JOURNAL OF TELECOMMUNICATIONS AND INFORMATION TECHNOLOGY

2/2012

(*Contents Continued on Back Cover*)

ISSN 1509-4553 on-line: ISSN 1899-8852

Warsaw 2012

Circulation: 300 copies

Sowa – Druk na życzenie, www.sowadruk.pl, tel. 22 431-81-40

JOURNAL OF TELECOMMUNICATIONS AND INFORMATION TECHNOLOGY

Preface

We present to the readers of the *Journal of Telecommunications and Information Technology* a special issue that mostly contains the papers from last year conference on Decision Support for Telecommunications and Information Society, which concerned diverse mathematical and computational techniques for supporting solutions of numerous problems in telecommunication networks, related markets and technology.

Włodzimierz Ogryczak and Paweł Olender in the paper *On MILP Models for the OWA Optimization* consider the ordered weighted averaging (OWA) aggregation that uses the weights assigned to the ordered values (i.e., to the largest value, the second largest and so on) rather than to the specific coordinates. It allows to evaluate solutions impartially when distribution of outcomes is more important than assignments of these outcomes to the specific criteria. This applies, for example, to systems with multiple independent users or agents whose objectives correspond to the criteria. The authors concentrate on the most efficient computationally approaches to solving mixed integer linear programming OWA problems and they present several computational experiments on such problems.

Marco Caserta, Silvia Schwarze, and Stefan Voß in the paper *Developing a Ring-Based Optical Network Structure with Glass-Through Nodes* consider a security in optical transport networks (OTN), in particular the concept of 1+1 protection that requires a connection of each origin-destination(OD)-pair by at least two node-disjoint paths. On a ring topology of a network, 1+1 protection is given naturally. Moreover, the routing effort is typically decreasing on rings. These observations motivate the investigation of ring structures for OTN. While developing a ring structure for telecommunication networks, several subtasks can be identified. Rings have to be designed, OD-pairs have to be assigned to rings, communication among rings has to be defined, a proper flow routing has to be chosen, and rings have to be dimensioned regarding the flow capacity. The authors address the first two issues, namely generation of rings and assignment of OD-pairs to rings. The authors proposed an algorithm for random generation of candidate rings and a mathematical model for assigning OD-pairs to rings in such a manner that active nodes are chosen accordingly. Examples of applications and problems for future research are also discussed.

Matthias Fricke, Andrea Heckwolf, Ralf Herber, Ralf Nitsch, Silvia Schwarze, Stefan Voß, and Stefan Wevering in the paper *Requirements of 4G-Based Mobile Broadband on Future Transport Networks* consider new standards in mobile communications regarding available bandwidth resulting from future long term evolution (LTE) technologies. It is expected that users of one radio cell will share more than 100 Mbit/s in future. To take advantage of the full feature set of next generation mobile networks, transport network design has to face new requirements, caused by the architectural changes of LTE technologies. The newly defined X2 interface especially has an impact on the transport network requirements. X2 enables direct communication between evolved base stations (eNBs) and thus, enforces local solutions. At the same time, there is a tendency to locate network elements at fewer, central sites in order to reduce operational expenditure, in particular concerning the transport layer. This leads to the question of how the direct X2 connection of eNBs on the logical layer can be accommodated with a general centralization of transport networks. The authors show that a centralized transport network is able to realize the local meshing between eNBs for LTE. However, for LTE advanced, the standards currently discussed by the 3GPP initiative could lead to enhanced the requirements on the X2 interface latency.

Rita Girão-Silva, José Craveirinha, and João Clímaco in the paper *Hierarchical Multiobjective Routing Model in MPLS Networks with Two Service Classes – A Comparison Case Study* consider a two-level hierarchical multicriteria routing model for multiprotocol label switching networks with two service classes (QoS, i.e., with quality of service requirements, and best effort services) and alternative routing. A heuristic resolution approach where non-dominated solutions are obtained throughout the heuristic run and kept in an archive for further analysis is also reviewed. An extensive analysis of the application of this procedure to two reference test networks for various traffic matrices is presented. A comparison of the results of the proposed method with a lexicographic optimization approach based on a multicommodity flow formulation using virtual networks is carried out. Finally, the results of a stochastic discrete event simulation model developed for these networks are presented in order to illustrate the effectiveness of the resolution approach and to assess the inaccuracies of the analytic results.

Piotr Rzepakowski and Szymon Jaroszewicz in the paper *Uplift Modeling in Direct Marketing* consider a precise targeting of marketing actions that can potentially result in a greater return on investment. Usually, response models are used to select good targets. They aim at achieving high prediction accuracy for the probability of purchase, based on a sample of customers, to whom a pilot campaign has been sent. However, to separate the impact of the action from other stimuli and spontaneous purchases, we should model not the response probabilities themselves, but instead, the change in those probabilities caused by the action. The problem of predicting this change is known as uplift modeling, differential response analysis, or true lift modeling. In the paper, tree-based classifiers designed for uplift modeling are applied to real marketing data and compared with traditional response models, and other uplift modeling techniques described in literature. Computational experiments show that the proposed approaches outperform existing uplift modeling algorithms and demonstrate significant advantages of uplift modeling over traditional response based targeting.

Andrzej Karbowski in the paper *Integrated Routing and Network Flow Control Embracing Two Layers of TCP/IP Networks – Methodological Issues* considers a cross-layer network optimization problem that involves network and transport layers, treating both routing and flows as decision variables. Due to the non-convexity of the capacity constraints when using Lagrangian relaxation method, a duality gap causes numerical instability. It is shown that the rescue preserving separability of the problem may be the application of the augmented Lagrangian method, together with Cohen's Auxiliary Problem Principle.

Kamil Kołtyś, Krzysztof Pieńkosz, and Eugeniusz Toczyłowski in the paper *Auction Models Supporting End-to-End Connection Trading* consider bandwidth allocation problem on the telecommunication market where there are many sellers and buyers. Sellers offer the bandwidth of telecommunication links. Buyers are interested in the purchase of the bandwidth of several links that makes up an end-to-end connection between two nodes of telecommunication network. The paper analyzes three auction models supporting such a bandwidth exchange: NSP (network second price), BCBT (model for balancing communication bandwidth trading) and BCBT-CG which is a modification of BCBT that applies column generation technique. All of these models concern divisible network resources, treat bandwidth of telecommunication links as an elementary commodity offered for sale, and allow for purchasing bandwidth along multiple paths joining two telecommunication nodes. All of them also aim at maximizing the social welfare. Considered auction models have been compared in the respect of economic and computational efficiency. Experimental studies have been performed on several test instances based on the SNDlib library data sets.

Piotr Arabas, Przemysław Jaskóła, Mariusz Kamola, and Michał Karpowicz in the paper *Analysis and Modeling of Domain Registration Process* consider the domain name reservation process for the polish `.pl` domain. Two models of various time scale are constructed and finally combined to build a long range high resolution model. The results of the prediction are verified by using real data.

Anna Felkner and Adam Kozakiewicz in the paper *RT_{+}^{T} – Time Validity Constraints in RT^{T} Language* maintain that most of the traditional access control models, such as mandatory, discretionary, and role based access control make authorization decisions based on the identity, or the role of the requester who must be known to the resource owner. Thus, they may be suitable for centralized systems but not for decentralized environments where the requester and service provider or resource owner are often unknown to each other. To overcome the shortcomings of traditional access control models, trust management models might use three different semantics (set-theoretic, operational, and logic programming) of RT^{T} a language from the family of role-based trust management languages (RT). RT^{T} is used for representing security policies and credentials in decentralized distributed access to control systems. A credential provides information about the privileges of users and the security policies issued by one or more trusted authorities. The core part of the paper is the introduction of time validity constraints to show how to make RT^{T} language more realistic. The new language, named RT_{+}^{T}, takes time validity constraints into account. The semantics for RT_{+}^{T} language is also shown. Inference system is introduced not just for specific moment, but also for time intervals. It evaluates maximal time validity when it is possible to derive the credential from the set of available credentials. The soundness and completeness of the inference systems with time validity constraints with respect to the set-theoretic semantics of RT_{+}^{T} language is proven.

Marcin Mincer and Ewa Niewiadomska-Szynkiewicz in the paper *Application of Social Network Analysis to the Investigation of Interpersonal Connections* present an application of social network analysis (SNA) to the investigation and analysis of social relationships of people. This application concerns data mining in the case of two social networks: Facebook and Twitter. The presented simulations illustrate how social analysis can be used to determine the interpersonal connections, importance of actors in a given social network and detect communities of people. The strengths and weaknesses of SNA techniques are discussed.

Jarosław Hurkała and Adam Hurkała in the paper *Effective Design of the Simulated Annealing Algorithm for the Flowshop Problem with Minimum Makespan Criterion* address the *n*-job, *m*-machine flowshop scheduling problem with minimum completion time (makespan) as the performance criterion. They describe an efficient design of the Simulated Annealing algorithm for solving approximately this NP-hard problem. The main difficulty in implementing the algorithm is no apparent analogy for the temperature as a parameter in the flowshop combinatorial problem. Moreover, the quality of solutions is dependent on the choice of cooling scheme, initial temperature, number of iterations, and the temperature decrease rate at each step as the annealing proceeds. The authors propose how to choose the values of all the aforementioned parameters, as well as the Boltzmann factor for the Metropolis scheme. Three perturbation techniques are tested and their impact on the solutions quality is analyzed. A heuristic and randomly generated solutions as initial seeds to the annealing optimization process are compared. Computational experiments indicate that the proposed design provides very good results – the quality of solutions of the Simulated Annealing algorithm is favorably compared with two different heuristics.

Rafał Kasprzyk in the paper *Diffusion in Networks* considers a concept of a method and its application to examine the dynamics of diffusion processes in networks. The proposed method was used as a core framework for the CARE (Creative Application to Remedy Epidemics) system.

Kamil Staszek, Jacek Kołodziej, Krzysztof Wincza, and Sławomir Gruszczyński in the paper *Compact Broadband Rat-Race Coupler in Multilayer Technology Designed with the Use of Artificial Right- and Left-Handed Transmission Lines* consider a compact broadband rat-race coupler for the first time designed and realized in a multilayer microstrip technology. To achieve both broad operational bandwidth and a compact size the 270° transmission line of a conventional rat-race, coupler has been replaced by a –90° left-handed transmission line realized with the use of a quasi-lumped element technique. Moreover, to achieve better compactness of the resulting coupler, all 90° right-handed transmission lines have been realized with the use of the same technique. It has been also proved that simple LC approximation of a left-handed transmission line can be successfully used for the design. Moreover, it has been shown that when appropriately chosen, the multilayer dielectric structure allows for realization of structures designed with the use of this simple approximation, for both right-handed and left-handed transmission lines, without losing too much of performance.

Andrzej P. Wierzbicki
Guest Editor

Paper

On MILP Models for the OWA Optimization

Włodzimierz Ogryczak[a] and Paweł Olender[a,b]

[a] *Institute of Control and Computation Engineering, Warsaw University of Technology, Warsaw, Poland*
[b] *National Institute of Telecommunications, Warsaw, Poland*

Abstract—The problem of aggregating multiple outcomes to form overall objective functions is of considerable importance in many applications. The ordered weighted averaging (OWA) aggregation uses the weights assigned to the ordered values (i.e., to the largest value, the second largest and so on) rather than to the specific coordinates. It allows to evaluate solutions impartially, when distribution of outcomes is more important than assignments these outcomes to the specific criteria. This applies to systems with multiple independent users or agents, whose objectives correspond to the criteria. The ordering operator causes that the OWA optimization problem is nonlinear. Several MILP models have been developed for the OWA optimization. They are built with different numbers of binary variables and auxiliary constraints. In this paper we analyze and compare computational performances of the different MILP model formulations.

Keywords**—*****location problem, mixed integer (linear) programming, multiple criteria, ordered weighted averaging (OWA).***

1. Introduction

Yager [1] introduced the so-called ordered weighted averaging (OWA) operator providing a parameterized family of aggregations that include the maximum, the minimum and the average criteria as special cases. Since its introduction, the OWA aggregation has been applied to many fields [2], including telecommunications [3], [4] and location analysis [5] among others.

In the OWA aggregation the weights are assigned to the ordered values (i.e., to the largest value, the second largest and so on) rather than to the specific coordinates. For a given weights vector $\mathbf{w} = (w_1, w_2, \ldots, w_m)$, $w_i \geq 0$ for $i = 1, 2, \ldots, m$, the OWA aggregation of an m-dimensional vector $\mathbf{x} = (x_1, x_2, \ldots, x_m)$ can be mathematically defined as follows. We introduce the ordering map $\Theta : R^m \to R^m$ such that $\Theta(\mathbf{x}) = (\theta_1(\mathbf{x}), \theta_2(\mathbf{x}), \ldots, \theta_m(\mathbf{x}))$, where $\theta_1(\mathbf{x}) \geq \theta_2(\mathbf{x}) \geq \ldots \geq \theta_m(\mathbf{x})$ and there exists a permutation τ of set $I = \{1, 2, \ldots, m\}$ such that $\theta_i(\mathbf{x}) = x_{\tau(i)}$ for $i = 1, 2, \ldots, m$. Further, we apply the weighted sum aggregation to ordered vectors $\Theta(\mathbf{x})$, i.e., the OWA aggregation function has the form:

$$a_{\mathbf{w}}(\mathbf{x}) = \sum_{i=1}^{m} w_i \theta_i(\mathbf{x}). \tag{1}$$

Note that formula (1) differs from that originally introduced by Yager [1], due to not necessarily normalized weights ($\sum_{i=1}^{m} w_i = 1$ in [1]).

When applying the OWA aggregation as an optimization criterion we get

$$\min \left\{ \sum_{i=1}^{m} w_i \theta_i(\mathbf{x}) : \mathbf{x} \in Q \right\}. \tag{2}$$

In this paper we analyze mathematical programming models for problem (2) with nonnegative weights ($w_i \geq 0$). The ordering operator Θ causes that the OWA optimization problem (2) is nonlinear even for the case of linear programming (LP) form of the original constraints. Yager [6] has shown that the nature of the nonlinearity introduced by the ordering operations allows us to convert the optimization (2) into a mixed integer programming problem. Ogryczak and Śliwiński [7] have shown that the OWA optimization with the monotonic weights can be formed as a standard linear program of higher dimension. Several models have been proposed for locations problems with the OWA criterion (the so-called ordered median problems) [8], [9], but their computational performance have never been compared. We have carried out such comparison and additionally compared it with linear formulation for specific cases.

The paper is organized as follows. In the next section various models with the OWA criterion are presented. Within these models different formulations are considered regarding the number of constraints. In Section 3 the experiment procedure is presented and obtained results with computational models efficiency are discussed.

2. Model Formulations

As usually the whole model can be divided into two parts: physical and preference model. The physical model is based on the discrete facility location problem. This problem can be formulated as mixed integer linear programming. The OWA operator constitutes the preference model. In general, it can also be formulated as mixed integer linear programming. However, as mentioned earlier, in specific cases it is possible to form it as linear programming. Certainly, the whole model will remain mixed integer due to underlying location problem. In the article we are focusing on the OWA optimization so we are referring to mixed integer and linear programming meaning the preference model only.

2.1. Location Problem

A standard formulation of facility location problem without the capacity limits was used. There is given a set of m sites (e.g., clients). We have to place n facilities to satisfy demands from the clients. Without loss of generality it can be assumed that the number of candidate sites is identical to the number of clients and additionally that $n \leq m$. Then each client is assigned to the facility that meets its demand. The assignment is done in such a way to optimize a given objective function. The objective function is usually based on distances (costs) between the clients and the facilities. Because we consider unlimited capacities each client is assigned the closest facility. Formally the model can be expressed in the following form:

$$\sum_{j=1}^{m} u_j = n, \tag{3}$$

$$\sum_{j=1}^{m} v_{ij} = 1 \quad \text{for} \quad i = 1,\ldots,m, \tag{4}$$

$$v_{ij} \leq u_j \quad \text{for} \quad i,j = 1,\ldots,m, \tag{5}$$

$$x_i = \sum_{j=1}^{m} c_{ij} v_{ij} \quad \text{for} \quad i = 1,\ldots,m, \tag{6}$$

$$u_j \in \{0,1\} \quad \text{for} \quad j = 1,\ldots,m, \tag{7}$$

$$v_{ij} \geq 0 \quad \text{for} \quad i,j = 1,\ldots,m, \tag{8}$$

where c_{ij} denotes the cost of satisfying the total demand of client i by facility j. There are used two groups of binary variables representing, respectively, the location and the allocation decisions:

- u_j – equal 1 if a facility is built at site j and 0 otherwise,
- v_{ij} – equal 1 if the demand of client i is satisfied by facility j and 0 otherwise.

The auxiliary variable x_i (6) expresses the cost of satisfying the demand of client i. The constraint (3) enforces that exactly n facilities are placed. The fact that each client is assigned to only one facility is modeled with constraint (4). Constraint (5) ensures that the clients are assigned to the existing facilities. Thus, above formulation defines a set of attainable values Q and the corresponding cost (outcome) vectors $\mathbf{x}$. On this basis preference models with OWA criterion can be defined.

2.2. The First MILP Model (M1)

The ordering operator Θ causes that the OWA optimization problem (2) is nonlinear, however, the nonlinearity can be transformed into discrete problem. Note that the quantity $\theta_1(\mathbf{x})$ representing the worst outcome can be easily computed directly by the LP minimization:

$$\theta_1(\mathbf{x}) = \min y_1 \tag{9}$$

subject to

$$y_1 \geq x_i \quad \text{for} \quad i = 1,\ldots,m. \tag{10}$$

Following Yager [6], similar formula can be given for any $\theta_k(\mathbf{x})$, although requiring the use of integer variables. Namely, for any $k = 1,2,\ldots,m$ the following formula is valid [7]:

$$\theta_k(\mathbf{x}) = \min y_k \tag{11}$$

$$\text{s.t.} \quad y_k + M z_{ki} \geq x_i \quad \text{for} \quad i = 1,\ldots,m, \tag{12}$$

$$\sum_{i=1}^{m} z_{ki} \leq k-1, \tag{13}$$

$$z_{ki} \in \{0,1\} \quad \text{for} \quad i = 1,\ldots,m, \tag{14}$$

where M is a sufficiently large constant (larger than any possible difference between various individual outcomes y_i). Note that for $k = 1$ all the binary variables z_{1i} are enforced to 0 thus reducing the optimization to the standard LP model for that case.

The entire OWA optimization model (2) can be formulated as the following mixed integer linear programming problem (MILP) [7]:

$$\min \sum_{k=1}^{m} w_k y_k, \tag{15}$$

$$y_k + M z_{ki} \geq x_i \quad \text{for} \quad i,k = 1,\ldots,m, \tag{16}$$

$$\sum_{i=1}^{m} z_{ki} \leq k-1 \quad \text{for} \quad k = 1,\ldots,m, \tag{17}$$

$$z_{ki} \in \{0,1\} \quad \text{for} \quad i,k = 1,\ldots,m, \tag{18}$$

$$\mathbf{x} \in Q. \tag{19}$$

This MILP model introduces $O(m^2)$ binary variables z_{ki} organized in m multiple choice constraints (special ordered sets) and m continuous variables y_k defined by the corresponding m inequalities. Actually, the original model introduced in [6] contains additional constraints

$$y_k \geq y_{k+1} \quad \text{for} \quad k = 1,\ldots,m-1, \tag{20}$$

representing ordering inequalities on variables y_k. Due to minimization with nonnegative weights w_k, these inequalities are redundant in the sense that they do not affect the optimal solution. However, we will examine if they influence the computational performance of the model. Additionally, we will also consider another redundant constraint,

$$\sum_{k=1}^{m} y_k = \sum_{i=1}^{m} x_i, \tag{21}$$

which balances the total sum of coordinates of basic cost vector $\mathbf{x}$ against sorted vector $\mathbf{y}$.

Eventually, we take into consideration three different formulations of model M1:

- M1_1 – formulation (15)–(19) without the redundant constraints,
- M1_2 – formulation (15)–(19) with one redundant constraint (20),
- M1_3 – formulation (15)–(19) with two redundant constraints (20) and (21).

2.3. The Second MILP Model (M2)

Several MILP models have been developed for the OWA optimization within the ordered median location problems [5]. Starting from quadratic MIP, through MILP models with $O(m^3)$ binary variables and finally setting the MILP model with $O(m^2)$ binary variables and constraints [8]. Adapting the model to our notation, it can be equivalently written as follows:

$$\min \sum_{k=1}^{m} w_k y_k , \tag{22}$$

$$y_k \geq y_{k+1} \quad \text{for} \quad k = 1,\dots,m-1, \tag{23}$$

$$y_k + M(1 - s_{ki}) \geq x_i \quad \text{for} \quad i,k = 1,\dots,m, \tag{24}$$

$$\sum_{i=1}^{m} s_{ki} = 1 \quad \text{for} \quad k = 1,\dots,m, \tag{25}$$

$$\sum_{k=1}^{m} s_{ki} = 1 \quad \text{for} \quad i = 1,\dots,m, \tag{26}$$

$$s_{ki} \in \{0,1\} \quad \text{for} \quad i,k = 1,\dots,m, \tag{27}$$

$$\sum_{k=1}^{m} y_k = \sum_{i=1}^{m} x_i, \tag{28}$$

$$\mathbf{x} \in Q, \tag{29}$$

where m^2 binary variables s_{ki} represent assignment of actual values x_i to the ordered ones y_k. That means, $s_{ki} = 1$ if the value of x_i is the k-th largest and zero otherwise. This model is based on a combination of assignment and sorting problems. The sorting part is realized by constraints (23), (25) and (26). For such a modeling approach the $m-1$ inequalities ordering variables y_k are necessary. On the other hand, due to minimization with the nonnegative weights w_k, the equation balancing vector $\mathbf{x}$ and $\mathbf{y}$ is redundant. Thus the model can be then considered without this (single) equation (28). Although, the full model containing the balance equation is applicable both for minimization and maximization cases. For these reasons we have analyzed two formulations of this model:

- M2_1 – formulation (22)–(29) with the redundant constraint (28),
- M2_2 – formulation (22)–(27), (29) without the redundant constraint.

2.4. LP Model

The ordering operator Θ used in the OWA aggregation is nonlinear and, in general, it is hard to implement. However, with decreasing weights the OWA aggregation is a piecewise linear convex function and its minimization can be expressed in the linear programming form [7]. This so-called deviational model is based on the linear programming representation of the cumulated ordered outcomes:

$$\overline{\theta}_k(\mathbf{x}) = \sum_{i=1}^{k} \theta_i(\mathbf{x}) \quad \text{for} \quad k = 1,\dots,m. \tag{30}$$

The quantities $\overline{\theta}_k(\mathbf{x})$ for $k = 1,\dots,m$ express, respectively: the worst (largest) outcome, the total of the two worst outcomes, the total of the three worst outcomes, etc. As it was shown in [7] each of these values can be found as the optimal value of the following LP problem:

$$\overline{\theta}_k(\mathbf{x}) = \min \left(kt_k + \sum_{i=1}^{m} d_{ik} \right) \tag{31}$$

subject to

$$d_{ik} \geq x_i - t_k, d_{ik} \geq 0 \quad \text{for } i = 1,\dots,m. \tag{32}$$

The ordered outcomes can be expressed as differences $\theta_k(\mathbf{x}) = \overline{\theta}_k(\mathbf{x}) - \overline{\theta}_{k-1}(\mathbf{x})$ for $k = 2,\dots,m$ and $\theta_1(\mathbf{x}) = \overline{\theta}_1(\mathbf{x})$. Hence, the OWA problem with weights w_k can be expressed in the form:

$$\min \sum_{k=1}^{m} (w_k - w_{k+1}) \left(kt_k + \sum_{i=1}^{m} d_{ik} \right) \tag{33}$$

$$d_{ik} \geq x_i - t_k \quad \text{for} \quad i,k = 1,\dots,m, \tag{34}$$

$$d_{ik} \geq 0 \quad \text{for} \quad i,k = 1,\dots,m, \tag{35}$$

$$\mathbf{x} \in Q. \tag{36}$$

For this model we also consider some redundant constraints. One of them represents ordering inequalities on variables t_k, thus taking the form:

$$t_k \geq t_{k+1} \quad \text{for} \quad k = 1,\dots,m-1. \tag{37}$$

The second constraint concerns ordering of deviations d_{ik} but in reverse order:

$$d_{ik} \leq d_{ik+1} \quad \text{for} \quad i = 1,\dots,m, \; k = 1,\dots,m-1. \tag{38}$$

The last redundant constraint is a relaxed form of the previous one:

$$\sum_{i=1}^{m} d_{ik} \leq \sum_{i=1}^{m} d_{ik+1} \quad \text{for} \quad k = 1,\dots,m-1. \tag{39}$$

We carry out computational analyzes of four following formulations:

- MLP1 – formulation (33)–(36) without the redundant constraints,
- MLP2 – formulation (33)–(36) with the redundant constraint (37),
- MLP3 – formulation (33)–(36) with the redundant constraint (38),
- MLP4 – formulation (33)–(36) with the redundant constraint (39).

3. Computational Tests

In order to analyze the computational efficiency of the presented models and their different formulations, we have applied them to various location problems and compared time needed to solve these tasks for specific formulations. The experiment procedure, including problem generation, is explained below. Next, results are presented and models comparison are discussed.

3.1. Experiments Design

The general scheme of experiments is analogous to that presented in [10]. To evaluate the models on different cases, basic parameters characterizing the location problem have been chosen and their sets of considered value were determined. Then, based on combinations of these parameters various instances of problem location have been defined. The parameters that have been considered are: the number of sites (locations), the number of service points to be placed and type of problem defined by the vector of weights in the OWA aggregation.

The number of sites is very important parameter because, in fact, it determines the size of the problem. We have considered smaller sizes for the mixed integer formulations, and bigger for the linear model:

- for the MILP models – SC1: $m = 8$, SC2: $m = 10$, SC3*: $m = 12$, SC4*: $m = 15$,
- additionally for the LP model – SC5: $m = 20$, SC6: $m = 25$, SC7: $m = 30$.

In cases SC3 and SC4 (marked by an asterisk) only one MILP formulation has been tested (for problems with monotonic weights) in order to compare it with the linear formulation.

The second parameter, the number of facilities, has been defined as proportional to the problem size (m value). Following cases have been assumed: $n = \lceil \frac{m}{4} \rceil$, $n = \lceil \frac{m}{3} \rceil$, $n = \lceil \frac{m}{2} \rceil$, $n = \lceil \frac{m}{2} + 1 \rceil$, where $\lceil a \rceil$ is the smallest integer value not smaller than a.

Type of problem defined by the vector of weights $\mathbf{w}$ plays an important role. It allows to represent a wide range of problems (strictly speaking the preferences), which is directly connected with a problem structure and thus with problem complexity. We have examined 12 problem types:

- TC1: N-median, i.e. $\mathbf{w} = (\underbrace{1, \ldots, 1}_{m})$,
- TC2: N-center problem, i.e. $\mathbf{w} = (1, \underbrace{0, \ldots, 0}_{m-1})$,
- TC3: k-centra problem, i.e. $\mathbf{w} = (\underbrace{1, \ldots, 1}_{k}, 0, \ldots, 0)$, where $k = \lfloor \frac{m}{3} \rfloor$,
- TC4: $k_1 + k_2$-trimmed mean problem, i.e., $\mathbf{w} = (\underbrace{0, \ldots, 0}_{k_1}, 1, \ldots, 1, \underbrace{0, \ldots, 0}_{k_2})$, where $k_1 = \lceil \frac{m}{10} \rceil$, and $k_2 = \lceil n + \frac{m}{10} \rceil$,
- TC5: $\mathbf{w}$ with binary entries alternating 0 and 1, and beginning with 1, i.e. $\mathbf{w} = (1, 0, 1, 0, 1, 0, \ldots)$,
- TC6: Such as TC5, but beginning with 0, i.e. $\mathbf{w} = (0, 1, 0, 1, 0, 1, \ldots)$,
- TC7: The repetition of the sequence $(1, 1, 0)$, i.e. $\mathbf{w} = (1, 1, 0, 1, 1, 0, \ldots)$,
- TC8: The repetition of the sequence $(1, 0, 0)$, i.e. $\mathbf{w} = (1, 0, 0, 1, 0, 0, \ldots)$,
- TC9: Beginning with m (size of the problem) and decreasing by 1, i.e., $\mathbf{w} = (m, m-1, \ldots, 2, 1)$,
- TC10: Such as TC9, but in reverse order (increasing), i.e., $\mathbf{w} = (1, 2, \ldots, m-1, m)$,
- TC11: Beginning with $3m$ and decreasing in a piecewise linear manner, k weights by 3, next k weights by 2 and rest by 1, i.e.,
$$\begin{aligned} \mathbf{w} = (3m, &\underbrace{3(m-1), \ldots, 3(m-k)}_{k},\\ &\underbrace{3(m-k)-2, \ldots, 3(m-k)-2k}_{k},\\ &3m-5k-1, 3m-5k-2, \ldots), \end{aligned}$$
where $k = \lfloor \frac{m}{3} \rfloor$,
- TC12: Such as TC11, but in reverse order (increasing), i.e.,
$$\begin{aligned} \mathbf{w} \;=\; (\;\ldots\; &, 3m-5k-2, 3m-5k-1,\\ &\underbrace{3(m-k)-2k, \ldots, 3(m-k)-2}_{k},\\ &\underbrace{3(m-k), \ldots, 3(m-1)}_{k}, 3m), \end{aligned}$$
where $k = \lfloor \frac{m}{3} \rfloor$.

The first two of the eight problems (TC1–TC8) are basic problems in the location theory [10]. The next two are not so popular but also used in this field. Problems TC5–TC8 are in some sense artificial and have been used particularly to test the computational efficiency. The last four problems have monotonic weights. Depending on the type of monotonicity, they are simpler (TC9, TC11 with decreasing weights) or harder (TC10, TC12 with increasing weights) problems. These types of the problems can be treated as extended versions of max min (TC9, TC11) and max max (TC10, TC12) objective functions, respectively.

For each size case we have generated 15 cost matrices, which have zero on the main diagonal and the remaining entries randomly generated from a discrete uniform distribution in the interval $[1, 100]$. These matrices have been assigned to each combination of the parameters with corresponding problem size. Thus, we have received a set of test problem instances.

3.2. Results

The efficiency comparison has been carried out based on the average computational time needed to solve a problem. We have compared computational time for specific sizes and problem types averaging over instances of cost matrices and cases of facilities number to be placed. The complete results for each model are presented in Tables 1–3.

First, we have examined the influence of redundant constraint on the computational efficiency of MILP models. In order to check the change of the performance we have compared different formulations within the individual models, which were presented in Subsections 2.2 and 2.3.

Table 1
Average solution time for MILP model M1

Formulation		M1_1	M1_2	M1_3
SC1	TC1	0.291	0.222	0.025
	TC2	0.011	0.014	0.034
	TC3	0.033	0.055	0.055
	TC4	0.115	0.152	0.186
	TC5	0.081	0.107	0.051
	TC6	0.094	0.126	0.163
	TC7	0.129	0.148	0.055
	TC8	0.049	0.073	0.056
	TC9	0.216	0.204	0.068
	TC10	0.500	0.411	0.145
	TC11	0.241	0.221	0.053
	TC12	0.302	0.241	0.052
SC2	TC1	1.728	1.378	0.047
	TC2	0.020	0.022	0.045
	TC3	0.126	0.225	0.141
	TC4	1.311	1.315	1.537
	TC5	0.289	0.456	0.173
	TC6	0.483	0.665	0.923
	TC7	0.512	0.699	0.135
	TC8	0.150	0.258	0.140
	TC9	1.143	0.969	0.185
	TC10	12.721	2.820	1.379
	TC11	1.217	1.002	0.153
	TC12	2.909	1.784	0.149
SC3	TC9	–	–	0.708
	TC11	–	–	0.543
SC4	TC9	–	–	2.515
	TC11	–	–	1.760

The results for the first model are given in Table 1. For better illustration of the differences we present it graphically for SC1: $m = 8$ in Fig. 1 (for SC2 the results are similar). One may notice that adding the redundant constraints leads to significant time reduction for the problems with all non-zero weights (TC1, TC9–TC12). The situation is different in the case of problems that are focused on minimizing larger values of the outcome vector components (TC2–TC4), where the time slightly increases. For the other problems (TC5–TC8) it is hard to define a clear trend of change. In particular, comparing the results for TC5 and TC6, for which weight vectors are alternating sequences of 0 and 1 (where in TC5 sequence begins with 1, and in TC6 with 0), one may notice a significant difference in the time change due to adding the redundant constraints. The same situation is for the second model, which can be

Table 2
Average solution time for MILP model M2

Formulation		M2_1	M2_2
SC1	TC1	0.120	6.553
	TC2	1.697	0.089
	TC3	2.317	0.176
	TC4	4.848	0.697
	TC5	0.742	1.806
	TC6	5.207	2.025
	TC7	0.561	3.478
	TC8	0.978	0.808
	TC9	1.045	4.988
	TC10	1.843	6.820
	TC11	0.574	5.481
	TC12	0.279	7.038
SC2	TC1	0.478	177.263
	TC2	14.421	0.426
	TC3	89.987	2.397
	TC4	225.014	6.421
	TC5	8.289	17.388
	TC6	192.945	31.407
	TC7	5.987	37.549
	TC8	16.010	6.949
	TC9	13.095	117.014
	TC10	63.474	197.318
	TC11	6.506	133.892
	TC12	1.993	222.029

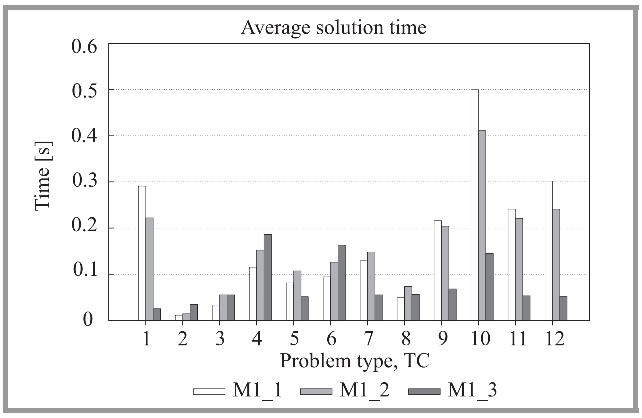


Fig. 1. Model M1 formulations comparison ($m = 8$).

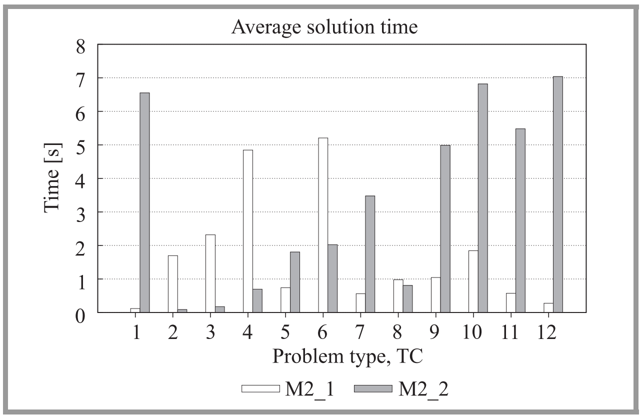


Fig. 2. Model M2 formulations comparison ($m = 8$).

seen in Table 2 and Fig. 2. The relationships discussed above are even more apparent here. In particular, there is a greater time increase after adding the redundant constraints in the case of problems TC2–TC4.

Next we have juxtaposed the results of model M1 with the results of model M2. For this purpose the corresponding formulations of these models have been confronted, namely the formulation with and without redundant constraints. In the former case the formulation M1_3 is compared with the formulation M2_1 (Fig. 3). As seen for all types of problems, the first model shows much better performance than the second one. There is a similar situation for the formulations without the redundant constraints, where the formulation M1_1 has been compared with the formulation M2_2 (Fig. 4). Here, also the solution time for the first model turns out to be much shorter than that for its counterpart for all types of problems. The scale of the differences are especially noteworthy and reaches one or even two (three for SC2) orders of magnitude. On this basis, model M1 seems to be more efficient than model M2.

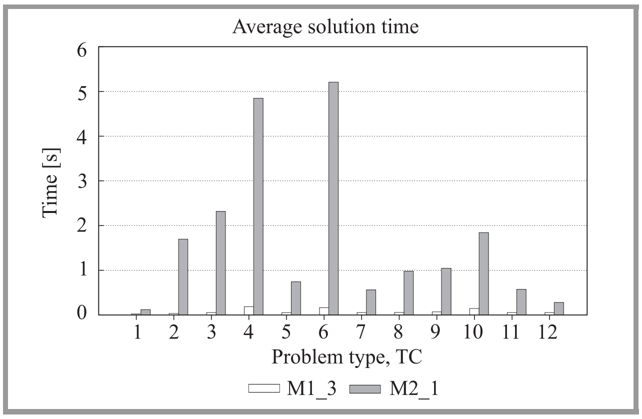


Fig. 3. The comparison of formulations with the redundant constraints ($m = 8$).

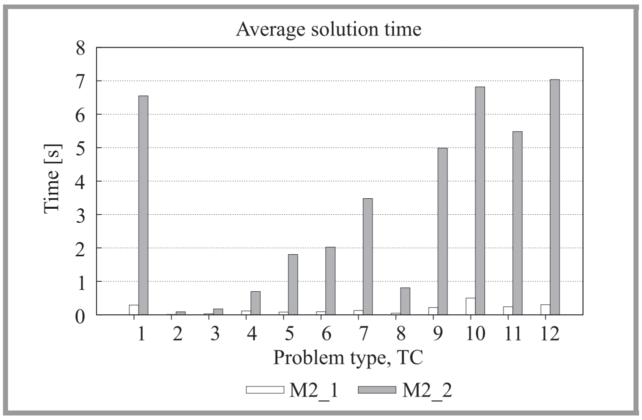


Fig. 4. The comparison of formulations without the redundant constraints ($m = 8$).

As mentioned earlier, the specific problems with appropriately monotonic (decreasing in the case of minimization) OWA weights can be formulated as the standard linear programming models. We have examined whether MILP models could also take advantage of this special structure. For this purpose we have compared their computational time for the problems with decreasing weights (easier problems – TC9, TC11) and increasing weights (harder problems – TC10, TC12).

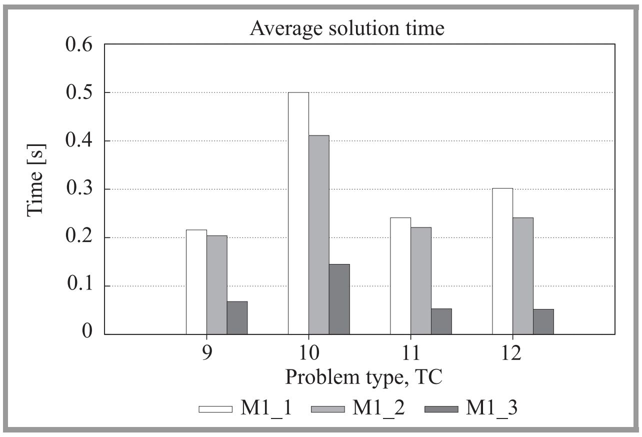


Fig. 5. Model M1 comparison with monotonic weights ($m = 8$).

The analysis shows that model M1 (Fig. 5), in the case of problems TC9 and TC10, demonstrates actually much better solution time for decreasing weights (TC9). In the case of problems TC11 and TC12, the differences are not so significant, and for the formulation M1_3 the solution times can be considered equal.

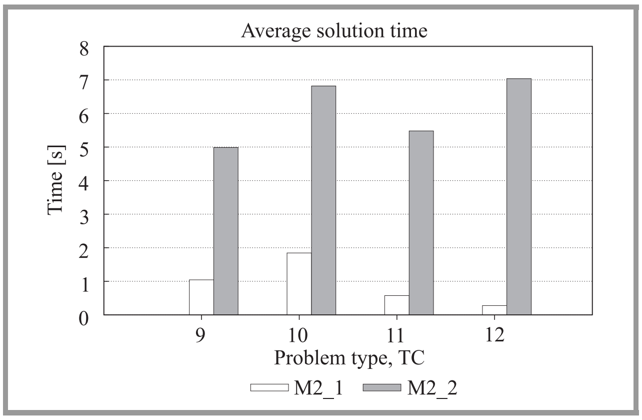


Fig. 6. Model M2 comparison with monotonic weights ($m = 8$).

Considering model M2 (Fig. 6) it can be seen that the formulation without the redundant constraints (M2_2) has also slightly shorter computational time for decreasing weights. The same is true for the formulation M2_1, when comparing TC9 with TC10 problem. However, the situation looks differently for the formulation M2_1 for TC11 and TC12 problem. Here, the problem with increasing weights has a shorter solution time. This suggests that the MILP models do not guarantee better performance for the problems with decreasing weights.

