cases of L = 2 and L = 4, as well as for the cases of L = 8. Table 2 summarizes the achievable data bit rates for all investigated schemes for the case of CDMA-DMT and for the cases of spreading factors L = 2, 4 and 8 when evaluated for links of 50m and 100m.

SI-POF (m)	50				100			
	Modulation Schemes							
	DMT	CDMA-DMT			DMT	CDMA-DMT		
Spreading Factor	L=1	L=2	L=4	L=8	L=1	L=2	L=4	L=8
32-QAM	1.57	1.63	1.69	1.70	0.84	0.95	0.99	1.01
64-QAM	1.97	1.52	1.66	1.68	1.97	1.00	1.03	1.05

3 Rate Adaptive Bit Loading

Rate adaptation refers to the process of maximizing the transmission rate subject to a power constraint and given a target BER. This is accomplished via bit-loading that is, by allocating the number of bits transmitted per subcarrier according to its corresponding channel SNR hence, by varying the modulation's constellation (e.g., M-QAM) as the channel indicates. In practice, the Chow's algorithm is usually implemented to allow for finite granularity over allocated bits per sub-channel [4]. Since DMT divides the available channel into sub-channels, bit-loading is directly applicable for each sub-channel (i.e., subcarrier). For the DFT-spread DMT and CDMA-DMT case, some modifications are essential prior to applying bit-loading [4].

3.1 Rate Adaptive Bit Loading

The rate-adaptive bit-loading maximization problem for DMT is formulated as,

$$\max(b)_{E_n} = \sum_{n=1}^{N} log_2\left(1 + \frac{SNR_n}{\Gamma}\right)$$

where, b is the achievable bit rate expressed as the sum of the individual bit rates of the N available subcarriers, is the SNR gap between the SNR needed for maximum capacity and the SNR to achieve this capacity at a given BER, SNRn = E_n□g_n is the signal to noise ratio of each subcarrier with gn being the sub-channel SNR when unit energy is applied and En is the allocated energy per subcarrier subject to the constraint of the total energy for transmission i.e $E_{tot} = \sum_{n=1}^{N} E_n$. subcarrier with gn being the sub-channel SNR when unit energy is applied and En is the allocated energy per subcarrier subject to the constraint of the total energy for transmission.

3.2 Rate Adaptive DFT-Spread DMT

Relative to the DMT case, the DFT-Spread DMT transceiver employs a precoding (spreading) stage. Hence, it has an extra L-point DFT stage at the transmitter combined with a Subcarrier Mapping module, as well as an extra L-point IDFT stage at the receiver combined with a Subcarrier De-Mapping module, with L < N/2 and N being the DFT/IDFT size of conventional DMT [4]. Generalizing for the case when

bit-loading is to be utilized, after bit allocation is derived as obtained for DMT and which determines the constellation order M (i.e., M-QAM) for a group of subcarriers with size Li, an extra Li-point DFT is performed for each group of these subcarriers. That is, an extra DFT stage is added for each group of subcarriers having the same QAM order M. Then, power loading as estimated by Chow's algorithm is applied to each spread subcarrier prior to applying the N-point IDFT which produces the final time domain DFT-spread DMT signal.

3.3 Rate Adaptive CDMA-DMT

CDMA-DMT, which is the baseband version of multicarrier (MC) CDMA involves a different approach for spreading the QAM symbols. It divides the initial bit stream into L parallel streams and after QAM mapping it copies and then spreads these symbols via point wise multiplication (i.e., chipping) with orthogonal spreading sequences (e.g., Walsh-Hadamard sequences). Then the parallel symbol streams are added prior to a DMT modulation. For the case of CDMA-DMT, bit-loading is achieved by considering the concept of equivalent sub-channel (i.e., subcarrier) which is introduced as in the case of MC-CDMA. This is based on the notion that a group of L spread data (where L is also the spreading factor and length of the spreading sequence) represent a single equivalent sub-channel having an effective channel function expressed as,

$$\left|h_{eff}\right|^2 = \frac{L}{\sum_{i=1}^{L} \frac{1}{|H_i|^2}}$$

where H_i is the i^{th} subcarrier's frequency response. Following [ ], Chow's algorithm uses the effective SNR, to produce the power and bit loading distribution [4].

3.4 Experimental results

In Fig. 13 the estimated SNR per sub-channel that is used to compute the rate-adaptive bit-loading Chow's algorithm for each modulation scheme is shown. This gives rise to the bit and power allocation shown in Fig. 14-16 for DMT, DFT-spread DMT and CDMA-DMT respectively.

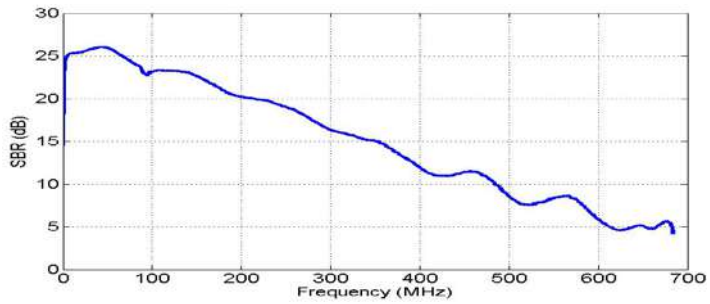


Fig. 13: Measured SNR per subchannel

When bit-loading is utilized, a transmission rate of ~2.55 Gbps is achieved for DFT-spread DMT whereas DMT and CDMA-DMT realizes a rate of ~2.37 Gbps and ~2.45 Gbps respectively. It is observed that DFT-spread DMT exhibits enhanced

performance compared to the other two schemes. A reason for this performance enhancement can be attributed to the fact that DFT-spread DMT has a lower PAPR compared to DMT hence, the average transmitted power is greater than that of DMT resulting in a SNR gain of approximately ~1.5 dB for bit-loading.

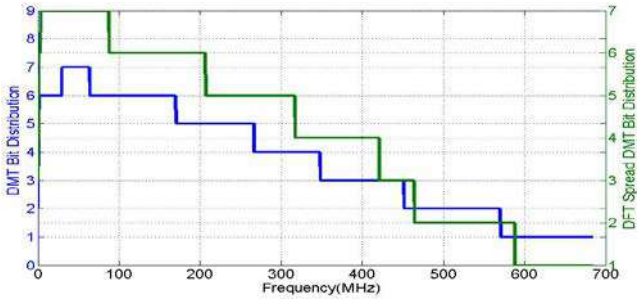


Fig. 14: Bit distribution per subchannel for the cases of DMT and DFT Spread DMT

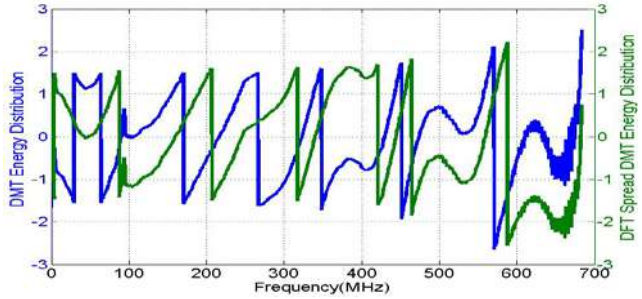


Fig. 15: Energy distribution per subchannel for the cases of DMT and DFT Spread DMT

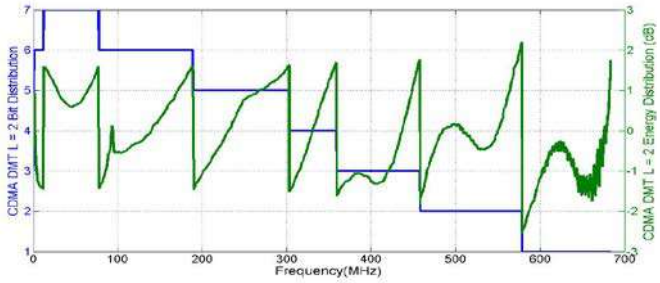


Fig. 16: Bit and Energy distribution per subchannel for the case of CDMA-DMT (L= 2)

4 Conclusions

In this dissertation, spreading multicarrier schemes have been proposed, providing performance improvement in transmission rates over SI-POF IM/DD high speed short range links.

The first modulation scheme, the DFT Spread DMT, spreads the symbols across subcarriers in the frequency domain. It exhibits improvement performance in comparison with conventional DMT because of its lower PAPR as well as due to its inherent single carrier nature. In this thesis, it is experimentally confirmed that DFT-spread DMT offers a performance improvement of 29.7% and 18% in achieved data rates given a target BER, for transmission over links of 50m and 100m respectively. In addition, when rate adaptation is utilized, the DFT-spread DMT offers a performance improvement of 12.6% compared to conventional DMT.

The second modulation scheme namely, CDMA-DMT also spreads the QAM symbols across subcarriers in the frequency domain, using specific spreading sequence. It is experimentally shown that as the spreading factor L increases, the CDMA-DMT performs better than DMT, especially for the cases of L = 2 and L = 4. Also, this technique outperforms DMT when combined with rate adaptive bit loading.

In conclusion, it is experimentally verified that both of the proposed modulation schemes outperform conventional DMT in all examined cases for IM/DD SI-POF transmission. Among them, the DFT-spread DMT exhibits the best performance. Taking into account that this scheme effectively combines the relative merits of multicarrier and single carrier transmission techniques, more elaborated research is motivated for optimizing its modulation properties and that could lead to a converged MC and SC modulation scheme for low cost short range high speed optical links based on plastic optical fibers.

Publications

1. S. Karabetsos, E. Pikasis, T. Nikas, A. Nassiopoulos, D. Syvridis, "A DFT-spread DMT modulation scheme for beyond 1Gbps transmission rate over 100m with 1mm SI-POF", Proceedings of the 20th International Conference on Plastic Optical Fibers (POF 2011), 14-16 September 2011, pp. 7-12, Bilbao, Spain
2. Karabetsos, S., Pikasis, E., Nikas, T., Nassiopoulos, A., Syvridis, D., "DFT-Spread DMT Modulation for 1-Gb/s Transmission Rate Over 100 m of 1-mm SI-POF," Photonics Technology Letters, IEEE, vol.24, no.10, pp.836-838, May15, 2012
3. Pikasis, E., Karabetsos, S., Raptis, N., Syvridis, D., "Performance Evaluation of CDMA-DMT for 1-mm SI-POF Short-Range Transmission Links," Photonics Technology Letters, IEEE, vol.24, no.22, pp.2042-2045, Nov.15, 2012
4. Pikasis, E., Karabetsos, S., Nikas, T., Syvridis, D., "Rate-Adaptive DFT-Spread DMT and CDMA-DMT for 1-mm SI-POF Short-Range Links," Photonics Technology Letters, IEEE , vol.25, no.16, pp.1574,1577, Aug.15, 2013

Retrieval of 3-Dimensional Rigid and Non-Rigid Objects

Konstantinos Sfikas*

National and Kapodistrian University of Athens,
Department of Informatics and Telecommunications,
ksfikas@di.uoa.gr

Abstract. This dissertation focuses on the problem of 3D object retrieval from large datasets in a near realtime manner. In order to address this task we focus on three major subproblems of the field: (i) pose normalization of rigid 3D models with applications to 3D object retrieval, (ii) non-rigid 3D object description and (iii) search over rigid 3D object datasets based on 2D image queries. Regarding the first of the three subproblems, 3D model pose normalization, two main novel pose normalization methods are presented, based on: (i) 3D Reflective Object Symmetry (ROSy) and (ii) 2D Reflective Object Symmetry computed on Panoramic Views (SymPan/SymPan+). Considering the second subproblem, a non-rigid 3D object retrieval methodology, based on the properties of conformal geometry and graph-based topological information (ConTopo++) has been developed. Furthermore, a string matching strategy for the comparison of graphs that describe 3D objects, is proposed. Regarding the third subproblem a 3D object retrieval method, based on 2D range image queries that represent partial views of real 3D objects, is presented. The complete 3D objects of the database are described by a set of panoramic views and a Bag-of-Visual-Words model is built using SIFT features extracted from them. The methodologies developed and described in this dissertation are evaluated in terms of retrieval accuracy and demonstrated using both quantitative and qualitative measures via an extensive consistent evaluation against state-of-the-art methods on standard datasets.

Keywords: 3D Objects, Rotation Normalization, Shape Modelling, Partial Matching, Range Images

1 Introduction

Information, commonly refers to a useful portion of data located among a collection of related entities. Recent advances in storage technologies and the widespread use of the Internet, have resulted in a vast increase of the amount of data stored in and distributed from large databases. Any attempt for manual annotation and information extraction is almost impossible, therefore rendering the need for an automated procedure, mandatory.

* Dissertation Advisor: Theoharis Theoharis, Professor

The process of extracting useful information from large amounts of data, in an automated manner and based on an example or descriptive query, is called information retrieval. Common types of information that can benefit from such a retrieval process are: textual, visual, audio and video data and most recently, 3D and 4D (3D over time) data.

In recent years, through the creation of inexpensive 3D scanners and the simplification of 3D modelling software, a large volume of 3D data has been created and stored in corresponding scientific and industrial/commercial repositories. Furthermore, 3D data can be processed in various, application dependent, ways and occasionally be combined with data of other types and modalities (e.g. textual annotation and/or thumbnails of 3D models). These data types can further be used as queries for the retrieval of 3D objects.

Some example applications that exploit the properties of 3D models and could greatly benefit from a retrieval process follow: in medicine large diagnostic 3D data are compared and researched in order to assist the process of making medical decisions. In biometrics a person's 3D facial model is searched over corresponding databases for identification purposes. Game development utilizes retrieval and reusability of 3D models in order to minimize production times and reduce the size of the final product. Other example application areas include engineering and archaeology. It can therefore be easily deduced, that 3D object retrieval is a key process, although in general it is complex and highly depended on the application.

2 Framework and problem statement

3D object retrieval applications can be classified into two major categories: inter-class and intra-class retrieval. Inter-class retrieval focuses on a generic domain of 3D objects and aims at finding the closest match among a set of 3D models that belong to a broad range of different classes. In this case, there is usually no prior knowledge regarding the characteristics or the nature of the 3D objects. Intra-class retrieval targets a specific 3D object domain (e.g. 3D faces, non-rigid 3D models, human action models, engineering models etc), where a match is sought between 3D models that belong to the same class but have their special characteristics defined differently. Intra-class 3D object retrieval methods usually exploit domain knowledge and shape characteristics of the 3D models, in order to attain higher performance.

For both categories, the generic framework of a 3D object retrieval system can be outlined as follows: preprocessing, pose normalization, shape descriptor extraction, feature matching.

At the first step of the 3D object retrieval pipeline, 3D models are preprocessed. In this step, the 3D models are cleaned up of any inconsistencies present due to the digitization process, i.e. double or reversed faces, structural gaps, etc. This step is highly dependent on the method/equipment used for the creation of the 3D models and may differ greatly from one application to another.

After basic preprocessing, *Pose Normalization* ensures that the geometric properties of the 3D models are defined in a uniform manner. The diversity of 3D object acquisition sources implies that 3D objects which may even be part of the same dataset, have their geometrical properties arbitrarily defined. Therefore, before any kind of processing is carried out, it must be ensured that the 3D objects have been normalized in terms of position, scaling and rotation (Fig 1 shows an example rotation normalization). Pose normalization of 3D objects is a common preprocessing step in various computer graphics applications [2, 22, 23, 27]. Visualization, broken fragment reconstruction, biometrics and 3D object retrieval are only a few examples of applications that benefit from a pose normalization procedure. To achieve pose normalization, for every 3D object, a corresponding set of normalization transformations in 3D space must be defined.

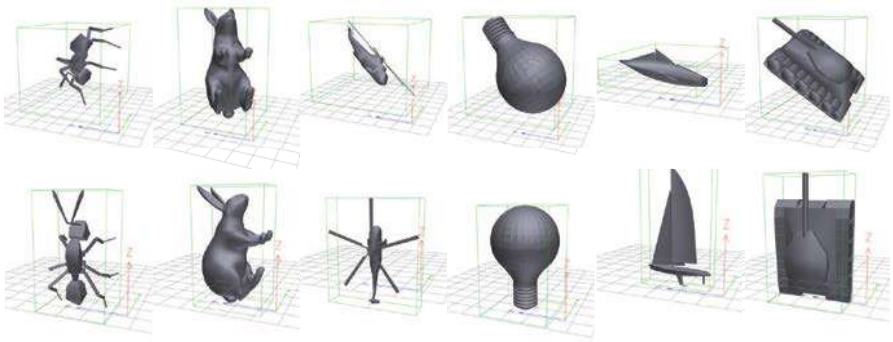


Fig. 1: Examples of non-aligned objects (top-row) and the corresponding rotation normalizations (bottom row).

The main step of a 3D object retrieval system is the computation of a feature set for each 3D model. In this step, the structural and/or other special characteristics of a 3D object are modelled and a shape descriptor that faithfully encodes the shape of the 3D model, in an efficient manner, is created. Feature selection is tightly connected to the corresponding application and can vary greatly for each 3D object retrieval system (e.g. intra-class retrieval exploits features that are more distinguishing within a specific domain, whereas inter-class retrieval uses more generic characteristics).

Finally, each 3D object's shape descriptor is used as a signature during the matching procedure. At this step, the signatures of the 3D models, stored in the database, are compared to the corresponding signatures of the query 3D model(s), using a specified metric. The selected metric is also dependent on both the features selected and the corresponding application. Finally, the response of

the 3D object retrieval system is the set of 3D object(s) that correspond to the closest match(es) of the given user query.

3 Contributions

This dissertation has made the following research contributions in the field of 3D object retrieval: two new 3D model pose normalization methods, a non-rigid 3D object retrieval methodology and a 3D object retrieval algorithm, based on range image queries. In detail, the contributions of this dissertation are the following:

3.1 ROSy Pose Normalization Method

A general purpose global pose normalization method, based on 3D object reflective symmetry.

In the ROSy method, the problem of pose normalization is described through the Surface-Oriented Minimum Bounding Box (SoMBB), a modified version of the Axis-Aligned Bounding Box (AABB) which is commonly used in collision detection techniques [24, 8].

The motivation behind the proposed method is to minimize the SoMBB of a 3D object so that the latter becomes aligned with its SoMBB and consecutively with the principal axes of space. Furthermore, to ensure that the 3D object's large planar areas are also in alignment with the principal planes of space, it is required that the average normal to the object's large planar areas become parallel to the box's face normals (Fig. 2).

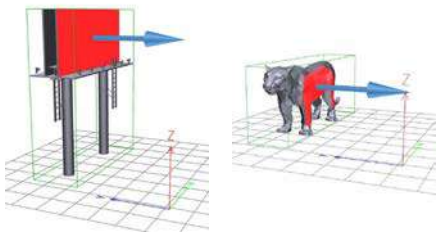


Fig. 2: 3D objects enclosed in their SoMBBs.

Initially, the axis-aligned minimum bounding box of a rigid 3D model is modified by requiring that the 3D model is also in *minimum angular difference* with respect to the normals to the faces of its bounding box.

To estimate the modified axis-aligned bounding box, a set of predefined principal planes of symmetry is used and the corresponding symmetric models are computed. Then, a combined spatial and angular distance, between the 3D model and its symmetric model, is calculated.

By minimizing this combined distance, through a set of rotations in space, the 3D model fits inside its modified axis-aligned bounding box and alignment with the coordinate system is achieved. [18]

3.2 SymPan+ Pose Normalization Method

A pose normalization method, based on panoramic views and reflective symmetry, is presented.

The motivation for the proposed method is that the use of reflective symmetry as a feature for pose normalization and 3D object retrieval seems to enhance the results [11], as most of the 3D objects exhibit symmetrical properties to some degree. These properties tend to be distinct between different classes and similar between objects of the same class, therefore enhancing the discrimination achieved by other commonly used characteristics, such as the spatial distribution and/or surface orientation of the 3D models. Qualitative and experimental investigation in 3D data-sets has led us to the observation that most objects possess at least a single plane of symmetry. Our approach is thus guided by this observation.

Initially, the surface of a 3D model is projected onto the lateral surface of a circumscribed cylinder, aligned with the primary principal axis of space. Based on this cylindrical projection, a *normals' deviation map* is computed.

Through an iterative procedure, the symmetry plane of the 3D model is parallelized with the axis of the projection cylinder, thus computing the first principal axis of the 3D model. This is achieved by rotating the 3D model and computing reflective symmetry scores on panoramic view images (Fig 3).

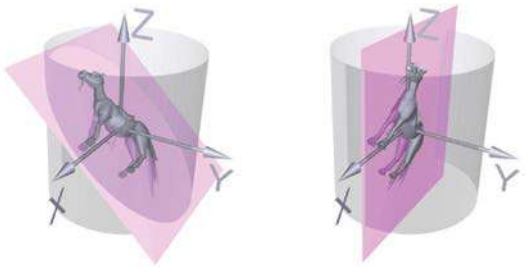


Fig. 3: Aligning the symmetry plane normal with the XY plane

The other principal axes of the 3D model are then estimated by computing the variance of the 3D model's panoramic views. [17, 21]

3.3 ConTopo++ Non-Rigid 3D Object Retrieval

Combining the properties of conformal geometry and graph-based topological information, a non-rigid 3D object retrieval methodology is proposed, which is

both robust and efficient in terms of retrieval accuracy and computation speed. While graph-based methods are robust to nonrigid object deformations, they require intensive computation which can be reduced by the use of appropriate representations, addressed through geometry-based methods. In this respect, a 3D object retrieval methodology, which combines the above advantages in a unified manner, is presented.

Initially, we define a graph, that captures the topological structure of an arbitrary 3D mesh. Each node of the graph represents a unique connected component, while each edge of the graph describes the relation between adjacent connected components. Each connected component is composed of 3D mesh faces that have the same label and are also pathwise-connected.

In this work we have used discrete conformal factors [1] as a labeling criterion due to their ability to identify the protrusive parts in a mesh. The faces of the 3D mesh are partitioned based on a linear multi-thresholding of the values of the discrete conformal factor, thus splitting the mesh into a set of connected components (see Fig. 4).

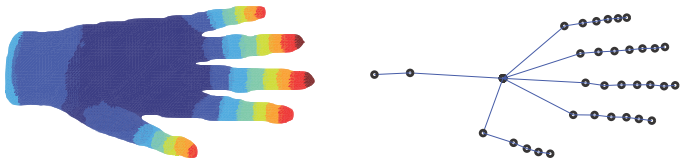


Fig. 4: Illustration of an eight-level quantized mesh and the corresponding graph.

Mesh matching compares both geometrical and topological features as a measure of similarity between two 3D meshes in a unified manner. During matching, the topological equivalence between the graphs of two 3D meshes is examined and enhanced by node-to-node comparison of geometrical features.

The matching procedure is based on string matching. Each ordered path, of graph nodes, that extends from the *core* partition of the 3D mesh down to each of its articulations is considered a *string*. Furthermore, besides the ordered connectivity of the string (graph) nodes, a number of features are also attached to them, which are used for the geometrical matching. [19]

3.4 3D Object Retrieval based on 2D Range Image Queries

A 3D object retrieval method, based on range image queries that represent partial views of real 3D objects, is presented.

The motivation behind the proposed method, is to use a 2D image in order to query a database of 3D objects and bridge the representation gap between the two in an efficient manner.

The complete 3D models of the database are described by a set of panoramic views and a Bag-of-Visual-Words model is built using SIFT features extracted

from them. To address the problem of partial matching, a spatial histogram computation scheme, on the panoramic views, that represents local information by taking into account spatial context, is suggested.

Furthermore, a number of optimization techniques are applied throughout the process, for enhancing the retrieval performance. [20]

4 Experimental Results

The experimental evaluation is based on the Precision-Recall curves and five quantitative measures: Nearest Neighbor (NN), First Tier (FT), Second Tier (ST), E-measure (E) and Discounted Cumulative Gain (DCG) [9, 22] for the classes of each corresponding dataset.

4.1 ROSy and SymPan+ Pose Normalization Methods

For the evaluation of the ROSy pose normalization method, we have chosen a state-of-the-art 3D object retrieval methodology, by Papadakis et al. [14], as the evaluation vehicle.

Papadakis' 3D object retrieval system, in its original form, uses a combination of the CPCA and NPCA algorithms to achieve pose normalization of a 3D model. ROSy itself has similar performance to CPCA and NPCA. However, the combination of the three pose normalization methods (namely ROSy+) gives a significant boost to the discriminative power of the retrieval process, outperforming the original hybrid (CPCA, NPCA) approach.

Similar to the way that the ROSy+ system has been used for the quantitative evaluation of the ROSy pose normalization method, for SymPan+ we have chosen the PANORAMA state-of-the-art 3D object retrieval system, by Papadakis et al. [15] as the evaluation vehicle. The proposed method replaces the NPCA pose normalization method in the existing hybrid scheme.

The direct effect of the proposed alignment methods can be evaluated by comparing against the original 3D object retrieval methods' performance. In terms of object retrieval performance, we compared against DLA [3], GSMD+SHD+R [12], Lightfield [4], SH-GEDT [10] and DESIRE [25].

In Fig. 5, using the experimental results given in [18, 17], we illustrate the P-R scores for the test subset of the PSB dataset, for the proposed pose normalization methods.

ROSy+ is able to achieve an average performance gain of about 3% over the original hybrid approach (mean value over the quantitative measures used). Furthermore, it is clear that ROSy+ performs better than state-of-the-art methods by an average of 2% - 5%. SymPan+ improves the discriminative power of the PANORAMA 3D object retrieval system by an average of 7% over the original approach. Furthermore, the SymPan+ method exhibits improved performance over ROSy+ by an average of 2 - 3%.

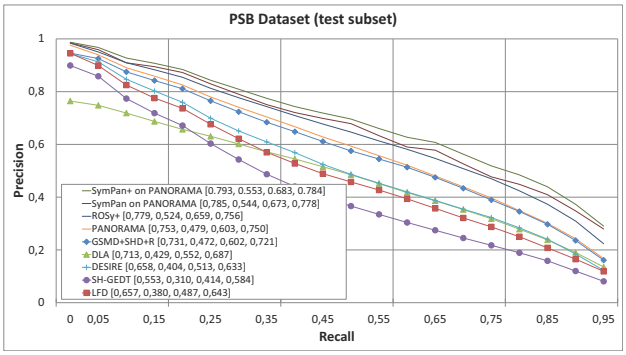


Fig. 5: Precision-Recall plot for the Princeton Shape Benchmark test dataset. SymPan, SymPan+ 3D model pose normalization methods on PANORAMA retrieval results are compared against state-of-the-art 3D object retrieval techniques.

4.2 ConTopo++ Non-Rigid 3D Object Retrieval Method

In the sequel, we compare the proposed non-rigid 3D object retrieval method ConTopo++ against other state-of-the-art methods on standard datasets.

In Fig. 6 we illustrate the P-R scores of the proposed method against the published results of the SHREC'10 *Non-rigid 3D Models* dataset. It is clear that the proposed method outperforms the track contestants, even though the published results were already of high performance.

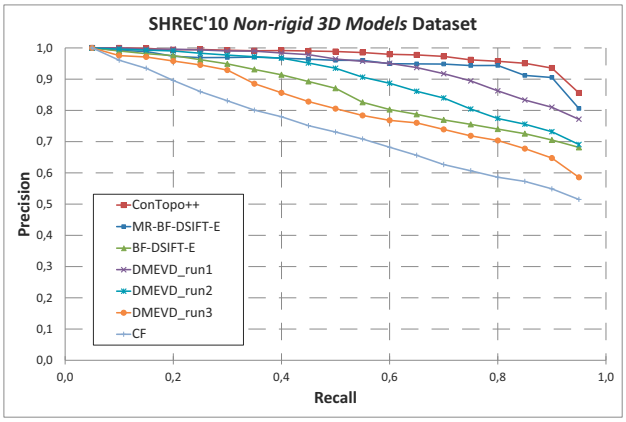


Fig. 6: Comparative results based on the average P-R scores for the SHREC'10 *Non-rigid 3D Models* dataset.

4.3 3D Object Retrieval Based on 2D Range Image Queries

The datasets that we used for the experimental evaluation of our proposed 3D object retrieval, based on 2D image query method are the following: (i) SHREC'09 *Querying with Partial Models* [6] and (ii) SHREC'10 *Range Scan Retrieval* [7]. We compared against existing results of the participating contestants.

More specifically, on the SHREC'09 *Querying with Partial Models* we compared against the variations of CMVD (Compact MultiView Descriptor) by Daras and Axenopoulos [5] and the BF-SIFT and BF-GridSIFT methods by Furuya and Ohbuchi. The P-R scores of Fig. 7 illustrate that the proposed method achieves superior performance compared to the variations of the CMVD, as well as both the BF-SIFT and the BF-GridSIFT retrieval methods.

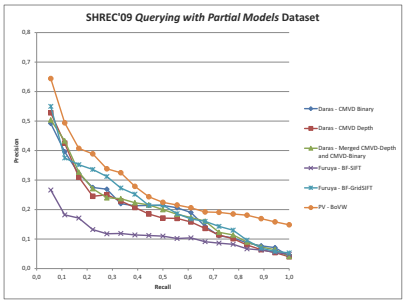


Fig. 7: Comparative results based on the average P-R scores for the SHREC'09 *Querying with Partial Models* dataset.

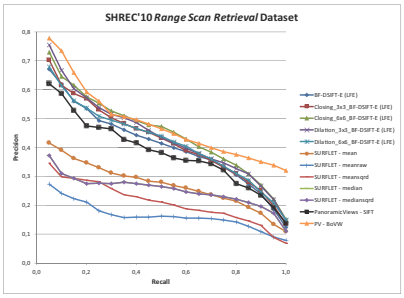


Fig. 8: Comparative results based on the average P-R scores for the SHREC'10 *Range Scan Retrieval* dataset.

On the SHREC'10 *Range Scan Retrieval* dataset we compared against the variations of the BF-DSIFT-E method proposed by Ohbuchi and Furuya [13] and the variations of the SURFLET method proposed by Hillebrand et al. [26]. The P-R scores of Fig. 8, illustrate that the proposed method outperforms the track contestants.

5 Conclusions

To address the problems of 3D model pose normalization, 3D object retrieval with applications to rigid and non-rigid models, as well as image based 3D object retrieval, four novel methodologies have been developed.

In the field of 3D model pose normalization two main novel methods, based on the reflective symmetry properties of 3D objects, have been proposed. All the proposed methods are able to produce high quality alignments of 3D objects, regardless of their originating class or morphology. These alignments are both stable and consistent.

To address the problem of non-rigid 3D object retrieval the ConTopo++ descriptor has been proposed. This non-rigid 3D object retrieval methodology is able to achieve high levels of retrieval accuracy and outperform many of the competing descriptors at a low computational cost. Fig. 9 illustrates some retrieval samples from the SHREC'10 *Non-rigid 3D Models* dataset.

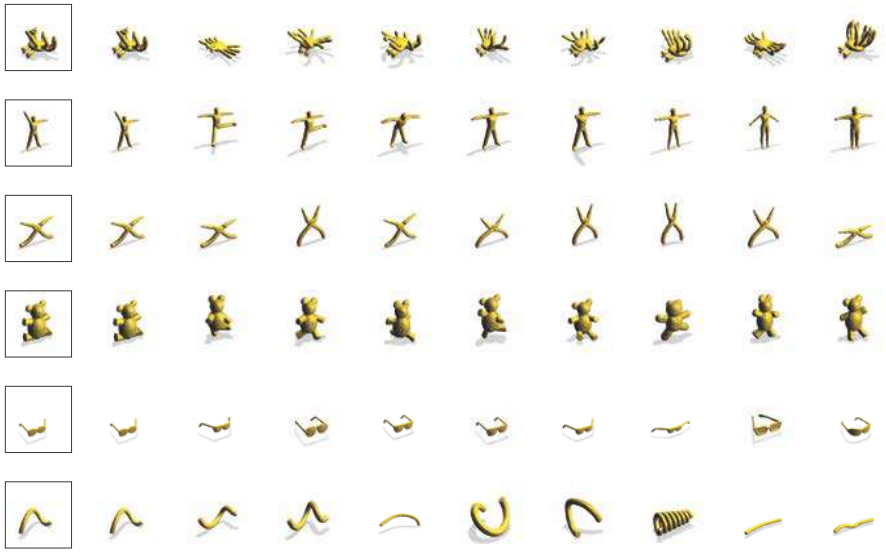


Fig. 9: Sample queries from the SHREC'10 *Non-rigid 3D Models* dataset. First column indicates the query model and results are illustrated in ranking order. The thumbnails have been taken from the SHREC'10 *Non-rigid 3D Models* dataset.