Because of the above results obtained for MILP models for monotonic weights we have carried out the direct comparison between the mixed integer and the linear (Subsection 2.4) OWA formulations. The results for the linear OWA formulations are given in Table 3. We have compared the most efficient (in the sense of considered problem types) MILP formulation (M1_3) and basic formulation of linear model (MLP1) for the problems with decreasing weights (TC9, TC11). As shown in the graphical comparison (Figs. 7 and 8), even the best considered mixed integer programming model has much worse performance than the linear formulation of OWA. The differences reach two orders of magnitude. Therefore, even if sometimes MILP models solve the OWA optimization with appropriate monotonic weights (simpler problems) more effectively than the general case OWA, they are still much less efficient than the linear OWA formulation for these specific cases.

Table 3
Average solution time for LP model

Formulation		MLP1	MLP2	MLP3	MLP4
SC3	TC1	0.008	0.007	0.008	0.008
	TC2	0.033	0.032	0.033	0.031
	TC3	0.042	0.041	0.042	0.042
	TC9	0.047	0.048	0.054	0.054
	TC11	0.036	0.036	0.042	0.041
SC4	TC1	0.012	0.012	0.013	0.012
	TC2	0.058	0.056	0.057	0.057
	TC3	0.067	0.064	0.065	0.067
	TC9	0.076	0.075	0.090	0.089
	TC11	0.059	0.057	0.071	0.071
SC5	TC1	0.018	0.019	0.019	0.018
	TC2	0.179	0.181	0.182	0.185
	TC3	0.199	0.199	0.203	0.205
	TC9	0.207	0.210	0.261	0.270
	TC11	0.139	0.138	0.182	0.183
SC6	TC1	0.024	0.024	0.029	0.024
	TC2	0.400	0.396	0.405	0.403
	TC3	0.528	0.529	0.531	0.534
	TC9	0.485	0.482	0.613	0.648
	TC11	0.305	0.285	0.390	0.395
SC7	TC1	0.032	0.030	0.035	0.032
	TC2	1.383	1.376	1.383	1.399
	TC3	1.163	1.157	1.170	1.164
	TC9	1.271	1.302	1.709	1.733
	TC11	0.750	0.727	0.979	1.017

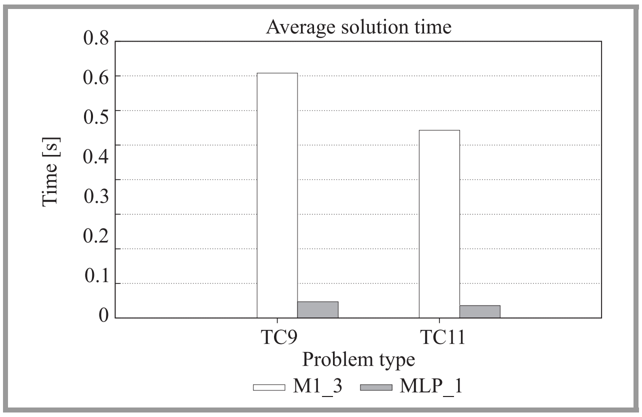


Fig. 7. MILP and LP models comparison with decreasing weights ($m = 12$).

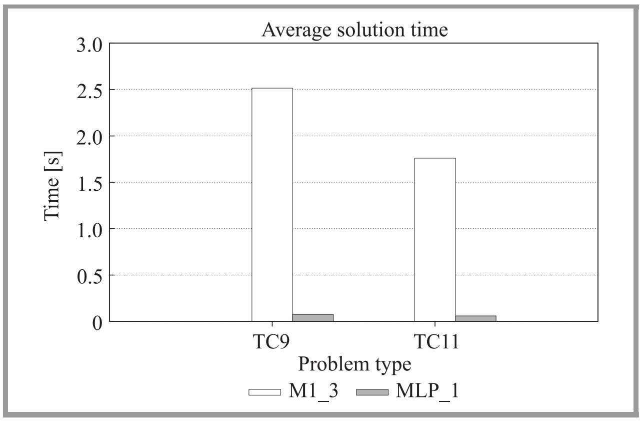


Fig. 8. MILP and LP models comparison with decreasing weights ($m = 15$).

Knowing that the linear programming formulation of the OWA has better computational performances than the mixed integer linear programming formulation and that the redundant constraints can significantly improve efficiency of the latter one, we have tested the influence of the redundant constraints on the linear programming model. We have considered four formulations from Subsection 2.4 for the problems with decreasing weights (TC9, TC11) and additionally TC1–TC3 as non-increasing. The results are presented in Table 3. In Fig. 9 the case for $m = 30$ is shown. One may notice no difference in the case of problems TC1–TC3. For formulation MLP1 and MLP2 there is also no difference in case of the other problems. However, for MLP3 and MLP4 formulations there is about 30% performance deterioration in the case of problems TC9 and TC11. Similar situation occurs for other size cases. Thus, it seems that the redundant constraints do not improve the performance of the linear programming model of the OWA optimization.

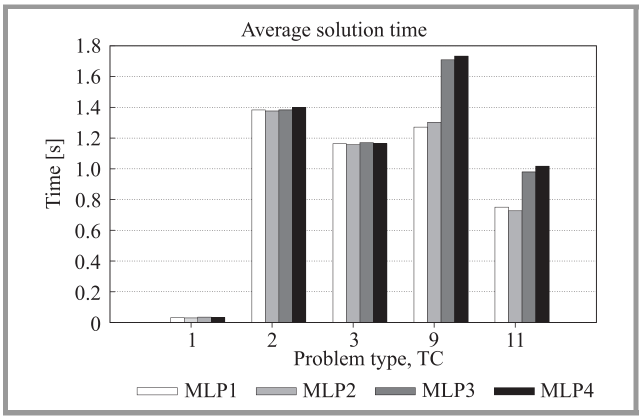


Fig. 9. Model LP formulations comparison ($m = 30$).

4. Conclusions

The paper analyzes two models of mixed integer programming and one linear programming model for optimization of the OWA criterion. Experiments were conducted to compare the computational efficiency of different formulations of these models. Based on the obtained results it can be concluded that the redundant constraints added to MILP models of OWA can significantly shorten the computational time for certain types of localization problems (certain classes of OWA weights vectors). Secondly, the model M1 appears to be much more efficient than the model M2. Besides, if the problem has special structure, which allows one to formulate OWA criterion as standard linear formulation, this should be exploited, as it greatly increases its computational efficiency. However, adding the redundant constraints to the linear programming OWA formulation does not help and may increase the computational time.

Because the results presented here are based on an average solution time, it seems desirable to conduct a more detailed statistical analysis (e.g., minimum, maximum, variance) of the results. Perhaps it will allow to find new dependencies and determine more detailed model characteristics. Better efficiency of the model M1 suggests also an opportunity to apply it to quadratic assignment problem (QAP), from which some transformations for the model M2 have been exploited [8].

Acknowledgment

The research was partially supported by the National Science Centre (Poland) under grant 2011/01/B/ST7/02967 “Integer programming models for joint optimization of link capacity assignment, transmission scheduling, and routing in fair multicommodity flow networks.”

References

[1] R. R. Yager, “On ordered weighted averaging aggregation operators in multicriteria decision making”, *IEEE Trans. Sys., Man and Cyber.*, vol. 18, pp. 183–190, 1988.

[2] R. R. Yager, J. Kacprzyk, and G. Beliakov (Eds.), *Recent Developments in the Ordered Weighted Averaging Operators: Theory and Practice*. Springer, 2011.

[3] W. Ogryczak, T. Śliwiński, and A. Wierzbicki, “Fair resource allocation schemes and network dimensioning problems”, *J. Telecom. Inform. Technol.*, no. 3, pp. 34–42, 2003.

[4] M. Köppen, K. Yoshida, and M. Tsuru, Y. Oie, “Annealing heuristic for fair wireless channel allocation by exponential ordered-ordered weighted averaging operator maximization”, in *Proc. IEEE/IPSJ Int. Symp. Appl. Internet*, Munich, Germany, 2011, pp. 538–543.

[5] S. Nickel and J. Puerto, *Location Theory: A Unified Approach*. Berlin, Springer, 2005.

[6] R. R. Yager, “Constrained OWA aggregation”, *Fuzzy Sets and Systems*, vol. 81, pp. 89–101, 1996.

[7] W. Ogryczak and T. Śliwiński, “On solving linear programs with the ordered weighted averaging objective”, *Eur. J. Oper. Res.*, vol. 148, pp. 80–91, 2003.

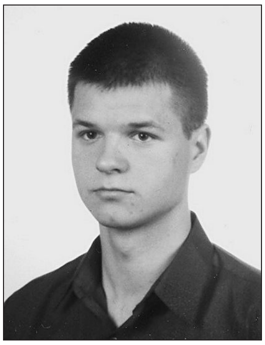
[8] N. Boland, P. Dominguez-Marin, S. Nickel, and J. Puerto, “Exact procedures for solving the discrete ordered median problem”, *Comput. Oper. Res.*, vol. 33, pp. 3270–3300, 2006.

[9] W. Ogryczak and A. Tamir, “Minimizing the sum of the *k* largest functions in linear time”, *Inform. Process. Letters*, vol. 85, pp. 117–122, 2003.

[10] P. Dominguez-Marin, *The Discrete Ordered Median Problem: Models and Solution Methods*. Springer, 2003.

Włodzimierz Ogryczak is a Professor and Deputy Director for Research in the Institute of Control and Computation Engineering (ICCE) at the Warsaw University of Technology, Poland. He received both his M.Sc. (1973) and Ph.D. (1983) in Mathematics from Warsaw University, and D.Sc. (1997) in Computer Science from Polish Academy of Sciences. His research interests are focused on models, computer solutions and interdisciplinary applications in the area of optimization and decision making with the main stress on: multiple criteria optimization and decision support, decision making under risk, location and distribution problems. He has published three books and numerous research articles in international journals.

E-mail: wogrycza@ia.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland



Paweł Olender received the M.Sc. in Computer Science from the Warsaw University of Technology, Poland, in 2008. Currently, he is a Ph.D. student in computer science at the Institute of Control and Computation Engineering at the Warsaw University of Technology. He is employed by the National Institute of Telecommunications in Warsaw. He has participated in projects related to data warehousing and analysis for a telecommunication operator. His research interests are focused on modeling, decision support, optimization, data mining and multi-agent systems.

E-mail: P.Olender@ia.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

E-mail: P.Olender@itl.waw.pl
National Institute of Telecommunications
Szachowa st 1
04-894 Warsaw, Poland

Paper

Developing a Ring-Based Optical Network Structure with Glass-Through Nodes

Marco Caserta[a], Silvia Schwarze[b], and Stefan Voß[b]

[a] *IE Business School, Madrid, Spain*
[b] *Institute of Information Systems, University of Hamburg, Hamburg, Germany*

Abstract—An important issue in designing optical transport networks (OTN) is security. The concept of 1+1 protection requires to connect each origin-destination(OD)-pair by at least two node-disjoint paths. In the case of a single edge or node failure, the connection of all OD-pairs is maintained under 1+1 protection. On a ring, 1+1 protection is given naturally. Moreover, on rings, the routing effort is typically decreasing. These observations motivate the investigation of ring structures for OTN. When developing a ring structure for telecommunication networks, several subtasks can be identified. Rings have to be designed, OD-pairs have to be assigned to rings, communication among rings has to be defined, a proper flow routing has to be chosen, and rings have to be dimensioned regarding flow capacity. In this paper, we address the first two issues, namely generation of rings and assignment of OD-pairs to rings. Our approach allows to distinguish active and non-active (glass-through) nodes in OTN. Active nodes are equipped with active routing hardware that weakens the optical signal and has impact on feasible ring lengths. Non-active nodes do not influence the optical signal. Although a consideration of active/non-active nodes is important in ring design, only a few references address this issue. We propose an algorithm for generating random ring candidates. Moreover, we present a mathematical model for the assignment of OD-pairs to rings subject to a feasible choice of active nodes. We test our methods using a case of Deutsche Telekom.

Keywords—active nodes, interring traffic, optical transport network, ring network design.

1. Introduction

We consider the development of optical transport networks (OTN) based on ring structures. This research has been motivated by a case of Deutsche Telekom and it has been carried out in cooperation with Deutsche Telekom and T-Systems, Darmstadt, Germany. The ring-based approach is motivated mainly by two facts. In telecommunication networks, the rings provide a stable environment. On a ring, each origin-destination(OD)-pair is connected by two node-disjoint paths. We say, 1+1 protection is ensured. In the case of a single node failure all connections are maintained, i.e., network stability is enhanced by 1+1 protection. Secondly, routing efforts are usually decreasing in rings due to the generally given node degree of two. Physically, this node degree is realized by *optical add drop multiplexer* (OADM). These hardware devices are located at ring nodes and have basically two tasks: the connection of two edges within a ring and the communication to external sources, i.e., sending and receiving of data from outside the ring. Using OADMs, a ring can be connected to an external network, or, two rings can be linked.

Ring creation is subject to physical limitations. In general, we assume the ring length to be non-bounded. However, the number of nodes on a ring is limited and depends on the ring length. The longer a ring, the less nodes are feasible, as each node and its established hardware does weaken the optical signal. However, we do not have a signal weakening in the case of glass-through nodes, i.e., nodes where the fiber is directly linked without interconnection by active hardware. Thus, in order to compute feasible ring lengths, we have to distinguish *active* and *non-active* (glass-trough) nodes.

A node has to be active if the node creates demand itself, i.e., if it is part of an OD-pair. Active hardware is required to send information from that node to the ring. Moreover, there is a second reason for a node to be active. Due to the physical limitations regarding ring size, typically a set of rings will have to be established to cover all demands in telecommunication networks of realistic size. In the case of multiple rings, communication among rings might be requested. This leads to the discussion of *interring* traffic. Two rings can be connected via joint active nodes using the installed OADMs. Thus, even if a node does not create demand, it might be active to act as interface between two rings, i.e., interring traffic is routed via this node. Note that two rings have to be connected by at least two active nodes as otherwise 1+1 protection would be violated by the single *transit node*.

We consider cost only regarding purchase and installation of equipment. Maintenance is not considered in our approach. The hardware cost arising for equipment of an OTN network is twofold. First, the cost appears for the optical fiber to be installed and it depends on the ring length. Second, active hardware has to be established dependent on the number of active nodes in the network. Moreover, nodes that act as interface for interring traffic require additional hardware that enables the connection of two OADMs on different rings.

Summarizing, we address the following optimization problem. Given the physical layer of an optical fiber network consisting of nodes and edges, as well as traffic demand for certain OD-pairs, generate a cost-minimizing ring-based network structure containing active and non-active nodes, such that all demand is satisfied.

The list of publications concerning ring network design for OTN is very extensive. See [1] for an overview on optical network design in general and [2] for a mesh-based approach. Closely related to our approach is the cycle cover problem (CCP) [3], or, ring cover problem (RCP) [4]. The CCP aims to find a least cost selection of simple rings such that for a given network all edges are covered. Variations of the CCP are the bounded cycle cover problem (BCCP) [5], the constrained cycle cover problem (CCCP) [6], and the lane covering problem [7]. The BCCP treats problems where the number of edges in a ring is bounded, whereas in the CCCP the number of edges as well as the flow capacity in rings is limited. For the lane covering problem, only a subset of edges in the graph has to be covered. The main difference to our model is that in the CCP and its variants, a cover of *edges* is required. This is due to the fact that for the CCP, network flows have been already fixed in a preprocessing step. Thus, demand is given on the network edges and has to be met by the ring selection. In our setting, however, flows are not fixed in advance. Thus, our demands are still given with respect to OD-pairs. Consequently, we require to cover and connect demand nodes only. In fact, we do not even need to cover all network nodes, but only active ones.

There are a lot more approaches on ring network design, see, e.g., [8]–[13] as a selection. However, in most references a discussion of active and non-active nodes is not considered. This includes the literature on the CCP mentioned above. The number of publications dealing with active nodes in combination with ring network design is scarce. References [14], [15] consider the location of active nodes, when rings are already given. Moreover, [16], [17] present *foundation design*, a model which allows non-active nodes on rings. In this approach, locations of active nodes, as well as the demand loaded onto the ring are already fixed for the candidate ring structures. Then, a cost minimizing set of rings is chosen by an integer programming formulation. However, there is no consideration of detailed interring flows. A survey on ring-based networks is given by [18].

In our research, we focus on a more general approach to include active/non-active nodes in OTN design. We do not predefine demand nor active nodes in a preprocessing step. Rather we combine the selection of rings with the selection of active nodes in an enhanced ring cover approach.

The paper is structured as follows. We present preliminaries on technical notations, input data and solution structure in Section 2. In Section 3 we discuss a separation of the problem into subtasks. We address the first subtask, namely random ring generation in Section 4 and proceed with the second subtask, the coverage of OD-pairs by rings, in Section 5. We finish by presenting a computational study in Section 6 and by giving concluding remarks in Section 7.

2. Problem Settings

In this section, we describe the problem settings in more details. It seems to be worthwhile to start with a short introduction to technical aspects of optical networks. An *optical transmission system* connects transmitters and receivers to an optical transmission medium. In particular, an electrical signal arriving at a transmitter is transformed to a light signal, then it is transmitted over an optical fiber edge and afterwards converted back to an electrical signal at the receiver station. To increase capacity utilization of the optical fiber, wavelength division multiplexing (WDM) is introduced. WDM is a technique that allows to send multiple signals simultaneously over one optical fiber by transferring them to light signals of different wavelength. The necessary hardware components are *optical multiplexer and demultiplexer (short: mux and demux)*, which allow electrical-optical (E-O) and optical-electrical (O-E) conversion. Whenever an optical signal is routed over a *lightpath*, i.e., a sequence of optical edges, signal routing on the traversed nodes has to be organized. Wavelength cross-connects (WXC) handle the routing at nodes. Typically, these hardware components allow to connect two, three or four edges at one node. If a higher node degree is necessary, cross-connects can be joined to systems offering a node degree greater than four. As a particular cross-connect allowing node degree two, we have OADMs. Dependent on length and type of the optical fiber, the optical signal looses power during the transmission. Whenever a certain distance between two cross-connects is reached, *amplifiers* can be used to reshape the signal. However, in this study, we exclude the consideration of this technique and focus on pure OADM installation. In terms of security, for each OD-pair, 1+1 protection has to be ensured. Origin and destination are connected by at least two node-disjoint paths. This ensures maintenance of the connection even in the case of one edge (or even one node) failure. If an OD-pair is covered by a single ring, 1+1 protection is given naturally. However, if traffic is routed over more than one ring it is necessary to ensure that the rings are disjoint. Moreover, rings have to be connected via two transit nodes, at least.

Next, we provide a formal presentation of the given input data. First, we are given a directed network $G = (N,E)$. G is defined by the nodes $n \in N$ and the edges $e \in E$. Moreover, for all edges we define the edge weight $\ell(e) \geq 0$ as the physically given length of the optical fiber edge $e \in E$. Network G presents a macroscopic view on the real telecommunication network: Nodes are estates where hardware is physically installed and a single node might represent more than one hardware unit. Edges correspond to tunnels and each tunnel connects two estates. Thus, a single edge might represent more than one fiber.

Second, point-to-point demands are given for OD-pairs. We define Q to be the set of OD-pairs generating traffic. OD-pairs are symmetric, demand of equal size occurs in both directions. Moreover, demand arises as 1 Gbit and 10 Gbit traffic. Thus, for all $q \in Q$ we introduce d_q^1 and d_q^{10} to denote 1 Gbit and 10 Gbit demand, respectively.

Finally, costs are specified in terms of node cost (dependent on installed active hardware) and edge cost (dependent on edge length). More precisely, we have costs of c_e per kilometer of optical fiber and costs of c_n for each installed OADM unit. Moreover, at transit nodes, additional costs arise for each installed mux-demux, namely c_{tn1} for handling 1 Gbit flow and c_{tn10} for handling 10 Gbit flow. Note that only in this case we have to distinguish between 1 Gbit and 10 Gbit flow with respect to cost. In terms of expenses for installing OADMs or optical fiber, the hardware costs are not influenced by bandwidth.

We proceed by defining required properties of network and solution structures. A *ring* r is defined as a node-disjoint (and consequently, edge-disjoint) closed path. Different rings may share nodes and we denote a set of *candidate rings* by R. Formally, a ring r is given as sequence of its edges: $r = \{e_{r_1}, e_{r_2}, \dots, e_{r_m}\}$ where m denotes the number of edges in r. The *length of a ring* is given by

$$\ell(r) = \sum_{e \in r} \ell(e) \ . \tag{1}$$

The length of a ring (in kilometers) is in principle nonbounded, but, dependent on the number of active nodes on the ring. The more nodes are active on a ring, the smaller this ring has to be due to a weakening of the optical signal. Or, the longer a ring, the less active nodes are allowed for this ring. Given a certain ring r with length $\ell(r)$, the number of active nodes in this ring is limited and we denote the *upper bound on the number of active nodes* in r as $\bar{a}(r)$.

We distinguish active and non-active nodes in the network. A node n is active on a ring r if n contains routing hardware. The only active routing hardware to be considered are OADMs as we generally assume to have node degree of two for establishing rings. A single OADM is able to connect a node only to a single ring. Thus, if a node is active in different rings, one OADM for each of these rings has to be established at the node. On each ring, only a subset of its nodes need to be active. On the other hand, a single node which is part of several rings might be active on some rings and non-active on others.

Interring traffic becomes possible via transit nodes. At these nodes, installation of multiplexer and demultiplexer allow traffic to leave one ring and enter the other ring. To ensure the 1+1 property, each pair of connected rings has to have at least two transit nodes in common. Traffic demand of a certain OD-pair might be routed via several rings, using established transit points. Due to connection of rings via transit points, it appears that flow is using only some parts of a ring. Thus, the edges of a single ring may carry different loads. This has to be respected when dimensioning the rings.

3. Solution Approaches

Solving the described problem includes a number of different subproblems. Rings have to be designed, active and transit nodes have to be chosen and a proper flow routing has to be established. These decisions have to be taken under the light of maintaining feasibility with respect to ring lengths, bandwidth and 1+1 protection. Clearly, these issues do influence each other. For instance, the ring design has impact on a feasible choice of active nodes, or, the choice of transit nodes influences the flow routing. However, to handle this complex problem it will be inevitable to divide the task into smaller portions. We propose the following partitioning into subtasks.

1. Generation of a *candidate ring set* R based on the physically given optical fiber network layer. Each ring is given as a sequence of its edges. Additional ring information like ring length $\ell(r)$ and upper bound on active nodes $\bar{a}(r)$ can be derived throughout the ring generation process.

2. An *extended ring cover problem under consideration of active nodes* has to be addressed. Given the ring candidates, choose a cost-minimizing subset of rings together with active nodes such that all OD-pairs are covered by rings. Interring traffic is not yet considered. That is, OD-pairs have to share a common ring. The limitation of the ring length might lead to feasibility problems if long distances have to be covered for some OD-pairs. Thus, for the extended ring cover, OD-pairs where no common ring is existing in the candidate set are excluded from consideration. These cases are postponed and addressed through a repair approach, see item 5.

3. Redefine the ring structure by allowing *interring traffic* to obtain cost savings. Interring traffic becomes possible for all *adjacent* rings, i.e., rings that share at least two nodes. Interring traffic allows to cover a single OD-pair by a set of connected rings instead of covering it by a single ring.

4. Given the ring structure, choose a proper *ring dimensioning* such that traffic demand is covered by the provided ring capacities. This includes to determine a proper flow routing.

5. *Repair and improvement*. Check for non-covered OD-pairs. Utilize repair methods such as generation and adding of new, suitable rings. Interring traffic can be introduced for this purpose. Moreover, establish improvement techniques, e.g., by shifting nodes between neighbored rings or by generating new rings based on experience on what a “good” ring is.

This comprehensive list of subproblems illustrates the complex nature of the problem. Addressing the complete problem within a single one-step solution procedure looks not very promising for this setting. Rather we focus on a multistep method that handles separately the subtasks described above. In the remainder of this paper, we investigate the first two issues, namely items 1 and 2. The remaining issues are left for future and ongoing work.

4. Ring Generation

The ring generation process is a preprocessing step that provides a set of candidate rings as input data for the subsequent optimization procedures. These rings are only given by their edges and do not carry any information on active nodes nor demand.

We follow the approach presented by [4]. In this reference, there is a suggestion for a ring generator based on the fundamental set of rings, resulting from a spanning tree in the network. By joining rings taken from the fundamental set, this method allows to generate all rings in a network. However, the number of rings in a network is growing exponentially and, for realistic settings, a complete enumeration would exceed the computational limits of subsequent optimization processes. Thus, we focus on a random approach that generates a selection of rings.

Next, we present the approach of [4] in more details. Given a network, generate an arbitrary spanning tree T (not necessarily minimal). Such a spanning tree consists of $|N|-1$ edges. Thus, we have $p=|E|-|N|+1$ remaining edges in $E\setminus T$. Moreover, whenever an edge $e\in E\setminus T$ is added to the spanning tree T, the resulting 1-tree does contain a unique ring. It is easy to see that for each of the p edges in $E\setminus T$ we obtain a new ring. Thus, p different rings can be obtained from one spanning tree T. The resulting set of rings is called a *fundamental set of rings*. Two rings r_1 and r_2 can be combined into a new subgraph $\bar{r}$ by the following procedure. Include all edges into $\bar{r}$ that are contained in exactly one of the two rings, either r_1 or r_2. Leave away all other edges of r_1 and r_2, i.e., all edges that appear in both rings. The resulting subgraph $\bar{r}$ might be not a ring. However, it can be shown that all rings of the network can be generated using one fundamental set of cycles and generating all combinations.

Algorithm 1: **Ring generator**

```
Require: Fiber network $G=(N,E)$, number of iterations $k$
Ensure: Set of rings $R$
 1: Set $R := \emptyset$
 2: Set number of rings per iteration: $p := |E|-|N|+1$
 3: for $i=1,\dots,k$ do
 4:     Generate a random spanning tree $T$ in $G$
 5:     for j = 1, ..., p do
 6:         Set $e :=$ $j$th edge in $E\setminus T$.
 7:         Set $r :=$ unique ring in $T\cup\{e\}$, see Algorithm 2
 8:         if $r\notin R$ then
 9:             $R := R\cup\{r\}$
10:         end if
11:     end for
12: end for
13: Output: $R$
```

As we are not interested in obtaining the complete set of rings in G, we propose the following approach. Generate an arbitrary number of random spanning trees. For each of these trees, build the fundamental set of rings as described above. Add these rings to the candidate ring set R unless they are not already stored in R.

Algorithm 2: **Detect unique ring in subnetwork**

```
Require: $T\cup\{e\}$ containing one unique ring
Ensure: Unique ring $r$
Set $\bar{T} := T\cup\{e\}$
Set $\bar{N} :=$ set of nodes adjacent to $\bar{T}$
Set $\bar{G}=(\bar{N},\bar{T})$
Set $N^1 := \{n\in\bar{N}$ : node degree of $n$ w.r.t. $\bar{G}$
                              equals one $\}$
Set $E^1 := \{e\in\bar{T}$ : edge $e$ adjacent to a node $n\in N^1\}$
while $N^1\neq\emptyset$ do
    $\bar{T} := \bar{T}\setminus E^1$
    $\bar{N} := \bar{N}\setminus N^1$
    Update $N^1$ and $E^1$
end while
Output: $r := \bar{T}$
```

A formal description of the ring generator is presented in Algorithm 1. In this procedure, step 7 needs further specification. How to detect a unique ring in a network $T\cup\{e\}$? We propose the following simple approach, see Algorithm 2. Check the node degree of each node in $\bar{T} := T\cup\{e\}$. As long as there are nodes with node degree one, remove from $\bar{T}$ each of these nodes together with the single adjacent edge. The procedure terminates with the unique ring.

Note that for each ring r obtained through Algorithm 1 we do already have the information on ring length $\ell(r)$ by Eq. (1) and on the maximal number of active nodes $\bar{a}(r)$.

5. Ring Cover Problem

In a second stage, based on the set of candidate rings R, an *extended ring cover problem under consideration of active nodes* (RCP-A) is addressed. RCP-A is the following. Given a set of OD-pairs as well as a set of candidate rings, choose a cost-minimizing set of rings together with active nodes such that each OD-pair $q\in Q$ shares (at least) one ring, where origin and destination of q are active.

Note that the RCP-A is related to the general CCP and its variants BCCP, CCCP, and lane covering. However, there are two main differences between these models. First, it is possible in RCP-A, to distinguish active from non-active nodes, which is not the case for the CCP and its variants. Second, in the CCP demand is already given for the edges, i.e., the flow routing has been already carried out in advance. In RCP-A, demand is still given for OD-pairs, routing of flow is not fixed.

The RCP-A requires that each OD-pair is covered by (at least) one ring. So far, we do not give the possibility for joining rings and for interchanging traffic between them. The advantage of this approach is obvious: 1+1 protection is ensured naturally. Nonetheless, we might run into troubles if there are OD-pairs that can not be covered by one single ring due to limitations in ring length. Moreover, this approach might be costly as it is likely that we end up with a large number of rings. To deal with the first issue, we recommend a repair algorithm that generates suitable rings, see item 5 in Section 3. Finding such rings is possible for each OD-pair q, unless there is just a single path connecting the origin and destination of q. In this case, 1+1 protection cannot be established and the considered OD-pair has to be treated separately. Addressing the second issue, we refer to item 3 in Section 3. Interring traffic will allow to serve a single OD-pair using a set of adjacent rings. Algorithmic approaches following item 3 will potentially reduce the number or rings and generate a leaner ring network structure.

For each OD-pair q, let nodes $o(q)$ and $d(q)$ denote the origin and destination of q, respectively. Recall that $\bar{a}(r)$ is the upper bound on active nodes on a ring r. Furthermore, let $n^Q = |Q|$, $n^R = |R|$, and $n^N = |N|$ be the number of OD-pairs, rings and nodes, respectively. We denote by the $n^Q \times n^R$-matrix B the relation between OD-pairs and rings. More precisely, let $b_{qr} = 1$ if ring r is able to cover OD-pair q, i.e., if nodes $o(q)$ and $d(q)$ are adjacent to edges $e \in r$. Finally, let $n^Q(r)$ be the number of OD-pairs that ring r is potentially able to cover, i.e., $n^Q(r) = \sum_{q=1}^{n^Q} b_{qr}$.

The following logical tests will serve as a preprocessing step to clean up the input data for the mathematical model.

- For all $r \in R$, check whether r covers at least one OD-pair, i.e, if $n^Q(r) \geq 1$. If not, ring r can be diminished.
- For all $q \in Q$, check whether an OD-pair q is covered by at least one ring, i.e., if $\sum_{r=1}^{n^R} b_{qr} \geq 1$. If not, shift the OD-pair q into a pool of *uncovered OD-pairs* (to be treated later).

The mathematical description of RCP-A is the following. There are three classes of variables. Variables x_{qr} indicate that an OD-pair q is covered by a ring r, y_{nr} indicate that a node is set active on a ring (i.e., hardware that connects n to r is established at n), and z_r indicate that a ring r is used, i.e., if some OD-pair is assigned to r. More precisely we have

$$x_{qr} = \begin{cases} 1, & \text{if OD-pair } q \text{ is assigned to ring } r, \\ 0, & \text{otherwise;} \end{cases}$$

$$y_{nr} = \begin{cases} 1, & \text{if node } n \text{ is active on ring } r, \\ 0, & \text{otherwise;} \end{cases}$$

$$z_r = \begin{cases} 1 & \text{if ring } r \text{ is chosen,} \\ 0 & \text{otherwise.} \end{cases}$$

We establish a set of constraints that ensure a proper interaction of the variables and ensure feasibility.

1. Each OD-pair has to be assigned to exactly one ring:

$$\sum_{r=1}^{n^R} b_{qr} x_{qr} = 1 \quad , \quad \forall\ q = 1, \ldots, n^Q . \tag{2}$$

2. If any OD-pair is assigned to a ring r, r has to be chosen.

$$z_r n^Q(r) \geq \sum_{q=1}^{n^Q} x_{qr} \quad , \quad \forall\ r = 1, \ldots, n^R . \tag{3}$$

3. If an OD-pair is assigned to a ring r then the origin and destination of that OD-pair have to be active on r.

$$x_{qr} \leq y_{o(q)r} \quad , \quad \forall\ q = 1, \ldots, n^Q,\ r = 1, \ldots, n^R , \tag{4}$$

$$x_{qr} \leq y_{d(q)r} \quad , \quad \forall\ q = 1, \ldots, n^Q,\ r = 1, \ldots, n^R . \tag{5}$$

4. Do not violate the maximal number of active nodes per ring.

$$\sum_{n=1}^{n^N} y_{nr} \leq \bar{a}(r) \quad , \quad \forall\ r = 1, \ldots, n^R . \tag{6}$$

Regarding the objective function, we address a minimization of total cost given by the sum of cost for optical fiber (dependent on ring length) and cost for the installation of OADMs (dependent on the number of active nodes).

$$\min c_e \sum_{r=1}^{n^R} \ell(r) z_r + c_n \sum_{n=1}^{n^N} \sum_{r=1}^{n^R} y_{nr} . \tag{7}$$

Alternative objectives might be discussed. For instance, more detailed cost values c_{nr} could be introduced to model costs for OADMs in dependency of the ring. Another approach could be to remove the hard constraint of covering all OD-pairs, see Eq. (2), and punish violation of (2) by introducing a corresponding term to the objective function. Thus, the selection of expensive rings could be avoided by accepting uncovered OD-pairs.

6. Computational Results

We tested our approach on a real world instance provided by Deutsche Telekom. The input fiber graph $G = (N, E)$ consisted of $|N| = 8,349$ nodes and $|E| = 12,397$ edges. In addition, a total number of $n^Q = 5,132$ OD-pairs has been given; 2761 nodes have a positive demand.

The numerical test included three phases. First, a ring generation has been carried out, see Algorithm 1. Afterwards the resulting data has been cleaned up by the logical tests described in Section 5. Finally, the adjusted candidate ring set has been fed into the mathematical model to solve RCP-A. The ring generator has been coded in C++ and compiled with the GNU compiler selection gcc 4.3 on a Pentium 4 1.3 MHz Linux workstation with 1 GB of RAM. The mathematical model has been implemented and solved using ILOG OPL 4.2.

Table 1
Results of the ring generator phase

Spanning tree	Newly generated unique rings	Total number of unique rings
1	4049	4049
2	2754	6803
3	2303	9106
4	1990	11096
5	1876	12972

The generation of rings by Algorithm 1 provided a set R of candidate rings with a total of $12,972$ unique rings out of $k = 5$ different spanning trees. Computational time was less than five minutes (wall-clock time). The detailed description of the five iterations produced by the ring generator is summarized in Table 1. With increasing number of spanning trees, the probability to produce duplicate rings is increasing and consequently, the number of new unique rings is decreasing.

Table 2
Rings chosen for the candidate ring set

	Min	Max	Av.	Stdev
Number of nodes	3	80	10.49	10.33
Ring length [km]	3	2182	186.54	277.79

Table 2 provides statistical information about ring length and number of nodes of rings created throughout the ring generation phase. Moreover, Fig. 1 illustrates the frequency of rings with a given number of nodes per ring.

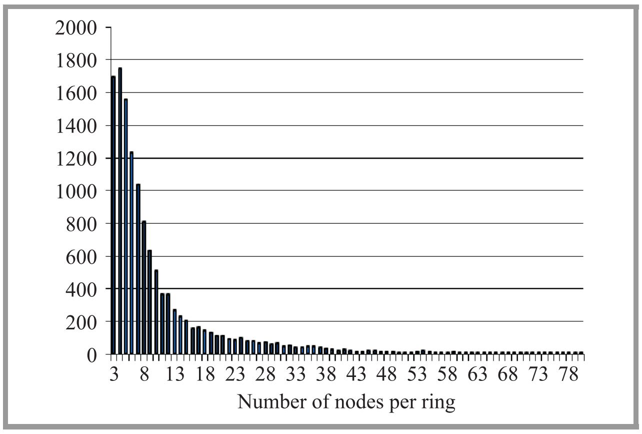


Fig. 1. Ring generation solution: frequency of rings by number of nodes.

In a second step, we clean up the candidate ring set R and the set of OD-pairs Q by logical tests. An OD-pair $q \in Q$ can be covered, i.e., assigned to a ring, only if there exists at least one ring $r \in R$ such that both origin and destination $o(q)$ and $d(q)$ are located on ring r. Moreover, we consider only rings that are able to cover at least one OD-pair.

In our experiments, 2002 rings had to be removed from R by logical preprocessing. We obtained a reduced candidate ring set $R' \subseteq R$ with $|R'| = 10,970$. In addition, the set of OD-pairs needed to be adjusted. We ended up with a reduced set of OD-pairs $Q' \subseteq Q$ of size $|Q'| = 4,420$.

We defined the *level of coverage* of a candidate ring set R with respect to a set of OD-pairs Q as $cov(R,Q) = |Q'|/|Q|$. For our test instance, we achieved $cov(R,Q) = 0.86$.

Finally, the implementation of RCP-A, see Eqs. (2)–(7), is the last phase of our computational test. RCP-A aims to detect a feasible and cost minimizing ring cover. The input data is given by the candidate ring set R and the set of OD-pairs Q'. We stopped the solver after a computational time of around one week and took the best feasible solution found so far. Note that optimality of this solution is not proved. The feasible solution contained 784 rings and 1048 active nodes. An active node was active on 2.3 rings at an average.

Table 3
Rings chosen by RCP-A

	Min	Max	Av.	Stdev
Number of nodes	3	80	16.50	14.50
Number of active nodes	2	7	3.07	1.39
Share of active nodes	0.03	1.00	0.30	0.20
Ring length [km]	5	2128	305.90	369.99

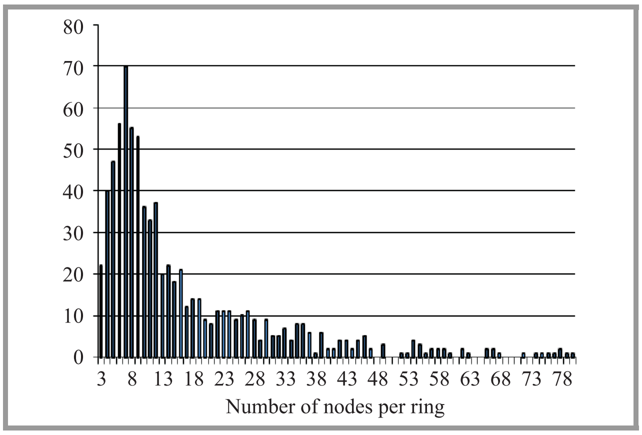


Fig. 2. RCP-A solution: frequency of rings by number of nodes.

See Table 3 for statistical information on the extended ring cover solution regarding the chosen rings. For an illustration of frequencies of rings with given number of nodes and given number of active nodes, see Figs. 2 and 3, respectively.

On average, longer rings with more nodes have been selected by RCP-A than given by the random candidate ring generator. Regarding active nodes, the solution de-

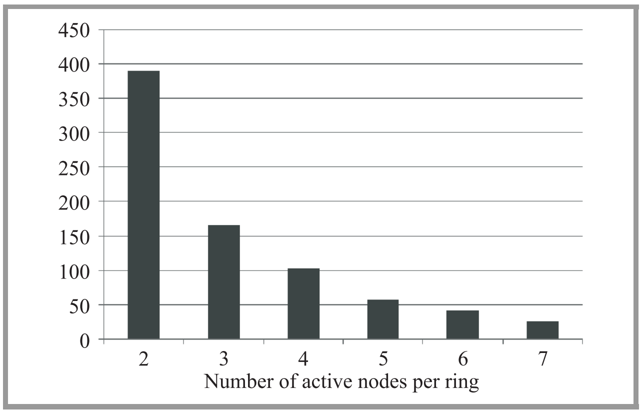


Fig. 3. RCP-A solution: frequency of rings by number of active nodes.

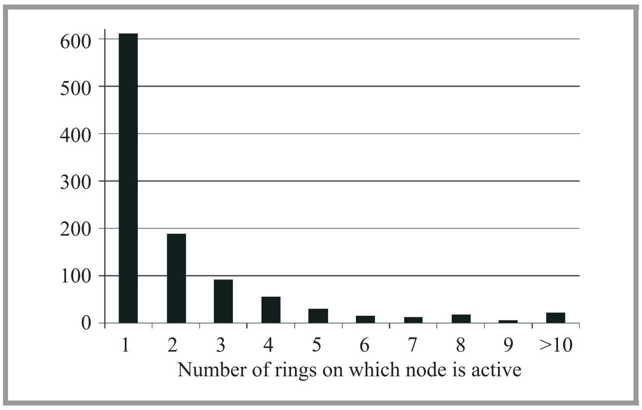


Fig. 4. RCP-A solution: frequency of nodes by number of assigned rings.

livered 2,409 out of 126,888 potential active-node/ring assignments. These assignments refer to unique 1048 active nodes. The majority of these nodes are assigned only to one ring, see Fig. 4. At maximum, one node has been assigned to 39 rings.