In the field of image-based 3D object retrieval, we proposed a spatial histograms strategy in a Bag-of-Visual-Words context that fits the information present in panoramic views of 3D objects to the task of partial matching. This improved 3D object retrieval methodology, was evaluated on the SHREC'09 *Querying with Partial Models* and SHREC'10 *Range Scan Retrieval* tracks against the corresponding state-of-the-art 3D object retrieval methodologies. In every case, the proposed method outperforms competing descriptors.

The described methodologies have proven to be robust in terms of retrieval accuracy and outperformed previous state-of-the-art methods in the corresponding evaluation tests. These tests were conducted on publicly available datasets.

Bibliography

[1] Mirela Ben-Chen and Craig Gotsman. Characterizing shape using conformal factors. In Perantonis et al. [16], pages 1–8.

[2] Benjamin Bustos, Daniel A. Keim, Dietmar Saupe, Tobias Schreck, and Dejan V. Vranic. An experimental comparison of feature-based 3D retrieval methods. In *3DPVT*, pages 215–222. IEEE Computer Society, 2004.

[3] Mohamed Chaouch and Anne Verroust-Blondet. Alignment of 3D models. *Graphical Models*, 71(2):63–76, 2009.

[4] Ding-Yun Chen, Xiao-Pei Tian, Yu-Te Shen, and Ming Ouhyoung. On visual similarity based 3D model retrieval. *Comput. Graph. Forum*, 22(3):223–232, 2003.

[5] Petros Daras and Apostolos Axenopoulos. A compact multi-view descriptor for 3D object retrieval. In Stefanos D. Kollias and Yannis S. Avrithis, editors, *CBMI*, pages 115–119. IEEE Computer Society, 2009.

[6] Helin Dutagaci, Afzal Godil, Apostolos Axenopoulos, Petros Daras, Takahiko Furuya, and Ryutarou Ohbuchi. SHREC'09 track: Querying with partial models. In Michela Spagnuolo, Ioannis Pratikakis, Remco C. Veltkamp, and Theoharis Theoharis, editors, *3DOR*, pages 69–76. Eurographics Association, 2009.

[7] Helin Dutagaci, Afzal Godil, Chun Pan Cheung, Takahiko Furuya, Ulrich Hillenbrand, and Ryutarou Ohbuchi. SHREC'10 track: Range scan retrieval. In Mohamed Daoudi, Tobias Schreck, Michela Spagnuolo, Ioannis Pratikakis, Remco C. Veltkamp, and Theoharis Theoharis, editors, *3DOR*, pages 109–115. Eurographics Association, 2010.

[8] Jeffrey Goldsmith and John Salmon. Automatic creation of object hierarchies for ray tracing. *IEEE Comput. Graph. Appl.*, 7(5):14–20, 1987.

[9] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, 2002.

[10] Michael M. Kazhdan, Thomas A. Funkhouser, and Szymon Rusinkiewicz. Rotation invariant spherical harmonic representation of 3D shape descriptors. In Leif Kobbelt, Peter Schröder, and Hugues Hoppe, editors, *Symposium on Geometry Processing*, volume 43 of *ACM International Conference Proceeding Series*, pages 156–164. Eurographics Association, 2003.

[11] Michael M. Kazhdan, Thomas A. Funkhouser, and Szymon Rusinkiewicz. Symmetry descriptors and 3D shape matching. In Jean-Daniel Boissonnat and Pierre Alliez, editors, *Symposium on Geometry Processing*, volume 71 of *ACM International Conference Proceeding Series*, pages 115–123. Eurographics Association, 2004.

[12] Zhouhui Lian, Paul L. Rosin, and Xianfang Sun. Rectilinearity of 3D meshes. *International Journal of Computer Vision*, 89(2-3):130–151, 2010.

[13] Ryutarou Ohbuchi and Takahiko Furuya. Scale-weighted dense bag of visual features for 3D model retrieval from a partial view 3D model. In *Computer*

Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on, pages 63 – 70, 2009.

[14] Panagiotis Papadakis, Ioannis Pratikakis, Theoharis Theoharis, Georgios Passalis, and Stavros J. Perantonis. 3D object retrieval using an efficient and compact hybrid shape descriptor. In Perantonis et al. [16], pages 9–16.

[15] Panagiotis Papadakis, Ioannis Pratikakis, Theoharis Theoharis, and Stavros J. Perantonis. Panorama: A 3d shape descriptor based on panoramic views for unsupervised 3d object retrieval. *International Journal of Computer Vision*, 89(2-3):177–192, 2010.

[16] Stavros J. Perantonis, Nickolas S. Sapidis, Michela Spagnuolo, and Daniel Thalmann, editors. *Eurographics Workshop on 3D Object Retrieval, 3DOR 2008, Crete, Greece, 2008. Proceedings*. Eurographics Association, 2008.

[17] Konstantinos Sfikas, Ioannis Pratikakis, and Theoharis Theoharis. SymPan: 3D Model Pose Normalization via Panoramic Views and Reflective Symmetry. In Umberto Castellani, Tobias Schreck, Silvia Biasotti, Ioannis Pratikakis, Afzal Godil, and Remco C. Veltkamp, editors, *3DOR*, pages 41–48. Eurographics Association, 2013.

[18] Konstantinos Sfikas, Theoharis Theoharis, and Ioannis Pratikakis. ROSy+: 3D object pose normalization based on PCA and reflective object symmetry with application in 3D object retrieval. *International Journal of Computer Vision*, 91(3):262–279, 2011.

[19] Konstantinos Sfikas, Theoharis Theoharis, and Ioannis Pratikakis. Non-rigid 3D object retrieval using topological information guided by conformal factors. *The Visual Computer*, 28(9):943–955, 2012.

[20] Konstantinos Sfikas, Theoharis Theoharis, and Ioannis Pratikakis. 3D object retrieval via range image queries in a Bag-of-Visual-Words context. *The Visual Computer*, 29(12):1351–1361, 2013.

[21] Konstantinos Sfikas, Theoharis Theoharis, and Ioannis Pratikakis. Pose normalization of 3D models via reflective symmetry on panoramic views. submitted to, 2014.

[22] Philip Shilane, Patrick Min, Michael M. Kazhdan, and Thomas A. Funkhouser. The princeton shape benchmark. In *SMI*, pages 167–178. IEEE Computer Society, 2004.

[23] Johan W. H. Tangelder and Remco C. Veltkamp. A survey of content based 3D shape retrieval methods. *Multimedia Tools Appl.*, 39(3):441–471, 2008.

[24] Gino van den Bergen. Efficient collision detection of complex deformable models using AABB trees. *J. Graph. Tools*, 2(4):1–13, 1998.

[25] Dejan V. Vranic. DESIRE: a composite 3D-shape descriptor. In *ICME*, pages 962–965. IEEE, 2005.

[26] Eric Wahl, Ulrich Hillenbrand, and Gerd Hirzinger. Surflet-pair-relation histograms: A statistical 3D-shape representation for rapid classification. In *3DIM*, pages 474–482. IEEE Computer Society, 2003.

[27] Titus B. Zaharia and Françoise J. Prêteux. 3D versus 2D/3D shape descriptors: a comparative study. In Edward R. Dougherty, Jaakko Astola, and Karen O. Egiazarian, editors, *Image Processing: Algorithms and Systems*, volume 5289 of *SPIE Proceedings*, pages 47–58. SPIE, 2004.

Spatial spectrum reuse in heterogeneous wireless networks: interference management and access control

Dimitrios K. Tsolkas*

National and Kapodistrian University of Athens
Department of Informatics and Telecommunications
Dtsolkas@di.uoa.gr

Abstract. The scope of this dissertation is to deal with challenges arising from the introduction of femtocells and D2D communications in cellular networks standardized by 3GPP Release 8 and beyond, i.e., Long Term Evolution (LTE) and LTE-Advanced (LTE-A) networks. More specifically, for the case of femtocells, the interference management problem is studied, while for the D2D communications the radio resource management and the spectrum access challenges are addressed. First, different control channel interference management schemes for femtocell-overlaid LTE/LTE-A networks are studied, while an innovative power control scheme for the femtocell downlink transmissions is proposed, utilizing the end user's quality of experience. Considering the much more dynamic environment defined by the D2D communications in a cellular network, two D2D spectrum sharing approaches are proposed, one based on resource allocation, and one based on contention. In both of the cases, the main requirement is the solution of the device discovery problem. To this end, enhancements in the 3GPP standardized access network are proposed, enabling a resource request/allocation procedure for device discovery transmissions, while a spatial spectrum reuse scheme is designed and evaluated, as an effort to reduce the consumption of radio resources for discovery transmissions.

1 Dissertation Summary

1.1 Motivation and scope

Nowadays, wireless communication services and broadband Internet access have converged to deal with the present requirements for ubiquitous and highly reliable communications. International Mobile Telecommunications-Advanced (IMT-Advanced) quantifies these requirements promising an all-Internet Protocol (IP) packet switched network with data rates analogous to those provided by wired communication systems.

* Dissertation Advisor: Lazaros Merakos, Professor

In this direction, a new architecture of cellular networks introduces a major paradigm shift from wide-range cells with high transmit power (macrocells) to low-power small-sized cells. The success of this shift relies on capitalizing on the performance improvements derived by increasing the spatial spectrum utilization and enhancing the indoor coverage. Historically, spatial spectrum reuse has been, by far, the most efficient approach in improving cellular system capacity, compared to approaches such as the adoption of efficient modulation schemes. Also, the enhancement of indoor coverage will be essential in the near future, since the majority of the voice and data traffic will originate from indoor users. To this end, the 3rd Generation Partnership Project (3GPP) standardized a novel type of small-sized cells called femtocells or femtos. Since the licensed spectrum resources are expensive and scarce, femtocells are expected to spatially reuse licensed spectrum under the so-called co-channel deployment. Installed by the consumers in an unplanned manner, they also provide the option to serve only a limited set of subscribed users through the so-called closed subscriber group (CSG) mode, changing in that way the landscape of cellular networks in the following years. However, before operators and consumers reap the benefits provided by femtocells, several challenges must be addressed, including the mitigation of the generated interferences.

In parallel to the femtocell proliferation, one of the prominent topics considered toward achieving IMT-Advanced requirements is the Device-to-Device (D2D) communications, i.e., direct communications in a cellular network, without the intervention of the base station, when the transmitter and the receiver are in close proximity. Differing from conventional approaches, such as Bluetooth and WiFi-direct, D2D communications utilize licensed spectrum, while no manual network detection-selection is needed. Comparing to the very appealing cognitive radio communications, where secondary transmissions are allowed in parallel with primary cellular transmissions, D2D communications are established by standard/primary cellular users, reaping the benefits of being synchronized and controlled by the central (primary) base station. The introduction of D2D communications in cellular networks is expected to be beneficial from a variety of perspectives. The short distance between D2D users results in better channel conditions, leading to higher data rates, lower delays and lower energy consumption. Additionally, D2D users are connected through a direct link and the intermediate transmission to a base station is avoided, saving network resources and processing effort from the network. Also, the coexistence of cellular and D2D links can lead to more efficient spectrum utilization and higher spatial spectrum reuse, while new business models, probably with a new charging policy for users, may be designed. However, D2D communications do not come without a cost. On the one hand, interference-free conditions between D2D and cellular transmissions, as well as among D2D pairs are required, while on the other hand, the device discovery problem should be faced, i.e., the need for a D2D transmitter to know whether the target receiver is in its vicinity and, thus, in valid distance to start a D2D communication.

The 3GPP (3rd Generation Partnership Project) already provides the fundamental specifications for the femtocells in Release 8 and Release 10 for LTE (Long Term Evolution) and LTE-A (LTE-Advanced) networks, respectively, while the first efforts

for standardizing D2D communications begun in Release 12, under the term Proximity Services (ProSe).

Taking all the above into account, the scope of this dissertation is to deal with challenges arising from the introduction of femtocells and D2D communications in LTE/LTE-A cellular networks. More specifically, the focus is on the following challenges:

- **Interference management in femtocell-overlaid networks**
  This problem refers to femtocells that reuse the cellular spectrum under the co-channel deployment, raising new types of interference. The problem is more severe when femtocells operate under the closed subscriber group mode, and, thus, deny the access to non-subscribed users. In this case, the interference perceived during the downlink by unsubscribed users in femtocell proximity needs more investigation. Special investigation is needed for the interference in control channels, which carry vital information for the connection maintenance.
- **Spectrum access and management for D2D communications**
  This problem refers to the management of the radio resources used for the direct transmissions in a cellular area. The main issue is how the available radio resources will be shared between cellular and direct communications, and also how the radio resources that are used for direct transmissions will be allocated to the D2D transmitters. The target is to guarantee interference-free conditions to cellular and D2D users.
- **Spectrum access and management for device discovery**
  This problem is quite similar to the previous one. However, it referred to discovery transmissions, i.e., frequent, low range direct transmissions with no QoS requirements that are used by a device in order to discover another device in its vicinity. Device discovery is an important procedure and is required prior the establishment of a D2D communication. The nature of these transmissions calls for designing different spectrum access and management schemes, than that used for the D2D communications.

A comprehensive study of the above mentioned problems is provided, while some innovative solutions and working directions are proposed.

1.2 Dissertation contributions

In this dissertation, the reader can find a comprehensive description of the architectural and physical layer aspects of the LTE/LTE-A networks, as well as, the current standardization efforts for D2D communications and the main specifications for solving the device discovery problem. However, the dominant contributions of this dissertation are summarized below:

- A thorough study of fundamental and emerging interference schemes for femtocell-overlaid LTE-A networks, and a qualitative and quantitative performance comparison from the perspective of control channel protection is provided. The focus is on the downlink control channel interference caused by femtocells to macrocell users located in the total macrocell area or in a target femtocell area,

while the impact of the femtocell deployment density on such interference is assessed.

- An examination on whether and in what extent the interferences in a femtocell-overlaid network are reflected as variations in the end-users' satisfaction is provided. Additionally, the relation between the SINR (signal to interference plus noise ratio) and the perceived Quality of Experience (QoE) at an interference-victim is studied and formulated towards designing a QoE-aware power control interference management scheme.
- A graph-coloring secondary resource allocation scheme for D2D communications is proposed. Under this scheme, interference information together with the primary resource allocation (for the cellular uplink transmissions) are represented by an enriched node contention graph (eNCG), which is utilized by graph-coloring algorithms to provide a secondary allocation for D2D communications.
- A contention-based spectrum access scheme for D2D communication in an LTE network is proposed, providing the performance analysis in terms of normalized throughput, access delay and energy consumption. The solution adapts the distributed coordination function (DCF) of the IEEE 802.11 standard to the LTE UL physical layer structure. For the analysis, Euclidian geometry is used to estimate the access and discovery probabilities (i.e., the probability a D2D transmitter to be outside the interfering area of a cellular transmitter and the probability the target D2D receiver to be located in transmitter's range) in an interference isolated cell.
- A set of enhancements is proposed in the conventional resource request/allocation procedure of an LTE-A access network towards allowing the allocation of spectrum resources for discovery transmissions. The proposed enhancements abide by the specification for device discovery provided by 3GPP.
- The spatial spectrum reuse opportunities posed by the FFR technique in the UL period of a multi-cellular LTE network are analytically studied and a D2D coordinator is proposed towards exploiting these opportunities for discovery transmissions. Simulations are used to validate the results of the theoretical study.

2 Results and discussion

2.1 Interference management in femtocell-overlaid networks

The interference problem in femtocell-overlaid networks is very challenging for the following reasons:

- Femtocell proliferation creates a highly dense network,
- femtocells are deployed by the end-consumers i.e., in a random/unplanned deployment manner,
- a heterogeneous network is created (two-tier network) and femtocells are expected to spatially reuse licensed spectrum defining the co-channel deployment, and
- femtocells can provide the option to serve only a limited set of subscribed users through the closed subscriber group (CSG) mode (no handover option for non-subscribed users)

In a femtocell-overlaid (or femto-overlaid) LTE-A network, where femtocells operate under the co-channel deployment and the closed subscriber group (CSG) mode, multiple types of interference can be found. A possible classification divides them into data and control channel interferences. Our main focus was on control channel IM in LTE-A networks overlaid by CSG femtocells. A categorization of fundamental and emerging IM schemes and a performance comparison from the perspective of control channel protection are provided. Four different categories of IM approaches are considered (frequency-domain, time-domain, power control, and resource allocation), and the advantages and limitations of each approach are presented. The evaluation focuses on the downlink (DL) control channel interference caused by femtocells to macrocell users located in the total macrocell area or in a target femtocell area; also, the impact of the femtocell deployment density on such interference is assessed. Considering the data channel protection in femto-overlaid LTE-A networks, we move beyond the conventional approaches, and design and evaluate a QoE-aware power control interference management scheme, revealing the importance of involving QoE in IM procedures.