7. Conclusion

We investigated two stages of OTN ring network design, namely the generation of rings and the assignment of OD-pairs to rings. The consideration of active and non-active nodes has been included into our approaches. We proposed an algorithm for random generation of candidate rings. Moreover, we presented a mathematical model for assigning OD-pairs to rings such that active nodes are chosen accordingly. While recent discussion on protection also considers issues beyond 1+1 (like 1:n, m:n, see, e.g., [19]) our research addresses important issues with real-world relevance. We tested our approaches using data of Deutsche Telekom, yielding a 86% coverage of OD-pairs by pure rings without interring traffic.

A next step would be to redefine the network structure obtained by RCP-A by enabling interring traffic. The aim is to reduce the number of rings and to obtain a leaner network structure. Moreover, OD-pairs that have been sorted out before RCP-A, as coverage by a single ring was not possible, could now be covered by a set of joined rings. For this matter, satisfaction of 1+1 protection has to be carefully checked. In particular, it has to be ensured that the flow is routed via node-disjoint rings to maintain node-disjoint paths. Putting the ideas forward to more general protection mechanisms could be another step.

Moreover, there is a need to develop the repair and improvement techniques. For instance rings of the candidate set could be merged to create new rings to include uncovered OD-pairs. In addition, the shifting of nodes between existing rings could lead to improvement methods. Finally, a proper ring dimensioning is dependent on a flow routing. A distinction of 1 GBit and 10 GBit demands should be addressed there. This line of research could also re-consider older work making the case for partitioning the overall network design problem into subproblems (see, e.g., [20]). This may also support the idea of solving one model for OD pairs that can be on a common ring and another for pairs that must route traffic over multiple rings.

References

[1] P. Soriano, C. Wynants, R. Séguin, M. Labbé, M. Gendreau, and B. Fortz, “Design and dimensioning of survivable SDH/SONET networks”, in *Telecommunications Network Planning*, B. Sansò and P. Soriano, Eds. Norwell, MA: Kluwer, 1999, pp. 147–167.

[2] H. Höller, B. Melián, and S. Voß, “Applying the pilot method to improve VNS and GRASP metaheuristics for the design of SDH/WDM networks”, *European J. Oper. Res.*, vol. 191, no. 3, pp. 691–704, 2008.

[3] M. Labbé, G. Laporte, and P. Soriano, “Covering a graph with cycles”, *Comput. Oper. Res.*, vol. 25, no. 6, pp. 499–504, 1998.

[4] J. L. Kennington, V. S. S. Nair, and M. H. Rahman, “Optimization based algorithms for finding minimal cost ring covers in survivable networks”, *Comput. Optimiz. Appl.*, vol. 14, pp. 219–230, 1999.

[5] D. S. Hochbaum and E. V. Olinick, “The bounded cycle-cover problem”, *INFORMS J. Comput.*, vol. 13, no. 2, pp. 104–119, 2001.

[6] G. Pesant and P. Soriano, “An optimal strategy for the constrained cycle cover problem”, *Annals Mathema. Artif. Intell.*, vol. 34, pp. 313–325, 2002.

[7] N. Immorlica, M. Mahdian, and V. S. Mirrokni, “Cycle cover with short cycles”, in *STACS 2005*, LNCS 3404, V. Diekert and B. Durand, Eds. Berlin: Springer, pp. 641–653, 2005.

[8] M. Henningsson, K. Holmberg, and D. Yuan, “Ring network design”, in *Handbook of Optimization in Telecommunications*, M. G. C. Resende and P. M. Pardalos, Eds. New York: Springer, pp. 291–311, 2006.

[9] C. Arbib and F. Rossi, “An optimization problem arising in the design of multiring systems”, *European J. Oper. Res.*, vol. 124, pp. 63–76, 2000.

[10] J. G. Klincewicz, H. Luss, and D. C. K. Yan, “Designing tributary networks with multiple ring families”, *Comput. Oper. Res.*, vol. 25, no. 12, pp. 1145–1157, 1998.

[11] A. Fink, G. Schneidereit, and S. Voß, “Solving general ring network design problems by meta-heuristics”, in *Computing Tools for Modeling, Optimization and Simulation*, M. Laguna, J. L. González Velarde, Eds. Boston: Kluwer, pp. 91–113, 2000.

[12] D. Rajan and A. Atamtürk, “A directed cycle-based column-and-cut generation method for capacitated survivable network design”, *Networks*, vol. 43, no. 4, pp. 201–211, 2004.

[13] M. I. Eiger, H. Luss, and D. F. Shallcross, "Network restoration under a single link or node failure using preconfigured virtual cycles", *Telecomm. Sys.*, vol. 46, pp. 17–30, 2011.

[14] A. Sutter, F. Vanderbeck, and L. Wolsey, "Optimal placement of add/drop multiplexers: heuristic and exact algorithms", *Oper. Res.*, vol. 46, no. 5, pp. 719–728, 1998.

[15] T. Y. Chow and P. J. Lin, "The ring grooming problem", *Networks*, vol. 44, no. 3, pp. 194–202, 2004.

[16] G. D. Morley and W. D. Grover, "Comparison of mathematical programming approaches to optical ring network design", in *Proc. of CCBR'99*, Ottawa, Canada, 1999, pp. 137–184.

[17] G. D. Morley and W. D. Grover, "Current approaches in the design of ring-based optical networks", in *Proc. IEEE Canadian Conf. Electrical and Comput. Eng.*, Edmonton, Canada, 1999, pp. 220–225.

[18] G. D. Morley and W. D. Grover, "A Comparative Survey of Methods for Automated Design of Ring-based Transport Networks", Tech. Rep., TR-97-04, TRLabs Network Systems, Canada, 1998.

[19] ITU-T G.808.1, "Generic protection switching – Linear trail and subnetwork protection", International Telecommunication Union Recommendation ITU-T G.808.1, 2010 [Online]. Available: http://www.itu.int/

[20] S. Cosares, D. N. Deutsch, I. Saniee, and O. J. Wasem, "SONET toolkit: a decision support system for designing robust and cost-effective fiber-optic networks", *Interfaces*, vol. 25, no. 1, pp. 20–40, 1995.

Marco Caserta received his Ph.D. in Industrial Engineering and Operations Research from the University of Illinois (USA), after earning a MSc in Management Engineering from the Politecnico di Milano (Italy). He is currently an associate professor at the IE Business School, Madrid, Spain. He teaches optimization related courses to graduate students within the International MBA program. His main research interest is concentrated on the design and development of metaheuristic-based algorithms for very large scale real-world optimization problems, with a special focus on logistics, telecommunication and transportation related problems. He has published a number of papers in journals in the area of operations research/management science.
E-mail: mcaserta@faculty.ie.edu
IE Business School
Maria de Molina st 12, Piso 5
28006 Madrid, Spain

Silvia Schwarze is researcher at the Institute of Information Systems at the University of Hamburg. She holds a degree in Mathematics (diploma) of the Leipzig University of Applied Science and a Ph.D. in Mathematics from the University of Göttingen. Her research interests are in the fields of logistics and telecommunications with a focus on network optimization, games on networks, mathematical programming and metaheuristics.
E-mail: schwarze@econ.uni-hamburg.de
Institute of Information Systems
University of Hamburg
Von-Melle-Park 5
20146 Hamburg, Germany

Stefan Voß is professor and director of the Institute of Information Systems at the University of Hamburg. Previous positions include full professor and head of the department of Business Administration, Information Systems and Information Management at the University of Technology Braunschweig (Germany) from 1995 up to 2002. He holds degrees in Mathematics (diploma) and Economics from the University of Hamburg and the Ph.D. and the habilitation from the University of Technology Darmstadt. His current research interests are in quantitative/information systems approaches to supply chain management and logistics including public mass transit and telecommunications. He is author and co-author of several books and numerous papers in various journals. Stefan Voß serves on the editorial board of some journals including being Editor of Netnomics, Editor of Public Transport, Associate Editor of INFORMS Journal on Computing and Area Editor of Journal of Heuristics. He is frequently organizing workshops and conferences.
E-mail: stefan.voss@uni-hamburg.de
Institute of Information Systems
University of Hamburg
Von-Melle-Park 5
20146 Hamburg, Germany

Paper

Requirements of 4G-Based Mobile Broadband on Future Transport Networks

Matthias Fricke[a], Andrea Heckwolf[a], Ralf Herber[b], Ralf Nitsch[c],
Silvia Schwarze[d], Stefan Voß[d], and Stefan Wevering[e]

[a] *Deutsche Telekom Netzproduktion GmbH, Fixed Mobile Engineering Deutschland, Darmstadt, Germany*
[b] *Deutsche Telekom AG, Group Technology, Darmstadt, Germany*
[c] *T-Systems International GmbH, Darmstadt, Germany*
[d] *Institute of Information Systems, University of Hamburg, Hamburg, Germany*
[e] *Nokia Siemens Networks GmbH & Co. KG, München, Germany*

Abstract—Long term evolution technologies provide new standards in mobile communications regarding available bandwidth. It is expected that users of one radio cell will share more than 100 Mbit/s in future. To take advantage of the full feature set of next generation mobile networks, transport network design has to face new requirements, caused by the architectural changes of LTE technologies. Especially the newly defined X2 interface impacts on the transport network requirements. X2 enables direct communication between evolved base stations (eNBs) and thus, enforces local solutions. At the same time a tendency of locating network elements at fewer, central sites to reduce operational expenditure can be observed, in particular concerning the transport layer. This leads to the question of how the direct X2 connection of eNBs on the logical layer can be accommodated with a general centralization of transport networks. Our considerations show that for LTE, a centralized transport network is able to realize the local meshing between eNBs. However, for LTE Advanced, the standards currently discussed by the 3GPP initiative could lead to enhanced requirements on the X2 interface latency. Consequently, the implications for the network architecture have to be analyzed in more detail.

Keywords—backhauling, LTE Advanced, mobile network design, X2 interface.

1. Introduction

In recent years the evolution of mobile communication proceeded mainly under the influence of the 3GPP initiative (3rd Generation Partnership Project [1]). 3GPP is a consortium, or collaboration, of standardization bodies in mobile communications. An important movement is the standardization of advanced mobile communication systems, in particular of the new technologies long term evolution (LTE) and LTE advanced (LTE-A) [2]. LTE technologies set new standards in mobile communication concerning bandwidth. In future, users of one radio cell will share more than 100 Mbit/s of bandwidth. Moreover, on the countryside, where neither appropriate DSL-based technology nor fiber-to-the-home technology is available, LTE offers new possibilities to provide flexible broadband solutions. For instance, in August 2010, Deutsche Telekom turned on the first LTE node in Kyritz, which is located in a rural area approximately 100 km north east of Berlin. To take advantage of next generation mobile networks, an adjustment and optimization of basic transport layers is inevitable. It will be necessary to analyze, which influence LTE and LTE-A take on traffic in access networks and on aggregation issues. The recent developments in telecommunication networks show the growing tendency that important network elements are concentrated at few locations. The number of sites with active hardware in access networks is reduced from tens of thousands to a few hundreds, by utilizing the optical transmission technology in combination with the increasing growth of the optical fiber network.

Concerning the current universal mobile telecommunications system (UMTS) environment, there is a star-shaped network connecting tens of thousands of antenna locations with some tens of radio network controllers (RNCs). A local meshing between base stations beyond the RNC-locations is not given in this setting. Thus, the current UMTS architecture supports the objective of reducing the number of sites and to hold complex technologies at a few locations. To design efficient fixed-mobile convergent networks, we have to answer the question which impact LTE and, in particular, the future standards of LTE-A have on network design. Will it be possible to realize the requirements of the X2 interface in terms of bandwidth and latency by using today's technology concepts? Will LTE-A lead to a trend reversal in network design? Do the antenna sites have to be connected via a local mesh? Is there active transmission hardware needed at the base stations? Are there applications for passive wavelength division multiplexing (WDM) technology?

In this paper, we present existing results found in literature and summarize these findings in order to highlight research challenges given by LTE-A. Based on this investigation, we analyze whether it is possible to meet the requirements of LTE and LTE-A, and, at the same time, reduce the number of sites in telecommunication networks. In particular, we give a brief introduction into the basics of LTE and

LTE-A in Section 2. In Section 3, we describe the development of mobile communication networks throughout the last decade. Afterwards, we analyze the requirements of LTE and LTE-A in Section 4. The results indicate that the current network structure suffices to enable the performance necessary for LTE. However, regarding LTE-A, the network architecture might have to be revisited to enable all required features. We close with a brief summary in Section 5. See [3]–[7] for recent surveys on LTE-A. Moreover, see [8] for an earlier version of this paper.

2. Basics of LTE and LTE-A

LTE has been developed as a successor of the UMTS radio network. The main features of LTE are increased bandwidth, support of multiple antennas at single base stations and the focus on packet switching (IP) protocol. In LTE, the local base stations are equipped with advanced functionality that enables them to take over tasks that have been carried out by central entities in a UMTS. The renaming of Node B (NB) to evolved Node B (eNB) illustrates the advanced abilities of the base stations. For instance, in the case of a moving user terminal, an eNB carries out independently the handover of the radio connection to a neighbored base station. In UMTS, an RNC has been responsible for this task. This modification of the network structure triggers the discussion of centralized vs. decentralized network design. In LTE, the network structure is flattened by the removal of the RNCs. However, the decentralization of important features, like handovers, implies the need for decentralizing related functionalities, like security operations. In turn, this decentralization contradicts the recent development in telecommunication networks to reduce the number of sites.

Physically, an eNB is equipped with two new interfaces. The X2 interface connects neighbored eNBs directly to support mobility [2]. For instance, handovers are enabled via X2. The S1 interface establishes a backhaul connection from an eNB to the core network. Via this connection, information is send on the user plane, as well as on the control plane.

While the standardization of LTE is finished and the first LTE sites are already established within Germany, the specification of LTE-A is still under discussion. LTE-A is designed to meet the requirements of the ITU (International Telecommunication Union) declared within the International Mobile Telecommunication (IMT)-Advanced specifications. The main design criteria for LTE-A are cost per delivered bit and system scalability. Moreover, reduction of latency, consistent area performance, and energy efficiency are addressed [9]. LTE-A shall provide a set of features to meet these requirements. These features are carrier aggregation, advanced multiple input multiple output (MIMO), coordinated multipoint (CoMP), relaying, and support of heterogeneous networks, see Section 3 for details.

The traffic growth in mobile communication is pushed by an increasing number of broadband subscribers, in particular due to a rising number of new devices, like smartphones and tablets. In addition, the number of devices is supposed to increase by newly developed machine-to-machine applications that are expected to establish machine devices in large numbers. In addition, new applications, like 3D services, establish demand for low latencies and high data rates. Consequently, those trends require the evolution of the current mobile communication network towards the standards of LTE-A.

3. The Evolution of Mobile Networks

By establishing the standardization of UMTS within the Rel-99 specification, the 3GPP initiative created a basis for increased data rates and an optimal implementation of packet based transmission. Table 1 gives details on the 3GPP standardization process and lists the 3GPP releases, the time of functional freeze, and the main radio features.

Table 1
Evolution of 3GPP specifications according to [10]

Release	Functional freeze	Main radio features of the release
Rel-99	March 2000	UMTS 3.84 Mcps (W-CDMA FDD & TDD)
Rel-4	March 2001	1.28 Mcps TDD (aka TD-SCDMA)
Rel-5	June 2002	HSDPA
Rel-6	March 2005	HSUPA (E-DCH)
Rel-7	Dec 2007	HSPA+ (64QAM DL, MIMO, 16QAM UL), LTE & SAE feasibility study, EDGE evolution
Rel-8	Dec 2008	LTE work item – OFDMA air interface, SAE work item, new IP core network, 3G femtocells, dual carrier HSPA
Rel-9	Dec 2009	Multi-standard radio (MSR), dual cell HSUPA, LTE-A feasibility study, SON, LTE femtocells
Rel-10	March 2011	LTE-A (4G) work item, CoMP study, four carrier HSDPA
Rel-11	Dec 2012	CoMP, inter-band carrier aggregation, enhanced ICIC, eight carrier HSDPA

There is a long history of standardization of advanced mobile communication systems. GSM, the first global system for digital mobile communication was specified in the late eighties to early nineties. From Rel-99 up to Rel-7, 3GPP has specified the UMTS network architecture with its packet-switched domain. On the radio side, the main focus was on increasing the data rates for the end users by means of high speed packet access (HSPA) technologies, both on the down- and uplink. Rel-7 included the HSPA+ technology and an LTE and system architecture evolution (SAE) feasibility study was started. Releases Rel-8 and 9 referred to LTE and included the orthogonal frequency-division multiple access (OFDMA) air interface specification, as well as the new SAE-based network architecture. SAE tries to simplify the architecture with an all-IP approach and it supports the requirements, like higher throughput and lower latency. Furthermore, Rel-9 also included an LTE-A feasibility study. LTE-A in its first release was frozen in spring 2011 within 3GPP Rel-10. Thus, the main building blocks of LTE-A technology are fixed. Rel-11 is targeted for December 2012 and shall include enhancements with respect to CoMP transmission, inter-band carrier aggregation and enhanced inter-cell interference coordination (ICIC) mechanisms. In this context, new requirements on backhauling networks are expected.

We proceed by discussing the fundamental change related to the SAE for LTE and LTE-A. Figure 1 illustrates the 3G network architecture for the packet switched domain as it has been specified by 3GPP. The NBs are connected to an RNC via an Iub interface. The RNC, e.g., takes care of the management of the NBs, the supervision of radio resources, and the handover control. The RNC is connected to the serving GPRS support node (SGSN) via an Iu packet switched (IuPS) interface. The SGSN manages subscriber access to the radio access network and it is controlling handover processes that cannot be handled by the RNC itself. Via the core network the SGSN connects to the gateway GPRS support node (GGSN). The GGSN is the mobility anchor point for the end user IP connections and implements the gateway functions towards internal service platforms and external data platforms. Therefore, it performs AAA functions, authentication, authorization and accounting, and enforces subscriber policy. The realization of a 3G network architecture is typically centralized. For instance, in the German 3G network we have tens of thousands NB sites, some tens of RNC sites, and only a fistful of GGSN sites.

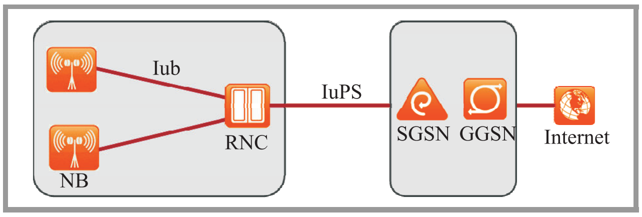


Fig. 1. 3G mobile service architecture for packet switched domain since 3GPP Rel-99 (GPRS/UMTS).

The network architecture given in Fig. 1 has remained unchanged until 3GPP Rel-7. Only after specifying LTE, the basic architecture has been modified, on the one hand to increase the efficiency in mobile networks and on the other hand to meet the demand for bandwidth. Figure 2 presents the newly defined SAE architecture used for LTE. One of the major goals of the 3GPP specification of the SAE was to completely shift towards IP technology on the one hand and to flatten the network architecture on the other hand. The latter has been achieved by removing the RNC network element and distributing its functionality to the eNB, and to the mobility management entity (MME) located in the core network. As a consequence, some of the handover functionality is implemented on the eNBs, which requires in turn an exchange of information between eNBs. This issue has been addressed by defining the X2 interface, which allows direct communication between eNBs. The former SGSN and GGSN packet core nodes are architecturally replaced by the MME and the SAE-gateway (SAE-GW). In this respect, it is important to note that the MME only handles the control plane, while the user plane is directly connected from the eNB to the SAE-GW.

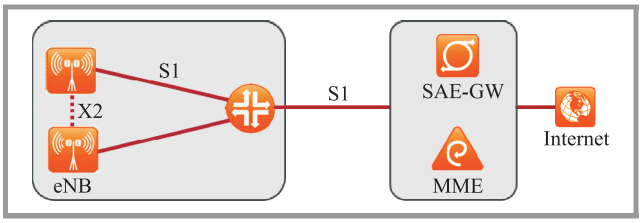


Fig. 2. 4G mobile service architecture since 3GPP Rel-8 (LTE).

We summarize the most relevant architecture differences between 4G and 3G with respect to the transport network:

- the 4G all-IP network architecture requires a packet-centric transport,
- more traffic has to be carried, since LTE and LTE-A will support up to 1 Gbit/s of traffic for a single user,
- the mobile network architecture between the eNBs and the core network sites is flat, which can also be reflected by the underlying transport network,
- X2 interfaces have been newly defined between eNBs, which needs to be covered by the transport network infrastructure.

This summary is true both for LTE and LTE-A.

The development of LTE-A focuses on providing higher bandwidths and improved performance for the users [11]. Next we consider technologies that enable LTE-A. The ITU provides clear requirements given by its IMT-Advanced (IMT-A) specifications, see Fig. 3. These requirements in-

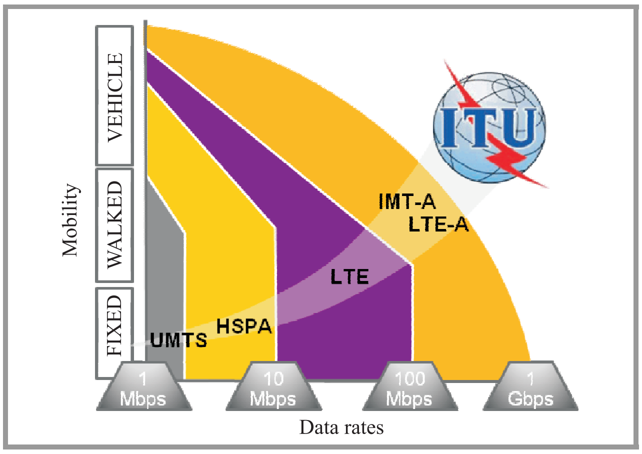


Fig. 3. LTE-A fulfills or exceeds the requirements of IMT-A defined by the ITU [2].

clude high mobility data rates of 100 Mbit/s, e.g., in trains and cars, and 1 Gbit/s for low mobility communications. The ITU requirements were taken up by 3GPP, as the basis for defining the LTE-A technology. In the mean time, the ITU has officially accepted LTE-A and IEEE 802.16m as IMT-A standards, because it was proven that these technologies can meet ITU's requirements. Figure 4 presents five approaches that enable LTE-A to achieve those high data rates.

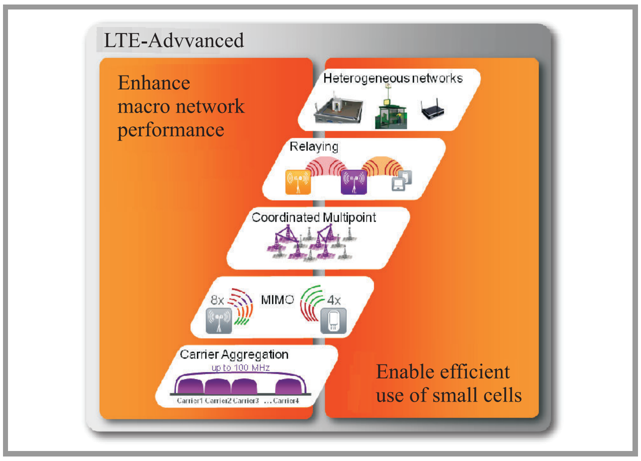

LTE-Advvanced
Enhance macro network performance
Heterogeneous networks
Relaying
Coordinated Multipoint
8x
MIMO
4x
Carrier Aggregation
up to 100 MHz
Carrier1 Carrier2 Carrier3 ... Carrier4
Enable efficient use of small cells

Fig. 4. Enabling technologies of LTE-A [2].

- LTE-A includes carrier aggregation of up to 100 MHz bandwidth, which is the basis in terms of frequency resources to enable very high data rates in a cell, see, [12], [13].
- Advanced MIMO antenna schemes are necessary to implement high data rates. Simulations indicate that up to eight eNB and eight user equipment antennas can be utilized efficiently [14] by MIMO.
- With CoMP it becomes possible to not only achieve better performance at the cell edges, it enables also enhanced interference cancellation mechanisms to improve the overall network performance, see, [15].
- Relaying through the radio access network avoids unnecessary investments in fiber infrastructure, especially for smaller cell diameters. This is because the available LTE spectrum can be used to transmit traffic directly between eNBs via the air interface. See [16], [17], [18] for studies on coverage extension through relaying technique.
- Heterogeneous networks will become an important matter and have to be better supported in future. Radio network deployments will include macro cells, but also micro, femto and pico cells [4], [19]. In this case overlapping of different signals at the user equipment can become a serious performance limitation. Therefore, LTE-A addresses the question how interferences can be avoided. One possibility is the coordinated elimination of interferences between eNBs, the ICIC.

4. Optimizing Transport Networks for 4G

One major aspect of transport network design for LTE is to deal with the increased bandwidth due to an increase of peak and average data rate. However, another important issue is to reduce latency. The users' quality of experience is affected not only by the data rate but also by latency. In addition, low latency is an important precondition to achieve high data rates due to throttling mechanisms of TCP/IP.

The roundtrip time can be crucial for network performance and thus affects the customers' quality of service. Not only for voice, but also for data communication a low latency, or, low roundtrip time is desirable. Figure 5 shows typical round trip times of different mobile access technologies, and, as reference, of DSL access. Note that the provided values are achievable in networks in low load condition and for a server that is located near to the radio access network. It can be observed that already with HSPA and HSPA+ technology the roundtrip time is strongly improved. However, these values are again clearly outperformed by LTE technologies. Even DSL technology can no longer compete with LTE in terms of latency, at least as long as interleaving is enabled for the sake of correcting errors. Therefore, there should be no doubt that LTE technology, providing a value of 20 ms roundtrip time is suitable for providing all kinds of real-time applications to end users.

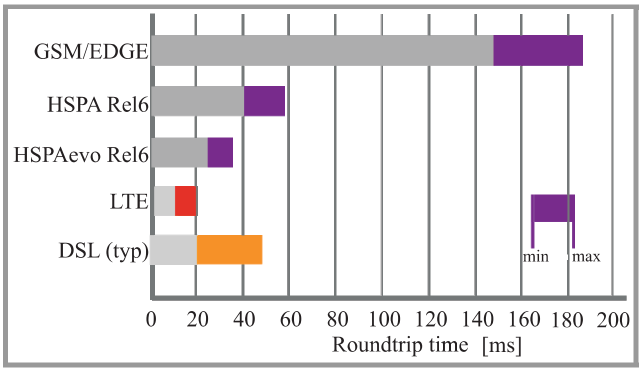


Fig. 5. Typical roundtrip times for different access technologies, server near RAN (data based on measures from Siemens, Nokia as well as [14], [20]).

We already pointed out earlier that the X2 interface, representing the logical interface between two eNBs is a very important conceptual building block of LTE, because the handover process is now controlled by the eNBs themselves. The question remains how to implement the X2 interface by means of the transport network. Before we analyze this in more details, we next provide a description of the main task of the X2 interface, namely the handover process in LTE networks. Figure 6 provides a schematic view on the handover process in LTE networks. The graphic depicts a user terminal, two eNBs and a gateway. The data streams are given as S1-u or X2-u where 'u' stands for the user

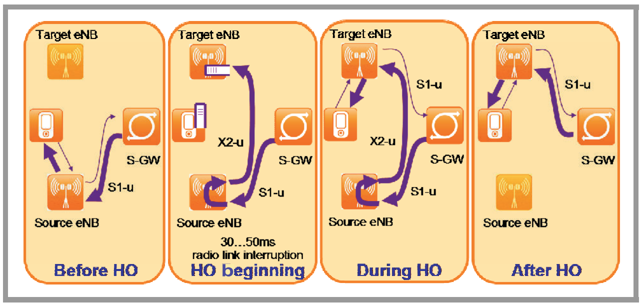


Fig. 6. Handover process in LTE networks.

plane. On the left, a situation is given where a terminal is connected to the source eNB, but is moving into the direction of another eNB. This is the starting point and denoted as the situation before the handover starts. As the terminal moves closer to the target eNB, the handover starts. For a duration of between 30 and 50 ms the radio link is interrupted as the user data is transferred via the X2 interface from the source to the target eNB. During the handover, the terminal is already connected to the target eNB via the air interface, but the target eNB still receives the user traffic via the X2 interface connected to the source eNB. Only when the handover is completed also on the MME, the SAE-GW redirects the traffic directly to the target eNB. From these observations we can conclude that it will be sufficient to have a latency of less than 30 to 50 ms on the X2 interface to maintain the service quality. This fact will be important when defining the requirements for an optimized transport network architecture for LTE.

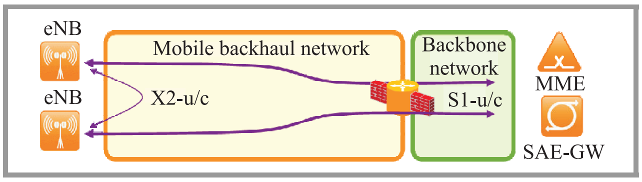


Fig. 7. Direct X2 connectivity between eNBs.

We proceed by considering backhauling in mobile networks. In general, there are two potential backhauling alternatives from a high-level point of view. These alternatives are given in Figs. 7 and 8. The notations ‘u’ and ‘c’ should indicate that we consider traffic from the user plane, as well as from the control plane. As illustrated in Fig. 7, X2 traffic can be routed directly between the eNBs. This might include packet functions in the transport network. As a consequence, we have a local meshing between the eNBs for realizing the X2 interface. As distances are short for direct X2 connectivity, we have an improved transport latency. However, to some extent, this is contradicting today’s 3G security architecture, as indicated by the firewall symbol. Typically, all traffic from and to the eNB is encrypted by means of Internet protocol security (IPSec). Today, the IPSec gateways are located centrally in the network, in order to ease operations. Thus, in the case of direct X2 connectivity, the traffic no longer passes through the central IPSec gateways. As a consequence, the security architecture would need to be adapted, too. This could be done either by decentralizing the IPSec gateways, or, by implementing a fully decentralized security architecture, where the target eNBs can decrypt traffic themselves.

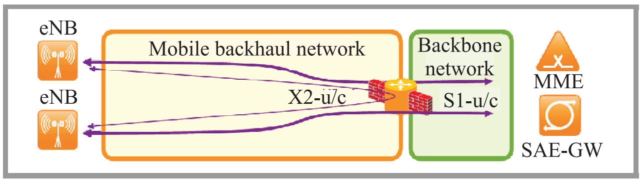


Fig. 8. Indirect X2 Connectivity Via Core Network.

Figure 8 depicts an alternative where the X2 traffic is routed via the core network and still passing through the centralized IPSec gateways. This scenario increases transport latency, but it allows to keep the current centralized security architecture. To realize this alternative, it is important to analyze whether the additional transport latency is jeopardizing the users’ quality of experience. Remember that for LTE, a roundtrip time of about 20 ms is given, see Fig. 5. On the other hand, due to the handover process, we have to deal with interruptions of the connection up to 50 ms anyway. Thus, it is fair to say that indirect X2 connectivity will not harm the quality of experience. In consequence, we can state that there is no reason and no need to implement direct X2 connectivity in case of LTE. The question remains whether this important result still holds true for LTE-A?

The specification of LTE-A is currently in an important stage [21]. New approaches have to be developed to enable 1 Gbit/s bandwidth, and at the same time, a decreased latency. Under this light, an extended usage of the X2 interface is under discussion. It is planned to design the extended X2 interface not only for the handover process, but also for information exchange in order to improve network performance. The most prominent example is the CoMP transmission where an end user terminal can receive traffic from multiple eNBs simultaneously. This approach aims to increase service quality at cell edges and to increase bandwidth.

Three CoMP-methods are under discussion:

- Coordinated scheduling / Coordinated beamforming (CS/CB),
- Joint processing / Dynamic cell selection (JP/DCS),
- Joint processing / Joint transmission (JP/JT).

Figure 9 shows, exemplary for CoMP JP/JT, how data transmission is carried out simultaneously from different eNBs. Important for the realization of CoMP is the ICIC. Some ICIC methods use the X2 interface for the exchange of information concerning interferences among the eNBs. Other methods base on a strict synchronization of eNBs, in particular if there are no X2 interfaces available in heterogeneous networks. For carrying out CoMP and ICIC, we have restrictive requirements on the transport network infrastructure, i.e., on the X2 interface bandwidth and on latencies. Simulations show very well that the lower the delay on the X2 interface, the more efficient the mechanisms work. Currently, there are latency values of about 1 ms under discussion [22].

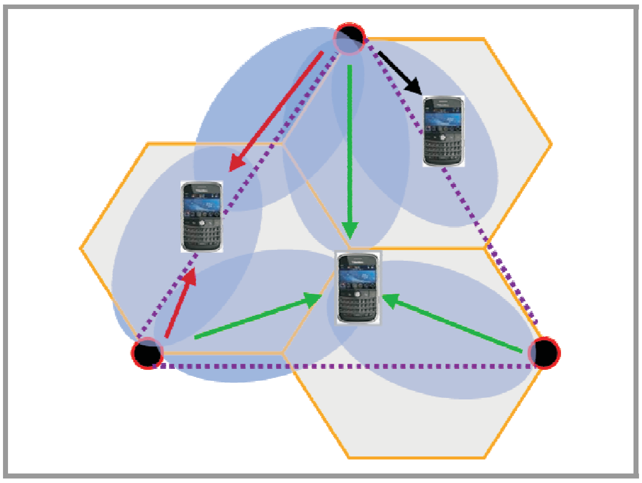


Fig. 9. Simultaneous data transmission from different eNBs to user equipment for CoMP JP/JT.

As stated before, the realization of X2 connectivity via core networks is realizable for LTE. In fact, for LTE, there is no benefit regarding latency when choosing the direct X2 connectivity. On top of that, a modified security architecture would be necessary when choosing direct X2 connections.

However, if latencies of 1 ms become standard for LTE-A, those issues will have to be reconsidered. The speed of light in the fiber provides a transport latency of 0.5 ms per 100 km. On top, processing latency has to be added for the central network element providing X2 connectivity. As a result, by rule of thumb one derives a maximum distance of 50 km between an eNB and a central network element. Thus, for implementing LTE-A, a direct X2 connectivity would become necessary, see Fig. 10. Summarizing, we could still support even very stringent latency requirements on the X2 interface by means of a direct transport connection between the eNBs. However, in this case the security architecture need to be implemented differently, as discussed previously.

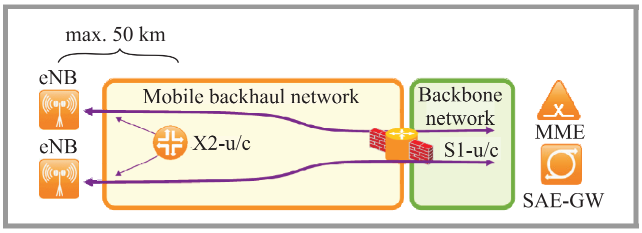


Fig. 10. CoMP may require direct X2 connectivity between eNBs due to stringent latency requirements.

Next we present an alternative approach. In today's typical deployments the NBs are located at the antenna sites. However, there are approaches of separating the more complex NB functions from their radio functions. The different functions of a base station are defined by OBSAI [23], the Open Base Station Architecture Initiative. Figure 11 shows this kind of separation of NB functions and radio functions in more details. Only the radio frequency (RF) modules of the base station are located on the antenna sites, the system modules or baseband modules are physically separated and deployed in centralized locations. Optical fiber and passive WDM technology can be deployed to transmit the OBSAI signals between RF and system modules. Such kind of separation is not only attractive from an operational perspective. This is due to the fact that more complex and error prone components of the base station can be placed centrally. Also the logical X2 interface can be implemented locally in the central locations between the system modules. It is even possible that one system module can serve multiple RF modules, so that synchronization information for CoMP is available naturally in one device.

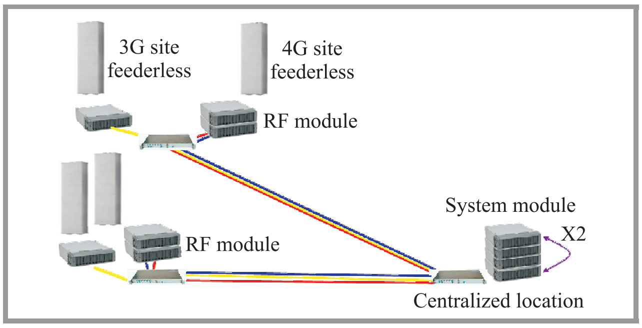


Fig. 11. Separation of RF and baseband modules.

However, this future option still requires further analysis and development. Another aspect makes the concept of functional separation interesting: if one moves towards smaller and smaller cells one might end up with femtocells, which might reside in a traffic light or a street cabinet. The reduction of size and power consumption at the antenna site will provide new flexibility in mobile network design.

5. Conclusion

LTE and LTE-A are exciting and important technologies for the future of communications. This is true not only in the case of mobility applications, but for special fixed broadband applications in the countryside, too. LTE is installed today in first locations and LTE-A is on its way regarding standardization. Technology for the user terminal, e.g., modems for the laptop or smartphones with LTE functionality is already available today. It is important to analyze the requirements of future access technologies already at an early stage, in order to optimize the underlying transport network architectures. Currently, not all requirements of LTE-A are specified, especially with respect to CoMP and the impact of the X2 interface. Our first analysis shows that the current transport network evolution strategies do not compromise any future roll-out of new broadband wireless access technologies.

References

[1] "3rd Generation Partnership Project" [Online]. Available: http://www.3gpp.org/

[2] "The advanced LTE toolbox for more efficient delivery of better user experience", *Technical White Paper*, Nokia Siemens Networks, 2011.

[3] S. Parkvall, E. Dahlman, A. Furuskär, Y. Jading, M. Olsson, S. Wänstedt, and K. Zangi, "LTE-advanced – evolving LTE towards IMT-advanced", in *Proc. IEEE Veh. Technol. Conf.*, Marina Bay, Singapore, 2008.

[4] P. E. Mogensen, T. Koivisto, K. I. Pedersen, I. Z. Kovacs, B. Raaf, K. Pajukoski, and M. J. Rinne, "LTE Advanced: the path towards gigabit/s in wireless mobile communications", in *Proc. Wireless VITAE'09*, Aalborg, Denmark, 2009, pp. 147–151.

[5] A. Kumar and Y. Liu, "LTE-Advanced: The roadmap to 4G mobile wireless networks", *Global J. Comp. Sci. Technol.*, vol. 10, no. 4, pp. 50–53, 2010.

[6] A. Ghosh, R. Ratasuk, B. Mondal, N. Mangalvedhe, and T. Thomas, "LTE-Advanced: Next-generation wireless broadband technology", *IEEE Wirel. Communi.*, vol. 17, no. 3, pp. 10–22, 2010.

[7] J. Parikh and A. Basu, "LTE Advanced: The 4G mobile broadband technology", *Int. J. Comp. Appl.*, vol. 13, no. 5, pp. 17–21, 2011.

[8] M. Fricke, A. Heckwolf, R. Herber, R. Nitsch, and S. Wevering, "Anforderungen von mobilem Breitband auf der Basis von 4G an zukünftige Transportnetze", in *Photonische Netze* (ITG-FB 228), ITG, Ed. Berlin: VDE Verlag, 2011, pp. 116–121.

[9] "2020: Beyond 4G: Radio Evolution for the Gigabit Experience", *White Paper*, Nokia Siemens Networks, 2011.

[10] "Introducing LTE Advanced", *Application Note*, Agilent Technologies, Nov. 2010.

[11] "Requirements related to technical performance for IMT-Advanced radio interface(s)", Report ITU-R M.2134, 2008.

[12] G. Yuan, X. Zhang, W. Wang, and Y. Yang, "Carrier aggregation for LTE-Advanced mobile communication systems", *IEEE Commun. Mag.*, vol. 48, no. 2, pp. 88–93, 2010.

[13] K. I. Pedersen, F. Frederiksen, C. Rosa, H. Nguyen, L. G. U. Garcia, and Y. Wang, "Carrier aggregation for LTE-Advanced: Functionality and performance aspects", *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 89–95, 2011.

[14] H. Holma and A. Toskala, *LTE for UMTS – OFDMA and SC-FDMA Based Radio Access*. Chichester: Wiley, 2009.

[15] M. Sawahashi, Y. Kishiyama, A. Morimoto, D. Nishikawa, and M. Tanno, "Coordinated multipoint transmission/reception techniques for LTE-advanced [Coordinated and Distributed MIMO]", *IEEE Wirel. Commun.*, vol. 17, no. 3, pp. 26–34, 2010.

[16] T. Beniero, S. Redana, J. Hamalainen, and B. Raaf, "Effect of relaying on coverage in 3GPP LTE-Advanced", in *Proc. IEEE Veh. Technol. Conf.*, Barcelona, Spain, 2009.

[17] E. Lang, S. Redana, and B. Raaf, "Business impact of relay deployment for coverage extension in 3GPP LTE-Advanced", in *Proc. Int. Worksh. LTE Evolution ICC 2009*, Dresden, Germany, 2009.

[18] R. Irmer and F. Diehm, "On coverage and capacity of relaying in LTE-Advanced in example deployments", in *Proc. IEEE 19th Int. Sympo. Personal, Indoor and Mobile Radio Communications*, Las Vegas, USA, 2008.

[19] A. Khandekar, N. Bhushan, J. Tingfang, V. Vanghi, "LTE-Advanced: Heterogeneous networks", in *Eur. Wirel. Conf.*, Lucca, Italy, 2010, pp. 978–982.

[20] M. Sauter, *Grundkurs Mobile Kommunikationssysteme*. Wiesbaden: Vieweg+Teubner, 2011.

[21] "3GPP TR 36.913, Requirements for further advancements for Evolved Universal Terrestrial Radio Access (E-UTRA)", (LTE Advanced), (Release 9), 2009.

[22] R. Ballentin, M. Doll, "Downlink coordinated scheduling", *VDE ITG* 5.2.4, Heidelberg, 2010.

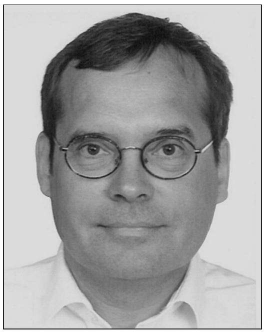
[23] "Open base station architecture initiative", *BTS System Reference Document*, Version 2.0, 2006.

Matthias Fricke has been deeply involved in network development and evolution at Deutsche Telekom in Darmstadt for more than 18 years. His current focus is the evolution of the nationwide next-generation packet optical network. Accordingly he also leads techno-economic studies concerning packet-optical transport systems and is the author and co-author of several papers on this and related topics. He holds a degree in Electrical Engineering from the Bremen University of Applied Sciences.

E-mail: matthias.fricke@telekom.de
Deutsche Telekom Netzproduktion GmbH
Fixed Mobile Engineering Deutschland
Heinrich-Hertz-Str. 3-7
64295 Darmstadt, Germany

Andrea Heckwolf had been an integral driving force at Deutsche Telekom in Darmstadt for over 30 years. Over the last 10 years her main focus was network planning and deployment. Her most recent responsibility was for the evolution and design of the nationwide SDH platform. She also led techno-economic studies on optical networks to include IP over DWDM, OTN and Gigabit Ethernet. Her final professional contribution was the analysis and development of efficient Back/Fronthauling structures for LTE Advanced technology. She sadly passed away in January 2012 after a short illness. In addition to her esteemed professional accomplishments, it was her personal approach which truly set her apart. Her unique ability to bring people and ideas together for the benefit of Deutsche Telekom, as well as the entire telecommunications industry, will be sorely missed.

E-mail: andrea.heckwolf@telekom.de
Deutsche Telekom Netzproduktion GmbH
Fixed Mobile Engineering Deutschland
Heinrich-Hertz-Str. 3-7
64295 Darmstadt, Germany



Ralf Herber started his career at Deutsche Telekom in the Research Institute FTZ in Darmstadt. He then worked for T-Systems as a senior consultant in various Telekom projects, both in the area of aggregation networks, as well as, backbone networks. In 2011, he joined Global Network Factory, the international branch of Deutsche Telekom. He works in the area of strategic development of international transport networks. He holds a Ph.D. in Electrical Engineering from the University of Technology Darmstadt, Germany.

E-mail: r.herber@telekom.de
Deutsche Telekom AG, Group Technology
Heinrich-Hertz-Str. 3-7
64295 Darmstadt, Germany

Ralf Nitsch is currently working in the area of strategic network development, network analysis and optimization at T-Systems International in Darmstadt, Germany. His main interest is the techno economical evaluation and effective use of new network technologies and is covering multi-layer aspects from optical transport up to the IP layer as well as topological and architectural aspects of aggregation and backbone networks. He holds a diploma degree and a doctorate in Applied Physics from University of Cologne, Germany.

E-mail: ralf.nitsch@t-systems.com
T-Systems International GmbH
Deutsche-Telekom-Allee 7
64295 Darmstadt, Germany

Stefan Wevering is currently working as Customer Solution Manager for the Customer Business Team Deutsche Telekom at Nokia Siemens Networks in Munich. In this position he is responsible for analyzing the major trends in telecommunication network evolution and for engineering solutions supporting those trends most effectively. His main areas of interest are IP Transformation, Fixed Mobile Convergence, Optical Transport Networks, Carrier Ethernet Technologies, Multi-Layer Optimization, and Mobile Backhauling. Formerly he worked in several different areas in Sales and R&D at Siemens Communications and Siemens Information & Communication Networks. Stefan Wevering holds a diploma degree in Physics and a Ph.D. in Applied Optics, both from the University of Osnabrück, Germany. He is an author and co-author of several international publications in renowned professional journals.

E-mail: stefan.wevering@nsn.com
Nokia Siemens Networks GmbH & Co. KG
St.-Martin-Str. 76
81541 München, Germany

Silvia Schwarze, Stefan Voß – for biographies, see this issue, p. 20.

Paper

Hierarchical Multiobjective Routing Model in MPLS Networks with Two Service Classes – A Comparison Case Study

Rita Girão-Silva[a,b], José Craveirinha[a,b], and João Clímaco[b]

[a] *Department of Electrical Engineering and Computers, University of Coimbra, Coimbra, Portugal*
[b] *Institute of Computers and Systems Engineering of Coimbra, University of Coimbra, Coimbra, Portugal*

Abstract—A two-level hierarchical multicriteria routing model for multiprotocol label switching networks with two service classes (QoS, i.e., with quality of service requirements, and best effort services) and alternative routing is reviewed in this paper. A heuristic resolution approach, where non-dominated solutions are obtained throughout the heuristic run and kept in an archive for further analysis is also reviewed. In this paper, an extensive analysis of the application of this procedure to two reference test networks for various traffic matrices is presented. Also a comparison of the results of our method with a lexicographic optimization approach based on a multicommodity flow formulation using virtual networks is carried out. Finally, results of a stochastic discrete event simulation model developed for these networks will be shown to illustrate the effectiveness of the resolution approach and to assess the inaccuracies of the analytic results.

Keywords—multiobjective optimization, routing models, simulation, telecommunication networks.

1. Introduction and Motivation

The routing calculation and optimization problems in modern multiservice networks are quite challenging, as the performance requirements in these networks are multidimensional, complex and sometimes contradictory. Routing problems in communication networks consist of the selection of a sequence of network resources (i.e., paths or routes) that will seek the optimization of some objective functions (o.f.), while satisfying a set of constraints. According to the route related metrics that are chosen, the performance of different routing decisions may be measured and quantified.

There are different classes of traffic with different service requirements in multiservice networks. With multiple and heterogeneous QoS (quality of service) routing requirements being taken into account, the routing models are designed to calculate and select one (or more) sequence of network resources (routes), with the aim of seeking the optimization of route related objectives and satisfying certain QoS constraints. There are potential advantages in formulating routing problems in these types of networks as multiple objective optimization problems, because the trade-offs among distinct performance metrics and other network cost function(s) (potentially conflicting) can be achieved in a consistent manner.

An in-depth methodological discussion of applications of multicriteria analysis in telecommunications seen from a knowledge theory broad perspective, is in [1], while [2] proposes a systematized conceptual framework for multiple criteria routing in QoS/IP networks, using a reference point-based approach.

The authors have presented a meta-model for hierarchical multiobjective network-wide routing optimization in MPLS networks in [3], along with a discussion on some key methodological and modeling issues associated with route calculation, and selection in MPLS networks. The application of this routing model framework is adequate to core or metro-core networks with a limited number of nodes. Two different classes of traffic flows are considered in this optimization approach, QoS (regarded as first priority flows) and BE – best effort (regarded as second priority flows). While the QoS flows have a guaranteed QoS level, related to the required bandwidth, the BE flows are routed with the best possible quality of service but without deteriorating the QoS of the QoS traffic flows. With this approach, the different traffic flows are treated according to their specific features. The routing model considered here is hierarchical, with two different priority levels. In the first priority level, the o.f. are concerned with network level objectives of QoS flows; in the second priority level, the o.f. are related to performance metrics for the different types of QoS services and to a network level objective for the BE traffic flows.

A heuristic approach (HMOR-S2_{PAS} – hierarchical multiobjective routing considering 2 classes of service with a Pareto archive strategy) devised to find "good" solutions (in the sense of multiobjective optimization[1]) to this hierarchical multiobjective routing optimization problem was

[1] In multiobjective optimization problems, see [4], one seeks to find non-dominated solutions since optimal (ideal) solutions are usually unfeasible. A non-dominated solution can be defined as a feasible solution such that (in minimization problems) it is not possible to decrease the value of an o.f. without increasing the value of at least one of the other o.f.

proposed in [5]. Its application to two reference test networks, $\mathscr{M}$ and $\mathscr{E}$, used in a benchmarking case study for various traffic matrices was also described. The evaluation of the performance of the proposed heuristic, by using an analytical model and stochastic discrete-event simulation was presented.

In this work, the heuristic approach HMOR-S2$_{\text{PAS}}$ is applied to two new networks, denoted by $\mathscr{G}$ and $\mathscr{H}$, obtained by a transformation of an original network in [6], by a redimensioning of the links. An extensive analysis of the application of this procedure to these networks for various traffic matrices is presented. A major objective of this experimental study is to test the developed routing method in new networks with different structure and increased connectivity, as compared with the ones in [5]. Furthermore, the results were obtained in these networks, using analytic and stochastic discrete-event simulation models in order to confirm the effectiveness of this heuristic approach to route calculation and selection in multiservice networks, and to assess the inaccuracies of the analytic results.

Furthermore, for comparison purposes we also implemented a network-wide optimization routing method based on a MCF (multicommodity flow) programming approach with two-path traffic splitting, using lexicographic optimization for dealing with the two main o.f. associated with QoS and BE traffic. This routing method (designated hereafter as MCF-lex-W) is a particular variant of the one proposed in [7] and from our point of view, this type of approach (among the ones previously developed) can be broadly comparable in terms of underlying objectives to our approach. This type of alternative method uses the concept of virtual residual networks whereby, in a first step, the routing calculation is performed for the QoS traffic (seeking to optimize a relevant o.f.) and in a second step the routing calculation for the BE traffic is performed considering only the remaining capacity in the links (resulting from the occupation of the QoS flows). This results in a virtual residual network and it is a classical form of dealing with routing problems in networks with two classes of services of different priority, as in [8], [9]. Since this type of models using MCFs assume deterministic flows (this is an intrinsic limitation of this type of approaches), the comparison with the results of our multiobjective model requires an adaptation to a stochastic environment of the type proposed in [7] and adapted to the developed models as described in Subsection 3.3.

The paper is organized as follows: the two-level hierarchical multiobjective alternative routing model with two service classes is reviewed in Section 2. The main features of the heuristic resolution approach are also reviewed. In the following section, after an explanation on the application of the model to a network case study and the description of the test networks considered in the experimental study, the MCF-lex-W method used for comparison purposes is described. Still in Section 3 the results obtained with this procedure by using analytic results and discrete-event stochastic simulations, for the two new test networks, considering three load scenarios are presented. The paper ends with a section on conclusions and an outline of future work.

2. Review of the Multiobjective Routing Model and the Heuristic Resolution Approach

In this section we will make a review of the essential aspects of the multiobjective routing model and of the heuristic resolution approach. Due to the complex nature of the model and of the resolution approach, we refer the readers to further details in [3].

2.1. The Multiobjective Routing Model

The model described here is an application of the multiobjective modeling framework (or "meta-model") for MPLS networks proposed in [3]. In this model, two classes of services are considered: QoS and BE. The sets $\mathscr{S}_Q$ and $\mathscr{S}_B$ include the different service types of each class, that may differ in important attributes, namely the required bandwidth.

The network is represented in this model through a capacitated directed graph, with an assigned capacity of C_k to every arc (or 'link') $k \in \mathscr{A}$. The traffic flows are represented in a stochastic form, based on the use of the concept of effective bandwidth[2] for macro-flows and on a generalized Erlang model for estimating the blocking probabilities in the arcs, as in the model used in [12].

A traffic flow is specified by $f_s = (i, j, \overline{\gamma}_s, \overline{\eta}_s)$ for $s \in \mathscr{S} = \mathscr{S}_Q \cup \mathscr{S}_B$ and a stochastic process (usually, a marked point process) is assigned to it. This process describes the arrivals and basic requirements of micro-flows[3], originated at the MPLS ingress node i and destined for the MPLS egress node j, using some LSP (label switched path). The characteristics of the traffic flows are expressed by $\overline{\gamma}_s$, the vectors of traffic engineering attributes of flows of service type s, and by $\overline{\eta}_s$, the vectors containing the description of mechanisms of admission control to all arcs k in the network by calls of flow f_s. The traffic engineering attributes associated with f_s calls and all the links, which may be used by f_s, including priority features, include information on the required effective bandwidth d_s and the mean duration $h(f_s)$ of each micro-flow in f_s.

[2] The effective bandwidth can be defined (see [10]) as the minimum amount of bandwidth that can be assigned to a flow or traffic aggregate in order to deliver 'acceptable service quality' to the flow or traffic aggregate. This concept may be used to approximate nodal behavior at the packet level and simplify the analysis at the connection level. Kelly [11] developed a formal mathematical definition of effective bandwidth in a network with stochastic traffic sources and statistical multiplexing. According to this definition, the effective bandwidth can be viewed as a specific stochastic measure of the utilization of transmission network resources by certain packet flow(s). With this concept, the traffic behavior at packet level may be "encapsulated" in a simplified manner.

[3] A micro-flow corresponds in this model to a 'call', that is, a node to node connection request with certain traffic engineering features.

The hierarchical multiobjective routing optimization model considered here has two levels with several o.f. in each level. At the first level, the first priority o.f. include W_Q, the total expected network revenue associated with QoS traffic flows, and $B_{Mm|Q}$, the worst average performance among QoS services, represented by the maximal average blocking probability among all QoS service types. These o.f. are formulated at the network level for the QoS traffic. At the second level, the second priority o.f. include $B_{ms|Q}$, the mean blocking probabilities for flows of type $s \in \mathscr{S}_Q$, and $B_{Ms|Q}$, the maximal blocking probability defined over all flows of type $s \in \mathscr{S}_Q$, as well as the total expected network revenue associated with BE traffic flows, W_B. The o.f. related to blocking probabilities in this second level are average performance metrics of the QoS traffic flows associated with the different types of QoS services. At both levels of optimization, 'fairness' objectives are explicitly considered in the form of min-max objectives: $\min_{\overline{R}}\{B_{Mm|Q}\}$ at the first level, and $\min_{\overline{R}}\{B_{Ms|Q}\}, \forall s \in \mathscr{S}_Q$ at the second level.

Hence the considered two-level hierarchical optimization problem for two service classes P-M2-S2 ('problem – multiobjective with 2 optimization hierarchical levels – with 2 service classes') is:

Problem P-M2-S2

- 1^{st} level $\left\{\begin{array}{ll} \text{QoS: Network obj.} & \max_{\overline{R}}\{W_Q\}, \\ & \min_{\overline{R}}\{B_{Mm|Q}\}; \end{array}\right.$

- 2^{nd} level $\left\{\begin{array}{ll} \text{QoS: Service obj.} & \min_{\overline{R}}\{B_{ms|Q}\}, \\ & \min_{\overline{R}}\{B_{Ms|Q}\}, \\ & \forall s \in \mathscr{S}_Q, \\ \text{BE: Network obj.} & \max_{\overline{R}}\{W_B\}; \end{array}\right.$

subject to equations of the underlying traffic model, with

$$W_{Q(B)} = \sum_{s \in \mathscr{S}_{Q(B)}} A_s^c w_s\,, \tag{1}$$

$$B_{Mm|Q} = \max_{s \in \mathscr{S}_Q} \{B_{ms}\}\,, \tag{2}$$

$$B_{ms|Q} = \frac{1}{A_s^o} \sum_{f_s \in \mathscr{F}_s} A(f_s) B(f_s)\,, \tag{3}$$

$$B_{Ms|Q} = \max_{f_s \in \mathscr{F}_s} \{B(f_s)\}\,, \tag{4}$$

where A_s^o is the total traffic offered by flows of type s, A_s^c is the carried traffic for service type s, $A(f_s)$ is the mean traffic offered associated with f_s (in Erlang), $B(f_s)$ is the node to node blocking probability for all flows f_s, and w_s is the expected revenue per call of service type s. For further details on the calculation of these o.f. see [3].

There are possible conflicts between the o.f. in P-M2-S2. In fact, in many routing situations, the maximization of W_Q may cause a deterioration on some $B(f_s), s \in \mathscr{S}_Q$, for certain traffic flows $A(f_s)$ with low intensity, which tends to increase $B_{Ms|Q}$ and $B_{ms|Q}$, and consequently $B_{Mm|Q}$. This justifies the interest and potential advantage in using multiobjective formulations in this context.

It is important to remark that in the formulation of P-M2-S2, W_Q is a first priority o.f. (together with $B_{Mm|Q}$), while W_B is a second level o.f. This formulation assures that the routing of BE traffic, in a quasi-stationary situation, will not be made at the expense of a decrease in QoS traffic revenue or of an increase in the maximal blocking probability of QoS traffic flows.

The traffic modeling approach used in the routing model is fully described in [3]. In the framework of the basic teletraffic model considered here, the blocking probabilities B_{ks}, for micro-flows of service type s in link k, are calculated by

$$B_{ks} = \mathscr{B}_s\left(\overline{d_k}, \overline{\rho_k}, C_k\right)\,, \tag{5}$$

with $\mathscr{B}_s$ representing the basic function (implicit in the teletraffic analytical model) that expresses the marginal blocking probabilities, B_{ks}, in terms of $\overline{d_k} = (d_{k1}, \ldots,\ d_{k|\mathscr{S}|})$ (vector of equivalent effective bandwidths d_{ks} for all service types), $\overline{\rho_k} = \left(\rho_{k1}, \ldots, \rho_{k|\mathscr{S}|}\right)$ (vector of reduced traffic loads ρ_{ks} offered by flows of type s to k) and the link capacity C_k. For simplifying purposes, the links are modeled through a multidimensional Erlang system with multirate Poisson traffic inputs. With this type of approximation, the calculation of $\{B_{ks}\}$ can be performed through efficient and robust numerical algorithms, which are essential in a network-wide routing optimization model of this type, for tractability reasons. The classical Kaufman (or Roberts) algorithm [13], [14] was used to calculate the functions $\mathscr{B}_s$ for small values of C_k; for larger values of C_k, approximations based on the uniform asymptotic approximation (UAA) [15] were used, having in mind its efficiency.

The decision variables $\overline{R} = \cup_{s=1}^{|\mathscr{S}|} R(s)$ represent the network routing plans, that is, the set of all the feasible routes (i.e., node to node loopless paths) for all traffic flows, with $R(s) = \cup_{f_s \in \mathscr{F}_s} R(f_s), s \in \mathscr{S}_Q \cup \mathscr{S}_B$ and $R(f_s) = (r^p(f_s)), p = 1, \ldots, M$ with $M = 2$ in this model. An alternative routing principle is used: for each flow f_s the first choice route $r^1(f_s)$ is attempted and if it is blocked the call will try the second choice route $r^2(f_s)$. A request will be blocked only if $r^2(f_s)$ is also blocked.

This routing optimization approach is of network-wide type, which means that the main o.f. of a given service class depend explicitly on all traffic flows in the network. Therefore, a full representation of the relations between the o.f. is achieved, taking into account the interactions between the multiple traffic flows associated with different services. This is accomplished by the features of the traffic model used to obtain the blocking probabilities $B(f_s)$, as the contributions of all traffic flows, which may use every link of the network are considered according to the approach in [3]. The focus is on the routing optimization from a global perspective (i.e., considering an explicit representation of all the traffic flows in the network and their interactions), which is the closest to reality. This is a major difference in comparison with other routing models that have been proposed for networks with two service classes, based on some form of decomposition of the network representation, leading to the consideration of 'virtual networks', one for each service class (e.g. in [7]).

The routing problem P-M2-S2 is highly complex, mainly because of two factors: all o.f. are strongly interdependent (via the $\{B(f_s)\}$), and all the o.f. parameters and (discrete) decision variables $\overline{R}$ (network route plans) are also interdependent. All these interdependencies are defined explicitly or implicitly through the underlying traffic model. Even in a simplest degenerated case, considering single service with single-criterion optimization and no alternative routing, the problem is NP-complete in the strong sense (see [16]). Considering the form of P-M2-S2, one may conclude on the great intractability of this problem.

2.2. The Heuristic Resolution Approach

The heuristic procedure HMOR-S2$_{\text{PAS}}$ (fully described in [5] and references therein) used to solve (in a multicriteria analysis sense) the routing problem P-M2-S2 is reviewed here. Using the theoretical foundations described in [17], this heuristic is based on the recurrent calculation of solutions to an auxiliary constrained bi-objective shortest path problem $\mathscr{P}_{s2}^{(2)}$, formulated for every end-to-end flow f_s,

$$\min_{r(f_s)\in\mathscr{D}(f_s)} \left\{ m^n(r(f_s)) = \sum_{k\in r(f_s)} m_{ks}^n \right\}_{n=1;2} .$$

The path metrics m^n to be minimized are the marginal implied costs[4] $m_{ks}^1 = c_{ks}^{Q(B)}$ and the marginal blocking probabilities $m_{ks}^2 = -\log(1-B_{ks})$; $\mathscr{D}(f_s)$ is the set of all feasible loopless paths for flow f_s, satisfying specific traffic engineering constraints for flows of type s. The efficiency of different candidate routes can be compared, considering both path metrics: the loss probabilities experienced along the candidate routes and the knock-on effects upon the other routes in the network (effects related to the acceptance of a call on that given route). It is important to remark that these network metrics are associated with the first level o.f. of P-M2-S2: the minimization of the metric blocking probability tends, at a network level, to minimize the maximal node-to-node blocking probabilities $B(f_s)$, while the minimization of the metric implied cost tends to maximize the total average revenue W_T.

In the heuristic, the auxiliary constrained shortest path problem $\mathscr{P}_{s2}^{(2)}$ is solved by the algorithm MMRA-S2 (modified multiobjective routing algorithm for multiservice networks, considering 2 classes of service) described in [18]. Generally, there is no feasible solution minimizing the two o.f. simultaneously. Therefore, the aim of the resolution of this problem is finding a 'best' compromise path from the set of non-dominated solutions, according to a system of preferences embedded in the working of the algorithm MMRA-S2. The implementation of this system of preferences relies on the definition of preference regions in the o.f. space obtained from aspiration and reservation levels (preference thresholds), defined for the two o.f.

The generation and selection of candidate solutions $(r^1(f_s), r^2(f_s))$ by MMRA-S2 for each f_s is based on rules that consider the network topology and the need to make a distinction between real time and non-real time QoS services, and BE services. An instability phenomenon may arise in the path selection procedure, as shown by a theoretical analysis of the model and confirmed by extensive experimentation: the route sets $\overline{R}$ (obtained by successive application of MMRA-S2 to every flow f_s) often tend to oscillate between certain solutions, some of which may lead to poor global network performance under the prescribed metrics. To avoid this instability, not all the paths of all the flows are liable to change on each iteration. A set of candidate paths for possible routing improvement are chosen by increasing order of a function $\xi(f_s)$ of the current $(r^1(f_s), r^2(f_s))$, as proposed in [18]. With this function $\xi(f_s)$ preference (concerning the calculation of new routes) is given to the flows, for which the route $r^1(f_s)$ has a low implied cost, and the route $r^2(f_s)$ has a high implied cost or to the flows, which currently have worse end-to-end blocking probability. A variation on the selected paths is performed, leaving the others unaltered.

In the dedicated heuristic HMOR-S2, each new solution is obtained by 'processing' the current best solution: routing solutions $R(s)$ for each service $s\in\mathscr{S}$ are sought which dominate the current one in terms of the so-called o.f. of interest for the service (the first level o.f. and the second level o.f. $B_{ms|Q}$ and $B_{Ms|Q}$ if $s\in\mathscr{S}_Q$, or W_B if $s\in\mathscr{S}_B$). This strategy leads to strict limitations being imposed on the acceptance of a new solution, and consequently some interesting solutions to the routing problem may not be further pursued. Therefore, instead of simply discarding every solution that does not dominate the current one, we have devised the PAS variant where some possibly interesting solutions are stored throughout the execution of the heuristic, and later checked in order to try and find the "best" possible solution to the problem in hand. The management rules of the archive (that is, addition and removal of solutions from the archive) and the evaluation of the solutions stored in the archive after the end of the outer cycle of the algorithm (in order to choose the "best" possible solution to the problem under analysis) are fully described in [5].

The analysis of the solutions stored in the archive relies on a system of priority regions in the bidimensional o.f. space, defined by preference thresholds (*requested* (or aspirational) and *acceptable* (or reservation) thresholds for each network function W_Q and $B_{Mm|Q}$). As an example of the definition of priority regions in the bidimensional o.f. space of the solutions in the archive, see Fig. 1.

The ideal optimum is represented by $O*$ and is obtained when both first level o.f. W_Q and $B_{Mm|Q}$ are optimized.

[4] The *marginal implied cost* for QoS(BE) traffic, $c_{ku}^{Q(B)}$, associated with the acceptance of a connection (or "call") of traffic f_u of any service type $u\in\mathscr{S}$ on a link k represents the expected value of the traffic loss induced on all QoS(BE) traffic flows resulting from the capacity decrease in link k (see [17]).

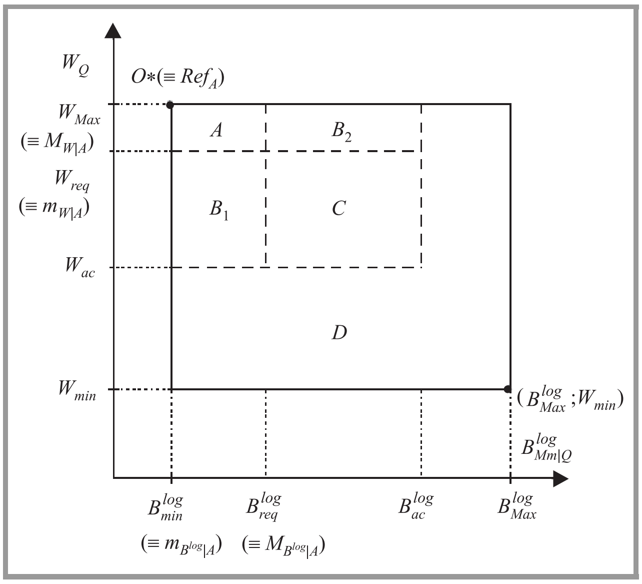


Fig. 1. QoS requirements used to define priority regions in the bidimensional o.f. space.

The region, for which the requested levels are satisfied for both o.f. is the first priority region A; the regions, for which only one of the requested values is satisfied and an acceptable value is guaranteed for the other metric are the second priority regions B_1 and B_2 (note that B_2 will be considered preferable to B_1 because, for solutions in any second priority region, preference is given to the one with greater W_Q even if with greater $B_{Mm|Q}$); the region where only acceptable values are guaranteed for both metrics is the third priority region C. Beyond the acceptable values, there lies the least priority region D. The preference thresholds used to define the priority regions are calculated in a fully automated manner (see [5]).

The approach chosen to select the "best" solution in the best possible priority region relies on the minimization of a weighted Chebyshev distance to a reference point. In this approach, reference (aspiration and reservation) levels are specified for each criterion. Let $W_{av} = \frac{W_{min}+W_{max}}{2}$, where $W_{min}(W_{max})$ is the minimal(maximal) value of W_Q in all the solutions in the archive, and $B_{av} = \frac{B_{min}+B_{max}}{2}$, where $B_{min}(B_{max})$ is the minimal(maximal) value of $B_{Mm|Q}$ in all the solutions in the archive. The reference levels are defined by $W_{req} = \frac{W_{av}+W_{max}}{2}$ and $W_{ac} = \frac{W_{min}+W_{av}}{2}$ for the QoS traffic revenue and $B^{log}_{req} = -\log\left(1 - \frac{B_{min}+B_{av}}{2}\right)$ and $B^{log}_{ac} = -\log\left(1 - \frac{B_{av}+B_{max}}{2}\right)$ for the blocking probability $B_{Mm|Q}$. The weighted Chebyshev distance of a non-dominated solution in a given preference region to the associated aspiration point is calculated, and the "best" solution will be the one in the best possible priority region that minimizes that distance.

Defining $\mathscr{R}$ as the best possible priority region in the o.f. space where at least one solution ρ can be found, a specific reference point $\left(\mathscr{C}^*_{1|\mathscr{R}};\mathscr{C}^*_{2|\mathscr{R}}\right)$ can be chosen in $\mathscr{R}$ as the ideal point in that region. The ideal point in each rectangular region is the top left corner of that region. As an example, see the reference point for region A (Ref_A) in Fig. 1. For a non-rectangular region such as D, the reference point is the ideal point of the whole o.f. space $O*$.

Other parameters that must be defined are the minimum $m_{i|\mathscr{R}}$ and maximum $M_{i|\mathscr{R}}$ values of each metric i for region $\mathscr{R}$. As an example, see the minimum and maximum values for both metrics in region A in Fig. 1.

The problem of selection of the final solution considers a weighted Chebyshev norm:

$$\min_{\rho\in\mathscr{R}} \max_{i=1,2} \left\{ w_{i|\mathscr{R}} \left| \mathscr{C}_i(\rho) - \mathscr{C}^*_{i|\mathscr{R}} \right| \right\},$$

where $\mathscr{C}_1(\rho) = B^{log}_{Mm|Q}(\rho)$ and $\mathscr{C}_2(\rho) = W_Q(\rho)$ are the metrics for solution ρ. The weights in the weighted Chebyshev distance, $w_{i|\mathscr{R}} = \frac{1}{M_{i|\mathscr{R}}-m_{i|\mathscr{R}}}$, allow the Chebyshev metrics $\left\{ w_{i|\mathscr{R}} \left| \mathscr{C}_i(\rho) - \mathscr{C}^*_{i|\mathscr{R}} \right| \right\}$ to be dimension free and proportional to the size of the rectangular region. This weighted Chebyshev norm is more adequate to the adopted technique of search and selection of non-dominated solutions in rectangular preference regions. In fact, the use of the weights (as defined in the method) makes the contour of the rectangle a isocost Chebyshev line for each particular region.

3. Experimental Results

3.1. Application of the Model to a Network Case Study

The network case study considered here is obtained from changes on the network models in [7] and [6]. An overview of the relevant features of the model proposed in this reference is provided here for a better understanding of the case study.

In [7], a model is proposed for traffic routing and admission control in multiservice, multipriority networks supporting traffic with different QoS requirements. Deterministic models are used in the calculation of paths, in particular mathematical programming models based on MCFs, rather than stochastic traffic models. The MCF models are only a rough approximation in this context and, in fact, they tend to under-evaluate the blocking probabilities. Therefore, the authors of [7] propose an adaptation of the original model, so as to obtain 'corrected' models, which provide a better approximation in a stochastic traffic environment. A simple technique to adapt the MCF model to a stochastic environment is the compensation of the requested values of the flows bandwidths in the MCF model with a factor $\alpha \geq 0.0$. With this compensation technique, the effect of the random fluctuations of the traffic that are typical of stochastic traffic models can be modeled. The higher the variability of the point processes of the stochastic model, the higher is the need for compensation and therefore the higher should α be. In the application example in [7], three values of α are proposed: $\alpha = 0.0$ corresponds to the deterministic approach; $\alpha = 0.5$ is the compensation parameter when calls arrive according to a Poisson process, service times follow an exponential distribution and the network is critically loaded; and $\alpha = 1.0$ for traffic flows with higher 'variability'.

The o.f. of the routing problems in [7] are the revenues W_Q and W_B, associated with QoS and BE flows, which should be maximized. A bi-criteria lexicographic optimization formulation is considered, so that the improvements in W_B are to be found under the constraint that the optimal value of W_Q is maintained.

In the deterministic flow-based model [7], a base matrix $T = [T_{ij}]$ with offered bandwidth values from node i to node j [Mbit/s] is given. A multiplier $m_s \in [0.0; 1.0]$ with $\sum_{s \in \mathscr{S}} m_s = 1.0$ is applied to these matrix values to obtain the offered bandwidth of each flow f_s of service type s to the network. In our stochastic traffic model, a matrix of offered traffic $A(f_s)$ is obtained by transforming the base matrix T:

$$A(f_s) \approx \frac{m_s T_{ij}}{d_s u_0} - \alpha \sqrt{\frac{m_s T_{ij}}{d_s u_0}} \text{ [Erl]}, \tag{6}$$

if $\frac{m_s T_{ij}}{d_s u_0} > \alpha^2$ and both $T(f_s) = m_s T_{ij}$ and $A(f_s)$ are high. Otherwise,

$$A(f_s) \approx \frac{m_s T_{ij}}{d_s u_0} \text{ [Erl]}, \tag{7}$$

where u_0 is a basic unit of transmission [bit/s].

In the original traffic routing model in [7], traffic splitting is used. This technique is not used in the model considered here.

3.2. Application of the Model to Two Different Test Networks

The routing model was applied to the test networks $\mathscr{G}$ and $\mathscr{H}$, for which the topology is depicted in Fig. 2. It has $|\mathscr{N}| = 10$ nodes, with 16 pairs of nodes linked by a direct arc and a total of $|\mathscr{A}| = 32$ unidirectional arcs, which means their average node degree is $\delta = 3.2$. As their average node degree is higher, these two networks $\mathscr{G}$ and $\mathscr{H}$ have more connectivity than networks $\mathscr{M}$ and $\mathscr{E}$ ($\delta_{\mathscr{M}} = 2.5$ and $\delta_{\mathscr{E}} = 2.4$), studied in [5]. Each bandwidth C'_k [Mbit/s] of each arc k for each of the networks is shown in Tables 1 and 2, and it was obtained by employing a very simple network dimensioning algorithm, explained below.

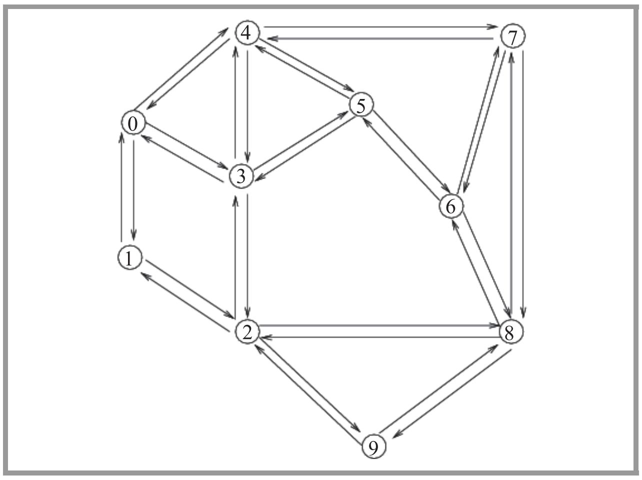


Fig. 2. Network topology for test networks $\mathscr{G}$ and $\mathscr{H}$ [6].

The test networks $\mathscr{G}$ and $\mathscr{H}$ were obtained after a redimensioning of the original network $\mathscr{O}$ given in [6]. This network $\mathscr{O}$ has a topology similar to the one in Fig. 2, with a capacity of $C'_k = 50$ Mbit/s for each arc, which is an equivalent to a capacity of $C_k = \frac{C'_k}{u_0} = 3125$ channels, as $u_0 = 16$ kbit/s. The offered traffic matrix is also provided in [6]. A routing solution using only shortest path direct routing, typical of Internet conventional routing algorithms is taken into account. In this routing solution, only one path for each flow (i.e., without an alternative path) is considered. The initial solution is the same for all services $s \in \mathscr{S}$ and the unidirectional paths for any given pair of nodes are symmetrical. The path for every flow f_s is the shortest one (that is, the one with minimum number of arcs); if there is more than one shortest path, the one with maximal bottleneck bandwidth (i.e., the minimal capacity of its arcs) is chosen; if there is more than one shortest path with equal bottleneck bandwidth, the choice is arbitrary.

The dimensioning of link capacities was made as follows. A value β_s for the mean blocking probabilities for flows of type s, B_{ms}, is defined with a possible variation of Δ_B. The matrix of offered traffic $A(f_s)$ is obtained from the traffic matrix T in [6] with $\alpha = 0.0$ (the value of α for which the load is higher). Considering the routing solution for network $\mathscr{O}$, the mean blocking probabilities B_{ms} are calculated and compared with the prescribed values at the beginning of the algorithm. If $B_{ms} > \beta_s$ for service s, then the links in paths for flows of service s will have their capacity increased; if $B_{ms} < \Delta_B \beta_s$ for service s, then the links in paths for flows of service s will have their capacity decreased. The algorithm proceeds iteratively until it converges (i.e., $\Delta_B \beta_s < B_{ms} < \beta_s, \forall s \in \mathscr{S}$). In some of the performed experiments, the algorithm oscillated between two different solutions, which prevented it from converging. Therefore, a maximum number of runs was also established, so as to avoid this situation.

The test networks $\mathscr{G}$ and $\mathscr{H}$ were dimensioned using this very simple network dimensioning algorithm, for $\beta_s = 0.1$ and $\beta_s = 0.12$ respectively, with $\Delta_B = 0.9$. This means that a situation of very high blocking, associated with traffic overload for all services, was considered (for $\alpha = 0.0$) in the dimensioning operation. The aim was a comparison of the performance of the considered static routing methods in overload conditions ($\alpha = 0.0$) and in low, and very low blocking conditions for the QoS traffic for $\alpha = 0.5$ and $\alpha = 1.0$. The original network $\mathscr{O}$ was not used in this study because it was dimensioned for extremely low blocking probabilities.

The traffic matrix $T = [T_{ij}]$ with offered total bandwidth values from node i to node j [Mbit/s] provided in [6] is used as an input to the routing model considered here. The routing model and other features proposed by [6] were not taken into account.

Table 1
Bandwidth of each arc C'_k, in Mbit/s, for the test network $\mathscr{G}$

↱	0	1	2	3	4	5	6	7	8	9
0		40.64		44.384	40.64					
1	40.64		35.024							
2		35.024		35.024					36.896	38.768
3	44.384		35.024		42.512	38.768				
4	40.64			42.512		44.384		40.64		
5				38.768	44.384		38.768			
6						38.768		46.256	40.64	
7					40.64		46.256		38.768	
8			36.896				40.64	38.768		44.384
9			38.768						44.384	

Table 2
Bandwidth of each arc C'_k, in Mbit/s, for the test network $\mathscr{H}$

↱	0	1	2	3	4	5	6	7	8	9
0		39.6		43.76	39.6					
1	39.6		33.36							
2		33.36		33.36					35.44	37.52
3	43.76		33.36		41.68	37.52				
4	39.6			41.68		43.76		39.6		
5				37.52	43.76		37.52			
6						37.52		45.84	39.6	
7					39.6		45.84		37.52	
8			35.44				39.6	37.52		43.76
9			37.52						43.76	

For both networks, the number of channels C_k is $C_k = \left\lceil \frac{C'_k}{u_0} \right\rceil$, with basic unit capacity $u_0 = 16$ kbit/s. There are $|\mathscr{S}| = 4$ service types with the features displayed in Table 3.

Table 3
Service features on the test networks $\mathscr{G}$ and $\mathscr{H}$

Service	Class	d'_s [kbit/s]	d_s [channels]	w_s	h_s [s]	D_s [arcs]	m_s
1 – video	QoS	640	40	40	600	3	0.1
2 – Premium data	QoS	384	24	24	300	4	0.25
3 – voice	QoS	16	1	1	60	3	0.4
4 – data	BE	384	24	24	300	9	0.25

The values of the required bandwidth d'_s in kbit/s are also in the table. The expected revenues for calls of type s, w_s, are equal to the required effective bandwidths $d_s = \frac{d'_s}{u_0}$ [channels]: $w_s = d_s, \forall s \in \mathscr{S}$. The average duration of a type s call is h_s and the maximum number of arcs for a type s call is D_s.