Main results

- The interference protection of control channels is a severe problem which poses the design of interference management schemes tailored to these channels. A thorough study on LTE-A standardized tools used for control channel protection is provided, while qualitative and quantitative comparison reveals the special characteristics of each scheme.
  - The carrier aggregation (CA) with cross carrier scheduling and the almost blank subframes (ABS), require hard coordination among femtocells and between femtocells and macrocells. It is shown that due to this requirement the above mentioned interference protection schemes cannot follow the dynamic nature of femtocell deployment posing for fast, reliable, and efficient coordination algorithms.
  - Extensive performance evaluation process shows that in dense femtocell-overlaid networks, the control channel protection through distributed uncoordinated interference protection schemes is preferable.
  - The power control (PC) approach is proved to be one of the most efficient uncoordinated interference protection schemes. The drawback of this scheme is that the interferences at the victim users cannot be reduced further than a bound defined by the required signal strength/quality at the serving users.

  One of the main results is depicted in Fig. 1, where we present the cumulative distribution function (CDF) of SINR values that can be potentially perceived by MUEs in a femtocell-overlaid area served by a specific eNB sector. Fig. 1a shows the performance of CA with cross-carrier scheduling and ABS schemes compared to a baseline scenario with no use of IM and a conventional one where no femtocells are used. Both IM schemes shift the CDF curve to the right, pushing the low SINR values to overcome the control channel decoding threshold (CCDT). However, a tail of very low level SINRs remains in both the CA and ABSs curves, validating that the elimination of interference in interference-hot areas, such as

close to HeNBs or in macrocell edges, is very challenging. Comparing the two IM schemes, the use of ABSs seems to have slightly better performance than that of CA with cross-carrier scheduling, especially due to the existence of coordination with the eNB that cancels the interference caused by the coordinated HeNB (typically the stronger interferer). Applying a PC scheme supplementary to these schemes leads to further improvement (Fig. 1b), sufficiently increasing the number of SINR values that exceed the CCDT. As depicted in Fig. 1b, the dynamic nature of PC allows for efficient combination with other IM schemes, especially for low SINR values, shorting the tails of the performance curves.

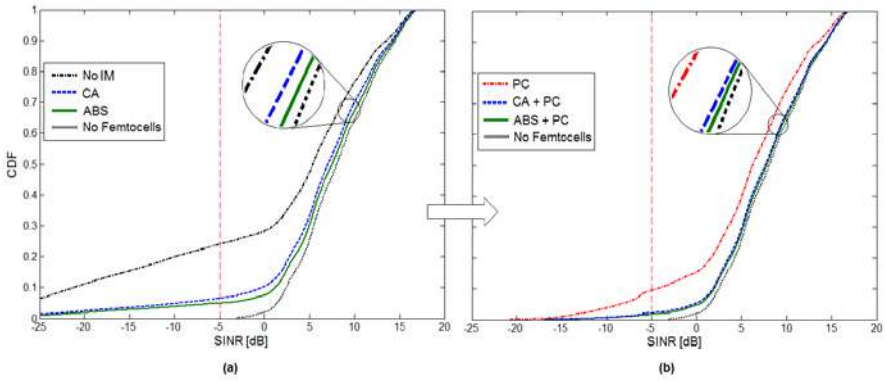


Fig. 1. Control channel interference in total cell area

- The involvement of QoE in network management is also studied, and a QoE-driven power control scheme for interference management is proposed. Simulation results show that the involvement of QoE criteria in interference management procedures is beneficial. More specifically, the proposed QoE-driven power control scheme decreases the transmission power of the femtocell base stations in lower levels that that achieved by QoS-based power control, reducing altruistically the interference at macrocell users and guaranteeing the QoE at serving users (Fig. 2).

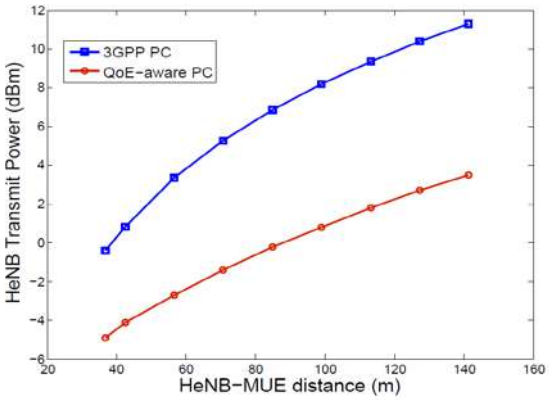


Fig. 2. Comparison of the proposed QoE-aware PC with the 3GPP PC formula

2.2 Spectrum access and management for D2D communications

Focusing on the problem of finding radio resources for D2D communications in an LTE network we propose the spatial reuse of UL cellular spectrum for D2D communications. More specifically, two different approaches are proposed. In the first one, eNBs are responsible for collecting interference information and allocating resources to D2D pairs, while in the second one, eNBs allocate a spectrum portion for D2D communications and the D2D transmitters follow a contention-based approach to access the spectrum.

Main results

- It is shown that, under a full knowledge of the interference map in a network, graph coloring theory can be used for allocating radio resources to D2D communicating pairs, achieving high spatial spectrum reuse factors and sufficiently serve multiple requests for D2D communications (Fig. 3). Simulation results show that gathering and processing interference information at the base station is a very complex problem which also burdens the network with extra signaling. This result raises the investigation for schemes where unreliable interference information or part of the interference information is available.

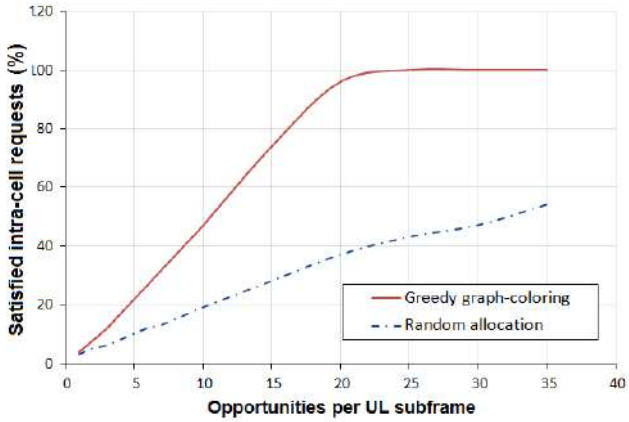


Fig. 3. Comparison of graph-coloring and random resource allocation

- Performance analysis is provided in terms of normalized throughput (see Fig. 4), access delay and energy consumption, of a contention-based scheme for D2D communications, where the D2D pairs compete to access the spectrum with no need for interference information collection and processing in a central node. For the analysis 4 scenarios are defined:
  - Scenario A1: D2D pair knows that there is no interference from the UL cellular transmission and the device discovery procedure has been applied and guarantees that D2D peers are in proximity.
  - Scenario A2: D2D pair knows that there is no interference from the UL cellular transmission. However, device discovery is not considered and a D2D transmitter tries to communicate without knowing whether the target D2D receiver is in its vicinity.
  - Scenario B1: shared spectrum is used by cellular UL and D2D communications and the D2D pair is not aware whether there is interference from the cellular UL transmission. Also, for this scenario a device discovery procedure has been applied and guarantees that D2D peers are in proximity.
  - Scenario B2: shared spectrum is used by cellular UL and D2D communications and the D2D pair is not aware whether there is interference from the cellular UL transmission. Also, the D2D transmitter tries to communicate without knowing whether the D2D receiver is in its vicinity.

  The performance of a contention-based scheme is highly correlated with the number of competing D2D pairs, and also the access and discovery probability (i.e., the probability a D2D transmitter to be outside the interfering area of a cellular transmitter and the probability the target D2D receiver to be located in transmitter's range). Euclidian geometry provides us with upper bounds for those probabilities in an interference isolated cell.

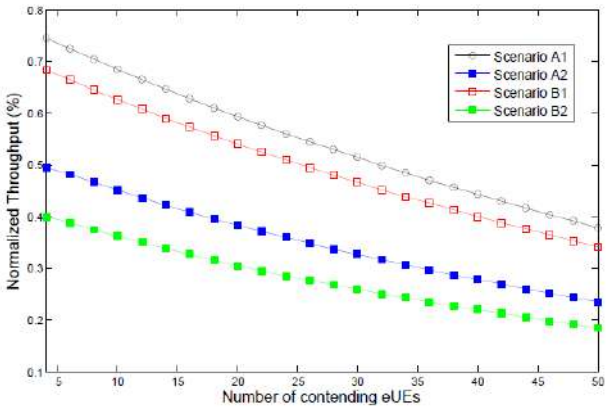


Fig. 4. Achievable normalized throughput of a single D2D pair

2.3 Spectrum access and management for device discovery

One of the main challenges that need to be addressed prior the introduction of D2D communications to cellular networks, is the discovery of devices in close proximity. The solution of this problem requires frequent transmission of discovery signals with which a UE announces its presence on a specific area or requests discovery information from a target UE. In both of the cases consumption of radio resources is needed. Moreover, considering the current efforts for launching D2D communications into the market, the discovery transmitters will rapidly increase, consuming in a more intense and massive way radio resources. Taking this into account, we examine whether the spatial reuse of the uplink (UL) cellular spectrum is a good candidate towards reducing the consumption of radio resources for discovery transmissions.

Main Results

- Thorough study of the LTE/LTE-A standardized access procedures shows that an application layer identity can be used for enabling resource allocation signaling for reactive device discovery transmissions. To this end, empty records allocated for future use in standardized (Radio Resource Control) RRC connection request and (Buffer Status Report) BSR messages, can be utilized to communicate the new identities to the serving base station.
- It is proved that, under certain conditions for the network density and the wireless environment, the fractional frequency reuse (FFR) technique can be exploited both as an inter-cell interference protector and an enabler of additional discovery transmissions.
  - Analytical results show that, for LTE/LTE-A network parameters, interference-unaware spatial spectrum reuse is suitable only for low-range low-demand transmissions, such as the discovery transmissions. Fig. 5 depicts the degradation of cellular transmissions for difference discovery transmission characteristics.

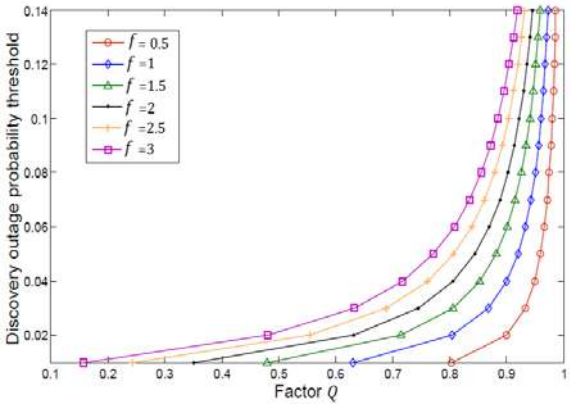


Fig. 5. Quantification of degradation factor Q for different discovery outage probabilities and target spatial spectrum reuse factors (f)

– Simulations validate that the theoretical bounds found through the analysis can be used in a realistic environment. More specifically, monitoring of the cellular transmission performance shows that the measured degradation is below the analytically calculated degradation threshold (Fig. 6).

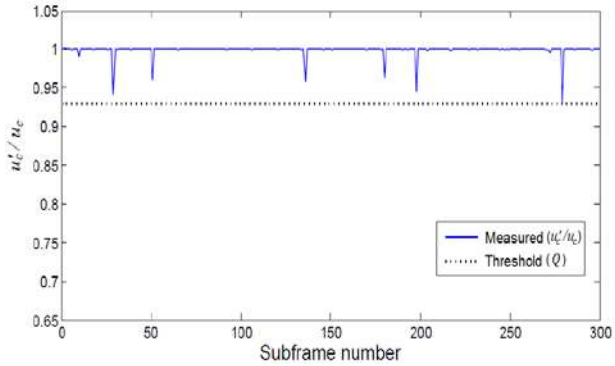


Fig. 6. Measured degradation ($\boldsymbol{u_c'/u_c}$ ratio) and comparison with the degradation threshold $\boldsymbol{Q}$

3 Conclusions

The introduction of femtocells and D2D communications in 3GPP cellular networks (LTE/LTE-A) shifts the conventional spatial spectrum reuse paradigm, to a more dynamic, flexible and distributed one. Femtocells are expected to be the most energy-efficient and cost-effective solution for improving spatial spectrum utilization and indoor coverage, while D2D communication defines an emerging technology which, in 4G networks, is expected to play an important role under the public safety concept, while in 5G networks will totally change the cellular communication land-

scape. In both of the cases, the design of more efficient radio resource management, interference management and spectrum access schemes is required. This dissertation focused on a subset of those challenges, contributing on the designing of the mobile communications of the future. Some solid and efficient solutions have been proposed, which abide by the most recent specifications defined by 3GPP. The main focus was on mitigating the interference in femtocell-overlaid networks and enabling D2D communications in LTE/LTE-A networks. Three general outcomes highlight the future research directions: first, the use of QoE promises a completely new network management and service provisioning model, second, a sharp change in the current spectrum management is required to accomplish a successful integration of cellular and D2D communications, and third, the solution of device discovery problem will launch a series of new proximity services, providing user devices with an augmented sense of their vicinity.

References

1. D. Tsolkas, N. Passas, and L. Merakos, “Enabling Device Discovery Transmissions in LTE Networks with Fractional Frequency Reuse”, IEEE Trans. on Mobile Computing, under review.
2. D. Tsolkas, N. Passas, and L. Merakos, “Alleviating Control Channel Interference in Femto-Overlaid LTE-Advanced Networks”, IEEE Communications Magazine, vol. 51, issue 10, Oct. 2013.
3. D. Tsolkas, N. Passas, and L. Merakos, “Increasing Spatial Spectrum Utilization through Opportunistic User-to-User Communications”, Springer International Journal of Wireless Information Networks, vol. 20, issue 1, May 2013.
4. D. Tsolkas, N. Passas, and L. Merakos, “Spatial Spectrum Reuse for Opportunistic Spectrum Access in Infrastructure-based Systems”, Springer Wireless Personal Communications Journal, vol. 69, issue 4, April 2013.
5. D. Xenakis, D. Tsolkas, N. Passas and L. Merakos, “Dynamic Resource Allocation in adaptive OFDMA systems”, Wiley Wireless Communications and Mobile Computing (WCMC), vol. 10, issue 10, Oct. 2010.
6. D. Tsolkas, N. Passas, and L. Merakos, “A Device Discovery Scheme for Proximity Services in LTE Networks”, The nineteenth IEEE Symposium on Computers and Communications (ISCC), Madeira, Portugal, June 23-26, 2014.
7. D. Tsolkas, E. Liotou, N. Passas, and L. Merakos, “The Need for QoE-driven Interference Management in Femtocell-Overlaid Cellular Networks”, 10th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (Mobiquitous 2013) , Tokyo, Japan, December 2013.
8. D. Tsolkas, E. Liotou, N. Passas, and L. Merakos, “Enabling D2D Communications in LTE Networks”, IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC) 2013, London, United Kingdom, September 2013.
9. D. Tsolkas, E. Liotou, N. Passas, and L. Merakos, “A Graph-Coloring Secondary Resource Allocation for D2D Communications in LTE Networks”, The 17th IEEE International Workshop on Computer-Aided Modeling Analysis and Design of Communication Links and Networks (IEEE CAMAD 2012), Barcelona, Spain, September 2012.