3.3. Routing Method Used for Comparison Purposes

Next we describe the MCF-lex-W routing model, based on MCFs with lexicographic optimization and considering two-path traffic splitting. This model is based on the one in [7] and it is used as an alternative benchmarking method for comparison with our multiobjective model. From a theoretical point of view, and considering the conceptual framework developed in [3], this type of model is also a network-wide optimization approach with features that make it an adequate alternative method for a ‘fair’ comparison with our model. The results with the method HMOR-S2$_{\text{PAS}}$ considered in this paper are compared with results from this routing procedure MCF-lex-W.

In this routing procedure, we first seek to route the QoS traffic flows in the given network. Next, we seek to route the BE traffic flows in a virtual network, where the arcs of the original network have a reduced capacity given by the original arc capacity minus the capacity used in the routing of the QoS flows. In the process of routing calculation, the aim is the maximization of the revenue of QoS and BE carried traffic (represented by the node-to-node offered bandwidth), using a lexicographic optimization approach. Traffic splitting is allowed, in situations where it is advantageous. There is the possibility of dividing the required bandwidth of each flow by multiple paths from source to destination, allowing for a better load distribution in the network. If the network is unable to accommodate all the traffic that is offered, a technique of admission control based on traffic thinning can be used, as proposed in [7].

Considering a traffic flow of the service s represented by f_s, originated at the MPLS ingress node i and destined for the MPLS egress node j, the bandwidth offered by that flow to the network is $T(f_s) = m_s T_{ij}$, as mentioned in Subsection 3.1. For each flow, a set of $L(f_s)$ feasible paths may be obtained, $\mathscr{L}(f_s) = \{p^0(f_s), p^1(f_s), \cdots, p^{L(f_s)-1}(f_s)\}$. Of all the possible paths between i and j, the ones with a number of arcs inferior to D_s (maximal number of arcs established for service s calls) are feasible. In the implemented model, the total bandwidth offered by flow f_s may be divided by $N_L = 2$ of these feasible paths, allowing for the possibility of traffic splitting. Let us define $x^l(f_s)$ as the amount of bandwidth of f_s that will be offered to the l-th path $p^l(f_s)$, and $y^l(f_s)$ as a binary variable, which is equal to 1 if the l-th path is actually used and 0 otherwise. Therefore, the following conditions have to be met, $\forall f_s \in \mathscr{F}_s, s \in \mathscr{S}$:

$$\sum_{l=0}^{L(f_s)-1} x^l(f_s) \leq T(f_s)\,, \tag{8}$$

$$0 \leq x^l(f_s) \leq T(f_s), \quad \forall l = 0, \cdots, L(f_s) - 1\,, \tag{9}$$

$$x^l(f_s) \leq T(f_s) y^l(f_s), \quad \forall l = 0, \cdots, L(f_s) - 1\,, \tag{10}$$

$$\sum_{l=0}^{L(f_s)-1} y^l(f_s) \leq N_L = 2\,. \tag{11}$$

The o.f. used in this routing method is the maximization of the network revenue $W_T = \sum_{s\in\mathscr{S}} \sum_{f_s\in\mathscr{F}_s} \sum_{l=0}^{L(f_s)-1} w_s x^l(f_s)$ that results from carrying the bandwidth offered by all the traffic flows to all the feasible paths, which are actually used. The possibility of traffic splitting should provide a flexible distribution of the load in the network, so as to maximize the carried traffic. This is particularly relevant in the context of this routing model since, after establishing the optimal routes of the QoS traffic (for which the whole average bandwidth demand is satisfied), it is necessary to calculate the routes for the BE traffic in the virtual residual network, so as to maximize the BE carried traffic.

The type of problem to be solved in this routing procedure is

Problem P-MCF-lex-W$_{\mathscr{S}}$

$$\max \left\{ \sum_{s\in\mathscr{S}} \sum_{f_s\in\mathscr{F}_s} \sum_{l=0}^{L(f_s)-1} w_s x^l(f_s) \right\}$$

subject to conditions (8)–(11) and

$$v_k \leq C'_k, \forall k \in \mathscr{A}\,,$$

$$v_k = \sum_{s\in\mathscr{S}} \sum_{f_s\in\mathscr{F}_s} \sum_{l=0}^{L(f_s)-1} a^l_k(f_s) x^l(f_s), \forall k \in \mathscr{A}\,,$$

where $a^l_k(f_s)$ is a binary variable equal to 1 if the link k belongs to $p^l(f_s)$, the l-th path for flow f_s, and 0 otherwise. The parameter v_k is the total load carried in each arc $k \in \mathscr{A}$.

The routing calculation approach in the case where QoS and BE traffic classes coexist uses a lexicographic formulation as the one in [7].

Firstly, the problem P-MCF-lex-$W_{\mathscr{S}_Q}$ is solved, and only the QoS traffic is considered. As a result, the values $x^l(f_s), \forall l = 0, \cdots, L(f_s) - 1, f_s \in \mathscr{F}_s, s \in \mathscr{S}_Q$ are obtained, which give the amount of bandwidth that is routed in each of the feasible paths for each of the QoS flows. Also, as a result of this problem, an information on v_k is obtained. Let this load be represented by $v_{k(Q)}$.

Secondly, the problem P-MCF-lex-$W_{\mathscr{S}_B}$ is solved, that is, only the BE traffic is considered. In this second problem, a virtual network consisting of the same links but with residual capacities $C'_k - v_{k(Q)}, \forall k \in \mathscr{A}$ is considered. The possibility of BE traffic thinning was considered, as the network has a reduced arc capacity and there is the possibility that not all the BE traffic flows may be carried.

After the resolution of the second problem, the values $x^l(f_s), \forall l = 0, \cdots, L(f_s) - 1, f_s \in \mathscr{F}_s, s \in \mathscr{S}_B$ are obtained, which gives us the amount of bandwidth that is routed in each of the feasible paths for each of the BE flows.

The resolution of both problems was performed by CPLEX 12.3.

Once both problems have been solved, the traffic representation model is transformed in order to obtain an approximation suitable to a stochastic traffic environment, hence enabling a comparison with the o.f. values obtained by HMOR-S2. This adaptation is performed as in [7] and considers three different values for the compensation parameter α (see explanation in Subsection 3.1). A matrix of offered traffic in Erlang is obtained by a transformation similar to Eqs. (6)–(7), that is

$$A^l(f_s) \approx \tfrac{x^l(f_s)}{d_s u_0} - \alpha \sqrt{\tfrac{x^l(f_s)}{d_s u_0}} \text{ [Erl] if } \tfrac{x^l(f_s)}{d_s u_0} > \alpha^2\,,$$

$$A^l(f_s) \approx \tfrac{x^l(f_s)}{d_s u_0} \text{ [Erl], otherwise}\,.$$

The arc capacity C'_k in Mbit/s (see Tables 1 and 2) is converted to a capacity of $C_k = \left\lceil \frac{C'_k}{u_0} \right\rceil$ channels. Once the offered traffic in Erlang and the arc capacities in circuits are known, the blocking probability for each offered flow in this stochastic environment may be calculated.

The blocking probabilities B_{ks}, for micro-flows of service type s in link k, are calculated as in Eq. (5). Afterwards, the blocking of each flow along its path is obtained, $B^l(f_s)$. As the offered traffic is also known, the calculation of the o.f. may be performed as in Eqs. (1)–(4). For further details, see Subsection 2.1.

3.4. Analytical Results

An analytical study was performed, where results using just a basic version of the heuristic without storage of current non-dominated solutions, HMOR-S2, were obtained. In these runs of the basic heuristic, the initial solution consists of the shortest path direct routing, typical of Internet

Table 4
Average o.f. values with 95% confidence intervals, for simulations with the routing plan obtained with the different heuristic strategies in network $\mathscr{G}$

Obj. func.	MCF-lex-W method solution	Routing method proposed by the authors						
		Initial solution	HMOR-S2 (Basis)		HMOR-S2$_{\text{PAS}}$(i)		HMOR-S2$_{\text{PAS}}$(f)	
			Analytical	Static routing model	Analytical	Static routing model	Analytical	Static routing model
$\alpha = 0.0$								
W_Q	**20907.71**◁	**20859.85**	**21686.92***	**21686.05**±37.51	**21688.63**⋄	**21688.42**±36.97	**21690.16**⋆	**21690.52**±37.22
$B_{Mm\|Q}$	**0.110**	**0.110**	**0.00661**	**0.00756**±0.000848	**0.00595**	**0.00679**±0.00105	**0.00545**	**0.00619**±0.00119
$B_{m1\|Q}$	0.110	0.110	0.00661	0.00756±0.000848	0.00595	0.00679±0.00105	0.00545	0.00619±0.00119
$B_{m2\|Q}$	0.0636	0.0689	0.000453	0.000892±0.000159	0.000480	0.000881±0.000152	0.000465	0.000828±0.000124
$B_{m3\|Q}$	0.00236	0.00308	0.000274	$0.000293\pm2.30\cdot10^{-5}$	0.000273	$0.000288\pm1.86\cdot10^{-5}$	0.000275	$0.000288\pm2.64\cdot10^{-5}$
$B_{M1\|Q}$	0.242	0.555	0.0684	0.0677±0.00869	0.0532	0.0653±0.0140	0.0613	0.0771±0.0117
$B_{M2\|Q}$	0.147	0.378	0.00302	0.00700±0.00134	0.00756	0.00869±0.00131	0.00699	0.00794±0.00174
$B_{M3\|Q}$	0.00622	0.0190	0.00312	0.00316±0.000298	0.00311	0.00317±0.000333	0.00287	0.00288±0.000312
W_B	*6918.87*	*6738.68*	*7167.15*	*7168.36*±12.05	*7163.81*	*7166.23*±10.85	*7158.14*	*7161.10*±11.67
$\alpha = 0.5$								
W_Q	**17678.97**▷	**17611.81**	**17685.88**†	**17683.53**±15.54	**17685.89**•	**17683.53**±15.54	**17685.89**⊙	**17683.53**±15.54
$B_{Mm\|Q}$	**0.00158**	**0.0160**	$\mathbf{1.13\cdot10^{-5}}$	$\mathbf{9.90\cdot10^{-7}}$ $\pm7.83\cdot10^{-7}$	$\mathbf{1.13\cdot10^{-5}}$	$\mathbf{9.47\cdot10^{-7}}$ $\pm7.67\cdot10^{-7}$	$\mathbf{1.04\cdot10^{-5}}$	$\mathbf{8.59\cdot10^{-7}}$ $\pm8.23\cdot10^{-7}$
$B_{m1\|Q}$	0.00158	0.0160	$1.13\cdot10^{-5}$	0	$1.13\cdot10^{-5}$	0	$1.04\cdot10^{-5}$	0
$B_{m2\|Q}$	0.000864	0.00926	$3.3\cdot10^{-9}$	0	$3.3\cdot10^{-9}$	0	$7.2\cdot10^{-9}$	0
$B_{m3\|Q}$	$2.68\cdot10^{-5}$	0.000371	$8.93\cdot10^{-7}$	$9.90\cdot10^{-7}$ $\pm7.83\cdot10^{-7}$	$8.14\cdot10^{-7}$	$9.47\cdot10^{-7}$ $\pm7.67\cdot10^{-7}$	$6.25\cdot10^{-7}$	$8.59\cdot10^{-7}$ $\pm8.23\cdot10^{-7}$
$B_{M1\|Q}$	0.00485	0.147	0.000143	0	0.000143	0	0.000128	0
$B_{M2\|Q}$	0.00273	0.0866	$1.03\cdot10^{-7}$	0	$1.03\cdot10^{-7}$	0	$4.45\cdot10^{-7}$	0
$B_{M3\|Q}$	$9.52\cdot10^{-5}$	0.00353	$4.52\cdot10^{-6}$	$2.24\cdot10^{-5}$ $\pm1.72\cdot10^{-5}$	$4.46\cdot10^{-6}$	$2.24\cdot10^{-5}$ $\pm1.72\cdot10^{-5}$	$4.46\cdot10^{-6}$	$2.24\cdot10^{-5}$ $\pm1.72\cdot10^{-5}$
W_B	*5275.03*	*5247.65*	*5296.56*	*5297.19*±12.83	*5296.56*	*5297.19*±12.83	*5296.57*	*5297.18*±12.84
$\alpha = 1.0$								
W_Q	**16028.11**×	**16025.69**	**16028.14**‡	**16077.61**±15.03	**16028.14**□	**16077.61**±15.03	**16028.14**⊗	**16077.61**±15.03
$B_{Mm\|Q}$	$\mathbf{6.45\cdot10^{-6}}$	**0.000577**	$\mathbf{5\cdot10^{-10}}$	**0**	$\mathbf{5\cdot10^{-10}}$	**0**	$\mathbf{5\cdot10^{-10}}$	**0**
$B_{m1\|Q}$	$6.45\cdot10^{-6}$	0.000577	$5\cdot10^{-10}$	0	$5\cdot10^{-10}$	0	$5\cdot10^{-10}$	0
$B_{m2\|Q}$	$3.79\cdot10^{-6}$	0.000334	$<1\cdot10^{-10}$	0	$<1\cdot10^{-10}$	0	$<1\cdot10^{-10}$	0
$B_{m3\|Q}$	$1.00\cdot10^{-7}$	$1.16\cdot10^{-5}$	$<1\cdot10^{-10}$	0	$<1\cdot10^{-10}$	0	$<1\cdot10^{-10}$	0
$B_{M1\|Q}$	$4.81\cdot10^{-5}$	0.00650	$1.27\cdot10^{-8}$	0	$1.27\cdot10^{-8}$	0	$1.27\cdot10^{-8}$	0
$B_{M2\|Q}$	$2.38\cdot10^{-5}$	0.00347	$<1\cdot10^{-10}$	0	$<1\cdot10^{-10}$	0	$<1\cdot10^{-10}$	0
$B_{M3\|Q}$	$7.62\cdot10^{-7}$	0.000123	$2\cdot10^{-10}$	0	$2\cdot10^{-10}$	0	$2\cdot10^{-10}$	0
W_B	*3340.47*	*3354.76*	*3355.88*	*3350.97*±24.92	*3355.88*	*3350.97*±24.92	*3355.88*	*3350.97*±24.92

MCF-lex-W method solution: ◁) 96.29%; ▷) 99.96%; ×) 99.75% of W_Q^{ideal} (the ideal revenue extracted from the data in [6]); HMOR-S2: *) 99.88%; †) 100%; ‡) 99.75% of W_Q^{ideal}; HMOR-S2$_{\text{PAS}}$(i): ⋄) 99.89%; •) 100%; □) 99.75% of W_Q^{ideal}; HMOR-S2$_{\text{PAS}}$(f): ⋆) 99.90%; ⊙) 100%; ⊗) 99.75% of W_Q^{ideal}.

conventional routing algorithms, such as the ones used in the network dimensioning algorithm. In further analytical studies, two different types of tests were conducted for the heuristic HMOR-S2$_{\text{PAS}}$:

- (i) tests: the initial solution is a solution typical of Internet conventional routing algorithms, such as the one used in the basic version runs.
- (f) tests: the initial solution of the HMOR-S2$_{\text{PAS}}$ heuristic is the routing plan obtained at the end of the basic heuristic runs for each specific α. With this experiment, it is possible to check whether that heuristic variant can improve the quality of the final solutions obtained with HMOR-S2 as an alternative to the direct use of the heuristic variant (as in the case of the (i) tests).

The multiobjective routing model in [6] is quite different from the one considered here, so no results concerning any of the o.f. considered here is provided in [6]. The only results that can be extracted from the proposed model in [6] are approximate ideal values for the QoS flows revenue, W_Q^{ideal}. The analytical results concerning the QoS flows revenue W_Q were compared with these approximate ideal values.

The experiments with the HMOR-S2$_{\text{PAS}}$ were conducted with an archive of size 5, chosen empirically after extensive experimentation. This experimentation showed that an increase in the archive size would not necessarily lead to better final results, because at the end of the heuristic run, when the final solution is actually chosen from those in the archive, the top 5 solutions tend to be the same regardless of the archive size.

Table 5
Average o.f. values with 95% confidence intervals, for simulations with the routing plan obtained with the different heuristic strategies in network $\mathscr{H}$

	MCF-lex-W	Routing method proposed by the authors						
Obj.	method	Initial	HMOR-S2 (Basis)		HMOR-S2$_{\text{PAS}}$(i)		HMOR-S2$_{\text{PAS}}$(f)	
Func.	solution	solution	Analytical	Static routing model	Analytical	Static routing model	Analytical	Static routing model
$\alpha = 0.0$								
W_Q	**20602.28**◁	**20358.90**	**21576.67***	**21559.04**±31.57	**21578.99**⋄	**21563.15**±32.06	**21616.01**⋆	**21597.91**±30.13
$B_{Mm\|Q}$	**0.147**	**0.169**	**0.0287**	**0.0306**±0.00147	**0.0299**	**0.0315**±0.00131	**0.0224**	**0.0245**±0.00175
$B_{m1\|Q}$	0.147	0.169	0.0287	0.0306±0.00147	0.0299	0.0315±0.00131	0.0224	0.0245±0.00175
$B_{m2\|Q}$	0.0891	0.111	0.00456	0.00696±0.000526	0.00410	0.00641±0.000370	0.00341	0.00580±0.000523
$B_{m3\|Q}$	0.00346	0.00536	0.00171	0.00170±$7.97\cdot10^{-5}$	0.00148	0.00147±$6.86\cdot10^{-5}$	0.000596	0.000601±$4.99\cdot10^{-5}$
$B_{M1\|Q}$	0.272	0.711	0.150	0.177±0.0194	0.330	0.312±0.0295	0.146	0.157±0.0231
$B_{M2\|Q}$	0.167	0.518	0.0162	0.0289±0.00289	0.0180	0.0304±0.00572	0.0145	0.0215±0.00229
$B_{M3\|Q}$	0.00699	0.0293	0.00828	0.00822±0.000302	0.00964	0.00959±0.000348	0.00362	0.00376±0.000476
W_B	*6724.15*	*6434.17*	*6877.69*	*6886.47*±9.03	*6905.99*	*6914.48*±11.01	*6927.67*	*6935.83*±10.55
$\alpha = 0.5$								
W_Q	**17670.25**▷	**17419.40**	**17685.66**†	**17683.32**±15.58	**17685.66**●	**17683.32**±15.57	**17685.82**⊙	**17683.45**±15.55
$B_{Mm\|Q}$	**0.00297**	**0.0558**	**0.000120**	**$9.19\cdot10^{-5}$** ±0.000133	**0.000120**	**$9.16\cdot10^{-5}$** ±0.000133	**$4.86\cdot10^{-5}$**	**$4.55\cdot10^{-5}$** ±0.000109
$B_{m1\|Q}$	0.00297	0.0558	0.000120	$8.52\cdot10^{-5}$ ±0.000138	0.000120	$8.52\cdot10^{-5}$ ±0.000138	$4.86\cdot10^{-5}$	$4.28\cdot10^{-5}$ ±0.000110
$B_{m2\|Q}$	0.00178	0.0335	$2.91\cdot10^{-7}$	0	$2.91\cdot10^{-7}$	0	$1.78\cdot10^{-7}$	0
$B_{m3\|Q}$	$5.69\cdot10^{-5}$	0.00143	$8.46\cdot10^{-6}$	$1.03\cdot10^{-5}$ ±$2.40\cdot10^{-6}$	$7.99\cdot10^{-6}$	$9.89\cdot10^{-6}$ ±$2.41\cdot10^{-6}$	$2.27\cdot10^{-6}$	$3.28\cdot10^{-6}$ ±$1.41\cdot10^{-6}$
$B_{M1\|Q}$	0.0139	0.327	0.00213	0.00392±0.00708	0.00213	0.00392±0.00708	0.000910	0.00273±0.00702
$B_{M2\|Q}$	0.00753	0.205	$1.69\cdot10^{-6}$	0	$1.69\cdot10^{-6}$	0	$9.58\cdot10^{-7}$	0
$B_{M3\|Q}$	0.000271	0.00906	$4.95\cdot10^{-5}$	$8.06\cdot10^{-5}$ ±$8.38\cdot10^{-6}$	$4.87\cdot10^{-5}$	$8.04\cdot10^{-5}$ ±$8.65\cdot10^{-6}$	$1.61\cdot10^{-5}$	$3.56\cdot10^{-5}$ ±$9.78\cdot10^{-6}$
W_B	*5243.12*	*5119.13*	*5295.37*	*5295.90*±13.14	*5295.37*	*5295.90*±13.14	*5295.76*	*5296.16*±13.30
$\alpha = 1.0$								
W_Q	**16024.58**×	**15998.35**	**16028.14**‡	**16077.61**±15.03	**16028.14**□	**16077.61**±15.03	**16028.14**⊗	**16077.61**±15.03
$B_{Mm\|Q}$	**$2.47\cdot10^{-5}$**	**0.00678**	**$2.19\cdot10^{-8}$**	**0**	**$2.19\cdot10^{-8}$**	**0**	**$1.78\cdot10^{-8}$**	**0**
$B_{m1\|Q}$	$2.47\cdot10^{-5}$	0.00678	$2.19\cdot10^{-8}$	0	$2.19\cdot10^{-8}$	0	$1.78\cdot10^{-8}$	0
$B_{m2\|Q}$	$1.25\cdot10^{-5}$	0.00416	$<1\cdot10^{-10}$	0	$<1\cdot10^{-10}$	0	$<1\cdot10^{-10}$	0
$B_{m3\|Q}$	$3.76\cdot10^{-7}$	0.000153	$1.7\cdot10^{-9}$	0	$1.6\cdot10^{-9}$	0	$1.6\cdot10^{-9}$	0
$B_{M1\|Q}$	0.000100	0.0530	$4.76\cdot10^{-7}$	0	$4.76\cdot10^{-7}$	0	$4.76\cdot10^{-7}$	0
$B_{M2\|Q}$	$7.07\cdot10^{-5}$	0.0298	$<1\cdot10^{-10}$	0	$<1\cdot10^{-10}$	0	$<1\cdot10^{-10}$	0
$B_{M3\|Q}$	$2.25\cdot10^{-6}$	0.00113	$1.10\cdot10^{-8}$	0	$1.10\cdot10^{-8}$	0	$1.10\cdot10^{-8}$	0
W_B	*3314.37*	*3341.90*	*3355.88*	*3350.97*±24.92	*3355.88*	*3350.97*±24.92	*3355.88*	*3350.97*±24.92

MCF-lex-W method solution: ◁) 94.89%; ▷) 99.91%; ×) 99.73% of W_Q^{ideal} (the ideal revenue extracted from the data in [6]); HMOR-S2: *) 99.37%; †) 100%; ‡) 99.75% of W_Q^{ideal}; HMOR-S2$_{\text{PAS}}$(i): ⋄) 99.39%; ●) 100%; □) 99.75% of W_Q^{ideal}; HMOR-S2$_{\text{PAS}}$(f): ⋆) 99.56%; ⊙) 100%; ⊗) 99.75% of W_Q^{ideal}.

The analytical results displayed in Tables 4 and 5 were obtained in approximately 47 s (on average) in a Linux environment on a Pentium 4 processor with 3 GHz CPU and 1 GB of RAM. In the tables, the values obtained for W_Q, $B_{Mm|Q}$ and W_B are highlighted, as they are the most interesting o.f. (from a traffic engineering perspective) in the two priority levels.

A comparison of the results obtained with the MCF-lex-W approach (described in Subsection 3.3) and the heuristic proposed by the authors shows that the latter approach provides consistently better values for all the o.f. in most cases. This improvement is particularly relevant concerning the ‘fairness’ QoS o.f. $B_{Mm|Q}$ as could be expected having in mind the nature of our model, which explicitly considers this parameter as an o.f. These results put also in evidence the superiority, especially concerning QoS related performance parameters, of a model such as ours, which has not only an imbedded stochastic representation of the traffic flows, but also a consistent (albeit approximate) and complete representation of the interactions among all traffic flows of all types. This is naturally something that the MCF-lex-W approach cannot provide, although leading to very similar values for the QoS traffic revenue. Notice that for lower traffic loads ($\alpha = 0.5$ and $\alpha = 1.0$), the values of W_Q and W_B are very similar in both methods. This can be explained by the fact that in these situations, corresponding to low and extremely low blocking probabilities the effects of the stochasticity of the traffic are attenuated or even negligible, as indeed reflected by the values of the blocking probability related parameters in Tables 4 and 5.

Regarding the analytical results with the heuristic variants considered by the authors, Tables 4 and 5 enable two different comparative analysis. Since in HMOR-S2$_{\mathrm{PAS}}$(i), the initial solution is the same as the one used in the basic heuristic HMOR-S2, the tables allow for a comparison of the final analytical results obtained with HMOR-S2 and HMOR-S2$_{\mathrm{PAS}}$. As for the PAS(f), the initial solution has the o.f. values shown in the table under HMOR-S2 (Basis), so that a comparison of the initial and the final analytical results with HMOR-S2$_{\mathrm{PAS}}$ can be made.

The final analytical results for the upper level o.f. are the same or show an improvement on the ones obtained with the basic heuristic, for all the values of α, for both versions of the heuristic HMOR-S2$_{\mathrm{PAS}}$. For this reason, and also taking into account that using the archive does not lead to an increase in the execution time, the heuristic HMOR-S2$_{\mathrm{PAS}}$ can be considered as a better approach for solving the routing problem. In particular, the (f) version (a run of the basic heuristic HMOR-S2 followed by a run of the PAS variant) provides improved results for W_Q and $B_{Mm|Q}$ for the routing problem under analysis especially for $\alpha = 0.0$, which corresponds to higher overload situations.

For $\alpha = 0.0$, the results for HMOR-S2$_{\mathrm{PAS}}$(f) show that there was a minor improvement in the QoS flows revenue obtained with HMOR-S2, of 0.02% and 0.18% in Table 4 and 5, respectively; as for the improvement in the $B_{Mm|Q}$ value, it was significant: 17.55% and 21.95% for networks $\mathscr{G}$ and $\mathscr{H}$, respectively. For $\alpha = 0.5$ and $\alpha = 1.0$, the results are practically the same for all the versions of the heuristic. However, note that for $\alpha = 0.5$ the HMOR-S2$_{\mathrm{PAS}}$(f) variant allowed for an improvement on the value of $B_{Mm|Q}$ in both networks.

The results presented in both tables confirm the advantages of using a Pareto archive strategy. In the situations of higher blocking ($\alpha = 0.0$), the use of this strategy leads to an improvement on the values of the first level o.f. of the routing model, especially for the blocking probability values $B_{Mm|Q}$. In the situations of lower blocking probability ($\alpha = 0.5$ and $\alpha = 1.0$), the main advantage of using the Pareto archive is the increased insensitivity to the initial solution, because for both networks the final solutions obtained with HMOR-S2$_{\mathrm{PAS}}$(i) and HMOR-S2$_{\mathrm{PAS}}$(f) are quite close or even the same. It should be noted that the difference between HMOR-S2$_{\mathrm{PAS}}$(i) and HMOR-S2$_{\mathrm{PAS}}$(f) is simply the initial solution.

3.5. Simulation Results

Simulation experiments with a static routing method using the solution provided by the heuristic were carried out. With this simulation study, the routing model results may be validated and the errors intrinsic to the analytical model, which provides the estimates for the o.f. may be evaluated.

A discrete-event stochastic simulation platform was used with the static routing model. The routing plan is the final solution obtained after one of the heuristic versions was run, and it does not change throughout the simulation, regardless of the random variations of traffic offered to the network. An initialization phase that lasts for a time $t_{warm-up}$ is followed by a phase of data collection: information on the number of offered calls and carried calls in the network for each flow $f_s, s \in \mathscr{S}$, is gathered, until the end of the simulation. Considering this information, $B(f_s), \forall s \in \mathscr{S}$ can be estimated. Subsequently, the values of the upper and lower level o.f. related to blocking probabilities can also be estimated. The number of carried calls in the network is used to estimate the expected revenues.

In Tables 4 and 5, the analytical values and the simulation results (average value ± half length of a 95% confidence interval, computed by the independent replications method, see e.g. [19]) of each o.f. are displayed. The simulation results displayed in the table were obtained with a total simulated time $t_{total} = 48$ h and a warm-up time $t_{warm-up} =$ 8 h. It took about 30 minutes of CPU time to get the results for both networks, in the computer mentioned earlier.

The analytical results and the corresponding static routing model simulation results have similar magnitude, with the analytical results slightly better than expected. The analytical and the simulation results for W_Q are close and the analytical result for that o.f. is inside the 95% confidence interval for all the heuristic versions for $\alpha = 0.0$ and $\alpha = 0.5$. For $\alpha = 1.0$, the analytical value of W_Q is actually worse than the corresponding simulation result. Notice that $\alpha = 1.0$ corresponds to a situation of lower traffic load, where in many instances all the offered calls of a certain service are actually carried. In these situations, the blocking estimate for that service is 0 and high values of the estimate of W_Q are obtained, surpassing the analytical values. Note that in lower traffic load situations ($\alpha = 0.5$ and $\alpha = 1.0$), the occurrence of blocking is a rare event. A well known result in statistics is that in these cases the uncertainty in the estimates is very high, as reflected in the very high relative half length of the calculated 95% confidence intervals of the blocking probabilities. Also for the situations of lower traffic load the simulation results for $B_{Mm|Q}$ are better than the corresponding analytical value, again because of the many instances throughout the executed simulations where the blocking estimate for a certain service is 0.

The simulation and analytic results are different mainly due to the imprecisions/inaccuracies intrinsic to the analytic/numerical resolution, in particular those associated with the simplifications of the traffic model, and the associated error propagation. In this model, the overflow traffic is treated as Poisson traffic and as a result, the analytical model is simplified and tends to underestimate the blocking probabilities in the network (and to overestimate the revenues). The errors resulting from this simplification propagate throughout the complex and lengthy numerical calculations associated with the resolution, for a great number of times, of the large systems of implicit non-linear equations used to calculate B_{ks} and $c_{ks}^{Q(B)}$. Another simplification assumed in the stochastic model for the traffic in the links is the superposition of independent Poisson flows and independent occupations of the links. However, we believe that the approximations in this model can be considered appropriate in this context for practical reasons. In fact, if more complex models were used to represent the traffic and to calculate the blockings, the computational burden would become too heavy. Plus, these errors do not compromise the inequality relations between the o.f. values, the comparison of which is at the core of the multiobjective routing optimization method. In fact, when the results obtained with the basic heuristic HMOR-S2 and with HMOR-S2$_{\text{PAS}}$ are compared, we observe a coherence in the analytical and simulation results, in the sense that whenever the analytical value of an o.f. is better for the (f) version than for the (i) version, the same tends to happen with the average values obtained with the static routing model simulation.

4. Conclusions and Further Work

This work began with a revision of a hierarchical bi-level multiobjective routing model in MPLS networks considering alternative routing, two classes of service (with different priorities in the optimization model) and different types of traffic flows in each class. The resolution of this very complex routing optimization model was performed by a heuristic, HMOR-S2$_{\text{PAS}}$, which was also reviewed. This procedure maintains the resolution framework of a previous heuristic, HMOR-S2, but introduces and treats in a special manner an archive of possible good solutions found throughout the execution of the heuristic.

The heuristic approaches HMOR-S2 and HMOR-S2$_{\text{PAS}}$ were applied to two new test networks, $\mathscr{G}$ and $\mathscr{H}$, obtained by a transformation of an original network in [6]. Various traffic matrices were considered, so as to include in the study different situations of higher and lower traffic load.

The analytical results for the different o.f. obtained with both heuristic variants (without and with the Pareto archive) were compared. The values of W_Q were also compared with the approximate ideal values obtained with the traffic matrix provided by [6] and offered to networks $\mathscr{G}$ and $\mathscr{H}$.

Furthermore, a comparison of the results obtained with the proposed heuristic HMOR-S2$_{\text{PAS}}$ with results from a routing method based on a MCF approach, with lexicographic optimization and the possibility of traffic splitting, similar to the one in [7] was carried out. The results show that the HMOR-S2$_{\text{PAS}}$ method provides consistently better values for all the o.f. in most cases. In particular, the results for the 'fairness' QoS o.f. $B_{Mm|Q}$ are significantly better with the proposed heuristic (where this parameter is explicitly considered as an o.f.).

Concerning QoS related performance parameters, we may conclude that the stochastic representation of the traffic flows and the complete representation of the interactions among all traffic flows of all types in our model allow for better results. Nevertheless, notice that the values of W_Q and W_B are very similar in both methods for lower traffic loads ($\alpha = 0.5$ and $\alpha = 1.0$), due to the attenuated effects of the stochasticity of the traffic in these situations, corresponding to low and extremely low blocking probabilities.

The results show that the heuristic with an archive of non-dominated solutions is always advantageous, both when the blocking is higher (in this situation HMOR-S2$_{\text{PAS}}$ tends to provide improved results for the routing problem) and lower (in this situation HMOR-S2$_{\text{PAS}}$ tends to give an increased insensitivity to the initial solution).

A more exact evaluation of the results of the heuristic was accomplished with a discrete-event simulation platform. In most cases, the analytical results and the static routing model simulation results have similar magnitude. The differences between them are due to inaccuracies intrinsic to the analytic/numerical resolution, but which have not any influence in the final routing solutions.

We conclude that the results obtained with analytic and stochastic discrete-event simulation models confirm the effectiveness of the HMOR-S2$_{\text{PAS}}$ approach to route calculation and selection in multiservice networks.

An important remark is that the PAS variant is not more complex than the basic heuristic. Nevertheless, the computational burden of either resolution approach is still heavy. This is the major limitation of this type of routing method and, as so, its potential practical application is currently restrained to networks with a limited number of nodes, such as the core and intermediate (metro-core) level networks of low dimension.

Further simplifications and improvements in the heuristic resolution approaches will be the focus of future work. The extension of the model to broader routing principles (such as probabilistic load sharing or traffic splitting) and an adaptation of the model, so that it can be applied to test networks based on actual MPLS networks are also possible subjects of future work.

Acknowledgements

This work was financially supported by programme COMPETE of the EC Community Support Framework III and cosponsored by the EC fund FEDER and national funds (FCT – PTDC/EEA-TEL/101884/2008 and PEst-C/EEI/UI0308/2011).

References

[1] A. P. Wierzbicki, "Telecommunications, multiple criteria analysis and knowledge theory", *J. Telecommun. Inform. Technol.*, no. 3, pp. 3–13, 2005.

[2] A. P. Wierzbicki and W. Burakowski, "A conceptual framework for multiple-criteria routing in QoS IP networks", *Int. Trans. Oper. Res.*, vol. 18, no. 3, pp. 377–399, 2011.

[3] J. Craveirinha, R. Girão-Silva, and J. Clímaco, "A meta-model for multiobjective routing in MPLS networks", *Central Eur. J. Oper. Res.*, vol. 16, no. 1, pp. 79–105, 2008.

[4] R. E. Steuer, *Multiple Criteria Optimization: Theory, Computation and Application*. Probability and Mathematical Statistics, Wiley, 1986.

[5] R. Girão-Silva, J. Craveirinha, and J. Clímaco, "Hierarchical multiobjective routing model in Multiprotocol Label Switching networks with two service classes – a Pareto archive strategy", *Engineering Optimization*, vol. 44, no. 5, pp. 613–635, 2012.

[6] S. C. Erbas and C. Erbas, "A multiobjective off-line routing model for MPLS networks", in *Proc. 18th Int. Teletraffic Congr. ITC-18*, J. Charzinski, R. Lehnert, and P. Tran-Gia, Eds., Berlin, Germany, 2003, Elsevier, Amsterdam, pp. 471–480.

[7] D. Mitra and K. G. Ramakrishnan, "Techniques for traffic engineering of multiservice, multipriority networks", *Bell Labs Technical J.*, vol. 6, no. 1, pp. 139–151, 2001.

[8] Q. Ma and P. Steenkiste, "Supporting dynamic inter-class resource sharing: A multi-class QoS routing algorithm", in *Proc. IEEE Infocom'99*, New York, USA, 1999, pp. 649–660.

[9] H. Kochkar, T. Ikenaga, and Y. Oie, "QoS routing algorithm based on multiclasses traffic load", in *Proc. IEEE Global Telecommun. Conf. GLOBECOM 2001*, San Antonio, TX, USA, 2001, pp. 2193–2198.

[10] D. Awduche, A. Chiu, A. Elwalid, I. Widjaja, and X. Xiao, "Overview and principles of Internet traffic engineering", RFC 3272, Network Working Group, May 2002.

[11] F. Kelly, "Notes on effective bandwidths", in *Stochastic Networks: Theory and Applications*, F. P. Kelly, S. Zachary, and I. Ziedins, Eds., vol. 4 of *Royal Statistical Society Lecture Notes Series*, Oxford University Press, 1996, pp. 141–168.

[12] D. Mitra, J. A. Morrison, and K. G. Ramakrishnan, "Optimization and design of network routing using refined asymptotic approximations", *Performance Evaluation*, vol. 36-37, pp. 267–288, 1999.

[13] J. S. Kaufman, "Blocking in a shared resource environment", *IEEE Trans. Communi.*, vol. COM-29, no. 10, pp. 1474–1481, 1981.

[14] J. W. Roberts, "Teletraffic models for the Telecom 1 integrated services network", in *Proce. 10th Int. Teletraffic Congr.*, Montreal, Canada, 1983.

[15] D. Mitra and J. A. Morrison, "Erlang capacity and uniform approximations for shared unbuffered resources", *IEEE/ACM Trans. Netw.*, vol. 2, no. 6, pp. 558–570, 1994.

[16] H. M. El Sayed, M. S. Mahmoud, A. Y. Bilal, and J. Bernussou, "Adaptive alternate-routing in telephone networks: Optimal and equilibrium solutions" *Information and Decision Technologies*, vol. 14, no. 1, pp. 65–74, 1988.

[17] J. Craveirinha, R. Girão-Silva, J. Clímaco, and L. Martins, "A hierarchical multiobjective routing model for MPLS networks with two service classes", in *Revised Selected Papers of the 23rd IFIP TC7 Conf. Sys. Model. Optimiz., Cracow, Poland, July 23-27, 2007*, A. Korytowski, K. Malanowski, W. Mitkowski, and M. Szymkat, Eds., vol. 312 of *IFIP Advances in Information and Communication Technology*, Springer, 2009, pp. 196–219.

[18] R. Girão-Silva, J. Craveirinha, and J. Clímaco, "Hierarchical multiobjective routing in Multiprotocol Label Switching networks with two service classes – A heuristic solution", *Int. Trans. Oper. Res.*, vol. 16, no. 3, pp. 275–305, 2009.

[19] A. M. Law and W. D. Kelton, *Simulation Modeling and Analysis*. Industrial Engineering and Management Science, 2nd edition. McGraw-Hill, 1991.

Rita Girão-Silva graduated in electrical engineering (telecommunications) in 1999 and received a Ph.D. diploma in electrical engineering (telecommunications and electronics) in 2009, both at the University of Coimbra, Portugal. She is an Assistant Professor at the Department of Electrical and Computer Engineering, University of Coimbra, and a researcher at INESC-Coimbra. Her research areas include routing models in telecommunications networks and multiobjective optimization.

E-mail: rita@deec.uc.pt

Department of Electrical Engineering Science
and Computers
University of Coimbra
Pólo II, Pinhal de Marrocos
P-3030-290 Coimbra, Portugal

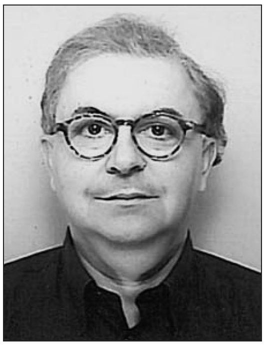
Institute of Computers and Systems Engineering
of Coimbra (INESC-Coimbra)
Rua Antero de Quental, 199
P-3000-033 Coimbra, Portugal

José Craveirinha is Full Professor in telecommunications at the Department of Electrical Engineering and Computers of the Faculty of Sciences and Technology of the University of Coimbra, Portugal, since 1997. He obtained the following degrees: undergraduate diploma in electrical engineering science (E.E.S.)-telecommunications and electronics at IST, Lisbon Technical University (1975), M.Sc. (1981) and Ph.D. in E.E.S. at the University of Essex (UK) (1984) and Doct. of Science (Agregado) in E.E.S.-telecommunications at the University of Coimbra (1996). Previous positions were: Associate Professor and Assistant Professor at the FCTUC, Coimbra University, telecommunication R&D engineer (at the CET-Portugal Telecom). He coordinated a research group in teletraffic engineering and network planning at the INESC-Coimbra R&D Institute since 1986 and was Director of this Institute during 1994–1999. He is author/co-author of more than 100 scientific and technical publications in teletraffic modeling, reliability analysis, planning and optimization of telecommunication networks. His main present interests are in multicriteria routing and reliability analysis models and algorithms for optical and multiservice-IP/MPLS networks.