10. D. Tsolkas, N. Passas, and L. Merakos, “Increasing Spectrum Utilization in Wireless Infrastructure-based Systems”, The 16th IEEE Symposium on Computers and Communications (IEEE ISCC), Corfu, Greece, June 2011.
11. D. Tsolkas, D. Xenakis, Nikos Passas and Lazaros Merakos, “Opportunistic Spectrum Access over Mobile WiMAX Networks”, IEEE CAMAD 2010, December 3-4, Miami, USA.
12. D. Xenakis, D. Tsolkas, Nikos Passas and Lazaros Merakos, “Dynamic Resource Allocation in Adaptive Wireless Multiuser – Multicarrier Systems”, European Wireless 2010 – PHYDYAS Special Session FP7, Italy 2010.
13. D. Xenakis, D. Tsolkas, Nikos Passas and Lazaros Merakos, “A Dynamic Subchannel Allocation Algorithm over IEEE 802.16e networks”, IEEE – International Symposium on Wireless Pervasive Computing 2008, Santorini - Greece, 7-9 May 2008.
14. D. Tsolkas, E. Liotou, N. Passas, and L. Merakos, “LTE-A Access, Core, and Protocol Architecture for D2D Communication”, book chapter in “Smart device to smart device communication”, Springer, Editors: S. Mumtaz and J. Rodriguez, ISBN: 978-3-319-04963-2, 2014.
15. D. Tsolkas, D. Xenakis, N. Passas, and L. Merakos, “Next Generation Cognitive Cellular Networks, LTE, WiMAX and Wireless Broadband Access”, book chapter in “Cognitive Radio for Wireless Cellular and Vehicular Networks”, Springer, Editors: Dr. Hrishikesh Venkataraman and Dr. Gabriel-Miro Muntean, ISBN 978-94-007-1827-2, 2012..
16. D. Tsolkas, D. Xenakis, D. Triantafyllopoulou, and N. Passas, “Medium Access Control Layer”, book chapter in “Advances on Processing for Multiple Carrier Schemes: OFDM & OFDMA”, Nova Science Publishers, Editors: F. Bader, and N. Zorba, ISBN: 978-1-61470-634-2, 4th quarter 2011.

Mining and Managing User-Generated Content and Preferences

Georgios Valkanas*

National and Kapodistrian University of Athens
Department of Informatics and Telecommunications
gvalk@di.uoa.gr

Abstract. Capturing users preferences in today's web ecosystem is essential to provide engaging services. In this thesis, we propose various ways that take user preferences into account. Skyline queries are a characteristic example. However, a problem is that the result size may be too large. To address this issue we propose techniques to diversify it, which is an NP-Hard problem and propose efficient approximate solutions. Alternatively, users describe their experiences with a product or service, and what aspects they liked (or did not) in online reviews. Using this information, we propose techniques to identify competitive products. We present a formal framework for the identification of competitors, which we evaluate extensively and demonstrated its efficiency and efficacy. Finally, users post their preferences and interests online, in social media platforms. Social media data can be used to identify events that occur in the physical world. However, given the domain's particularities, straightforward solutions do not perform well. Therefore, we resort to affective theories of emotion from psychology. Using these theories as a starting point, we monitor the aggregate emotional state of large, geographical groups of people and automatically identify abrupt changes, linking them to the underlying events. We develop *i*) custom geocoding techniques, *ii*) a classification framework mapping social data to emotions, *iii*) an online outlier detection algorithm to identify abrupt changes, and *iv*) a visual component to ease the presentation of events.

Keywords: web mining, review mining, event detection, user preferences

1 Dissertation Summary

The World Wide Web has changed dramatically over the years since its initial inception, and is still evolving as new technologies emerge. Online services and applications are more pervasive nowadays, allowing users to share online aspects of their everyday lives. More importantly, users feel *comfortable* with doing so, which is a major shift in their attitude regarding privacy in digital environments. This general change in behavior has made the boundaries between the physical and online world less transparent.

* Dissertation Advisor: Dimitrios Gunopulos, Professor

Sharing of information takes place in various forms. The common denominator is that users express their preferences and personal opinions on various topics, such as music, products, politics, etc. Although new technologies provide the necessary framework(s) for the users to express themselves, novel techniques are required to turn the available data into useful and actionable information. Such a need translates into interesting and challenging research questions, which we have to address, in order to provide the next generation services. For instance, more expressive query types are needed, whereby user preferences can be taken into account. At the same time, we should develop techniques that extract meaningful and insightful information from this high-volume, user-generated content.

Skyline queries are an indicative example. These queries support multi-objective optimization and are geared towards returning items with different trade-offs. Although easier to understand (from the user's perspective), given that preferences are defined on each attribute separately, the output size can become extremely large. It would, then, be very tedious for the user to inspect all the results manually and soem post-processing of the results is necessary.

One way to address this problem is to incorporate additional constraints in the result selection process. For example, we can select a subset of skyline points that meet certain criteria or that optimize a given objective function. *Diversifying* the skyline result is one such approach, which is also highly desirable considering that *skyline* queries aim to return points with trade-offs [26]. Selecting a subset of the skyline points has been looked into in the past, with different goals in mind. In particular, the work in [12] proposed the selection of k skyline points that maximize the coverage of the space. Another approach selected k points that best describe the skyline contour [17], which could be seen as a diversification approach of the skyline result. However, in this thesis we demonstrated that coverage solves a *different* problem from *diversification*. Moreover, the technique in [17] assumes an underlying Euclidean space, which is not always the case for skyline queries. Therefore, that technique becomes unusable in non-Euclidean spaces, in partially-ordered domains or in feature spaces with non numerical attributes, where skyline queries are still meaningful. In addition, the result of [17] may change depending on the weights of each feature, which negates the axis-weighting invariance property of skyline queries. Finally, as we demonstrated in our thesis, the subset of skyline points selected by [17] correlate more with coverage than with diversity.

Ranking the skyline points could be another approach, giving them a degree of importance and returning them in an ordered fashion. Ranking the skyline has been the focus of past research [29, 6, 31]. However, these techniques face one of two problems: *i*) they may return points that are **not** part of the skyline [31] or *ii*) they return skyline points with *extreme* values in a single dimension [29, 6]. The latter techniques also consider an exponentially large search space, given that they take into account all non-empty feature subspaces (at most $O(2^d)$, where d is the number of considered dimensions). To counter these problems, we propose to rank skyline points using a novel technique, inspired by Information Retrieval. In particular, we adapt the well known Term Frequency-Inverse Doc-

ument Frequency (TF-IDF) scheme to the skyline domain, and propose efficient techniques to rank skyline points accordingly [27].

User feedback can be provided in other formats as well, such as semi-structured and free text. Online reviews fall under the second category, and has received considerable attention in recent years. This increased attention is due to the impact that reviews have in the marketability of products. In fact, surveys have shown that users prefer products that have already been reviewed, so that they know the item's pros and cons, and can, therefore, make informed decisions. Through a combination of user feedback and product specifications, we can derive a rigid framework to analyze and compare such products.

More specifically, based on the users' needs - as expressed in their feedback - and the extent to which a product can cover similar needs - given by its characteristics -, we can identify how *competitive* two products are. This is extremely useful for both item producers (e.g., the companies), as well as item consumers (e.g., the end users). Despite its importance, a formal framework to identify *competitive* items had been largely missing until now. The recent availability of online reviews has allowed us to test both the efficiency and efficacy of techniques that return the top-k *most competitive* products, with respect to a given item of interest [10].

Previous works on *competitor* identification has focused on the retrieval of comparative expressions, such as "X is better / worse than Y", or "(product) vs (product) Y" [2, 11]. The underlying assumption is that if two products co-occur frequently in such expressions, they are more competitive, as opposed to products that occur less frequently. The problem with this approach is that, oftentimes, there is a scarcity of such expressions, making our confidence in the drawn results quite weak. Additionally, these approaches are only useful when a product is compared as a whole, whereas users may discuss certain features of a product in their reviews. For these reasons, in this thesis we propose a rigid framework, utilizing both product specifications as well as user feedback at the feature-level. Competitiveness is then measured as the degree to which two products fulfill the same needs of groups of people with similar requirements. We present techniques to efficiently retrieve the top-k competitors of a given item, and evaluate our method's efficacy using a user study. Our results demonstrate that our techniques are very efficient, and that our model aligns well with users' intuition of competitiveness.

Finally, despite the sharp increase in numbers of online reviews over the years, these are nowhere near the data volume produced in social media. Popular social media platforms have extremely high user adoption, with Facebook boasting more than 1.28 *billion* active users per month (as of March 31, 2014), and Twitter - a later founded company - having more than 255 million active users per month (as of July 2014). A driving force of these frameworks is their networking component, with people linking to one another, as a prerequisite to share information. Undoubtedly, social media is among the most prolific areas for research nowadays, not only because of the user adoption, but also due to the usefulness of the data in various diverse disciplines: computer science, psy-

chology, sociology and journalism to name a few. Moreover, there are practical applications where the data can be used. Advertising and community detection are typical use cases, whereas (real-time) event detection, interaction analysis, and user behavior understanding increasingly gain attention. Making sense of the user-genereated content in these mediums is also extremely challenging, because of the data volume and content diversity, which is as high as the underlying population and their interests.

As a first objective, we wanted to explore the properties of data posted in social media platforms, to better understand the kind of information we are dealing with. Towards this direction, a major outcome of this thesis is to show that elevated access is primarily needed for applications that rely heavily on up-to-date information and do not only focus on popular items. Applications that only deal with popular items, can be well served through default access [24].

The fast pace of social media platforms is a key factor in considering them as online news reporting tools. However, mining high volumes of data to identify (newsworthy) events is far from trivial. Previous techniques have focused on event monitoring, implying that the event is already identified or somehow known [13]. Others simplify the problem by searching for specific keywords, which can accurately describe the event [16]. Online clustering techniques have also been explored [3, 30], however, they do not perform well in fast-paced mediums [20]. It is easy to see that identifying events, regardless of type and without prior knowledge of any descriptive keywords, calls for a different approach.

For this reason, we resort to psychological theories, according to which events impact the user psychologically, and more specifically their affective state, compelling them to externalize their thoughts. We argue that newsorthy events will impact large groups of users, and by monitoring a group's aggregate affective / emotional state, we will be able to capture abrupt changes and trace them back to the source, i.e., the event.

Within this research question, however, there are several other issues to resolve. In particular, we must identify an event's location, so we develop custom geocoding techniques, that convert textual information to GPS coordinates [18]. Extracting the affective state of a single user is challenging on its own, let alone for an entire group. We solve this problem through a classification framework, mapping social media data to a set of predefined basic emotions [19]. Capturing abrupt changes requires a careful formulation of the problem, as well as efficient computation techniques, due to the high volumes of real-time data we are dealing with. We formulate this problem as an instance of online outlier detection and propose online techniques that approximate the Probability Density Function (PDF) of the aggregate emotional state [20]. Information visualization is also important in that domain, to better explain an occuring event. Therefore, we propose a User Interface that presents all of the information in an appropriate way [21]. Such an approach also requires a great deal of system and software engineering, and end-to-end solutions could also be used to facilitate the data harvesting process [25, 23, 28].

2 Results and Discussion

2.1 Skyline Diversification

Let $\mathcal{D}$ be a d-dimensional dataset, where w.l.o.g. smaller values are preferred, i.e., we are interested in *minimizing* each attribute. [1] We say that $p = (p.x_1, ..., p.x_d) \in \mathcal{D}$ *dominates* $q = (q.x_1, ..., q.x_d) \in \mathcal{D}$ (and write $p \prec q$), when: $\forall i \in \{1, ..., d\}, p.x_i \leq q.x_i \wedge \exists j \in \{1, ..., d\} : p.x_j < q.x_j$. The skyline $\mathcal{S} \subseteq \mathcal{D}$, is composed of all points in $\mathcal{D}$ that are not dominated by any other point.

To overcome the limitations of a Euclidean space assumption, we propose to use the *Jaccard distance* for diversity computation. Each skyline point p is associated with the set of points that it dominates, denoted by $\Gamma(p) = \{q \in \mathcal{D} | p \prec q\}$. The *domination score* of p is the cardinality of $\Gamma(p)$. The similarity between p and q is defined as the Jaccard similarity between the sets $\Gamma(p)$ and $\Gamma(q)$, i.e.,

$$J_s(p, q) = \frac{|\Gamma(p) \cap \Gamma(q)|}{|\Gamma(p) \cup \Gamma(q)|}$$

and ranges between 0 and 1. The corresponding distance measure is thus $J_d(p, q) = 1 - J_s(p, q)$ and it is well known that it satisfies all metric properties. We select the Jaccard distance as a measure of diversity because:

i) it relies solely on the dominance relations among points, therefore, no user-defined distance function or other input is required,
ii) the quality of the resulting set of points does not depend on the skyline $\mathcal{S}$ alone, but on the characteristics of $\mathcal{D}$ as well
iii) it leads to elegant ways of diversity computation by means of min-wise independent permutations, and
iv) it is the most widely accepted measure for set (dis)similarity.

We model k-diversity as a *k-dispersion* problem, which is NP-Hard [9]. In k-dispersion, the goal is to find k objects that optimize an objective function of their distance. The optimal solution is given by:

$$OPT = \arg \max_{\substack{\mathcal{A} \subseteq \mathcal{S} \\ |\mathcal{A}| = k}} f(\mathcal{A})$$

There are two basic alternatives for the objective function: *i*) maximize the sum of distances (k-MSDP) and *ii*) maximize the minimum distance (k-MMDP). Although both alternatives are valid, we choose to work with k-MMDP because *i*) it leads to 2-approximation algorithms, instead of the 4-approximation of k-MSDP [15], and *ii*) it intuitively returns results of better quality.Given the NP-Hardness of the problem, we resort to approximate solutions. In fact, the greedy approach can be quite inefficient, due to a large number of range queries. Therefore, we propose the use of the MinHashing technique [5], that transforms

[1] We focus on numerical attributes for ease of presentation. Our approach applies to categorical ones equally well.

the original space into a more compact one, where computations are much faster. In particular, we develop the SkyDiver framework, which operates in two phases.

Phase 1: Fingerprinting. This phase generates a *signature* of reduced size for each skyline point, based on MinHashing. Alternatively, we can use *Locality Sensitive Hashing* (LSH) as a memory efficient alternative.

Phase 2: Selection. This phase is responsible for selecting the k most diverse skyline points, and can be applied to either the MinHash or the LSH signatures.

Assume that the data set is viewed as a matrix M with $n - m$ rows and m columns, $m = |\mathcal{S}|$ and $n = |\mathcal{D}|$. Each skyline point is represented by a single column, whereas a dominated point is represented by a row. In this matrix, $M[i, j] = 1$ iff the j-th skyline point dominates the i-th data point and 0 otherwise. Let $\mathcal{H} = \{h_1, ..., h_t\}$ be a set of t min-wise independent hash functions, where each h_i performs a random permutation of the rows. The cardinality of $\mathcal{H}$ (i.e., the number of hash functions used) determines the size of each signature. To generate random permutations, each hash function $h_i \in \mathcal{H}$ is of the form

$$h_i(x) = a_i \cdot x + b_i \mod P$$

where P is a prime number larger than $n - m$ and a_i, b_i are randomly chosen constants taking integer values in $[1, P]$. According to [5], if $J_s(p, q)$ is the Jaccard similarity between skyline points p and q, then for each hash function h_i it holds

$$Prob[h_i(p) = h_i(q)] = J_s(p, q).$$

Recall that each row of the matrix M is a bit-array. If $M[i, j] = 1$ then the $\mathcal{S}_j \prec \mathcal{D}_i$. Each row is hashed t times, by every $h_i \in \mathcal{H}$ and the signature of each skyline point is updated accordingly. Therefore, each signature is composed of t integer values. According to [7], if $\Omega(\varepsilon^{-3}\beta^{-1}\log(1/\delta))$ is the signature size, where ε is the maximum allowed error ($0 < \varepsilon < 1$), then with probability at least $1 - \delta$ it holds

$$(1 - \varepsilon)J_s(p, q) + \varepsilon\beta \leq \widehat{J}_s(p, q) \leq (1 + \varepsilon)J_s(p, q) + \varepsilon\beta$$

where $0 < \beta < 1$ is the required precision.

Given the signature matrix $\widehat{M}$, we can select the k skyline points in the transformed space, which is a metric space. Therefore we can use the greedy approach on the signatures and acquire a 2-approximation solution. However, due to distance distortions, as a result of embedding the distances in lower dimensionality (through MinHashing), it is possible to obtain a sub-optimal solution. The following theorem relates the true optimal solution, to the one computed by working with MinHash signatures.