E-mail: jcrav@deec.uc.pt

Department of Electrical Engineering and Computers
University of Coimbra
Pinhal de Marrocos
3030-290 Coimbra, Portugal

Institute of Computers and Systems Engineering
of Coimbra (INESC-Coimbra)
Rua Antero de Quental, 199
P-3000-033 Coimbra, Portugal



João Clímaco is Full Professor at the Faculty of Economics of the University of Coimbra and President of the Scientific Committee of the INESC-Coimbra. He obtained the M.Sc. degree in Control Systems at the Imperial College of Science and Technology, University of London (1978); the "Diploma of Membership of the Imperial College of Science and Technology" (1978); the Ph.D. in Optimization and Systems Theory, Electrical Engineering Department, University of Coimbra (1982) and the title of "Agregação" at the University of Coimbra (1989). He served in the past as the Vice-President of ALIO – Latin-Iberoamerican Operational Research Societies Association and Vice-President of the Portuguese OR Society. He belongs to the editorial board of the following scientific journals: Investigação Operacional (Journal of the Portuguese OR Society), Journal of Group Decision and Negotiation, International Transactions in Operational Research (ITOR), ENGEVISTA and Rio's International Journal on Sciences of Industrial and Systems Engineering and Management. He is also member of the editorial board of the University of Coimbra Press. His current interests of research include multicriteria decision aiding, multiobjective mathematical programming, location analysis and telecommunication network planning and management.

E-mail: jclimaco@inescc.pt
Institute of Computers and Systems Engineering
of Coimbra (INESC-Coimbra)
University of Coimbra
Rua Antero de Quental, 199
P-3000-033 Coimbra, Portugal

Paper

Uplift Modeling in Direct Marketing

Piotr Rzepakowski and Szymon Jaroszewicz

National Institute of Telecommunications, Warsaw, Poland

Abstract—Marketing campaigns directed to randomly selected customers often generate huge costs and a weak response. Moreover, such campaigns tend to unnecessarily annoy customers and make them less likely to answer to future communications. Precise targeting of marketing actions can potentially results in a greater return on investment. Usually, response models are used to select good targets. They aim at achieving high prediction accuracy for the probability of purchase based on a sample of customers, to whom a pilot campaign has been sent. However, to separate the impact of the action from other stimuli and spontaneous purchases we should model not the response probabilities themselves, but instead, the change in those probabilities caused by the action. The problem of predicting this change is known as uplift modeling, differential response analysis, or true lift modeling. In this work, tree-based classifiers designed for uplift modeling are applied to real marketing data and compared with traditional response models, and other uplift modeling techniques described in literature. The experiments show that the proposed approaches outperform existing uplift modeling algorithms and demonstrate significant advantages of uplift modeling over traditional, response based targeting.

Keywords— decision trees, information theory, marketing tools, uplift modeling.

1. Introduction

When a customer is not completely anonymous, a company can send marketing offers directly to him/her. For example an Internet retailer's product offer can be sent by e-mail or by traditional post; telecommunication operators may advertise their services by SMS, voice calls or other communication channels.

However, to make campaigns effective they should be directed selectively to those who, with high probability, will respond positively (will, e.g., buy a product, or visit a web site). Properly targeted campaign will give a greater return on investment than a randomly targeted one, and, what is even more important, it will not annoy those who are not interested in the offer. It is well known in the direct marketing community that campaigns do put off some customers. There are however few methods available to identify them. See [1]–[4] for more detailed information.

In this paper we experimentally verify the above claims on real direct marketing data. The data is publicly available [5] and comes from an online retailer offering women's and men's merchandise; the next section gives a more detailed description. We test both standard, response based models, as well as uplift approaches described in literature and compare them with decision trees designed especially for uplift modeling, which we introduced in [6], [7]. The experiments verify that the uplift approach gives much better marketing results. Moreover, we demonstrate that our decision trees, designed especially for uplift modeling, outperform other uplift approaches described in literature.

2. Problem Statement

In this section, we describe the marketing data on which we have tested our models. The dataset [5], provided on Kevin Hillstrom's MineThatData blog, contains results of an e-mail campaign for an Internet based retailer. The dataset [5] contains information about 64,000 customers who last purchased within at most twelve months. The customers were subjected to a test e-mail campaign:

– 1/3 were randomly chosen to receive an e-mail campaign featuring men's merchandise,

– 1/3 were randomly chosen to receive an e-mail campaign featuring women's merchandise,

– 1/3 were randomly chosen to not receive an e-mail.

The data describes customer behavior for two weeks after the campaign. The details of the dataset are summarized in Tables 1 and 2.

Table 1
Hillstrom's marketing data: customers' attributes

Attribute	Definition
Recency	Months since last purchase
History_Segm	Categorization of dollars spent in the past year
History	Actual dollar value spent in the past year
Mens	1/0 indicator, 1 = customer purchased mens merchandise in the past year
Womens	1/0 indicator, 1 = customer purchased womens merchandise in the past year
Zip_Code	Classifies zip code as urban, suburban, or rural
Newbie	1/0 indicator, 1 = new customer in the past twelve months
Channel	Describes the channels the customer purchased from in the past year

Table 2
Hillstrom's marketing data: type of e-mail campaign sent and activity in the two following weeks

Attribute	Definition
Segment	E-mail campaign the customer received
Visit	1/0 indicator, 1 = customer visited website in the following two weeks
Conversion	1/0 indicator, 1 = customer purchased merchandise in the following two weeks
Spend	Actual dollars spent in the following two weeks

The author asked several questions to be answered based on the data. Here we address the problem of predicting the people who visited the site within the two-week period (attribute *Visit* in Table 2) because they received the campaign. The estimate was based by comparing customer behavior on the treatment and control groups, i.e., comparing customers who did and did not receive an e-mail.

During an initial analysis we have found that about 10.62% of the people visited the site spontaneously, but after the campaign (combined men's and women's) the visits increased to 16.7%. Men's merchandise campaign outperformed women's, as the increase in visits was about 7.66% (from 10.62% to 18.28%), while the women's merchandise campaign resulted in an increase of only 4.52% (from 10.62% to 15.14%).

Afterward, we used traditional response based targeting, as well as uplift modeling based targeting to select the customers for the campaign. Because there is a large difference in response between treatment groups who received advertisements for men's and women's merchandise, the two campaign types were analyzed, both jointly and separately. In the first case, the treatment group consists of all those who received an e-mail and the control group of those who did not. In the second case, there are two treatment groups, one for man's and one for women's merchandise campaign; both treatment groups are analyzed separately with respect to the same control group. Since the men's merchandise group showed little sensitivity to attribute values, our experiments focused primarily on the women's merchandise group.

The following two sections give the literature overview, describe the uplift modeling methodology used and compare it to the traditional predictive modeling. Section 5 presents experimental results.

3. Uplift Modeling

In this section we give a more detailed overview of uplift modeling and review available literature.

Traditionally used response models are built on a sample of data about the customers. Each record in the dataset represents a customer and the attributes describe his/her characteristics. In the *propensity* models, historical information about purchases (or other success measures like visits) is used, while in the *response* models, all customers have been subject to a pilot campaign. A distinguished class attribute informs on whether a customer responded to the offer or not. Afterward, the data is used to build a model that predicts conditional probability of response *after* the campaign. This model is then applied to the whole customer database to select people with high probability of purchasing the product. The process is illustrated in Fig. 1.

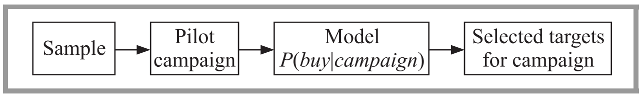


Fig. 1. Response model creation process.

However, in reality, we can divide the customers into four groups, i.e., those who:

- responded *because* of the action,
- responded *regardless* of the action (unnecessary costs),
- did not respond and the action had *no impact* (unnecessary costs),
- did not respond *because* the action had a *negative impact* (*e.g.* a customer got annoyed by the campaign, might even have churned).

Propensity models, as well as traditional response models are not capable of distinguishing those four groups, while uplift models can do that. This is because traditional models predict the conditional class probability

$$P(response|treatment),$$

while uplift models predict the change in behavior resulting from the action

$$P(response|treatment) - P(response|no\,treatment).$$

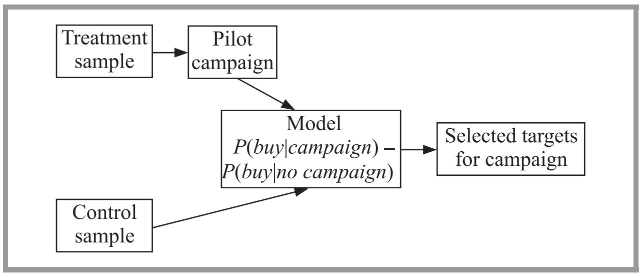

Treatment sample
Pilot campaign
Model $P(buy|campaign)$ – $P(buy|no\ campaign)$
Selected targets for campaign
Control sample

Fig. 2. Uplift model creation process.

To build an uplift model, a random sample (the *treatment* dataset) of customers is selected and subjected to the marketing action. Disjoint sample is also selected (the *control* dataset), to which the action is not applied, and which serves as the background against which the results of the action will be measured. The model is now built for predicting the *difference* between class probabilities on the two sets of data. The process is illustrated in Fig. 2.

3.1. Literature Overview

The problem of uplift modeling has received little attention in literature – a surprising fact, if one considers its practical importance.

There exist two overall approaches to uplift modeling. The first, obvious one is to build two separate classifiers. One on the treatment and another on the control dataset (as shown in Fig. 3). For each classified object class probabilities predicted by the control model are subtracted from those predicted by the treatment model, giving a direct estimate of the difference in behavior caused by the action.

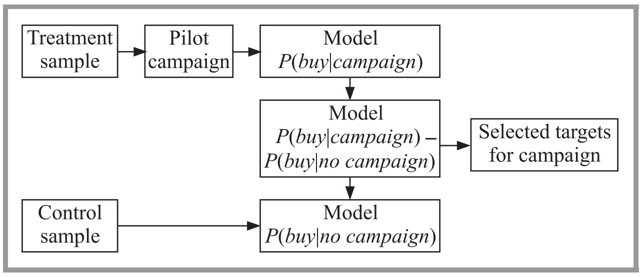

Treatment sample
Pilot campaign
Model $P(buy|campaign)$
Model $P(buy|campaign) - P(buy|no\ campaign)$
Selected targets for campaign
Control sample
Model $P(buy|no\ campaign)$

Fig. 3. Uplift model based on two separate classifiers

This approach has a major disadvantage: the behavior of the differences between class probabilities can be very different than the behavior of the class probabilities themselves. Thus, it is possible that the models will focus too much on modeling the class in both datasets, instead of focusing on the differences between them. The problem is exacerbated by the fact that the variation in the difference between class probabilities is usually much smaller than variability in class probabilities themselves. For example, in case of decision trees, the double model approach does not necessarily favor splits, which lead to *different* responses in treatment and control groups, just splits, which lead to predictable outcomes in each of the groups separately, wasting valuable data. See [1], [4], [8], [9] for details.

Let us now look at the second type of approaches, which attempt to model the difference between treatment and control probabilities directly.

One of the first 'native' uplift modeling approaches builds a single decision tree, by trying to maximize the uplift criterion at each step [1]. The splitting criterion used by the algorithm, called $\Delta\Delta P$, selects tests, which maximize the difference between the differences between treatment and control probabilities in the left and right subtrees. This corresponds to maximizing the desired difference, directly in the fashion of greedy algorithms. More formally, suppose we have a test A with outcomes a_0 and a_1. The $\Delta\Delta P$ splitting criterion is defined as

$$\Delta\Delta P(A) = \big|\big(P^T(y_0|a_0) - P^C(y_0|a_0)\big) - \big(P^T(y_0|a_1) - P^C(y_0|a_1)\big)\big|,$$

where y_0 is a selected (positive) class. The calculation of the criterion for subtree is illustrated in Fig. 4. While the original $\Delta\Delta P$ criterion works only for binary trees and two-class problems, we have generalized it in [6], [7] to multiway splits and multiclass problems to make comparisons with other methods easier.

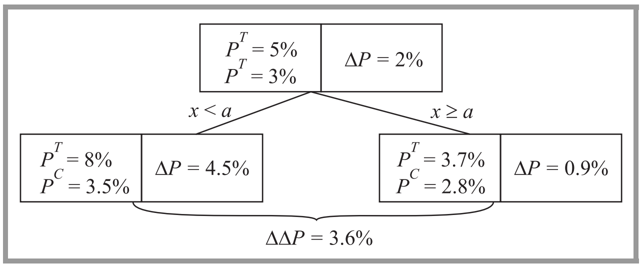


Fig. 4. An example calculation of the $\Delta\Delta P$ criterion

The first paper explicitly discussing uplift modeling was [3]. It presents an extensive motivation including several used cases. Recently, a detailed description of their decision tree learning algorithm has been published in [4]. The decision trees have been adapted to the uplift case by using a splitting criterion, based on statistical tests of the differences between treatment and control probabilities introduced by the split. There is also a variance based pruning technique. See [4] for more details.

Other approaches to uplift modeling include modifications of the naive Bayesian classifier and logistic regression [10], or different approaches to uplift decision tree learning, see e.g., [9].

In [6], [7] we have presented another algorithm for learning uplift decision trees. Our approach follows the more modern tree learning algorithms which use information theory for test selection. We describe it in the next section.

4. Information Theory Based Uplift Decision Trees

In [6], [7] we presented an approach to uplift decision tree learning more in the spirit of modern learning algorithms (such as Quinlan's C4.5 [11]) with tests selected based on information theoretical measures, and overfitting controlled by tree pruning. The first paper presented the case where all customers receive and identical offer, the second extended the approach to the case when multiple treatments are possible. In the remaining part of the paper we only deal with the single treatment case. This section provides a description of those algorithms, which, while being quite thorough, leaves out several details. The reader is referred to [6], [7] for a full description.

4.1. Notation

Let us now introduce the notation used in this section. Recall that nonleaf nodes in a decision trees are labeled with *tests* [11]. We create a single test for each categorical attribute, the outcomes of this test are all attribute's values. For each numerical attribute X we create tests of the form $X < v$, where v is a real number. Tests will be denoted with uppercase letter A and the class attribute with the letter Y. Values from the domains of attributes and test outcomes will be denoted by corresponding lowercase letters, e.g., a will denote an outcome of a test A, and y a specific class; $\sum_a$ denotes the sum over all outcomes of a test A, and $\sum_y$ the sum over all classes.

We need to introduce special notation reflecting the fact, that, contrary to the standard Machine Learning setting, we now have *two* training datasets: treatment and control. The probabilities estimated from the treatment dataset will be denoted by P^T and those estimated from the control dataset by P^C. We assume that Laplace correction is used while estimating the probabilities P^T and P^C.

Additionally, let N^T and N^C denote the number of records in the treatment and control samples respectively, and $N^T(a)$ and $N^C(a)$, the number of records in which the outcome of a test A is a. Finally let $N = N^T + N^C$ and $N(a) = N^T(a) + N^C(a)$.

4.2. Splitting Criteria

One of the most important aspects of a decision tree learning algorithm is the criterion used to select tests in the nodes of the tree. In this section we present two uplift specific splitting criteria. Instead of using the target quantity directly, we attempt to model the amount of *information* that a test gives about the difference between treatment and control class probabilities. In [6], [7] we stated several postulates which an uplift splitting criterion should satisfy, and proved that our criteria do indeed satisfy them.

The splitting criteria we propose are based on distribution divergences [12]–[15] – information theoretical measures of differences between distributions. We use two distribution divergence measures, the Kullback-Leibler divergence [12], [14] and the squared Euclidean distance [13]. Those divergences, from a distribution $Q = (q_1, \dots, q_n)$ to a distribution $P = (p_1, \dots, p_n)$, are defined respectively as

$$\begin{aligned} KL(P:Q) &= \sum_i p_i \log \frac{p_i}{q_i}, \\ E(P:Q) &= \sum_i (p_i - q_i)^2. \end{aligned}$$

Given a divergence measure D, our splitting criterion is

$$D_{gain}(A) = D\left(P^T(Y) : P^C(Y)|A\right) - D\left(P^T(Y) : P^C(Y)\right),$$

where A is a test and $D\left(P^T(Y) : P^C(Y)|A\right)$, the conditional divergence defined below. Substituting for D the KL-divergence and squared Euclidean distance divergence we obtain our two proposed splitting criteria, the KL_{gain} and E_{gain}.

To justify the definition, note that we want to build the tree, in which the distributions in the treatment and control groups differ as much as possible. The first part of the expression picks a test, which leads to most divergent class distributions in each branch; from this value we subtract the divergence between class distributions on the whole dataset in order to measure the increase or *gain* of the divergence resulting from splitting with test A. This is analogous to entropy gain [11] and Gini gain [16] used in standard decision trees. In fact, one of our postulates was that, when the control dataset is missing the splitting criteria should reduce to entropy and Gini gains respectively [6].

Conditional KL-divergences have been used in literature [14] but the definition is not directly applicable to our case, since the probability distributions of the test A differ in the treatment and control groups. We have thus defined conditional divergence as:

$$D(P^T(Y) : P^C(Y)|A) = \sum_a \frac{N(a)}{N} D\Big(P^T(Y|a) : P^C(Y|a)\Big). \tag{1}$$

The relative influence of each test value is proportional to the total number of training examples falling into its branch in both treatment and control groups.

4.3. Correcting for Tests with Large Number of Splits and Imbalanced Treatment and Control Splits

In order to prevent a bias towards tests with high number of outcomes decision, tree learning algorithms normalize the information gain dividing it by the information value of the test itself [11]. In our case the normalization factor is more complicated, as the information value can be different in the control and treatment groups. Moreover, it is desirable to punish tests, which split the control and treatment groups in different proportions, since such splits indicate that the test is not independent from the assignment of cases to the treatment and control groups.

The proposed normalization value for a test A is given by

$$\begin{aligned} I(A) = H\left(\frac{N^T}{N}, \frac{N^C}{N}\right) KL(P^T(A) : P^C(A)) \\ + \frac{N^T}{N} H(P^T(A)) + \frac{N^C}{N} H(P^C(A)) + \frac{1}{2}, \end{aligned} \tag{2}$$

for the KL_{gain} criterion, and

$$\begin{aligned} J(A) = Gini\left(\frac{N^T}{N}, \frac{N^C}{N}\right) E(P^T(A) : P^C(A)) \\ + \frac{N^T}{N} Gini(P^T(A)) + \frac{N^C}{N} Gini(P^C(A)) + \frac{1}{2}, \end{aligned}$$

for the E_{gain} criterion.

The first term is responsible for penalizing uneven splits. The unevenness of splitting proportions is measured using the divergence between the distributions of the test outcomes in the treatment and control datasets. However, penalizing uneven splits only makes sense if there is enough data in *both* treatment and control groups. The $KL(P^T(A) : P^C(A))$ term is thus multiplied by $H\left(\frac{N^T}{N}, \frac{N^C}{N}\right)$, which is close to zero when there is a large imbalance between the number of data in treatment and control groups (analogous, Gini based measures are used for E_{gain}). The following two terms penalize tests with large numbers of outcomes, just as in classification decision trees [11]. The final $\frac{1}{2}$ term prevents the division by very small normalization factors from inflating the value of the splitting criterion for tests with highly imbalanced outcome probabilities. Notice that when $N^C = 0$ the criterion reduces to $H(P^T(A)) + \frac{1}{2}$ which is identical to normalization used in standard decision tree learning (except for the extra $\frac{1}{2}$). After taking the normalization into account, the final splitting criteria become

$$\frac{KL_{ratio}(A)}{I(A)}, \quad \text{and} \quad \frac{E_{ratio}(A)}{J(A)}.$$

4.4. Application of the Tree

Once the tree has been built, its leaves correspond to subgroups of objects, for which the treatment and control class distributions differ. The question now is how to apply the tree to make decisions on whether the marketing action should be applied to customers falling into a given leaf. To this end, we annotate each leaf with an expected profit, which can also be used for scoring new customers.

The assignment of profits uses an approach similar to [1], [9]. Each class y is assigned to profit v_y, that is, the expected income if a given object (whether treated or not) falls into this class. If each object in a leaf l is targeted, the expected profit (per object) is equal to $-c + \sum_y P^T(y|l)v_y$, where c is the cost of performing the action. If no object in l is targeted, the expected profit is $\sum_y P^C(y|l)v_y$. Combining the two, we get the following expected gain from treating each object falling into l:

$$-c + \sum_y v_y \left(P^T(y|l) - P^C(y|l)\right). \tag{3}$$

4.5. Pruning

Decision tree pruning has decisive influence on the performance of the model. There are several pruning methods, based on statistical tests, Minimum Description Length principle, and others [11], [17]–[19].

We chose the simplest, but nevertheless effective pruning method based on using a separate validation set [17], [18]. For the classification problem, after the full tree has been built on the training set, the method traverses the tree bottom up and tests, for each node, whether replacing the subtree rooted at that node with a single leaf would improve accuracy on the validation set. If this is the case, the subtree is replaced, and the process continues.

Applying this method to uplift modeling required an analogue of classification accuracy. To this end we have devised a measure of improvement called the *maximum class probability difference*, which can be viewed as a generalization of classification accuracy to the uplift case. The idea is to look at the differences between treatment and control probabilities in the root of the subtree and in its leaves, and prune if, overall, the differences in leaves are not greater than the difference in the root. In each node we only look at the class, for which the difference was largest on the training set, and in addition remember the sign of that difference such that only differences, which have the same sign in the training and validation sets contribute to the increase of our criterion.

More formally, while building the tree on the *training* set, for each node t, we store the class $y^*(t)$, for which the difference $\left|P^T(y^*|t) - P^C(y^*|t)\right|$ is maximal, and also remember the sign of this difference $s^*(t) = \mathrm{sgn}(P^T(y^*|t) - P^C(y^*|t))$. During the pruning step, suppose we are examining a subtree with root r and leaves $l_1, \ldots, l_k$. We calculate the following quantities with the stored values of y^* and s^*, and all probabilities computed on the *validation* set:

$$\begin{aligned} d_1(r) &= \sum_{i=1}^{k} \frac{N(l_i)}{N(r)} s^*(l_i)\left(P^T(y^*(l_i)|l_i) - P^C(y^*(l_i)|l_i)\right), \\ d_2(r) &= s^*(r)\left(P^T(y^*(r)|r) - P^C(y^*(r)|r)\right), \end{aligned}$$

where $N(l_i)$ is the number of validation examples (both treatment and control) falling into the leaf l_i. The first quantity is the maximum class probability difference of the unpruned subtree and the second is the maximum class probability difference we would obtain on the validation set, if the subtree was pruned and replaced with a single leaf. The subtree is pruned if $d_1(r) \leq d_2(r)$.

The class y^* is an analogue of the predicted class in standard classification trees. In [7] we describe the relation of maximum class probability difference to classification accuracy.

5. Experimental Evaluation on Direct Marketing Data

We now present an application of uplift models, as well as traditional response models to the problem of selection of customers for an e-mail campaign based on the data described in Section 2. The target is to maximize the num-

Table 3
Models used in the experiments

Response models	
SingleTree.E	Decision tree model based on the E_{ratio} criterion
SingleTree.KL	Decision tree model based on the KL_{ratio} criterion
SingleTree.J48	Decision tree model based on J48 Weka implementation
Uplift models	
UpliftTree.E	Uplift decision tree based on the E_{ratio} criterion
UpliftTree.KL	Uplift decision tree based on the KL_{ratio} criterion
DoubleTree.J48	Separate decision trees for the treatment and control groups (J48 Weka implementation)

ber of visits to the web site that were *driven* by the campaign.

We compared six different models, three response models and three uplift models (Table 3).

The models were evaluated using 10×10 crossvalidation, all figures present results obtained on the test folds.

We begin by building models with both types of campaign e-mails treated jointly. The results for traditional response models are presented in Fig. 5. The figure shows cumulative percent of total page visits for customers sorted from the highest to the lowest score. The area under the curve for each model is included in the legend. The given value is the actual area under the curve, from which the area under the diagonal line corresponding to random selection is subtracted. The greater the area, the better. We can see that all traditional response models perform much better at predicting who will visit the site than random selection.

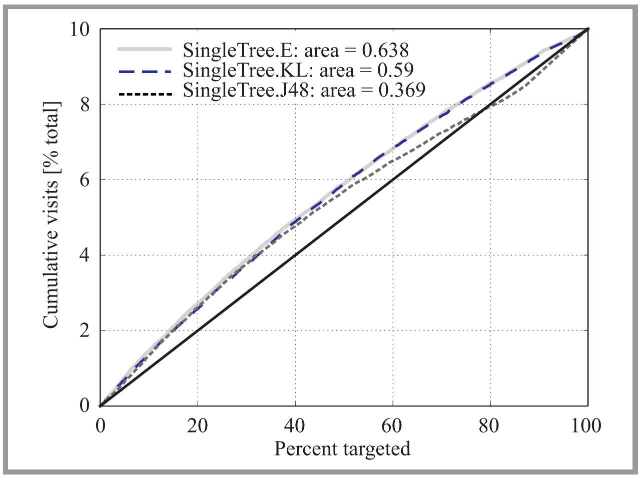


Fig. 5. Cumulative visits (lift) predicted by classification models built just on the treatment dataset.

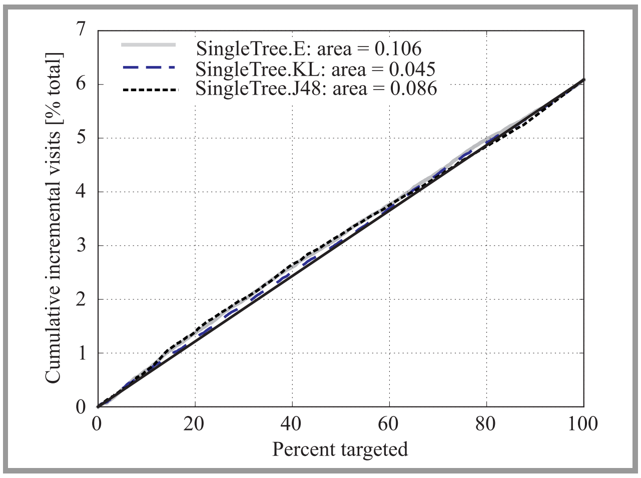


Fig. 6. Cumulative incremental visits (uplift) predicted by classification models built just on the treatment dataset.

Traditional models predict all possible visits, so they indicate as positive customers visit the site spontaneously, as well as those who visit as a result of the campaign. However, those models are not successful in predicting new visits. To indicate this, Fig. 6 shows the cumulative percentage (of the total population) of the *new visits*. The curve is obtained by subtracting two gain curves (such as those used in Fig. 5): the one obtained on the control dataset from the one obtained on the treatment dataset. Areas under those curves are also indicated. Fig. 7 includes the same results for dedicated uplift models.

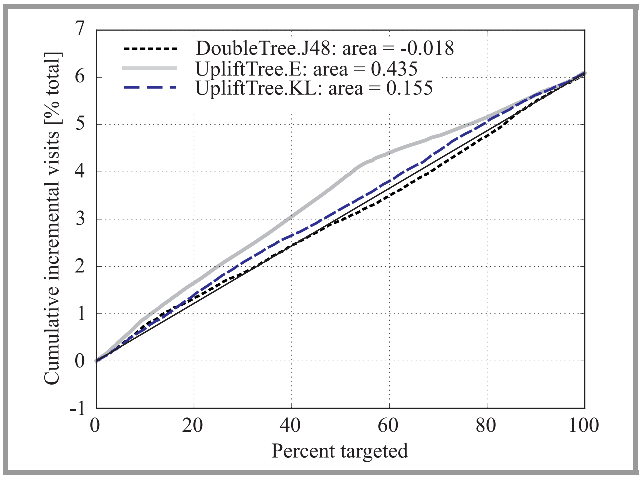


Fig. 7. Cumulative incremental visits (uplift) predicted by uplift models built on treatment and control datasets.

Results presented in Fig. 6 and Fig. 7 show that traditional response models are very poor in predicting uplift, i.e., which customers are likely to visit the site *because* of the campaign (areas under their uplift curves are practically equal to random selection), even though they are highly effective in predicting who will visit the site, i.e., combined spontaneous and campaign induced visits. This is not what a marketer is looking for, because targeting customers, which have high response scores does not generate a tangible increase in the number of visits.

In contrast, uplift models perform much better at predicting new visits. This is especially true for the model based on the E_{ratio} criterion, which very significantly outperformed all response based models. The KL_{ratio} based model performed much worse than the E_{ratio} based, but still outperforms traditional response models. The approach based on two separate models also performed poorly, confirming the superiority of dedicated uplift approaches.

Below, we show two top levels of an uplift decision tree for combined men's and women's merchandise campaigns (UpliftTree.E built on one of the crossvalidation folds). The *womens* attribute gives the most information about the increase in visits, and is placed in the root of the tree. It splits the data more or less in half. In a subgroup of 55.3% of the customers ($womens = 1$) we reached an uplift of 7.9% and in 45% of this subgroup ($zip\_code = Suburban$) an uplift of 8.4%. This is much more than the average uplift of 6.1%. In a small group ($womens = 0$, $history \geq 1621.49$) the uplift is negative (-17.3%); the campaign had a negative effect on this group (note that these are highly valuable customers who made large purchases before).

UpliftTree.E (Combined campaigns):
Total uplift = 6.1%

– [44.7%] $womens = 0$: $uplift = 3.8\%$
  - [0.1%] $history \geq 1621.49$: $uplift = -17.3\%$
  - [99.9%] $history < 1621.49$: $uplift = 3.9\%$
– [55.3%] $womens = 1$: $uplift = 7.9\%$
  - [14.8%] $zip\_code = Rural$: $uplift = 5.9\%$
  - [45.0%] $zip\_code = Suburban$: $uplift = 8.4\%$
  - [40.2%] $zip\_code = Urban$: $uplift = 8.1\%$

Next, new models were built on women's and men's merchandise campaign data separately. As the results for the men's merchandise campaign showed little dependence on customers' attributes, we show only the results for the women's merchandise campaign. The results are presented in Figs. 8, 9 and 10). The advantage of uplift models is much more pronounced than in the case of both campaigns treated jointly. The KL_{ratio} based model worked very well in this case, its performance was practically identical to that of the E_{ratio} based model, and much better than the performance of the model based on two separate decision trees. It is enough to target just about half of the customers to achieve results almost identical to targeting the whole database.

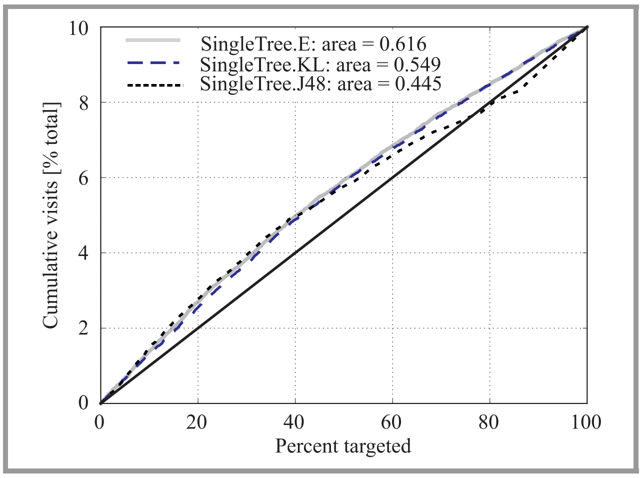


Fig. 8. Cumulative visits (lift) after the women's merchandise campaign predicted by classification models built just on the treatment dataset.

We now look at the top two levels of an uplift tree model build on the data from women's merchandise campaign. We can see that also for this group the *women's* attribute is very important. In a group of 55.3% of the customers ($womens = 1$) the uplift is 7.3%. It means that by directing the campaign to this group we can encourage $55.3\% \times 7.3\% = 4.04\%$ of the total population to visit our site.

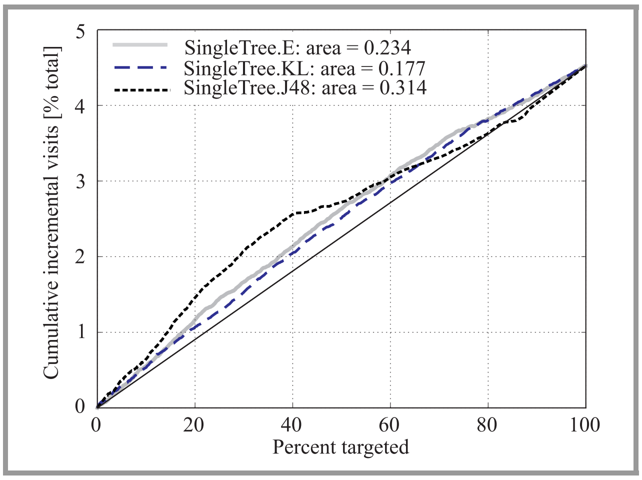


Fig. 9. Cumulative incremental visits (uplift) after women's campaign predicted by classification models built just on the treatment dataset.

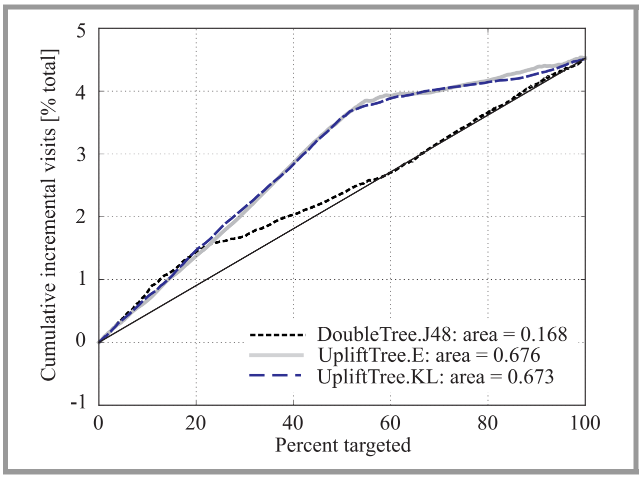


Fig. 10. Cumulative incremental visits (uplift) after women's campaign predicted by uplift models built on the treatment and control datasets.

UpliftTree.E (Women's merchandise campaign):
Total uplift = 4.5%

– [44.9%] $womens = 0$: $uplift = 1.1\%$
  - [0.2%] $history \geq 1618.85$: $uplift = -26.3\%$
  - [99.8%] $history < 1618.85$: $uplift = 1.1\%$
– [55.3%] $womens = 1$: $uplift = 7.3\%$
  - [0.9%] $history \geq 1317.02$: $uplift = -9.4\%$
  - [99.1%] $history < 1317.02$: $uplift = 7.5\%$

6. Conclusions

Our experiments confirm the usefulness of uplift modeling in campaign optimization. Using uplift models, we can predict new buyers much more precisely than using traditional response or propensity approaches. The effectiveness in predicting new visits by response models is low, even if

accuracy of predicting all visits is high. The reason for this is that the response models do not distinguish between spontaneous and new buyers. Quite often, the spontaneous hits are more frequent, and the models tend concentrate on them. Only if the uplift is correlated with the class itself, the response models are able to indicate new buyers.
Additionally, our experiments confirm that dedicated uplift modeling algorithms are more effective than the naive approach based on two separate models.

7. Acknowledgments

This work was supported by Research Grant no. N N516 414938 of the Polish Ministry of Science and Higher Education (Ministerstwo Nauki i Szkolnictwa Wyższego) from research funds for the period 2010–2012.

References

[1] B. Hansotia and B. Rukstales, “Incremental value modeling”, *J. Interactive Marketing*, vol. 16, no. 3, pp. 35–46, 2002.

[2] N. J. Radcliffe and R. Simpson, “Identifying who can be saved and who will be driven away by retention activity”, *White paper*, Stochastic Solutions Limited, 2007.

[3] N. J. Radcliffe and P. D. Surry, “Differential response analysis: modeling true response by isolating the effect of a single action”, in *Proc. Credit Scoring Credit Control VI*, Edinburgh, Scotland, 1999.

[4] N. J. Radcliffe and P. D. Surry, “Real-world uplift modelling with significance-based uplift trees”, Portrait Tech. Rep. TR-2011-1, Stochastic Solutions, 2011.

[5] K. Hillstrom, “The MineThatData e-mail analytics and data mining challenge”, MineThatData blog, 2008 [Online]. Available: http://blog.minethatdata.com/2008/03/minethatdata-e-mail-analytics-and-data.html, retrieved on 02.04.2012.

[6] P. Rzepakowski and S. Jaroszewicz, “Decision trees for uplift modeling”, in *Proc. 10th IEEE Int. Conf. Data Mining ICDM-2010*, Sydney, Australia, Dec. 2010, pp. 441–450.

[7] P. Rzepakowski and S. Jaroszewicz, “Decision trees for uplift modeling with single and multiple treatments”, *Knowledge and Information Systems*, pp. 1–25, 2011 [Online]. Available: http://www.springerlink.com/content/f45pw0171234524j

[8] C. Manahan, “A proportional hazards approach to campaign list selection”, in *Proc. Thirtieth Ann. SAS Users Group Int. Conf. SUGI*, Philadelphia, PA, 2005.

[9] D. M. Chickering and D. Heckerman, “A decision theoretic approach to targeted advertising”, in *Proc. 16th Conf. Uncertainty in Artif. Intell. UAI-2000*, Stanford, CA, 2000, pp. 82–88.

[10] V. S. Y. Lo, “The true lift model – a novel data mining approach to response modeling in database marketing”, *SIGKDD Explor.*, vol. 4, no. 2, pp. 78–86, 2002.

[11] J. R. Quinlan, *C4.5: Programs for Machine Learning.* Morgan Kauffman, 1992.

[12] I. Csiszár and P. Shields, “Information theory and statistics: A tutorial”, *Found. Trends in Commun. Inform. Theory*, vol. 1, no. 4, pp. 417–528, 2004.

[13] L. Lee, “Measures of distributional similarity”, in *Proc. 37th Ann. Meet. Associ. Computat. Linguistics ACL-1999*, Maryland, USA, 1999, pp. 25–32.

[14] T. S. Han and K. Kobayashi, *Mathematics of information and coding.* Boston, USA: American Mathematical Society, 2001.

[15] S. Jaroszewicz and D. A. Simovici, “A general measure of rule interestingness”, in *Proc. 5th Eur. Conf. Princ. Data Mining Knowl. Discov. PKDD-2001*, Freiburg, Germany, 2001, pp. 253–265.

[16] L. Brieman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Monterey, USA: Wadsworth Inc., 1984.

[17] T. Mitchell, *Machine Learning.* McGraw Hill, 1997.

[18] J. R. Quinlan, “Simplifying decision trees”, *Int. J. Man-Machine Studies*, vol. 27, no. 3, pp. 221–234, 1987.

[19] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann, 2005.

Piotr Rzepakowski received his M.Sc. degree in Computer Science from Warsaw University of Technology, Poland, in 2003. Currently, he is a Ph.D. student at the Faculty of Electronics and Information Technology at Warsaw University of Technology and a research assistant at the National Institute of Telecommunications in Warsaw, Poland. His research interests include data mining, data analysis and decision support. He has taken part in several industrial projects related to data warehousing and data analysis.
E-mail: P.Rzepakowski@itl.waw.pl
National Institute of Telecommunications
Szachowa st 1
04-894 Warsaw, Poland

Szymon Jaroszewicz is currently an Associate Professor at the National Institute of Telecommunications, Warsaw, Poland and at the Institute of Computer Science of the Polish Academy of Sciences. He received the Master's degree in Computer Science at the Department of Computer Science at the Szczecin University of Technology in 1998 and his Ph.D. at the University of Massachusetts Boston in 2003, where in 1998 and 1999 he was a Fulbright scholar. In 2010, he received his D.Sc. degree at the Institute of Computer Science, Polish Academy of Sciences. His research interests include data analysis, data mining and probabilistic modeling; he is the author of several publications in those fields. He has served as a program committee member for major data mining conferences and he is a member of the editorial board of Data Mining and Knowledge Discovery Journal.
E-mail: S.Jaroszewicz@itl.waw.pl
National Institute of Telecommunications
Szachowa st 1
04-894 Warsaw, Poland

Paper

Integrated Routing and Network Flow Control Embracing Two Layers of TCP/IP Networks – Methodological Issues

Andrzej Karbowski[a,b]

[a] *Institute of Control and Computation Engineering, Warsaw University of Technology, Warsaw, Poland*
[b] *Research and Academic Computer Network (NASK), Warsaw, Poland*

Abstract—A cross-layer network optimization problem is considered. It involves network and transport layers, treating both routing and flows as decision variables. Due to the nonconvexity of the capacity constraints, when using Lagrangian relaxation method a duality gap causes numerical instability. It is shown that the rescue preserving separability of the problem may be the application of the augmented Lagrangian method, together with Cohen's Auxiliary Problem Principle.

Keywords—decomposition, flow control, Lagrangian relaxation, networks, optimization, routing, TCP/IP.

1. Introduction

In the standard approach TCP congestion control together with active queue management (AQM) algorithms attempted to maximize aggregated utility over source rates, assuming that routing is given and fixed at the timescale of interest. However, it seems that it would be more profitable, when we will treat TCP and IP layers together and maximize cross-layer utility at the timescale of route changes. The integrated routing and network flow control problem was first addressed by Wang, Li, Low and Doyle [1] and independently by Jaskóła and Malinowski [2]. Unfortunately, due to the nonconvexity of the constraints' functions, the algorithm based on the price method (Lagrangian relaxation) is numerically unstable. Duality gap is the reason of problems [1]. The paper shows how this gap can be overcome, while not losing separability of the problem.

2. Problem Formulation

Our goal is to maximize the sum of utilities of all connections with respect to routing and flows over the whole network, taking into account the capacities of links.

Formally, the optimization problem can be described as follows:

$$\max_{x\in X, R\in\mathcal{R}} \sum_{s\in S} U_s(x_s)\,, \tag{1}$$

$$Rx \le c,\; R = [r_{ij}]_{\overline{\overline{L}}\times\overline{\overline{S}}}\,, \tag{2}$$

where:

x_s – flow from the source s to a (single) destination node;

$x \in X \subset \mathbb{R}^{\overline{\overline{S}}}$ – vector of all flows;

S – the set of all sources;

X – the set of admissible flows; it is a Cartesian product of intervals X_s belonging to nonnegative half lines;

U_s – the sources' (connections') utility functions; it is assumed, that they are strictly concave and continuous;

L – the set of all links;

R – the matrix of binary elements with the number of rows equal the number of links $\overline{\overline{L}}$ and the number of columns equal the number of sources (active connections at a given time); the element r_{ls} equals 1 when the link l belongs to a path from the source s to a given destination node;

R_s – s-th column of the matrix R;

$\mathcal{R}_s$ – the set of all possible vectors representing paths from s to a given destination node;

$\mathcal{R}$ – the set of all possible matrices, that is all possible combinations of vectors from the sets $\mathcal{R}_s$;

$c \in \mathbb{R}_+^{\overline{\overline{L}}}$ – links capacity vector.

3. The Standard Price Decomposition Method

The Lagrangian for the problem (1)–(2) is as follows:

$$\begin{aligned} L(x,R,\lambda) &= \sum_{s\in S} U_s(x_s) - \sum_{l\in L} \lambda_l \left(\sum_{s\in S} r_{ls}x_s - c_l \right) \\ &= \sum_{s\in S} \left(U_s(x_s) - x_s \sum_{l\in L} \lambda_l r_{ls} \right) + \sum_{l\in L} \lambda_l c_l\,, \end{aligned} \tag{3}$$

where λ_l are nonnegative Lagrange multipliers. Due to the duality theory, this Lagrangian will be further maximized with respect to x and R and minimized with respect to λ [3]. The iteration in the standard price method consists of two-steps [1]:

1. Solve the primal problem

$$(x(t),R(t)) = \arg \max_{x\in X, R\in \mathcal{R}} \sum_{s\in S} \left(U_s(x_s) - x_s \sum_{l\in L} \lambda_l(t) r_{ls} \right). \tag{4}$$

   Let us notice that owing to the specific structure and the nonnegativity of x_s for all s the overall optimization problem (4) can be decomposed in the following way:

$$\begin{aligned} &\max_{x\in X} \max_{R\in\mathcal{R}} \sum_{s\in S} \left(U_s(x_s) - x_s \sum_{l\in L} \lambda_l(t) r_{ls} \right) = \\ &= \sum_{s\in S} \max_{x_s\in X_s} \left[U_s(x_s) + \max_{R_s\in\mathcal{R}_s} \left(-x_s \sum_{l\in L} \lambda_l(t) r_{ls} \right) \right] \\ &= \sum_{s\in S} \max_{x_s\in X_s} \left[U_s(x_s) - x_s \min_{R_s\in\mathcal{R}_s} \left(\sum_{l\in L} \lambda_l(t) r_{ls} \right) \right]. \end{aligned} \tag{5}$$

   From the final form of Eq. (5) it is seen, that [1]:

   - the primal problem (4) can be decomposed into a family of problems assigned to subsequent sources s with local variables $x_s, r_{1s}, r_{2s}, \ldots$ which can be solved independently,
   - the inner optimization $\min_{R_s\in\mathcal{R}_s} \sum_{l\in L} \lambda_l(t) r_{ls}$ for the given source index s (and its connection to a destination node) is simply the shortest path problem with metrics defined by Lagrange multipliers $\lambda_l(t)$, $l = 1, 2, \ldots$.

   Summing up, the problem (4) may be solved by solving for every source $s \in S$:

   - The shortest path problem:

$$R_s(t) = \arg \min_{R_s\in\mathcal{R}_s} \sum_{l\in L} \lambda_l(t) r_{ls}. \tag{6}$$

     Let us denote the optimal value of the performance index in Eq. (6) as $d_s(t)$, that is:

$$d_s(t) = \sum_{l\in L} \lambda_l(t) r_{ls}(t). \tag{7}$$

   - The flow optimization problem:

$$\max_{x_s\in X_s} \left(U_s(x_s) - x_s d_s(t) \right). \tag{8}$$

2. Modify Lagrange multipliers so as to get a better approximation of the solution of the dual problem $\min_{\lambda\geq 0}[L_D(\lambda) = \max_{x\in X, R\in\mathcal{R}} L(x,R,\lambda)]$

$$\lambda_l(t+1) = \max\left(0, \lambda_l(t) + \rho \left(\sum_{s\in S} r_{ls}(t) x_s(t) - c_l \right)\right), \; l \in L, \tag{9}$$

   where $\rho > 0$ is a properly chosen step coefficient.

Unfortunately, this algorithm is unstable [1]. The reason is a duality gap caused by the nonconvexity of capacity constraint (2) and the discrete character of variables r_{ls}.

4. Augmented Lagrangian Approach and Auxiliary Problem Principle in Cross-Layer Optimization

In optimization problems where the duality gap is present, we use augmented Lagrangian or, in other words, shifted penalty function method [3], [4], [5]. For the problem (1)–(2) it will have the form:

$$\begin{aligned} L_a(x,R,\lambda) = &\sum_{s\in S} U_s(x_s) - \frac{1}{2} \sum_{l\in L} \rho_l \left\{ \left[\max\left(0, \left(\sum_{s\in S} r_{ls} x_s - c_l\right) + \right.\right.\right. \\ &\left.\left.\left. + \frac{\lambda_l}{\rho_l}\right)\right]^2 - \left(\frac{\lambda_l}{\rho_l}\right)^2 \right\} = \sum_{s\in S} U_s(x_s) + \\ &- \frac{1}{2} \sum_{l\in L} \frac{\rho_l}{\rho_l^2} \left\{ \left[\rho_l \max\left(0, \left(\sum_{s\in S} r_{ls} x_s - c_l\right) + \right.\right.\right. \\ &\left.\left.\left. + \frac{\lambda_l}{\rho_l}\right)\right]^2 - \rho_l^2 \left(\frac{\lambda_l}{\rho_l}\right)^2 \right\} = \sum_{s\in S} U_s(x_s) + \\ &- \frac{1}{2} \sum_{l\in L} \frac{1}{\rho_l} \left\{ \left[\max\left(0, \lambda_l + \rho_l \left(\sum_{s\in S} r_{ls} x_s + \right.\right.\right.\right. \\ &\left.\left.\left.\left. - c_l\right)\right)\right]^2 - \lambda_l^2 \right\}, \end{aligned} \tag{10}$$

where $\rho_l, l \in L$ are penalty coefficients.

The solution of the problem (1)-(2) is sought, as before, by solving the minimax problem:

$$\min_{\lambda\geq 0} \max_{x\in X, R\in\mathcal{R}} L_a(x,R,\lambda). \tag{11}$$

Augmented Lagrangians have one serious drawback – due to the quadratic terms (in our case – squares of the sums of products of variables) they are not separable, that is the optimization problem is not decomposable.

The easiest way to transform the augmented Lagrangian to a separable form consists in the application of so-called Auxiliary Problem Principle proposed by Cohen [6], [7]. This principle says, that if we want to solve the problem:

$$\max_{u\in U} J_1(u) + J_2(u), \tag{12}$$

where J_1 is an additive (that is separable), strictly concave functional, while J_2 is a differentiable, nonadditive, not necessarily strictly concave, functional, we may instead solve a sequence of auxiliary problems:

$$\begin{aligned} u(t+1) = \arg\max_{u\in U} \Big[G_\varepsilon^{u(t)}(u) = \\ \varepsilon J_1(u) + \varepsilon < J_2'(u(t)), u > - K(u) + < K'(u(t)), u > \Big]. \end{aligned} \tag{13}$$

In the above expression, $< .,. >$ denotes the scalar product, $\varepsilon > 0$ – a constant parameter, t is the index of iteration and

$$K(u) = ||u||_2^2. \tag{14}$$

In short, the idea of this transformation lies in the linearization of the nonseparable component and addition of a regularizing, strictly concave, proximal component (more precisely, the subtraction of a strictly convex proximal component $||u-u(t)||_2^2$, with accuracy to the constant $||u(t)||_2^2$, which does not influence the optimization).

In the case of our problem (1)–(2) with the augmented Lagrangian Eq. (10):

$$u = \begin{bmatrix} x \\ R \end{bmatrix}, \tag{15}$$

$$J_1(u) = \sum_{s\in S} U_s(x_s), \tag{16}$$

$$J_2(u) = -\frac{1}{2}\sum_{l\in L}\frac{1}{\rho_l}\left\{\left[\max\left(0,\lambda_l+\rho_l\left(\sum_{s\in S} r_{ls}x_s - c_l\right)\right)\right]^2 - \lambda_l^2\right\} \tag{17}$$

and

$$\begin{aligned} G_\varepsilon^{u(t)}(u) = \varepsilon\sum_{s\in S} U_s(x_s) - \varepsilon\sum_{s\in S}\Bigg\{\sum_{l\in L}\Bigg[\max\Bigg(0,\lambda_l+ \\ +\rho_l\left(\sum_v r_{lv}(t)x_v(t)-c_l\right)\Bigg) r_{ls}(t)x_s+ \\ +\max\left(0,\lambda_l+\rho_l\left(\sum_v r_{lv}(t)x_v(t)-c_l\right)\right)x_s(t)r_{ls}\Bigg]\Bigg\}+ \\ -\sum_{s\in S}\left(x_s^2+\sum_{l\in L} r_{ls}^2\right)+2\sum_{s\in S} x_s(t)x_s+2\sum_{s\in S}\sum_{l\in L} r_{ls}(t)r_{ls}. \end{aligned} \tag{18}$$

5. Decomposition Scheme and the Algorithm

Grouping together and rearranging terms dependent on the same variables in (18), we will get:

$$\begin{aligned} G_\varepsilon^{u(t)}(u) = \sum_{s\in S}\Bigg[\varepsilon U_s(x_s)-x_s^2+2x_s(t)x_s+ \\ -\varepsilon\sum_{l\in L}\max\left(0,\lambda_l+\rho_l\left(\sum_v r_{lv}(t)x_v(t)-c_l\right)\right)r_{ls}(t)x_s\Bigg]+ \\ -\sum_{s\in S}\sum_{l\in L}\Bigg[r_{ls}^2-2r_{ls}(t)r_{ls}+\varepsilon\max\Bigg(0,\lambda_l+ \\ +\rho_l\left(\sum_v r_{lv}(t)x_v(t)-c_l\right)\Bigg)x_s(t)r_{ls}\Bigg]. \end{aligned} \tag{19}$$

Let us notice that for $r_{ls}\in\{0,1\}$, $r_{ls}^2=r_{ls}$, so we will finally get:

$$\begin{aligned} G_\varepsilon^{u(t)}(u) = \sum_{s\in S}\Bigg\{\varepsilon U_s(x_s)-x_s^2+\Bigg[2x_s(t)+ \\ -\varepsilon\sum_{l\in L}\max\left(0,\lambda_l+\rho_l\left(\sum_v r_{lv}(t)x_v(t)-c_l\right)\right)r_{ls}(t)\Bigg]x_s\Bigg\}+ \\ -\sum_{s\in S}\sum_{l\in L}\Bigg\{\Bigg[1-2r_{ls}(t)+\varepsilon\max\Bigg(0,\lambda_l+ \\ +\rho_l\left(\sum_v r_{lv}(t)x_v(t)-c_l\right)\Bigg)x_s(t)\Bigg]r_{ls}\Bigg\}. \end{aligned} \tag{20}$$

Let us denote now:

$$\begin{aligned} V_s^{u(t)}(x_s,\lambda,\varepsilon,\rho) = \varepsilon U_s(x_s)-x_s^2+ \\ +\left[2x_s(t)-\varepsilon\sum_{l\in L}\max\left(0,\lambda_l+\rho_l\left(\sum_v r_{lv}(t)x_v(t)-c_l\right)\right)r_{ls}(t)\right]x_s, \end{aligned} \tag{21}$$

$$\begin{aligned} \varphi_{ls}^{u(t)}(\lambda_l,\varepsilon,\rho_l) = 1-2r_{ls}(t)+ \\ +\varepsilon\max\left(0,\lambda_l+\rho_l\left(\sum_v r_{lv}(t)x_v(t)-c_l\right)\right)x_s(t). \end{aligned} \tag{22}$$

With this notation the function $G_\varepsilon^{u(t)}(u)$ can be written as:

$$G_\varepsilon^{u(t)}(u) = \sum_{s\in S} V_s^{u(t)}(x_s,\lambda,\varepsilon,\rho) - \sum_{s\in S}\sum_{l\in L}\varphi_{ls}^{u(t)}(\lambda_l,\varepsilon,\rho_l)r_{ls}, \tag{23}$$

and the primal optimization problem $\max\limits_{x\in X,R\in\mathcal{R}} L_a(x,R,\lambda)$ with the augmented Lagrangian (10) is equivalent to the following auxiliary problem:

$$\begin{aligned} \max_{x\in X,R\in\mathcal{R}}\left[G_\varepsilon^{u(t)}(u)=\sum_{s\in S}V_s^{u(t)}(x_s,\lambda,\varepsilon,\rho)-\sum_{s\in S}\sum_{l\in L}\varphi_{ls}^{u(t)}(\lambda_l,\varepsilon,\rho_l)r_{ls}\right]= \\ =\max_{x\in X}\sum_{s\in S}V_s^{u(t)}(x_s,\lambda,\varepsilon,\rho)-\min_{R\in\mathcal{R}}\sum_{s\in S}\sum_{l\in L}\varphi_{ls}^{u(t)}(\lambda_l,\varepsilon,\rho_l)r_{ls}= \\ =\sum_{s\in S}\left[\max_{x_s\in X_s}V_s^{u(t)}(x_s,\lambda,\varepsilon,\rho)-\min_{R_s\in\mathcal{R}_s}\sum_{l\in L}\varphi_{ls}^{u(t)}(\lambda_l,\varepsilon,\rho_l)r_{ls}\right]. \end{aligned} \tag{24}$$

Let us notice that the structure of the problem (24) is very similar to the problem (4), but the decomposition scheme goes further, because actually for a given λ we got a complete separation of the shortest path problems (variables r_{ls}), from the flow optimization problems (variables x_s).

The simplest gradient steepest descent algorithm of modification of the Lagrange multipliers due to Eq. (10) will be the following:

$$\begin{aligned} \lambda_l(t+1)=\lambda_l(t)\left(1-\frac{\beta}{\rho_l}\right)+ \\ +\frac{\beta}{\rho_l}\max\left(0,\lambda_l(t)+\rho_l\left(\sum_{s\in S}r_{ls}(t)x_s(t)-c_l\right)\right). \end{aligned} \tag{25}$$

The values of parameters should be chosen from the intervals [7]:

$$0<\beta\le\min_{l\in L}\rho_l,\quad 0<\varepsilon<\frac{b}{\tau^2\max\limits_{l\in L}\rho_l}, \tag{26}$$

where b,τ are, respectively, Lipschitz constants of the function K (14) and the constraint function (2).

Summing up, the iteration of the modified, based on augmented Lagrangian approach, algorithm will be as follows:

1. Solve the primal problem, decomposed into the family of independent problems for every source $s\in S$:

$$x_s(t)=\arg\max_{x_s\in X_s}V_s^{u(t)}(x_s,\lambda(t),\varepsilon,\rho), \tag{27}$$

$$r_{ls}(t)=\arg\min_{R_s\in\mathcal{R}_s}\sum_{l\in L}\varphi_{ls}^{u(t)}(\lambda_l(t),\varepsilon,\rho_l)r_{ls}. \tag{28}$$

Functions $V_s^{u(t)}$ and $\varphi_{ls}^{u(t)}$ are defined by Eqs. (21) and (22).

2. Modify Lagrange multipliers for all links $l \in L$

$$\lambda_l(t+1) = \lambda_l(t)\left(1 - \frac{\beta}{\rho_l}\right) +$$

$$+\frac{\beta}{\rho_l}\max\left(0, \lambda_l(t) + \rho_l\Big(\sum_{s\in S} r_{ls}(t)x_s(t) - c_l\Big)\right). \quad (29)$$

The presented approach was implemented and thoroughly tested on many big networks generated by Netgen [8]. The results proved its high effectiveness [9], [10].

References

[1] J. Wang, L. Li, S. H. Low, and J. C. Doyle, "Cross-layer optimization in TCP/IP networks", *IEEE/ACM Trans. Networking*, vol. 13, no. 3, pp. 582–595, 2005.

[2] P. Jaskóła and K. Malinowski, "Two methods of optimal bandwidth allocation in TCP/IP networks with QoS differentiation", in *Proc. Summer Simulation Multiconf. SPECTS 2004*, San Jose, California, 2004, pp. 373–378.

[3] D. P. Bertsekas, *Lagrange Multiplier Methods in Constrained Optimization*. Academic Press, 1982.

[4] A. P. Wierzbicki, "A penalty function shifting method in constrained static optimization and its convergence properties", *Archiwum Automatyki i Telemechaniki*, vol. 16, pp. 395–416, 1971.

[5] R. T. Rockafellar, "Augmented Lagrange multiplier functions and duality in nonconvex programming", *SIAM J. Control*, vol. 12, no. 2, pp. 268–285, 1974.

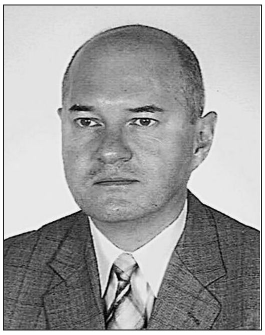
[6] G. Cohen, "Optimization by decomposition and coordination: a unified approach", *IEEE Trans. Autom. Control*, vol. AC-23, no. 2, pp. 222–232, 1978.

[7] G. Cohen, D. L. Zhu, "Decomposition coordination methods in large scale optimization problems. The nondifferentiable case and the use of augmented Lagrangians", in *Advances in Large Scale Systems, Vol. I*, J. B. Cruz, Ed. Greenwich, Connecticut: JAI Press, 1984, pp. 203–266.

[8] D. Klingman, A. Napier, and J. Stutz, "NETGEN: a program for generating large scale capacitated assignment, transportation, and minimum cost flow network problems", *Management Sci.*, vol. 20, no. 5, pp. 814–821, 1974.

[9] P. Paluch, "Cross-layer optimization in TCP/IP networks – the augmented Lagrangian approach assessment", B.Eng. thesis, Warsaw University of Technology, 2011.

[10] A. Karbowski and P. Paluch, "Integrated routing and network flow control embracing two layers of TCP/IP networks – implementation and tests results" in preparation.



Andrzej Karbowski received the M.Sc. degree in Electronic Engineering (specialization automatic control) from Warsaw University of Technology (Faculty of Electronics) in 1983. He received the Ph.D. in Automatic Control and Robotics, in 1990. He works as adjunct both at Research and Academic Computer Network (NASK) and at the Faculty of Electronics and Information Technology (at the Institute of Control and Computation Engineering) of Warsaw University of Technology. His research interests concentrate on data networks management, optimal control in risk conditions, decomposition and parallel implementation of numerical algorithms.

E-mail: A.Karbowski@ia.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Research and Academic Computer Network (NASK)
Wąwozowa st 18
02-796 Warsaw, Poland

Paper

Auction Models Supporting End-to-End Connection Trading

Kamil Kołtyś, Krzysztof Pieńkosz, and Eugeniusz Toczyłowski

Institute of Control and Computation Engineering, Warsaw University of Technology, Warsaw, Poland

Abstract—The paper concerns bandwidth allocation problem on the telecommunication market where there are many sellers and buyers. Sellers offer the bandwidth of telecommunication links. Buyers are interested in the purchase of the bandwidth of several links that makes up an end-to-end connection between two nodes of telecommunication network. We analyze three auction models supporting such a bandwidth exchange: NSP (network second price), BCBT (model for balancing communication bandwidth trading) and BCBT-CG which is a modification of BCBT that applies column generation technique. All of these models concern divisible network resources, treat bandwidth of telecommunication links as an elementary commodity offered for sale, and allow for purchasing bandwidth along multiple paths joining two telecommunication nodes. All of them also aim at maximizing the social welfare. Considered auction models have been compared in the respect of economic and computational efficiency. Experimental studies have been performed on several test instances based on the SNDlib library data sets.

Keywords*—*bandwidth auctions, divisible commodities, end-to-end connections, multi-commodity trade, multi-path routing.

1. Introduction

In this paper we consider a bandwidth market [1], [2] on which many sellers and many buyers are interested in the exchange of the links' bandwidth. The bandwidth of link is an elementary network resource that allows for transmitting some amount of data between two telecommunication nodes over given period of time. Telecommunication network consists of many nodes connected by numerous links. Therefore, on the bandwidth market there are typically many different network resources that may be offered for sale by different sellers, e.g., companies laying cables, network providers and other telecommunication link owners. Generally, the buyers, such as network providers, service providers and geographically spread organizations want to purchase the bandwidth of several links to realize specific network services.

Here, we focus on the case, in which buyers are interested in obtaining end-to-end connections. The end-to-end connection is a network service that allows for transmitting data between two arbitrary nodes in the telecommunication network. These nodes do not have to be directly connected by single link, but they may be joined by a path consisting of many links. Thus, in order to provide an end-to-end connection with predefined capacity, path(s) joining source and destination nodes of this end-to-end connection have to be determined and bandwidth of those path(s) has to be allocated to this end-to-end connection.

Each buyer or seller participates in the bandwidth exchange in order to achieve one's individual goals. The buyer derives utility from getting the end-to-end connection and he wants to purchase this end-to-end connection for the minimum price. The difference between buyer's utility and the end-to-end connection price defines one's net benefit from the trade. On the other hand the seller incurs the cost of network resource and he wants to sell the network resource for the maximum price. The difference between the network resource price and the seller's cost defines one's net benefit from the trade. Rational market participants aims at maximization of their net benefits.

The sum of all market participants' net benefits is called social welfare. From the global point of view it is desirable to ensure economic efficiency of bandwidth market in terms of social welfare. In other words, the problem is to determine the allocation of network resources offered for sale to end-to-end connections offered for purchase that will result in the maximum social welfare. This problem is complicated by the fact that the buyers' utilities and sellers' costs are their private information and market participants may not have interest in sincerely eliciting this information.

Currently, the bandwidth market is organized on the basis of bilateral agreements. This means that the trade is carried out by making deals between one seller and one buyer that have to negotiate with each other the contract terms. Because no one is encouraged to reveal one's private information, these negotiations are often very complex and time consuming. Also, the details about transactions are rather privately held than publicly announced information. This makes the whole trading process non-transparent and it limits access to important market related data (e.g., prices). If there are many paths realizing specific end-to-end connection, the buyer having limited information about links' prices may have difficulties in determining the cheapest one. Moreover, bilateral agreements do not support buyer in obtaining end-to-end connection when this end-to-end connection cannot be provided by a single seller. In such a case the buyer must independently negotiate with several owners of links to realize desired end-to-end connection. This leads to the risk of purchasing incomplete set of links if the trade negotiations fail with one of seller, whereas agreements with other sellers would be drawn up and signed.

Thus, there is a need for more sophisticated organization of bandwidth market. Some authors [3], [4] claim that the introduction of new forms of bandwidth trading is only a matter of time. One of the key aspects that may facilitate this process is a development and application of new market mechanisms, such as bandwidth auctions [4].

2. Bandwidth Auction Models

Auction is one of the oldest way of performing the trade on the market. It is characterized by the fact that it defines the kind of offers the auction participants may submit to report their willingness of selling or purchasing commodities being traded and that it defines the formal rules that allow on the basis of submitted offers to determine what commodities are sold or bought by each auction participant. The formal rules of the auction are typically divided into the allocation rule that determines the amount of commodities being exchanged between each buyer and seller, and the pricing rule that sets revenues of sellers and payments of buyers.

2.1. Overview of Bandwidth Auctions

In literature, there are many auction models that support end-to-end connection trading on the bandwidth market. These models make different assumptions about the bandwidth auction.

One of the most important assumptions relates to the number of network providers that may participate in the bandwidth auction as a seller. One group of the models [5]–[10] concerns one-sided auction, in which the bandwidth of all telecommunication links is offered for sale by one auctioneer to many buyers. Mostly, the auctioner is, or acts on behalf of, a provider that owns or manages the telecommunication network. Although some of these models [8]–[10] can be used to trade network resources owned by different providers, they require that all of them pass their true private information to the auctioneer. Therefore, these auction models do not take into account that the network resource providers may act strategically competiting with each other.

However, proceeding bandwidth market liberalization favors competition between providers and it seems that above assumption may be too restrictive. So, there is a need of market mechanisms that would allow to perform the bandwidth exchange between many sellers and many buyers. It is even believed that the development of double auctions supporting bandwidth trading is one of the most promising new research directions [11]. So far, there are a few models for double auction that supports end-to-end connection trading [12]–[17].

The another important assumption concerns the divisibility of network resources. Some auction models allow for trading bandwidth in modules of predefined capacity [10], [14]–[17] while others treat network resources as fully divisible commodities [5]–[9], [12], [13]. Both these assumptions may be resonable depending on the telecommunication technology. In the lower layers of telecommunication network, the links' capacities have often modular character, e.g., optic fibres, SDH modules. In the higher layers, the bandwidth of links may be divided in almost every real fraction of Mbit/s, e.g., ATM virtual paths, IP flows.

The last assumption, mentioned here, relates to the way of supporting end-to-end connection trading. Most of the proposed auction models require that the buyer specifies the single path to be used to realize desired end-to-end connection. These auction models ensure that the same amount of bandwidth will be allocated to the buyer at each link constituing specified path. A more flexible approach from the buyer point of view is applied in the Network Second Price model (NSP) [12]. The NSP model allows the buyer to specify many paths that can be used to realize desired end-to-end connection. The NSP model considers all this paths when allocating bandwidth to particular end-to-end connection, so the buyer can increase the chance of purchasing end-to-end connection with predefined capacity by specifying many paths. The Kelly's model [5] and the model for balancing communication bandwidth trading (BCBT) [13] are even more flexible than NSP, because they allow the buyer to submit an offer for commodity representing a demand for end-to-end connection. In these models buyer does not have to specify any paths, but only source and destination nodes of the end-to-end connection. Therefore, these models seem to be more convenient to the buyer as one does not have to know the network topology.

In paper [18] double auction models concerning bandwidth as a modular commodity have been analyzed. Here we compare double auction models supporting end-to-end connection trading that treat bandwidth as fully divisible commodity. To this group of auction models belong the BCBT and NSP models that are known from the literature and also the BCBT-CG model that is here proposed.

2.2. Comparision of the BCBT and NSP Models

The BCBT and NSP models relate to the auctions that can be classified as sealed-bid single-round double auctions. This means that in both auction models sellers and buyers submit their offers knowing nothing about the offers of other auction participants and the auction mechanism determines the allocation, and pricing only on the basis of submitted offers. In both models the bandwidth of telecommunication links is an elementary, fully divisible commodity. Let E be a set of all telecommunication links. In NSP model there is no information about which network nodes are connected by particular link $e \in E$. In BCBT model there is a set V denoting all telecommunication nodes. For each link $e \in E$ and each node $v \in V$ the parameter a_{ve} defines if node v is a source ($a_{ve} = 1$) or destination ($a_{ve} = -1$) node of link e or that the node v is not incident to link e ($a_{ve} = 0$). Thus BCBT model has a full information about the network topology.

The NSP and BCBT models allow the seller to submit the sell offer for particular link. With each link $e \in E$ there can be many sell offers involved that form the set $S(e)$. The set of all sell offers is denoted by $S = \cup_{e \in E} S(e)$. In both models each sell offer $l \in S$ is characterized by two parameters: S_l – the minimum unit price at which seller is willing to sell the bandwidth of link, and $x_l^{\max}$ – the maximum amount of bandwidth offered for sale by the seller.

Considered auction models differ substantially in the commodity type for which the buy offer can be submitted. Let D denote the set of all buy offers. In the NSP model, the buy offer is related with a set of possible paths. The buyer that submits the buy offer $d \in D$ specifies the set of possible paths P_d. For each path $p \in P_d$ and each link e he has to define a binary parameter b_{edp} that equals 1 if link e belongs to path p and equals 0 otherwise. This means that the buyer must know the network topology in order to correctly specify the paths. In the case of BCBT the buy offer is involved with a commodity that represents a demand for end-to-end connection. In the buy offer d the buyer specifies only a source node s_d and a destination node t_d. In both NSP and BCBT models the buyer also has to define in the buy offer d two parameters: E_d – the maximum unit price at which the buyer is willing to buy the bandwidth of end-to-end connection, and $x_d^{\max}$ – the maximum amount of bandwidth offered for purchase by the buyer.

The allocation rules of NSP and BCBT models decide, which offers are accepted aiming at social welfare maximization. In other words, both allocation rules match sell and buy offers allocating bandwidth of links offered for sale to end-to-end connections offered for purchase in order to achieve the maximum sum of all market participants' net benefits. Both allocation rules allow for multipath routing, i.e., each end-to-end connection may be realized by several paths. The essential difference between NSP and BCBT allocation rules is that the allocation rule of NSP has given the predefined paths for each end-to-end connection while the allocation rule of BCBT itself has to determine the paths for each end-to-end connection. Note that the BCBT model that has full information about network topology, considers all paths that can be used to realize particular end-to-end connection. In the case of NSP model the allocation rule is restricted to paths specified by the buyers. Therefore, for given submitted offers, the NSP model takes into account only some subset of all allowable paths that can be generated in the case of BCBT model. The allocation rules of NSP and BCBT models can be formulated as linear programming problems with the same objective functions. It is worth to mention that the LP problem defining NSP allocation rule is a restriction of LP problem defining BCBT allocation rule. Thus, assuming that for the buyers it is indifferent what paths are used to satisfy their demands for end-to-end connections, the allocation obtained by the NSP model cannot be better in terms of the social welfare than the one given by the BCBT model.

The NSP and BCBT auction models define also different pricing rules. NSP adapts VCG-style pricing [12] while BCBT determines the clearing prices on the basis of dual prices of LP formulation of its allocation rule. Further we will focus on the comparison of NSP and BCBT allocation rules, so we do not discuss here the details of both pricing rules.

2.3. The BCBT-CG Auction Model

In this paper we propose the BCBT-CG model. The BCBT-CG model is a modification of BCBT model that differs from BCBT only in the way of determining the optimal allocation. The allocation rules of BCBT-CG and BCBT models are equivalent in the respect of social welfare.

In BCBT model the allocation is determined by solving a LP problem, in which all possible paths for each end-to-end connection are considered at once. As opposed to this approach, in BCBT-CG the allocation problem is decomposed into the master problem and the subproblem using column generation technique.

The master problem is a restriction of the allocation problem defined by BCBT model in which for each buy offer a set of allowable paths is specified. Thus, in the master problem of BCBT-CG model for each buy offer d there is defined a set of paths P_d with each path described by binary parameters b_{edp} like in the NSP model. The aim of the master problem is to determine the optimal (i.e., providing maximum social welfare) bandwidth allocation assuming that end-to-end connection can be realized only by the predefined paths.

For given buy offer d let us denote by the variable x_{dp} the amount of bandwidth allocated at path $p \in P_d$. Then the variable $x_d = \sum_{p \in P_d} x_{dp}$ is the total amount of bandwidth allocated to end-to-end connection involved with buy offer d. Moreover, let us define the variable x_l that indicates the amount of bandwidth sold at link involved with sell offer l. Note that the variables x_d and x_l also denote the realization volume of buy offer d and sell offer l, respectively. Then, the master problem of BCBT-CG can be formulated as following LP problem:

$$\hat{Q} = \max \left(\sum_{d \in D} E_d x_d - \sum_{l \in S} S_l x_l \right), \quad (1)$$

$$\sum_{d \in D} \sum_{p \in P_d} b_{edp} x_{dp} \leq \sum_{l \in S(e)} x_l, \quad \forall_{e \in E}, \quad (2)$$

$$x_d = \sum_{p \in P_d} x_{dp} \quad \forall_{d \in D}, \quad (3)$$

$$0 \leq x_d \leq x_d^{\max}, \quad \forall_{d \in D}, \quad (4)$$

$$0 \leq x_l \leq x_l^{\max}, \quad \forall_{l \in S}, \quad (5)$$

$$0 \leq x_{dp}, \quad \forall_{d \in D}, \forall_{p \in P_d}. \quad (6)$$

The objective function (1) aims at the maximization of the difference between buyers' payments and sellers' revenues according to the buy and sell prices specified in the offers. So, assuming that the offers are sincere, i.e., they represent real private information of the auction participants, the objective function ensures that for given predefined set of paths P_d the optimal solution of the master problem gives the allocation with the maximum social welfare. Constraints (2) guarantee that for each link the total amount of bandwidth of this link allocated to all paths predefined for buy offers cannot be greater than the sum of realization volumes of all sell offers related to this link. Next group of constraints (3) state that each buy offer realization volume equals to the sum of bandwidth allocated at all paths specified for this offer. Constraints (4) and (5) define allowable realization volumes of buy and sell offers, respectively.

The column generation subproblem assumes that for each link there is defined a unit price, i.e., the price, at which one unit of link's bandwidth is sold/purchased. The subproblem relies on calculating for each end-to-end connection the cheapest path according to given links' prices. This can be done by an arbitrary shortest path algorithm.

Here we formulate the subproblem as a LP problem. Let the variable x_{ed} denote the amount of bandwidth of link e allocated to end-to-end connection involved with buy offer d. Let the parameter $\hat{\lambda}_e$ indicate the unit price of link e. Then the solution of subproblem can be obtained by solving following LP problem:

$$\min \sum_{d \in D} \sum_{e \in E} \hat{\lambda}_e x_{ed}\,, \tag{7}$$

$$\sum_{e \in E} a_{ve} x_{ed} \;=\; \begin{cases} 1 & v = s_d \\ 0 & v \neq s_d, t_d \\ -1 & v = t_d \end{cases}, \quad \forall_{v \in V, d \in D}\,, \tag{8}$$

$$0 \;\leq\; x_{ed}, \quad \forall_{d \in D}, \forall_{e \in E}\,. \tag{9}$$

The objective function (7) minimizes the total cost of the realization of all end-to-end connections according to link's prices $\hat{\lambda}_e$. Equations (8) define flow conservation constraints that must be met for each end-to-end connection. Simplex algorithm determines vertex solution $\hat{x}_{ed}$ of above LP problem giving for each buy offer d the cheapest path p defined by $b_{edp} = \hat{x}_{ed}$ with unit buy price equal to $\sum_{e \in E} \hat{\lambda}_e \hat{x}_{ed}$.

The complete allocation rule of the BCBT-CG is defined by following iterative algorithm based on the column generation technique that exploits above definitions of the master problem Eqs. (1)–(6) and the subproblem Eqs. (7)–(9):

1. For each buy offer d initialize a set of predefined paths P_d, e.g., set P_d may include one path p being a solution of the cheapest path subproblem with links' prices $\hat{\lambda}_e = \min_{l \in S(e)} S_l$.

2. Solve master problem for given P_d: determine optimal allocation $\hat{x}_l$, $\hat{x}_d$, $\hat{x}_{dp}$ and optimal values of dual prices $\hat{\lambda}_e$ i $\hat{\omega}_d$ corresponding to constraints (2) and (3), respectively.

3. Solve the cheapest path subproblem for given links' prices $\hat{\lambda}_e$: for each buy offer d determine the cheapest path realizing relevant end-to-end connection as a path σ such that $b_{ed\sigma} = \hat{x}_{ed}$ for each link e.

4. If for each buy offer d the cheapest path σ fulfills following condition $\sum_{e \in E} \hat{\lambda}_e b_{ed\sigma} \geq \hat{\omega}_d$, then the allocation $\hat{x}_l$, $\hat{x}_d$, $\hat{x}_{dp}$ detemined in step 2 is optimal. Otherwise, for each buy offer d, for which the cheapest path σ fulfills condition $\sum_{e \in E} \hat{\lambda}_e b_{ed\sigma} < \hat{\omega}_d$, add path σ to the set P_d and go to the step 2.

At first, the set of paths realizing relevant end-to-end connection is initialized for each buy offer. In the second step the master problem is solved. In this way the optimal allocation is determined and the unit prices of links and end-to-end connections are set according to the optimal values of dual prices. In the third step the subproblem is solved and the cheapest path is found for each buy offer. In the fourth step for each buy offer it is checked if the unit price of the cheapest path is greater or equal to unit price of end-to-end connections. If all paths fulfil this condition, the allocation determined in second step is optimal and the algorithm stops. Otherwise, for each buy offer, for which the condition is not met, the cheapest path is added to the set of predefined paths and the algorithm goes to the second step where the next iteration begins.

3. Experimental Studies

The allocation rules of NSP, BCBT and BCBT-CG models have been compared in the respect of economic and computational efficiency. The experimental studies have been performed on several test instances concerning allocation problems on the bandwidth market.

3.1. Test Instances

Test instances for the allocation problems on bandwidth market have been based on the SNDlib library [19]. Although this library contains data sets for survivable fixed telecommunication network design problems, the information derived from it was very usefull in the preparation of test instances for allocation problems on the bandwidth market. Some data such as network topology (nodes and links) and demands for end-to-end connection have been directly applied in prepared test instances. Other data such as links' capacities, demands' volumes and distances between nodes have been used for the generation of the offers' parameters. Three data sets from the SNDlib library have been used: *sun*, *janos-us* and *giul39*. Figure 1 shows the network topologies given by considered data sets. Table 1 contains information about the number of nodes, links and end-to-end connections for these data sets.