Theorem 1. *Let OPT be the value of the optimal solution to the k-diversity problem in the original space and let x, y denote the corresponding skyline points, i.e., $J_d(x, y) = OPT$. Similarly, let $\widehat{OPT}$ be the optimal value if the problem is solved using MinHash signatures and let a, b be the corresponding skyline points, i.e., $\widehat{J}_d(a, b) = \widehat{OPT}$. For a given ε and sufficiently small δ, it holds that: $J_d(a, b) \geq \frac{1+\varepsilon}{1-\varepsilon}OPT - \frac{2\varepsilon}{1-\varepsilon}$.*

2.2 Competitor Mining

Competitiveness is a challenge that every product or service provider has to face, regardless of the application domain. A significant amount of relevant work has demonstrated the strategic importance of identifying and monitoring an entity's competitors [14]. In this thesis we focus on a formal framework for competitor identification:

Problem 1. We are given a set of items $\mathcal{I}$, defined within the feature space $\mathcal{F}$ of a particular domain. Then, given any pair of items I, I' from $\mathcal{I}$ we want to define a function $C_{\mathcal{F}}(I, I')$ that computes the competitiveness between the two in the context of the domain.

Figure 1 provides a (simplified) overview of our approach, where we illustrate the competitiveness between three different items I_1, I_2 and I_3. Each item is mapped to the set of features that it can offer to the users. Three distinct features are considered in this example: A, B and C. Note that, for this simple example, we only consider binary features (i.e. available/not available). Our actual formalization accounts for a much richer space of binary, categorical and numerical features. The left side of the figure shows three groups of users (g_1, g_2, g_3). The example assumes that these are the only groups in existence. Users are grouped based on their preferences with respect to the features. For example, the users in group g_2 are only interested in features A and B. As can be seen by the figure, items I_1 and I_3 are not competitive to each other, since they simply do not appeal to the same groups of users. On the other hand, I_2 is in competition with both I_1 (for groups g_1 and g_2) and I_3 (for g_3). Finally, another interesting observation is that I_2 competes with I_1 for a total of 4 users, and with I_3 for a total of 9 users. In other words, I_3 is a stronger competitor for I_2, since it claims a much larger portion of I_2's market-share than I_1. In our work, we propose ways to deduce these user-groups from sources such as query logs and customer reviews, and describe methods to estimate the size of the market

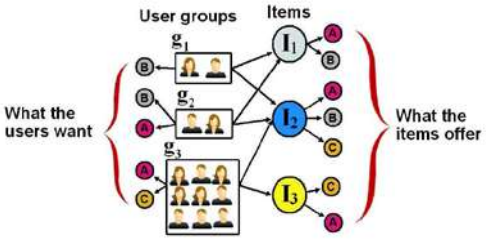


Fig. 1. Simplified example of our competitiveness paradigm

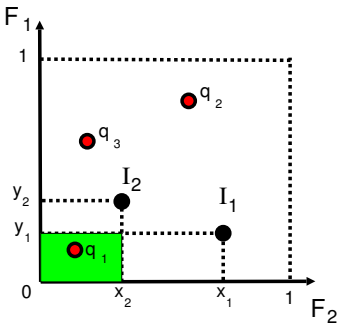


Fig. 2. Geometric interpretation of pairwise coverage

share that they represent. Our work is the first to utilize the opinions expressed in customer reviews as a resource for mining competitiveness.

In order to evaluate the competitiveness of two given items I_i, I_j in the context of a subset of features $\mathcal{F}'$, we need to compute the number of possible value assignments over $\mathcal{F}'$ that are satisfied by *both* items. Formally, we define pairwise coverage as follows:

Definition 1. [**Pairwise Coverage**] *Given the complete set of features $\mathcal{F}$ in a domain of interest, let $\mathcal{V}_{\mathcal{F}'}$ be the complete space of all possible value-assignments over the features in a subset $\mathcal{F}' \subseteq \mathcal{F}$. Then, the* coverage *$cov(\mathcal{V}_{\mathcal{F}'}, I_i, I_k)$ of a pair of items I_i and I_j with respect to $\mathcal{V}_{\mathcal{F}'}$ is defined as the portion of $\mathcal{V}_{\mathcal{F}'}$ that is covered by both items.*

Considering the above definition, we observe that the coverage of each dimension (i.e. each feature $F \in \mathcal{F}'$) is independent of the others. Therefore, we first compute the percentage of each dimension that is covered by the pair. We can then optimally compute the coverage of the entire space $\mathcal{V}_{\mathcal{F}'}$ as the product of the respective coverage values $\mathcal{V}_{\{F\}}$ for every $F \in \mathcal{F}'$. Formally:

$$cov(\mathcal{V}_{\mathcal{F}'}, I_i, I_j) = \prod_{F \in \mathcal{F}'} cov(\mathcal{V}_{\{F\}}, I_i, I_j) \tag{1}$$

This computation has a clear geometric interpretation: The portion of the space $\mathcal{V}_{\mathcal{F}'}$ that is covered by a pair of items can be represented as a hyper-rectangle in $|\mathcal{F}'|$-dimensional space. For each dimension F, $cov(\mathcal{V}_{\{F\}}, I_i, I_j)$ gives us the portion of the dimension that is covered by the two items. Finally, by multiplying the individual coverage values, we are essentially computing the volume of the hyper-rectangle that represents the entire space $\mathcal{V}_{\mathcal{F}'}$. This is graphically portrayed in Figure 2, which shows the common coverage of two items I_1, I_2 (green area) in the context of a dimensional space $\{F_1, F_2\}$.

Definition 1 allows us to evaluate the coverage provided by a pair of items to (the value space of) any subset of features $\mathcal{F}'$. Conceptually, $\mathcal{F}'$ captures the fraction of the population that is interested in the features included in $\mathcal{F}'$. Further, we define $\mathcal{Q}$ to be the collection of subsets with a non-zero weight. Formally: $\mathcal{Q} = \{\mathcal{F}' \in 2^{\mathcal{F}} : w(\mathcal{F}') > 0\}$. Taking the above into consideration, we formally define the competitiveness of two items I_i, I_i as follows:

Definition 2. [**Competitiveness**] *Given the complete set of features $\mathcal{F}$ of a domain of interest, let $\mathcal{Q}$ be the set of all subsets of $\mathcal{F}$ that have a non-zero popularity weight. Then, the competitiveness of two given items I_i and I_i is defined as:*

$$C_{\mathcal{F}}(I_i, I_j) = \sum_{\mathcal{F}' \in \mathcal{Q}} w(\mathcal{F}') \times cov(\mathcal{V}_{\mathcal{F}'}, I_i, I_j) \tag{2}$$

where $cov(\mathcal{V}_{\mathcal{F}'}, I_i, I_j)$ is the portion of $\mathcal{V}_{\mathcal{F}'}$ that is covered by both I_i and I_j.

Given this definition of competitiveness, we study the natural problem of finding the top-k competitors of a given item I^*:

Problem 2. We are given a set of items $\mathcal{I}$, defined within the feature space $\mathcal{F}$ of a domain. Then, given a single item $I \in \mathcal{I}$, we want to identify the k items from $\mathcal{I} \setminus \{I\}$, that maximize the pairwise competitiveness with I:

$$I^* = \underset{I' \in \mathcal{I} \setminus \{I\}}{\operatorname{argmax}} C_{\mathcal{F}}(I, I') \tag{3}$$

Instead of finding the top-k competitors using a naive solution that iterates over all items, computes their competitiveness with respect to I^* and finally orders them, we develop a more efficient technique, namely *CMiner*. Our algorithm makes use of an indexing scheme, called the *Dominance Pyramid*, which is built based on the dominance property of items, as discussed in the skyline section. It also applies very efficient pruning on the search space, by bounding the score of candidate points. Building upon a result from [4], in this thesis, we show that the algorithmic complexity of our technique is O($|\mathcal{I}|$ * $|\mathcal{Q}|$ * k^2), where $\mathcal{Q}$ is the set of feature subsets with non-zero weights [2].

2.3 Event Detection

Our objective with social media data [3] can be summarized as follows:

Problem 3. [**Event Detection**] Given a time ordered stream of tweets, identify those posts which *i*) alter significantly and abruptly the emotional state of a (potentially) large group of users, and *ii*) can be traced back to event e.

Event Extraction Workflow The problem we described fits nicely with an outlier detection definition. Figure 3 shows a schematic overview of our system [4]

[2] This is to retrieve the top-k competitors of every item in the dataset
[3] We focus on Twitter, http://twitter.com
[4] Storage image by Barry Mieny, under CC BY-NC-SA license.

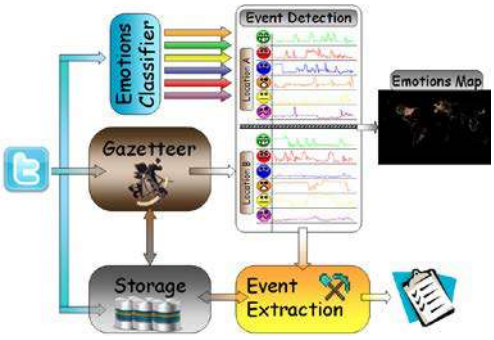


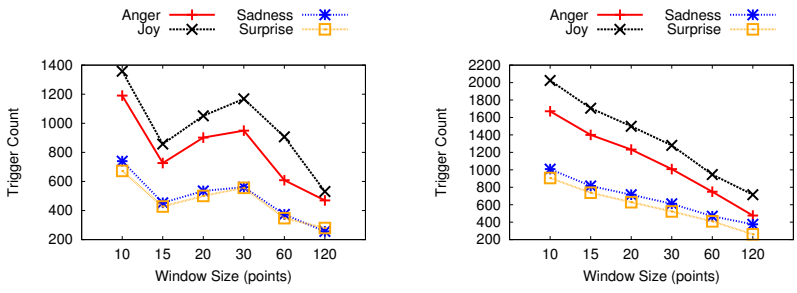
Fig. 3. Schematic interaction of our system's components

used to identify events. The Twitter stream is our input, feeding two components, namely the *emotions classifier* and the *location extraction* subsystem. Through a custom geocoding component [18], each incoming tweet is mapped to a GPS location, which will also be the location of the event (assuming one occurs).

We classify tweets to 6 basic emotions, proposed by American psychologist Paul Ekman [8]: *anger, fear, disgust, happiness, sadness, surprise*, plus a neutral one (*none*), to describe the absence of an emotion. Tweets with non-neutral emotions are further processed by virtual sensors, one per location and per emotion. Virtual sensors count how many tweets they have received during the last aggregation interval a (system parameter). These values form the aggregate emotional state, which are fed to the event detection mechanism. By operating over the w most recent values, the event detection module approximates their distribution using non-parametric models (kernel estimators in particular), estimating the Probability Density Function (*PDF*). The same models are used for event detection, identifying tuples which are outside of the norm over the most recent history of tuples, given by the combination of (a, w) parameters.

The Temporal Nature of Twitter Discussions Using a large crawl of Twitter data (2 month period), we evaluated the behavior of the medium and our event extraction approach. In particular, we computed the number of times a "trigger" was raised (which corresponds to an individual event in our case). Figure 4 shows the number of times our approach raised an event as a function of the history it maintains, when aggregating emotions over the past 1 minute and monitoring the entire stream at once (we use only one sensor).

Interestingly, a bigger sample size results in more triggers. This is due to maintaining outdated information compared to the fast pace of the medium. Increasing the history length results in fewer triggers (Figure 4(b)), because new points can be matched against more sampled data, and are less likely to be flagged as outliers. On the other hand, Figure 4(a) leads to a very interesting observation. Using a 50% sample, there is a dramatic drop in the number of triggers, when we increase the window size from 10 to 15; from that point, until



(a) Sample size = 50%

(b) Sample size = 100%

Fig. 4. #Times a trigger was raised, w.r.t. the window size. a = 1min, r = 0.01, p=0.1

a window size of 30, triggering events increases slightly, and begins decreasing from that point on. This means that for 1 minute aggregations, there are rapid changes in the observed emotions; therefore a window of 10 points may be too narrow, to maintain a representative "history". On the other hand, a window between 15 – 30 minutes seems like a better choice. This result correlates very well with the real time nature of the medium, where people tend to speak and respond very quickly to their tweets. It also means that events that are present in our data create some momentum over a mid-size period (∼30minutes), and then dissipate.

3 Conclusions

In this thesis, we considered various scenarios where user preferences can be taken into account, in order to improve the quality of a provided service. In particular, we focused on different use cases, where user preferences play a vital role, namely: *i*) skyline queries, *ii*) review mining and competitor identification, and *iii*) social media. The problems that arise in each of these domains are unique, and we proposed techniques to efficiently solve them, while providing quality guarantees on our solutions. Our analysis also revealed some interesting properties for the social media domain. Finally, we effectively modeled competitors in a formal framework. User preferences and feedback may come in different forms, such as mouse movements [1], or interaction with online content [22].

References

1. I. Arapakis, M. Lalmas, and G. Valkanas. Understanding within-content engagement through pattern analysis of mouse gestures. In *CIKM*, pages 1439–1448, 2014.
2. S. Bao, R. Li, Y. Yu, and Y. Cao. Competitor mining with the web. *IEEE Transactions on Knowledge Data Engineering*, pages 1297–1310, 2008.
3. H. Becker, M. Naaman, and L. Gravano. Learning similarity metrics for event identification in social media. In *WSDM*, pages 291–300, 2010.
4. J. L. Bentley, H. T. Kung, M. Schkolnick, and C. D. Thompson. On the average number of maxima in a set of vectors and applications. *Journal of the ACM*, 25(4):536–543, 1978.
5. A. Z. Broder, M. Charikar, A. M. Frieze, and M. Mitzenmacher. Min-wise independent permutations. *Journal of Computer and System Sciences*, 60(3):630–659, 2000.
6. C. Y. Chan, H. V. Jagadish, K.-L. Tan, A. K. H. Tung, and Z. Zhang. On high dimensional skylines. In *EDBT*, pages 478–495, 2006.
7. M. Datar and S. Muthukrishnan. Estimating rarity and similarity over data stream windows. In *ESA*, pages 323–334, 2002.
8. P. Ekman, W. Friesen, and P. Ellsworth. *Emotion in the human face: guide-lines for research and an integration of findings*. Pergamon Press, 1972.
9. C.-C. Kuo, F. Glover, and K. S. Dhir. Analyzing and modeling the maximum diversity problem by zero-one programming. *Decision Sciences*, 24(6):1171–1185, 1993.

10. T. Lappas, G. Valkanas, and D. Gunopulos. Efficient and domain-invariant competitor mining. In *SIGKDD*, pages 408–416, 2012.
11. R. Li, S. Bao, J. Wang, Y. Liu, and Y. Yu. Web scale competitor discovery using mutual information. In *ADMA*, 2006.
12. X. Lin, Y. Yuan, Q. Zhang, and Y. Zhang. Selecting stars: The k most representative skyline operator. In *ICDE*, pages 86–95, 2007.
13. M. Mathioudakis and N. Koudas. Twittermonitor: trend detection over the twitter stream. In *SIGMOD*, 2010.
14. M. E. Porter. *Competitive Strategy: Techniques for Analyzing Industries and Competitors*. Free Press, 1980.
15. S. S. Ravi, D. J. Rosenkrantz, and G. K. Tavyi. Heuristic and special case algorithms for dispersion problema. *Operations Research*, 42(2):299–310, 1994.
16. T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *WWW*, pages 851–860, 2010.
17. Y. Tao, L. Ding, X. Lin, and J. Pei. Distance-based representative skyline. In *ICDE*, pages 892–903, 2009.
18. G. Valkanas and D. Gunopulos. Location extraction from social networks with commodity software and online data. In *ICDM Workshops (SSTDM)*, 2012.
19. G. Valkanas and D. Gunopulos. Event detection from social media data. *IEEE Data Eng. Bull.*, 36(3):51–58, 2013.
20. G. Valkanas and D. Gunopulos. How the live web feels about events. In *CIKM*, 2013.
21. G. Valkanas and D. Gunopulos. A ui prototype for emotion-based event detection in the live web. In *SS-KDD-HCI @ SouthCHI*, pages 89–100, 2013.
22. G. Valkanas and D. Gunopulos. Predicting download directories for web resources. In *WIMS*, page 8, 2014.
23. G. Valkanas, D. Gunopulos, I. Galpin, A. J. G. Gray, and A. A. A. Fernandes. Extending query languages for in-network query processing. In *MobiDE*, pages 34–41, 2011.
24. G. Valkanas, I. Katakis, D. Gunopulos, and A. Stefanidis. Mining twitter data with resource constraints. In *IEEE/WIC/ACM International Conference on Web Intelligence*, pages 182–187, 2014.
25. G. Valkanas, A. Ntoulas, and D. Gunopulos. Rank-aware crawling of hidden web sites. In *WebDB*, 2011.
26. G. Valkanas, A. N. Papadopoulos, and D. Gunopulos. Skydiver: A framework for efficient skyline diversification. In *EDBT*, pages 406–417, 2013.
27. G. Valkanas, A. N. Papadopoulos, and D. Gunopulos. Skyline ranking à la IR. In *ExploreDB*, pages 182–187, 2014.
28. G. Valkanas, A. Saravanou, and D. Gunopulos. A faceted crawler for the twitter service. In *WISE*, pages 178–188, 2014.
29. A. Vlachou and M. Vazirgiannis. Ranking the sky: Discovering the importance of skyline points through subspace dominance relationships. *Data & Knowledge Engineering*, 69(9):943–964, 2010.
30. J. Weng and B.-S. Lee. Event detection in twitter. In *ICWSM*, 2011.
31. M. L. Yiu and N. Mamoulis. Efficient processing of top-k dominating queries on multi-dimensional data. In *VLDB*, pages 483–494, 2007.

Localization and Mobility Management in Heterogeneous Wireless Networks with Network-Assistance

Dionysis Xenakis

National and Kapodistrian University of Athens
Department of Informatics and Telecommunications
nio@di.uoa.gr

Abstract. Today's heterogeneous wireless network (HWN) is a collection of ubiquitous wireless networking elements (WNEs) that support diverse functional capabilities and networking purposes. In such a heterogeneous networking environment, localization and mobility management will play a key role for the seamless support of emerging applications, such as social networking, massive multiplayer online gaming, device-to-device (D2D) communications, smart metering, first-responder communications, and unsupervised navigation of communication-aware robotic nodes. Most of the existing wireless networking technologies enable the WNEs to assess their current radio status and directly (or indirectly) estimate their relative distance and angle with respect to other WNEs of the same Radio Access Technology (RAT); thus, the integration of such information from the ubiquitous WNEs arises as a natural solution for robustly handling localization between WNEs and mobility management of moving WNEs governed by resource-constrained operation. Under this viewpoint, we investigate how the utilization of such spatial information can be used to enhance the performance of localization and mobility management in the today's HWN. In this work we focus and contribute in the areas of: i) localization and peer-discovery between non-homogeneous WNEs, ii) network-assisted D2D discovery in cellular networks, iii) energy-efficient handover (HO) decision in the macrocell - femtocell network, and iv) network-assisted vertical handover decision (VHO) for the integrated cellular and WLAN HWN.

Keywords: Localization, Peer-to-peer discovery, Handover, Heterogeneous Wireless Networks, Device-to-Device Discovery, Femtocells.

1 Dissertation Summary

1.1 Introduction

Over the past few years, wireless networks have transformed from a set of single-tier operator-deployed circuit-switched systems, designed to support voice-centric services in wide geographical regions, to a set of multi-tier networking

[1] Dissertation Advisor: Lazaros Merakos, Professor

clusters of user-installed IP-based wireless networking elements (WNEs), designed to support heterogeneous communication capabilities and diverse networking requirements. The nowadays heterogeneous wireless network (HWN) is composed by tower-mounted cellular base stations (BSs) providing wide area coverage (a.k.a. macrocells), user-deployed low-power and small-sized base stations that boost the area spectral efficiency of the licensed spectrum [1] (e.g. femtocells), wireless local area network (WLAN) stations that enable high-data rate connections to the Internet over the unlicensed spectrum [2], as well as other low-cost low-power and battery-operated sensors that monitor, measure, and commute localized changes in nearby sink nodes (e.g. in the smart grid).

In such a heterogeneous wireless networking environment, the mobile terminals (or the WNEs) are required to discover the set of nearby WNEs that they can access and, if needed, to seamlessly transfer their ongoing services by associating with the one(s) that meet their particular communication requirements. Even though different terms are used among the different systems for the discovery, e.g. network discovery in IEEE-based systems or cell search in 3GPP-systems, and the association phase, e.g. handover for intra-system mobility in cellular systems and vertical handover for inter-system mobility in heterogeneous systems, the discovery and association phases are integral part of the mobility management (MM) process of all the existing wireless networking technologies.

Since the nowadays mobile terminals are equipped with numerous radio access interfaces, that enable them to access the Internet via multiple Radio Access Technologies (RATs), mobility management is a challenging issue for safeguarding the robustness of the nowadays HWN. Firstly, the support of multiple radio interfaces asks for increased complexity and battery consumption at the mobile terminal, due to the substantially increased number of (not necessarily homogeneous) WNEs that should be discovered and evaluated with respect to their ability to support the service requirements of the mobile terminal. Secondly, the recent surge of interest for the direct exchange of localized traffic between nearby devices without network involvement, a.k.a. peer-to-peer (P2P) communications, questions the scalability of the predominant user-assisted network-controlled mobility management model that is currently adopted by the vast majority of cellular networks. Social networking applications, massive multiplayer online gaming, device-to-device (D2D) communications, smart metering, first-responder communications, and unsupervised navigation of communication-aware robotic nodes, are only some of the emerging applications that motivate this disruptive communication paradigm [3][4][5]. Thirdly, the nowadays HWN is characterized by the unplanned deployment of densely overlapping (in coverage) WNEs that serve diverse communication purposes over the same spectrum. This feature not only dictates the employment of semi-autonomous terminal-based discovery, but also transforms the nowadays HWN to a stochastic system dominated by the spatial dependencies of the heterogeneous WNEs.

Aiming to improve the mobility, interference, and energy management at the WNEs, more and more wireless networking technologies incorporate a suite of measurement capabilities to their baseline operation. Such measurements en-

able the fixed (or moving) WNEs to assess the status of the ambient radio environment, e.g. interference level in a specific spectrum band, and directly (or indirectly) estimate their physical distance and angle with respect to other WNEs of the same RAT. The incorporation of knowledge on the radio status or the spatial dependencies between the WNEs, arises as the natural solution for effectively handling the unplanned and overlapping deployment of WNEs upon mobility management. The integration of such spatial information can also be the cornerstone for more accurate localization between WNEs that do not necessarily support the same RAT, i.e. heterogeneous WNEs. Besides, localization, which refers to the process by which a WNE estimates its physical distance (or connectivity) to another WNE, is currently considered as a vital component in the future 5G network where the estimation of proximity between the myriads of WNEs can be a limiting performance factor [6].

Under this viewpoint, in this doctoral thesis we investigate how knowledge of the radio status or the spatial distribution of the WNEs can be used to enhance the performance of localization, discovery, and association in the nowadays HWN. Besides, the exploitation of such spatial information is the common denominator for all algorithms and analytical models developed in this work. The remainder of this section is organized as follows. In section 1.1, we start with an illustrative example that motivates the utilization of radio/positioning measurements from the heterogeneous WNEs as means of improving the performance of localization, discovery, and association in the nowadays HWN. In section 1.2, we briefly summarize related works and our key contributions in the areas of localization and peer-discovery in HWN, device-to-device discovery in cellular network, energy-efficient handover decision in the macrocell - femtocell network, energy-efficient vertical handover decision in the cellular/Wi-Fi network, and mobility management in the LTE-Advanced Network with Femtocells.

1.2 Motivating Example and Research Areas

In Figure 1 we depict an instance of the nowadays HWN, which is composed by long-range cellular BSs, e.g. macrocells, numerous small-sized stations that operate in the licensed spectrum, e.g. picocells or femtocells, WLAN access points that operate in the unlicensed spectrum, e.g. Wi-Fi hotspots, dual-mode cellular/Wi-Fi hotspots (fourth generation (4G) hotspots), low-power sensors, e.g. ZigBee sensors, localized traffic aggregators/sink nodes, e.g. dual-mode Wi-Fi/ZigBee smart meters, D2D-enabled cellular devices, and communication-enabled robotic nodes, e.g. dual-mode Wi-Fi/cellular robots. To better comprehend the key research areas of this work, consider the scenario where the dual-mode robot (source peer) seeks to discover a malfunctioning ZigBee sensor (target peer) to replace it. The dual-mode robot is assumed to host active connections to the Internet. Firstly, the communication-enabled robot is required to (continuously re-)assess its physical distance to the malfunctioning ZigBee sensor by employing localization, i.e. estimate the distance Z1. Secondly, as the robot moves towards the malfunctioning ZigBee sensor, at some point it will have to choose between associating with Femto 1 (femtocell) or BS 2 (macrocell),

which refers to the scenario of intra-system mobility in the macrocell-femtocell network, i.e. horizontal handover. In the sequel, the robot will have to choose between associating with Femto 2 (femtocell), Wi-Fi 2 (WLAN access point), BS 2 (macrocell) or BS 1 (macrocell), which refers to the scenario of inter-system mobility between heterogeneous RATs, i.e. vertical handover. Since the robot is a battery-operated device, its ongoing services should be seamlessly transferred to the WNE that not only guarantees a prescribed Quality of Service (QoS) target, but also requires the minimum energy consumption overhead for communications, i.e. need for energy-efficient horizontal or vertical handovers. Finally, assuming that the ZigBee sensor (Sensor 1) is located in a difficult to access area, the dual-mode robot is required to discover a local D2D-enabled device (User 1) that will be responsible for remotely navigating the robot by exploiting visual signal from its on-board camera (localized real-time video traffic).

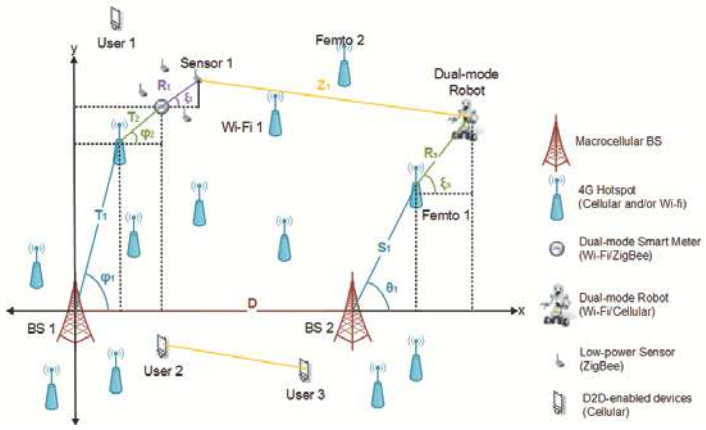


Fig. 1. Motivating example for localization and mobility management in wireless heterogeneous networks using network-assistance

Aiming to cover the first challenge, which refers to the localization between not necessarily homogeneous WNEs over large geographical areas, e.g. the robot and the ZigBee sensor, in this work we analyze how partial or full knowledge on the spatial dependencies between the nodes, e.g. relative distances and angles with respect to a reference direction, affect the localization precision and the peer discovery accuracy in the nowadays multi-tier clustered HWN. On the other hand, aiming to cover the challenge of energy-efficient horizontal handover in the macrocell - femtocell network, we propose an energy-efficient handover algorithm that exploits standard measurements on the radio status of nearby base stations (femtocells or macrocells) to identify the one that minimizes the transmit power at the mobile terminal given a prescribed mean Signal to Interference plus Noise Ratio (SINR) target (QoS indicator). To address the challenging issue of energy-efficient network selection between cellular and WLAN WNEs, we propose an energy-efficient vertical handover algorithm that utilizes standard measurements

on the radio status of nearby base stations or WLAN access points, in order to identify the point of attachment (PoA) that minimizes the power consumption at the mobile terminal given a prescribed mean SINR (QoS indicator). Finally, we also analyze how different combinations of location information on the cellular network layout can be used to enhance the performance of network-assisted D2D discovery. We note that even though we use the example in Figure 1 to allow a more easy understanding of the research areas addressed in this work, the proposed analytical models and algorithms apply to more generic scenarios and deployment layouts.

1.3 Related Works and Key Contributions

Localization and Peer-Discovery in Heterogeneous Wireless Networks Localization poses several challenges that span from mitigating (or exploiting) prominent effects of the wireless medium [5] [7] to employing multi-user detection (MUD) and cooperation for more accurate localization [8]. The Poisson point process (PPP) has been recently shown to be as accurate as the grid model and a good fit for modeling the locations of small-sized stations in multi-tier cellular networks with independent tiers [1]. Besides, the PPP model has been used to derive near-optimal strategies for random peer discovery in homogeneous networks [9]. In parallel, a considerable amount of works identify that the locations of short-range WNEs are not completely random, e.g. sensors, femtocells, or more generic WNEs [10][11], and typically form clusters around other WNEs of increased radii.

In our work, we derive closed-form expressions for the conditional probability distribution of the distance between two heterogeneous WNEs, given partial knowledge of the spatial relations between their upper-tier parent WNEs. We also show that the probability density function (pdf) expressions describe the statistical behavior of localization between heterogeneous WNEs. Moreover, we analyze the performance of location-aware peer discovery between heterogeneous WNEs given different knowledge of the HWN layout. We also analyze the impact of the key system parameters on the performance of location-aware peer discovery and derive optimal strategies for the placement of upper-tier WNEs as means of maximizing the peer discovery probability between two heterogeneous WNEs of interest. We conclude with valuable insights for the design of location-aware peer discovery in the today's HWN.

Device-to-Device Discovery in Cellular Networks Most of the related literature to our work deals with the analysis and optimization of D2D communications [12]. In parallel, PPPs, which have been extensively used for the analysis of multi-tier cellular networks [1], are increasingly used to model and analyze the performance of D2D communications [13]. Our work addresses the challenging issue of network-assisted D2D discovery in random spatial networks. Our key contributions can be summarized as follows. Firstly, we derive closed-form expressions for the conditional pdf and ccdf of the distance between two

D2D peers, given various combinations of location information parameters including at least the distance or the neighboring degree between their associated BSs. Secondly, we analyze the performance of network-assisted D2D discovery given the most prominent combinations of location information parameters. Our analysis readily quantifies how different levels of location knowledge affect the D2D discovery probability. Thirdly, we examine the behavior of the D2D discovery probability with respect to key system parameters, with the emphasis given on the BS density. We identify conditions under which the D2D discovery is optimized and provide analytical expressions for computing the optimal BS density. Finally, we provide useful design guidelines for network-assisted D2D discovery in cellular networks.

Handover Decision in the Macrocell - Femtocell Network Current literature also includes various algorithms and studies for the HO decision phase in the two-tier network [14]. However, current HO decision algorithms emphasize on reducing the number of HOs in the two-tier network mainly based on user mobility and traffic type criteria. In most of the cases, the impact of the proposed algorithms on the UE energy consumption, the RF interference, and the network signaling load, is not investigated.

In our work, we jointly consider the impact of interference, energy consumption, and user mobility during the HO decision phase in the two-tier LTE-A network. A strong innovation of our work is the exchange and utilization of standard LTE-A measurements to accurately estimate the mean UE transmit power on a per candidate cell basis, given a prescribed mean SINR target. The exclusion of candidate LTE-A cells which can compromise wireless connectivity, and the incorporation of the user's prescribed SINR target during the mean UE transmit power estimation, are two more important features of the proposed algorithm towards sustained wireless connectivity, enhanced QoS support, and reduced outage probability. We also provide comprehensive description of the network signaling procedure required for employing the proposed algorithm.

Energy-Efficient Vertical Handover Decision in the Cellular / Wi-Fi network Current literature includes a noteworthy amount of algorithms for horizontal and vertical handover decision for heterogeneous networks [14][15]. However, only a few works utilize the ANDSF functionality for efficient network discovery and seamless mobility at the MMT. In addition, even though the utilization of standard LTE-A measurements for handover has been recently proposed [16], the joint utilization of the enhanced radio measurement capabilities of the LTE-A and the IEEE 802.11-2012 systems is an unexplored research area [17]. In our work, we propose an Andsf-assisted eneRgy-effiCient vertical Handover decisiON (ARCHON) algorithm for the heterogeneous IEEE 802.11-2012 / LTE-A network. The proposed algorithm, referred to as ARCHON, enables a MMT to select the network PoA that minimizes its average overall power consumption and guarantees a mean SINR target for its ongoing connections.

Mobility Management in the LTE-Advanced Network with Femtocells
Current literature lacks of surveys and comparative studies engaged with the matter. In our work, we discuss the open issues for MM support in the presence of femtocells and overview the key aspects of MM in the LTE-A system. Moreover, we survey current state-of-the-art HO decision algorithms for the two-tier macrocell-femtocell network and overview their key features, main advantages and disadvantages under the viewpoint of the LTE-A system. We also evaluate the performance of the most prominent current state-of-the-art algorithms by providing both qualitative and quantitative comparisons by using the Small Cell Forum evaluation methodology.

2 Results and Discussion

In this section, we briefly introduce the system model (section 3.1), one of our main theorems (section 3.2), and a key result (section 3.3) of our work in the area of Localization and Peer Discovery in Heterogeneous Wireless Networks with location-assistance.

2.1 System Model

We consider a fairly general HWN of M tiers, where each tier consists of WNEs that serve similar communication purposes and support the same RAT. The WNEs belonging to the m-th tier are referred to as tier-m WNEs ($m = 1, \ldots, M$). We consider that the tier-1 WNEs form a homogeneous PPP Φ_1 with intensity λ in the Euclidean plane, e.g. medium to long range base stations, and that, for $m > 1$, the tier-m WNEs are clustered around *some of* the tier-$(m-1)$ WNEs. We emphasize on around *some of* and *not all* tier-$(m-1)$ WNEs, since in practical deployments we do not expect a tier-m cluster around every tier-$(m-1)$ WNE. Let Φ_m denote the complete point process (PP) of tier-m WNEs, i.e. the union of all tier-m clusters. Given that a tier-m cluster is present around the tier-$(m-1)$ WNE $v_i \in \Phi_{(m-1)}$, we assume it to be in the form $N_{v_i}^m = N_i^m + v_i$, where the point sets N_i^m are independently and identically distributed (i.i.d) and independent of the parent PP Φ_{m-1}. All tier-m clusters are modeled by the Thomas cluster process as follows [11]: a) the number of points in each tier-m cluster is Poisson distributed with mean $\bar{c}_m$, and b) the WNEs in a tier-m cluster are scattered independently according to a symmetric normal distribution around the parent tier-$(m-1)$ WNE with variance σ_m^2.

We now turn our attention to the two WNEs of interest, coined as *source* and *target* peers. We consider that the source peer associates with a tier-m_s WNE, coined as the *associated WNE of the source peer*, and that the target peer associates with a tier-m_t WNE, coined as the *associated WNE of the target peer*. The associated WNEs of the two peers can belong to different tiers in the HWN. Accordingly, the two peers do not necessarily support the same RAT. We assume that the source and the target peers are located around their associated WNEs according to a symmetric normal distribution with variances σ_s^2 and

Table 1. Location Information Parameters (Spatial Information)

Parameter	**Notation**	**Comments**
Inter-site distance between the tier-1 parent WNEs of the peers	D	Can be estimated by performing TD or RSRP measurements between the tier-1 parent WNEs of the two peers.
Neighboring degree between the tier-1 parent WNEs of the peers	k	Can be estimated in a similar manner with D (lower accuracy is required).
Distance between the source peer and its associated WNE	R_s	Can be estimated by performing TD, ToA, RSS, or RF power level, either at the source peer or its associated WNE.
Angle between the source peer and its associated WNE	ξ_s	Can be estimated by performing AoA measurements or by employing other indirect estimation methodologies depending on the RAT.
Distance between the target peer and its associated WNE	R_t	Can be estimated in a similar manner with R_s.
Angle between the target peer and its associated WNE	ξ_t	Can be estimated in a similar manner with ξ_s.
Distance between the tier-m and the tier-*(m-1)* parent WNEs of the source peer	S_{m-1}	Can be estimated by performing TD, ToA, RSS, or RF, either at the tier-m parent WNE or at the tier-*(m-1)* parent WNE, depending on the RAT.
Angle between the tier-m and the tier-*(m-1)* parent WNEs of the source peer	θ_{m-1}	Can be estimated in a similar manner with ξ_s, depending on the RAT. It is assumed to be measured with respect to the reference direction from the tier-1 parent of the source peer to the tier-1 parent of the target peer.
Distance between the tier-m and the tier-*(m-1)* parent WNEs of the target peer	T_{m-1}	Can be estimated in a similar manner with S_{m-1}.
Angle between the tier-m and the tier-*(m-1)* parent WNEs of the target peer	ϕ_{m-1}	Can be estimated in a similar manner with ξ_s. Measured with respect to the reference direction from the tier-1 parent of the source peer to the tier-1 parent of the target peer.

σ_t^2, respectively. The locations of the two peers are assumed to be mutually independent and independent of the locations of other WNEs.

Since we are interested on analyzing how different levels of location-awareness affect the performance of localization and peer discovery in HWNs, we assume the presence of a location information server (LIS) that maintains some basic knowledge of the HWN layout. We assume that the LIS is aware of the clustering relations between the WNEs and capable of identifying the parent WNEs of both peers up to tier-1. For brevity, we refer to the tier-m WNE in the sequence of parent WNEs for the source peer as the *tier-m parent of the source peer* ($m < m_s$), and use a similar terminology for the parents of the target peer ($m < m_t$).

Aiming to capture the different levels of location-awareness that the LIS can provide to the peers, we consider it capable of utilizing spatial information on the relative distance and angle between two tagged WNEs of interest. In Table 1, we list the spatial information considered in this paper and provide insights on how they can be estimated in existing systems. In the sequel, we denote by $\mathcal{L}_s$ and $\mathcal{L}_t$ the set of parent WNEs of the source and the target peer, respectively, for which the LIS has knowledge of their relative polar coordinates with respect to their upper-tier parent WNEs. The remainder set of parent WNEs are denoted by $\bar{\mathcal{L}}_s$ and $\bar{\mathcal{L}}_t$, respectively. Fig. 1 depicts all parameters and random variables (RVs) involved in our analysis.

Since the performance of localization and peer discovery is tightly coupled with the definition of proximity between the WNEs of interest, we define the peer discovery probability as follows:

$$\mathcal{A}_{\mathcal{J}} = P\left[Z \leq \left(\frac{c}{Z_{th}}\right)^{\frac{1}{a}} \middle| \mathcal{J}\right], \tag{1}$$

where $\mathcal{J}$ denotes the available knowledge of the HWN topology, c is a scaling factor, a is a decay exponent, Z is the distance between the two peers, and Z_{th} is a fixed threshold that guarantees proximity between the two peers.

2.2 Main Result

Theorem 1. *The conditional pdf $f_{Z|D}(z)$ of the distance Z between the source and the target WNEs in a multi-tier clustered random HWN, given a) the distance D between their tier-1 parent WNEs and b) the relative polar coordinates of their parent WNEs in $\mathcal{L}_s$ and $\mathcal{L}_t$, is given by*

$$f_{Z|D}(z) = \frac{z}{\sigma^2} e^{-\frac{\eta_x^2+\eta_y^2+z^2}{2\sigma^2}} I_0\left[\frac{z\sqrt{\eta_x^2+\eta_y^2}}{\sigma^2}\right], \tag{2}$$

where $I_0[x]$ is the modified Bessel function and the parameters η_x, η_y, and σ are given by:

$$\eta_x = D + \sum_{j\in\mathcal{L}_s} S_j \cos\phi_j - \sum_{i\in\mathcal{L}_t} T_i \cos\theta_i, \tag{3}$$

$$\eta_y = \sum_{i \in \mathcal{L}_t} T_i \sin\theta_i - \sum_{j \in \mathcal{L}_s} S_j \sin\phi_j, \tag{4}$$

$$\sigma^2 = \sum_{j \in \bar{\mathcal{E}}_s} \sigma_j^2 + \sigma_s^2 + \sum_{i \in \bar{\mathcal{E}}_t} \sigma_i^2 + \sigma_t^2. \tag{5}$$

The corresponding ccdf $\bar{F}_{Z|D}(z)$ is given by

$$\bar{F}_{Z|D}(z) = Q_1\left[\frac{\sqrt{\eta_x^2 + \eta_y^2}}{\sigma}, \frac{z}{\sigma}\right], \tag{6}$$

where $Q_1[a,b]$ is the Marcum-Q function of the first order. If the relative polar coordinates (R_s, ξ_s) of the source peer and/or (R_t, ξ_t) of the target peer are also given, (2) and (6) hold with η_x, η_y, and σ, as given in the PhD dissertation.

Theorem 1 can be used to analytically evaluate the peer discovery probability between two WNEs (1), given any combination of spatial information that includes the distance D. The requirement of knowing D can be readily met in practical HWNs, where the locations of tier-1 WNEs typically remain fixed over time. The results in Theorem 1 not only allow heterogeneous WNEs to handle the uncertainty on their proximity, but also enable them to employ different levels of location-awareness upon localization or peer discovery depending on the available spatial information.

2.3 On the Impact of Angles between the WNEs

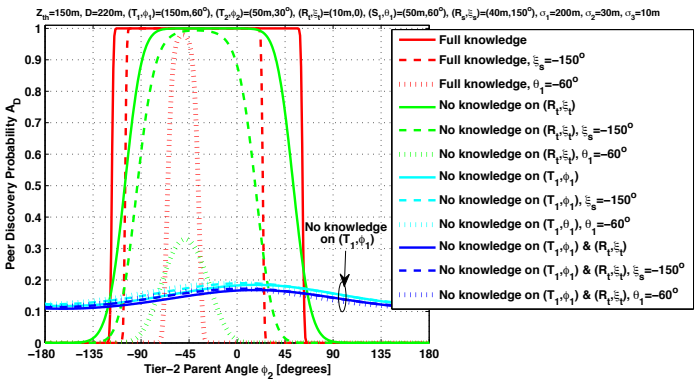


Fig. 2. Peer Discovery Probability given D vs. Tier-2 Parent Angle ϕ_2 [degrees]

The employment of accurate AoA measurements increases the complexity and processing requirements for the radio transceiver. With this in mind, in Fig.

2 we investigate the impact of the ϕ_2 angle between the tier-3 and the tier-2 parents of the target peer on the probability $\mathcal{A}_D$. As expected, when the LIS has full knowledge on the locations of the peers and their parent WNEs, the peer discovery can be either successful or not. Notably, there exists a ϕ_2 interval within which the peer discovery remains roughly unaffected. This interval is shown to be expanded, shifted, or compressed, in relation with the rest of the parameters governing the HWN topology. For example, if the angle ξ_t between the target peer and its parent tier-3 WNE is equal to -150^o (instead of 150^o), then the ϕ_2 interval for successful peer discovery is compressed and shifted to the left (red dashed line in Fig. 2. This effect is due the fact that for $\theta_2 = -150^o$ the two peers are separated by a higher distance. Even more evident is the compression of the ϕ_2 interval when the angle between the tier-1 and tier-2 parent WNEs of the source peer is set to $\theta_1 = -60^o$ instead of $(\theta_1 = 60^o)$ (red dotted line), since the distance between them is even higher. Such an effect is also expected in the nowadays HWN, where the distance between WNEs in higher tiers is (on the average) higher compared to the one between lower-tier WNEs.

Interestingly, a similar interval exists when the LIS is not aware of the relative coordinates of the target peer (green lines). Notice that the lack of such information prolongs the tail of the respective ϕ_2 interval with full knowledge in both directions. When the LIS has no knowledge of the coordinates (T_1, θ_1) of the tier-2 parent WNE, the probability $\mathcal{A}_D$ is shown to remain roughly unaffected by the values of ϕ_2 (blue and cyan lines). This relation indicates that the benefits from performing accurate measurements on the angles between low-tier WNEs are marginal when the relative coordinates of high-tier parent WNEs are not known to the LIS.

From the discussion above, we draw two important design guidelines. Firstly, the accurate estimation of the angle between low-tier WNEs and their parent WNEs is necessary only when an accurate estimation is required, e.g. the proximity threshold is low. Secondly, depending on the available spatial information, the low-tier WNEs can relax the accuracy of AoA measurements without significantly deteriorating the performance of peer discovery.

3 Conclusions

More and more WNEs are capable of estimating their radio-status as well as their relative position with respect to other nearby WNEs of the same technology. Integrating such spatial information from the ubiquitous WNEs of different RATs, is a key enabler for fine-grained localization and mobility management between the myriads of WNEs. In our work, we have investigated how knowledge of the radio status or the spatial distribution of the WNEs can be used to enhance the performance of localization, peer discovery, and association in the nowadays HWN. Our key contributions lie in the areas of localization and peer-discovery in HWN, device-to-device discovery in cellular network, energy-efficient handover decision in the macrocell - femtocell network, energy-efficient vertical handover decision in the cellular/Wi-Fi network, and mobility manage-

ment in the LTE-Advanced Network with Femtocells. We have provided both analytical and simulation means to evaluate the performance of the proposed frameworks, models, and algorithms.

References

1. H. S. Dhillon, R. K. Ganti, F. Baccelli, and J. G. Andrews, "Modeling and Analysis of K-tier Downlink Heterogeneous Cellular Networks," IEEE J. Sel. Areas Commun., vol. 30, no. 3, pp. 550-560, Mar. 2012.
2. IEEE Std 802.11-2012 (Revision of IEEE 802.11-2007), "Part 11: Wireless LAN Medium Access Control and Physical Layer Specifications", March 2012.
3. IEEE Std 802.15.4 (Revision of IEEE Std 802.15.4-2006), "Part 15.4: Low-Rate Wireless Personal Area Networks (LR-WPANs)", June 2011.
4. X. Lin, J.G. Andrews, A. Ghosh, R. Ratasuk, "An overview of 3GPP device-to-device proximity services", IEEE Commun. Mag., vol.52, no.4, pp.40-48, Apr. 2014.
5. Y. Yan, Y. Mostofi, "Co-Optimization of Communication and Motion Planning of a Robotic Operation under Resource Constraints and in Fading Environments", IEEE Trans. on Wirel. Commun., vol.12, no.4, pp.1562-1572, April 2013.
6. J. G. Andrews, S. Buzzi, W. Choi, S. Hanly, A. Lozano, A. Soong, J. C. Zhang, "What will 5G be?", to appear in IEEE J. on Sel. Areas in Commun., July 2014.
7. S. Marano, W. M. Gifford, H. Wymeersch, M. Z. Win, "NLOS identification and mitigation for localization based on UWB experimental data", IEEE J. Sel. Areas Commun., vol.28, no.7, pp.1026-1035, September 2010.
8. G. Garcia, S. Muppirisetty, E. Schiller, H. Wymeersch, "On the Trade-off Between Accuracy and Delay in Cooperative UWB Localization: Performance Bounds and Scaling Laws", IEEE Trans. on Wirel. Commun., accepted, 2014.
9. T. Kwon, J. Choi, "Spatial Performance Analysis and Design Principles for Wireless Peer Discovery", IEEE Trans. on Wirel. Commun., accepted, 2014.
10. Y. Zhong, W. Zhang, "Multi-Channel Hybrid Access Femtocells: A Stochastic Geometric Analysis", IEEE Trans. on Commun., Vol. 61, No. 7, pp. 3016-3026,2013.
11. R. K. Ganti, M. Haenggi, "Interference and Outage in Clustered Wireless Ad Hoc Networks", IEEE Trans. on Inform. Theory, vol.55, no.9, pp.4067-4086, Sept. 2009.
12. B. Kaufman, J. Lilleberg, and B. Aazhang, "Spectrum Sharing Scheme between Cellular Users and ad-hoc Device-to-Device Users", IEEE Trans. on Wirel. Comm., vol. 12, no. 3, pp. 10381049, Mar. 2013.
13. X. Lin, R. Ratasuk, A. Ghosh, and J. G. Andrews, "Modeling, Analysis and Optimization of Multicast Device-to-Device Transmission", submitted to IEEE Trans. on Wirel. Comm., Sep. 2013.
14. D. Xenakis, N. Passas, L. Merakos, and C. Verikoukis, "Mobility Management for Femtocells in LTE-Advanced: Key Aspects and Survey of Handover Decision Algorithms", IEEE Comm. Surv. & Tut., vol. 16, no. 1, pp. 64-91, 2014.
15. X. Yan, Y. A. Sekercioglu, S. Narayanan, "A survey of vertical handover decision algorithms in Fourth Generation heterogeneous wireless networks", Comp. Networks, Elsevier, col. 54, no. 11, pp. 1848-1863, Aug. 2010.
16. D. Xenakis, N. Passas, L. Merakos, and C. Verikoukis, "Energy-Efficient and Interference-Aware Handover Decision for the LTE-Advanced Femtocell Network", IEEE Intern. Commun. Conf. (ICC), June 2013.
17. D. Xenakis, N. Passas, L. Merakos, and C. Verikoukis, "ARCHON: An ANDSF-Assisted Energy-Efficient Vertical Handover Decision Algorithm for the Heterogeneous IEEE 802.11/LTE-Advanced Network", IEEE ICC, June 2014.