Table 1
Number of nodes, links and end-to-end connections for each data set

Data set	Nodes	Links	End-to-end connections
sun	27	102	67
janos-us	26	84	650
giul39	39	172	1471

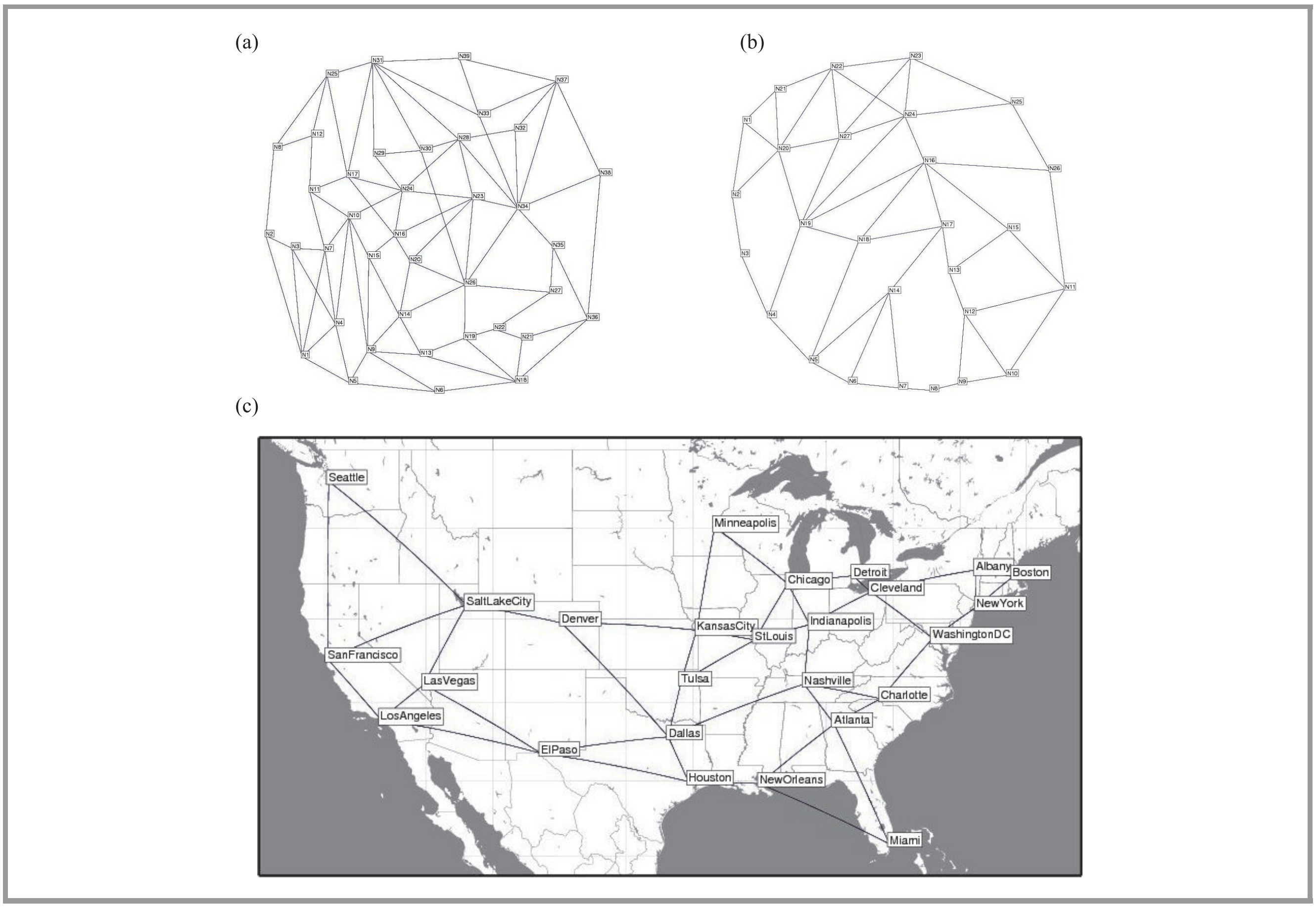


Fig. 1. Topologies for considered data sets: (a) sun; (b) giul39; (c) janos-us.

Data sets derived from the SNDlib do not include all data required to form full test instance of the allocation problem on the bandwidth market. Some data involved with offers has had to be generated. It has been assumed that:

- the unit price specified in offer concerning link/end-to-end connection is proportional to the distance between nodes connected by this link/demand,
- the total amount of bandwidth offered for sale (purchase) at particular link (end-to-end connection) is proportional to the link capacity (demand volume).

Table 2
Number of sell and buy offers for each test instance

Test instance	Sell offers	Buy offers
sun-2	206	129
sun-4	407	274
sun-6	619	411
janos-us-2	165	1287
janos-us-4	340	2666
janos-us-6	500	3916
giul39-2	330	2938
giul39-4	708	5869
giul39-6	1043	8803

Moreover, in the NSP model it is required that for each buy offer a set of paths is specified. It should be noted that the path specification has significant influence on the economic and computational efficiency of NSP model. Thus, in the case of NSP model we have decided to consider 100 variants of each test instance. In *k*-th variant for each buy offer the *k* cheapest paths (calculated according to the minimum unit sell prices specified in sell offers) is specified as a set of possible paths.

On the basis of each considered data set, a three test instances have been prepared with respectively 2, 4 and 6 offers on average submitted for each link/end-to-end connection. Table 2 presents the number of sell and buy offers generated for each test instance. As mentioned above, in the case of NSP model, for each test instance 100 variants of path specification for buy offers have been considered.

3.2. Computational Efficiency Analysis

For all test instances the allocation have been determined using NSP, BCBT and BCBT-CG auction models. All LP problems have been solved by means of CPLEX 12.1 on the computer with processor Intel Core2 Duo T8100 2.1 GHz, main memory 3 GB and 32-bit operating system MS Vista.

Table 3
The time of determining the optimal allocation

Test instance	NSP		BCBT	BCBT-CG
	$k = 5$	$k = \min\{k^*, 100\}$		
sun-2	0.11	0.02 (k^*=9)	0.23	3.36
sun-4	0.05	0.03 (k^*=4)	0.73	2.96
sun-6	0.09	0.41 (k^*=32)	2.42	3.31
janos-us-2	0.44	2.68 (k^*=38)	8.36	5.01
janos-us-4	0.64	8.71 (k^*=33)	38.92	11.82
janos-us-6	1.19	46.18 (k^*=84)	88.92	20.56
giul39-2	1.26	7.79 (k^*=8)	74.79	32.54
giul39-4	4.26	172.71 (k^* >100)	34.51	18.25
giul39-6	10.87	402.83 (k^* >100)	1162.69	176.33

Iterative algorithm constituing allocation rule of BCBT-CG model has been implemented in AIMMS 3.10.

In the case of NSP model, 100 variants of buy offers have been considered for all test instances. For each test instance there has been determined the smallest variant k^*, for which NSP gives the same value of social welfare as BCBT and BCBT-CG models. For each variant $k < k^*$ the allocation obtained by NSP is worse in terms of social welfare than the allocation given by BCBT (BCBT-CG) model. On the other hand, for variants $k \geq k^*$ the allocation obtained by NSP is equivalent in the respect of economic efficiency to allocation given by the BCBT (BCBT-CG) model.

Table 3 presents the time of determining the optimal allocation by each auction model. In the case of NSP model the results for two variants are shown: $k = 5$ and $k = k^*$. The values of k^* vary for different test instances and are given in the Table 3. As it can be seen, if the NSP model is applied then obtaining as efficient allocation as in the case of BCBT (BCBT-CG) model, it requires that the buyers specify quite many paths in their buy offers. The minimum is four paths, but there are test instances that requires 30 or more paths to be specified for each buy offer. For last two test instances, namely *giul39-4* and *giul39-6*, even 100 paths specified for each buy offer have not been enough to ensure that the NSP will result in the same social welfare as BCBT or BCBT-CG model. For these test instances the time for the NSP model in Table 3 is given for $k = 100 < k^*$.

From the obtained experimental results it follows that in the most of test instances the NSP model is faster than BCBT and BCBT-CG. If we consider only the variants of test instances with $k = 5$ then it turns out that NSP model is undoubtedly the fastest one. However, it should be noted that for variants with $k = 5$ only for the test instance *sun-4* the NSP model provide as efficient allocation as the BCBT or BCBT-CG model. Also, for the NSP model, the time of determining the k-th cheapest paths by the buyers is here not taken into account.

Considering variants of test instances with $k = k^*$, we can see that there are test instances for which BCBT or BCBT-CG requires less time than NSP in order to compute the optimal allocation. The allocation time for BCBT is shorter than in the case of NSP model for the test instance *giul39-4*. In turn, the BCBT-CG model is faster than NSP for three test instances: *janos-us-6*, *giul39-4* and *giul39-6*.

Comparing the allocation times for the BCBT and BCBT-CG models, we can see that for the larger test instances based on data sets *janos-us* and *giul39*, the BCBT-CG model requires less time to determine the optimal allocation than the BCBT model, and the difference is significant especially for the last test instance *giul39-6*. Thus, the computational efficiency of determining the optimal allocation can be improved by applying the BCBT-CG model instead of BCBT.

3.3. Economic Efficiency Analysis

The BCBT and BCBT-CG models are equivalent in the respect of economic efficiency as they give the allocations providing the same social welfare. In the case of NSP model, the value of resulting social welfare depends on the path specifications made by the buyers in their buy offers. The allocation obtained by NSP for variant k of given test instance will be at least as efficient as the one obtained for variant $k - 1$. Moreover, the social welfare provided by the NSP model cannot be higher than the one given by the BCBT or BCBT-CG model. Here, we analyze how the social welfare obtained by the NSP model changes in the relation to the opitmal allocation given by BCBT (BCBT-CG) considering the first 10 variants of each test instance.

Table 4 presents the ratio between the value of social welfare provided by NSP and the one given by the BCBT (BCBT-CG). Experimental results show that this ratio for all considered test instances is at least 95% just for k greater than 3. If $k = 5$, then the ratio is 99% or higher for almost all test instances except for two (*giul39-4* and *giul39-6*). So, if all buyers can anticipate and specify in the buy offers the five cheapest path realizing demanded end-to-end

Table 4
The ratio between the value of social welfare provided by NSP and BCBT (BCBT-CG)

Test instance	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$	$k=9$	$k=10$
sun-2	0.93	0.97	0.99	0.99	0.99	0.99	0.99	0.99	1	1
sun-4	0.93	0.98	0.99	1	1	1	1	1	1	1
sun-6	0.93	0.95	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
janos-us-2	0.89	0.95	0.98	0.98	0.99	0.99	0.99	0.99	0.99	0.99
janos-us-4	0.95	0.98	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
janos-us-6	0.82	0.97	0.98	0.99	0.99	0.99	0.99	0.99	0.99	0.99
giul39-2	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1	1	1
giul39-4	0.8	0.9	0.93	0.95	0.96	0.97	0.97	0.98	0.98	0.98
giul39-6	0.83	0.91	0.94	0.96	0.97	0.97	0.98	0.98	0.98	0.99

connections then it is highly possible that the NSP would determine the allocation, which is almost as economically efficient as the allocation given by BCBT.

However, it should be noted that the NSP model does not guarantee as efficient allocation as BCBT or BCBT-CG model. For some test instances, e.g., *giul39-4* and *giul39-6*, even if the byuers would know the sell prices of each link and specify the 100 least expensive paths in their offers the NSP model leads to the social welfare that is lower than the one that can be provided by the application of the BCBT (BCBT-CG) model (see Table 3).

4. Conclusions

In this paper we compare three auction models that treat bandwidth as divisible commodity and support end-to-end connection trading. One of these models, called BCBT-CG, is proposed here as a modification of BCBT model. The BCBT-CG allows for determining the allocation that is equivalent to the one obtained by the BCBT model. However, it applies as an allocation rule the iterative algorithm based on the column generation technique that has better computational efficiency than the BCBT allocation rule. Experimental studies verify that the BCBT-CG model can be used instead of BCBT to reduce the time of determining the optimal allocation.

Compared to BCBT and BCBT-CG models, the NSP model is less convenient from the buyer point of view as it requires from him to specify the paths realizing demanded end-to-end connection. In the case of NSP model, the buyer has to know the network topology and he bears the responsibility for choosing the appropriate set of paths, which would give him the best payoff. On the other hand, in the BCBT and BCBT-CG models the buyer must only specify the source and destination nodes of the desired end-to-end connection. The allocation rules of these auction models are responsible for determining the optimal paths.

From the experimental results it follows that the NSP model does not guarantee as efficient allocation as the BCBT and BCBT-CG models, even if all buyers specify the 100 least expensive paths in their buy offers. The merit of NSP is that it allows for the fast determination of almost optimal allocation requiring just a five cheapest paths to be specified by each buyer. However, the time of determining the allocation by the NSP increases with the number of paths specified in the buy offers and if there is many such a paths, the NSP may be slower than the BCBT and BCBT-CG models.

Acknowledgments

The authors acknowledge the Ministry of Science and Higher Education of Poland for supporting the research. The research was supported by the Polish National Budget Funds 2010–2013 for science under the grant N N514 044438.

References

[1] D. Upton, “Modelling the market for bandwidth”, Ph.D. thesis, University of Cambridge, 2002.

[2] G. Cheliotis, “Structure and dynamics of bandwidth markets”, Ph.D. thesis, N.T.U. Athens, 2001.

[3] A. Iselt, A. Kirstadter, and R. Chahine, “Bandwidth trading – a business case for ASON?”, in *Proc. 11th Int. Telecomm. Netw. Strategy Planning Symp. NETWORKS 2004*, Viena, Austria, 2004, pp. 63–68.

[4] R. Rabbat and T. Hamada, “Revisiting bandwidth-on-demand enablers and challengers of a bandwidth market”, in *Proc. 10th IEEE/IFIP Netw. Oper. Manag. Symp.*, Vancouver, Canada, 2006, pp. 1–12.

[5] F. Kelly, “Charging and rate control for elastic traffic”, *Eur. Trans. Telecommun.*, no. 8, pp. 33–37, 1997.

[6] B. Hajek and S. Yang, “Strategic buyers in a sum bid game for flat networks” *manuscript*, 2004.

[7] R. Johari and J. N. Tsitsiklis, “Efficiency loss in a network resource allocation game”, *Mathem. Opera. Research*, no. 29, pp. 407–435, 2004.

[8] A. Lazar and N. Semret, “Design, analysis and simulation of the progressive second price auction for network bandwidth sharing”, in *Proc. 8th Int. Symp. Dynamic Games Appl.*, Vaals-Maastricht, The Netherlands, 1998.

[9] N. Semret, “Market mechanisms for network resource sharing”, Ph.D. thesis, Columbia University, 1999.

[10] M. Dramitinos, G. D. Stamoulis, and C. Courcoubetis, “An auction mechanism for allocating the bandwidth of networks to their users”, *Comput. Netw.*, vol. 51, no. 18, pp. 4979–4996, 2007.

[11] I. Koutsopoulos and G. Iosifidis, “Auction mechanisms for network resource allocation”, in *Proc. 8th Int. Symp. Model. Optimiz. Mobile, Ad Hoc and Wirel. Netw. WiOpt*, Avignon, France, 2010, pp. 554–563.

[12] R. Jain and J. Walrand, "An efficient mechanism for network bandwidth auction", in *Proc. IEEE/IFIP Network Oper. Manag. Symp.Workshops 2008*, Salvador, Brazil, 2008, pp. 227–234.

[13] W. Stańczuk, J. Lubacz, and E. Toczyłowski, "Trading links and paths on a communication bandwidth markets", *J. Universal Comput. Sci.*, vol. 14, no. 5, pp. 642–652, 2008.

[14] R. Jain and P. Varaiya, "Combinatorial exchange mechanisms for efficient bandwidth allocation", *Commun. Inf. Syst.*, vol. 3, no. 4, pp. 305–324, 2004.

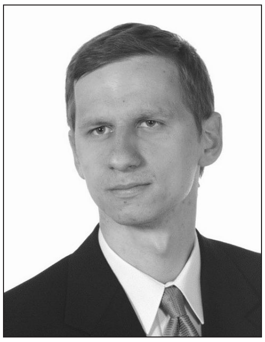
[15] R. Jain and P. Varaiya, "An efficient incentive-compatible combinatorial market mechanism", in *Proc. 42nd Allerton Conf. Commun. Control. Comput.*, Monticello, Il, USA, 2004.

[16] R. Jain and P. Varaiya, "The combinatorial seller's bid double-auction: an asymptotically efficient market mechanism", *J. Econom. Theory*, 2006.

[17] K. Kołtyś, P. Pałka, E. Toczyłowski, and I. Żółtowska, "Multicommodity auction model for indivisible network resource allocation", *J. Telecommun. Inform. Technol.*, no. 4, pp. 25–31, 2008.

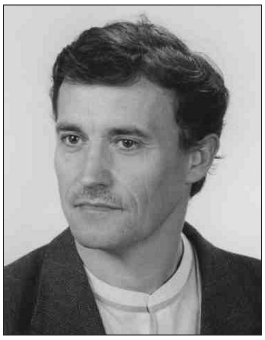
[18] K. Kołtyś, P. Pałka, E. Toczyłowski, and I. Żółtowska, "Bandwidth trading: a comparison of the combinatorial and multicommodity approach", *J. Telecom. Inform. Technol.*, no. 2, pp. 67–72, 2010.

[19] "Survivable network design library" [Online]. Available: http://sndlib.zib.de/



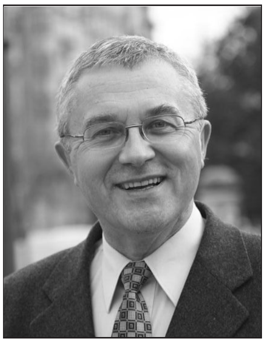
Kamil Kołtyś received the M.Sc. degree in computer science in 2007 from the Warsaw University of Technology, Poland. Currently, he prepares his Ph.D. thesis in computer science at the Institute of Control and Computation Engineering at the Warsaw University of Technology. His research interests include decision support, optimization and bandwidth trading. His current research is focused on application of multicommodity turnover models to network resource allocation.

e-mail: K.J.Koltys@elka.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland



Krzysztof Pieńkosz received the M.Sc. degree in 1984, Ph.D. in 1992 and D.Sc. in 2011 from the Warsaw University of Technology. In 1984–1986 he worked at the Oil and Gas Institute developing simulation and analysis methods for gas distribution networks. Since 1986 he is with the Institute of Control and Computation Engineering at the Warsaw University of Technology. His current interests are focused on the operations research in particular discrete optimization models and methods applied to the resource allocation problems.

E-mail: K.Pienkosz@ia.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland



Eugeniusz Toczyłowski is a Professor, the head of Operations and Systems Research Division, at the Institute of Control and Computation Engineering at the Warsaw University of Technology (WUT). He received the M.Sc. degree in 1973, Ph.D. in 1976, D.Sc. in 1989, and the title of Full Professor in 2004. His main research interests are centered around the operations research models and methods, including structural approaches to large scale and discrete optimization, auction theory and competitive market design under constraints, multicommodity trading models, and design of management information systems.

e-mail: E.Toczylowski@ia.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Paper

Analysis and Modeling of Domain Registration Process

Piotr Arabas[a,b], Przemysław Jaskóła[b], Mariusz Kamola[a,b], and Michał Karpowicz[b]

[a] *Institute of Control and Computation Engineering, Warsaw University of Technology, Warsaw, Poland*
[b] *Research Academic Computer Network (NASK), Warsaw, Poland*

Abstract—The paper presents analysis of the domain name reservation process for the polish `.pl` domain. Two models of various time scale are constructed and finally combined to build long range high resolution model. The results of prediction are verified using real data.

Keywords*—*domain market, domain registration, forecasting, time series modeling.

1. Introduction

In time of rapid growth of Internet, domain names became an important commodity [1]. In consequence, the volume of DNS market became dependent on overall economic conditions and expectedly follows standard laws of demand, and supply. Furthermore, as the number of attractive domain names is limited, there exists possibility of investing and earning relatively high profits. For all these reasons domain registration statistics present interesting set of data to be analyzed. The aim of this article is to present results of analysis and modeling of domain registration process.

Similar analysis were presented in [2], [3], [4], while the secondary market was studied in [5], however none of these papers covered Poland. Much broader literature is devoted to semantic analysis of domain names, which can be used to assess their qualities [6], [7] or pricing [8], [9]. As far as some of the results of these works have direct connection with demand modeling and pricing domain names, they, in our opinion, neglect the most basic behavior of domain users.

In this paper, we concentrate on primary market (registration) modeling. We try to find out some specific characteristics of this process using abundant data of Polish domain registry. First, we try to identify its general properties by analyzing basic statistics in various time scales and applying harmonic analysis to determine characteristic periods. We show that data conform to some patterns, two of them – weekly and yearly – being most obvious. Following this observation, we propose to construct specialized models on both time scales and, possibly, compose more complex models of them. It must be noted that even short horizon modeling may provide valuable predictions, e.g., for planning of an advertising campaign.

The rest of the paper is organized as follows: in Section 2 we describe the problem, which is subject of this research. Next, in the Section 3 we show related work and draw our solution. In the Section 4 the data and basic characteristics of the process are presented along with results of a preliminary analysis. The Section 5 presents the model built to reflect long-range behavior of registration process together with the results of one-year ahead prediction. The short range model and results of its verification are described in the Section 6. Then, in Section 7 we combine both models into a composite model allowing one year prediction with resolution of one day. We conclude in Section 8.

2. Description of the Problem

The domain names are organized in a hierarchical manner, with the last part of each name being a name of top level domain (TLD). Important portion of TLDs are national domains with `.pl` being polish TLD. The registry of each TLD is kept by some institution designated by ICAN, being responsible for domains worldwide. In Poland, such registry for `.pl` domain, together with various regional, functional etc. sub-domains is managed by NASK (Research and Academic Computer Network). The interest in analyzing and modeling of the domain registration process is caused by several factors. First of all, registration is a commercial activity with fees paid for registration and then, repetitively, each year for prolonging domain activity. NASK sells domains mostly on the wholesale market to the number of companies offering various other network services to end users. It must be noted that domains are not only bought by companies or individuals who need to establish a new internet service, e.g., webpage, but also (as mentioned earlier) as a kind of investment, for future resale on the secondary market.

The result of this segmentation are different behaviors of various groups of clients – big companies are possibly less price sensitive than individual users, however, most sensitive and in the fact chimeric group may be the investors. This group may also have different strategies of renewing domains – some domains which are not needed (e.g., then turned out to be unprofitable) may be dropped and some may be re-registered after short time. Although we do not analyze renewal of domains here and neglect influence of its price on registration process it is important to realize that periodic expiration of a large number of domains may result in apparently spontaneous accumulation of re-registrations.

3. Proposed Solution

As it was mentioned in the introduction, the body of work related to modeling of domain registration process is relatively scarce left aside papers devoted to semantic analysis of domain names. What we try to do is to analyze of the registration process as a whole – we do not distinguish more and less valuable domains, as NASK sells them on the wholesale basis without such differentiation. We also decided not to model price factor to simplify the model. In a fact, we tried modeling price–demand correlation using some basic economic models, e.g., Cobb-Douglas or Gutenberg [10], however we found it ineffective and possible unnecessary. The reason was relatively scarce amount of data resulting from rare and usually too small changes in the pricing strategy. In the analyzed period, only one price change had clearly visible effect – it was lowering the registration price in 2008 (see Fig. 1). Furthermore, our aim was to construct models that could be used for prediction on some clearly defined horizon, which application allows considering external factors as constant and recalculating models if necessary.

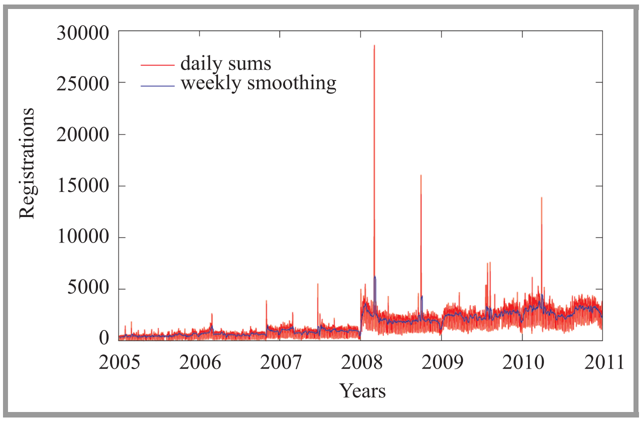


Fig. 1. Domain registrations before removing anomalies.

With such assumptions the process may be modeled as pure time series, which allows the use of well known methodology (see, e.g., [11], [12]). The approach is well grounded for modeling economic and sociology data, and we suppose that our case does not differ much from, e.g., air travel frequency [13], or real estate prices [14]. The basic assumption, which we adopted after, e.g., [12], [15] is that the base process (domain registration in this case) may be decomposed in the following way:

$$x_t = p_t + s_t + e_t, \tag{1}$$

where p_t is trend, s_t is the seasonal and e_t is the irregular component. The approach is natural since trend can be easily observed in the registration data (see Fig. 1), it will be also shown in the next section that the seasonal component is even stronger.

In economic modeling the seasonality is typically defined as periodic process corresponding to yearly cycle (see, e.g., [15], [16]), however the same technique may be used to other, longer or shorter periods. In a fact, it is typical for many processes to exhibit seasonality on several timescales, the best example being presence of short and long economy cycles (waves) [12], [17], or even infinite number of time scales like for self-similar processes [18].

The models used to describe seasonality range from relatively simple periodic (e.g., trigonometric) functions to complex formulas involving regression and relying on expert knowledge, some of them being recognized standards, like X-12 or STL [12], [15], [19]. Other techniques incorporate some approximation methods like, e.g., wavelet analysis [20]. Although using such complex models allows attain precision and draw from rich experience of other researchers, we limited our work to application of the simplest models based on calculation of seasonal means [12], while we tried exploring various time-scales of analyzed process, and finally constructing a model covering all time scales. We did it for two reasons: first, the results of such modeling are simpler to interpret so it is possible to assess the most important properties of registration process clearly. Next, as the aim of the work was prediction, it is easier to build stable forecasts using simpler (i.e., having less parameters, but also needing less restricting assumptions) models.

4. Data

Data were made available by Polish domain registry and consisted of daily sums of registered domains in years 2005–2010. All kinds of domains in polish `.pl` domain, i.e., regional, functional, etc. were summed up. The data were in raw format, as directly dumped from system logs and contained some irregularities. There were two sorts of them:

- missing or duplicated samples of extremely low value,
- samples of anomalously high value.

The first group may be associated with malfunction of the infrastructure, mainly the database software. The second kind of anomalies is mainly caused by some extraordinary promotions, resulting in higher than usual sales; it can be easily observed in the Fig. 1. Fortunately, there were only two gaps in data, which we decided to interpolate. Also, some additional data cleaning had to be performed.

Anomalously high values pose much more problems, as we cannot precisely isolate them by analyzing registration history only. Another important question is what value should be inserted instead of anomalous sample. We decided to be very conservative and deal only with these samples, which we can associate with known marketing campaigns. With help of marketing division staff we identified two such events in 2008, and another two in 2009. Furthermore, we were able to assess number of domains registered during these campaigns, which in turn allowed us to subtract them from appropriate samples. We did not eliminate one possible anomaly in the beginning of 2010, as we could not identify its cause. The data after cleaning are depicted in Fig. 2.

The filtered data contain some likely anomalies still, however they are not so high like those removed, and do not

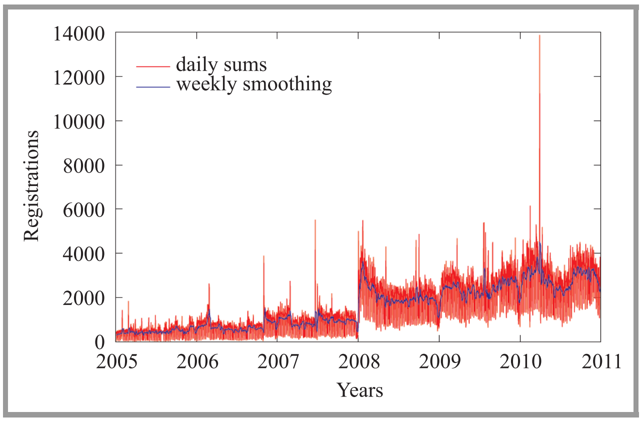


Fig. 2. Domain registrations after removing identified anomalies.

influence smoothed data visibly (except mentioned earlier and left unfiltered anomaly in 2010). Thanks to this it is possible to make some important observations – first of all, the number of registered domains grows, the trend is however disturbed by one rapid rise in the beginning of year 2008. The phenomenon may be easily explained by significant lowering of registration fee in that year. It must be noted that after a change in pricing strategy in 2008, registration price was much lower than renewal fee. In the result, many domains, which were probably bought as a kind of investment, are dropped after one year, while another are re-registered in the beginning of next year, and give cause to some rise in first months of each year.

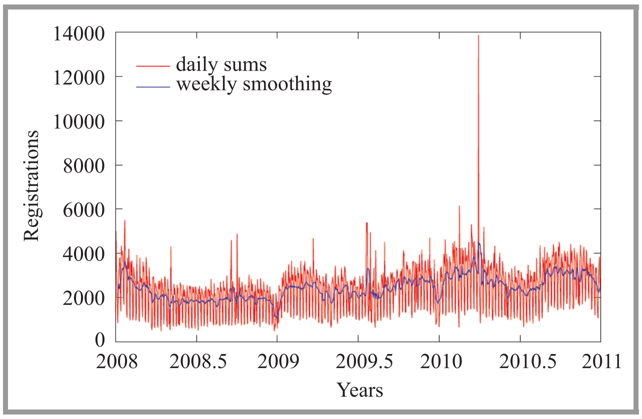


Fig. 3. Domain registrations for last three years after removing identified anomalies.

To observe yearly changes it is better to have a look at graph presenting only 3 years (Fig. 3). The data show visible yearly pattern – manifesting mainly in very low number of registrations during winter holidays and also some higher frequency variations, which may be easily identified as weekly cycles. To emphasize these variations, another set of graphs depicting each year separately is presented in Fig. 4. Yearly patterns may be observed in monthly aggregated data presented in the analogous set of diagrams – see Fig. 5. Summing information from both set of graphs, it must be said that weekly pattern is clearly visible and

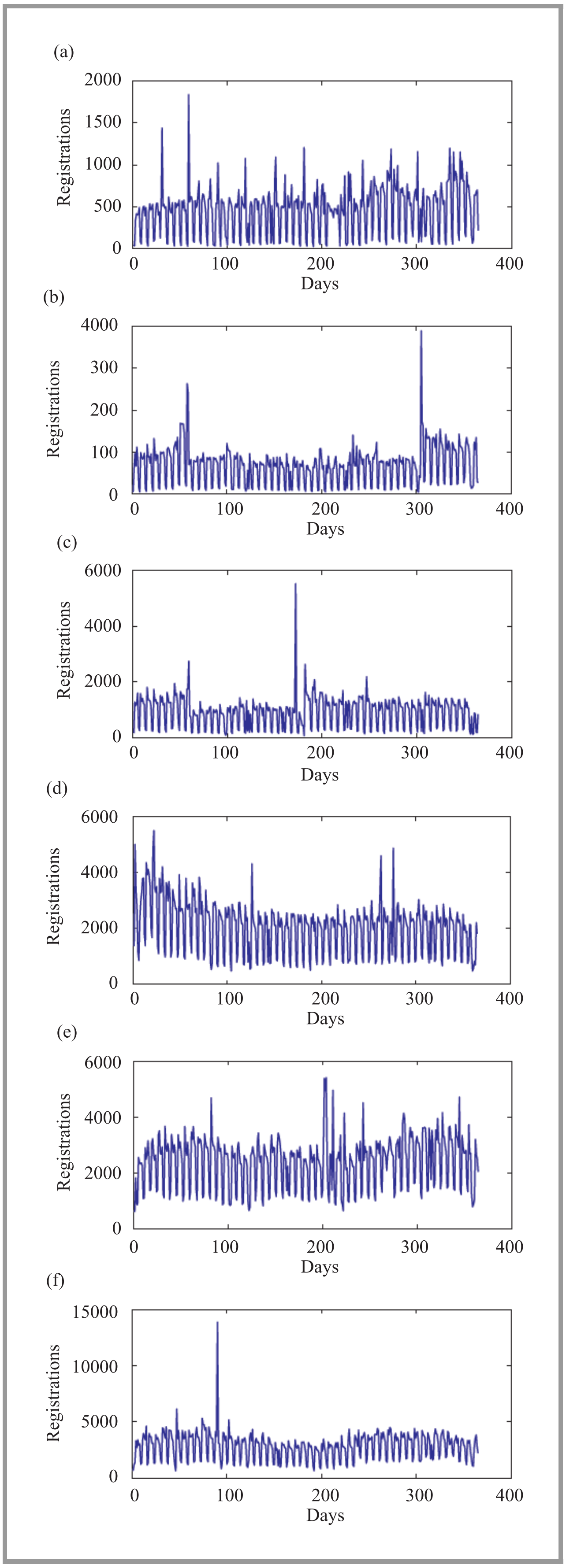


Fig. 4. Registrations in years: (a) 2005; (b) 2006; (c) 2007; (d) 2008; (e) 2009; (f) 2010.

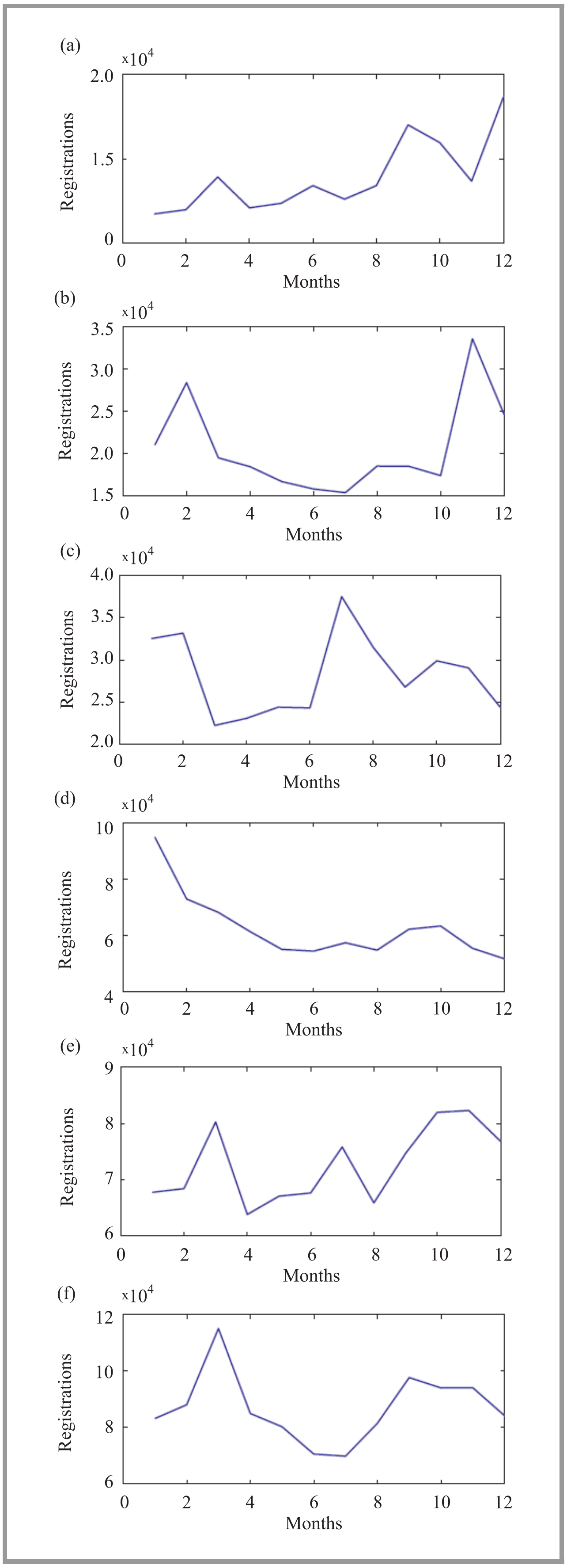


Fig. 5. Monthly sums of registrations in years: (a) 2005; (b) 2006; (c) 2007; (d) 2008; (e) 2009; (f) 2010.

relatively regular, while yearly pattern has rather vague character. It is possible to identify two periods of lower sales during the year – first, more noticeable and easier to locate is winter holidays. The second could be associated with summer holidays, however it tends to move around.

To check for existence of other characteristic periods, we applied spectral analysis by computing power spectrum for period 2005–2009 (see Fig. 6). In this case, we skipped last year as it is used for verification of models presented in the next sections. The number of analyzed samples is too small to gain significant results for longer periods (e.g., one year), however period of one week is again clearly visible. Another period equals approximately to half a week may be treated as a kind of harmonic frequency, and can be explained by the shape of weekly pattern.

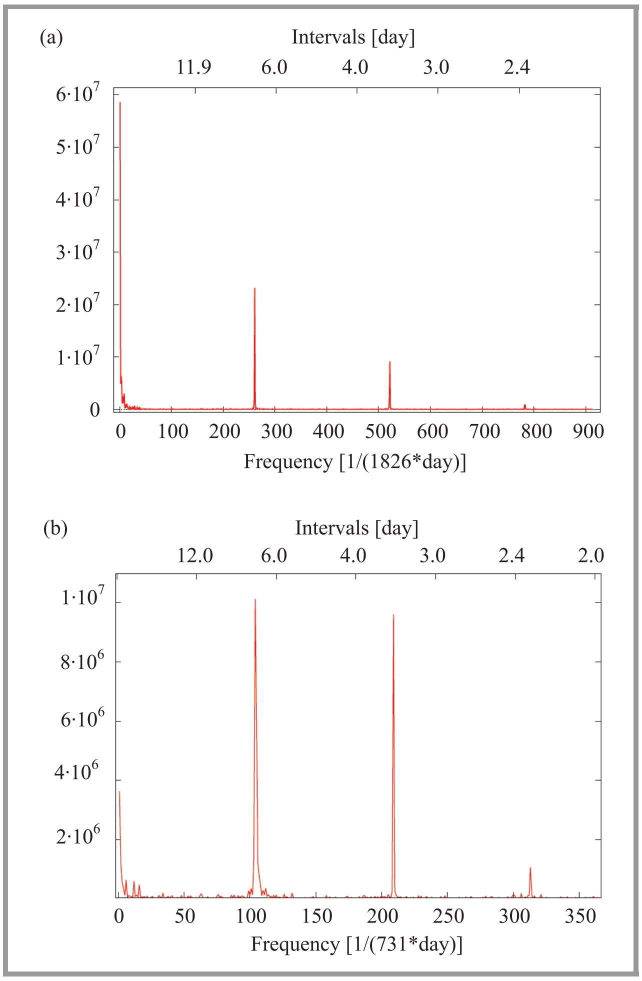


Fig. 6. Power spectrum of the registration process in years 2005–2009.

The most important result of these preliminary analysis is identification of two characteristic periods of the registration process: shorter with length of one week and longer associated with yearly variations. Following these observations, we decided to build two models describing longer and shorter cycles separately to simplify their construction and to allow further analysis.

5. Long Range Modeling

When making strategic decisions like setting new prices or planning a capacity of DNS servers, it is useful to have an estimation of the future sales. Such a prediction can be built upon appropriately designed model and, the most important requirement is a prediction horizon long enough – at least one year. On the other hand, there is no need for high temporal resolution – predicting sales in subsequent months is typically sufficient. After initial analysis we decided not to model influence of domain prices on registrations. There were two reasons for this: first the price changes are relatively rare so it is difficult to gather data necessary to identify any model. But the situation is even more complex, as end users do not observe NASK prices being wholesale prices for dealers. Every dealer has his/her own pricing strategy, furthermore domain names are often sold as a part of a bundle – together with Internet access, web service or mailbox.

5.1. Seasonality and Trend

For the above reasons, we decided to treat the registration process as a time series and build a model using the most classical approach, i.e., to estimate the trend and seasonality first. Then, having as we hoped stationary residuals, we planned to fit an autoregressive process to them. For identification we used monthly aggregated data from period of 2005–2009, and then 2008–2009, while we used data from 2010 for verification.

Such shortening of the learning period is the result of a rapid jump in registrations after lowering prices in 2008, what can be best seen in the graph in Fig. 7 showing two trends fitted to deseasonalized data. Values for the last twelve months in the graph Fig. 7 are predictions for year 2010 – it can be easily seen that including rise in 2008 in unfiltered form results in excessive rate of growth.

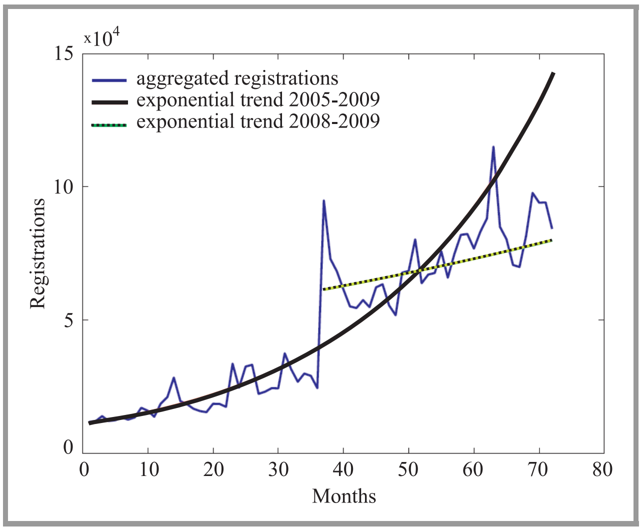


Fig. 7. Exponential trends fitted to deseasonalized registrations: longer (thick) line is trend fitted to the whole 2005–2009 period, shorter (dotted) – 2008–2009 period.

Similarly, the seasonal changes are more regular in last two years (although it can be hardly seen in Fig. 5), so they can be also better identified using shorter period.

The model was constructed by averaging registrations in subsequent months. This way we constructed average registrations sums for January, February, etc., which in connection with the trend provides important information about registration process, and when extrapolated can be used as a simplest prediction (see Fig. 8). Similarly to what can be observed in the Fig. 7, a prediction using the trend and the seasonality fitted to shorter period is much better, in fact it follows the general shape of the line. The greatest discrepancy – in the begging of the predicted period is caused by possible anomaly, which was left unfiltered due to the lack of information – cf. discussion in Section 4 and Fig. 2.

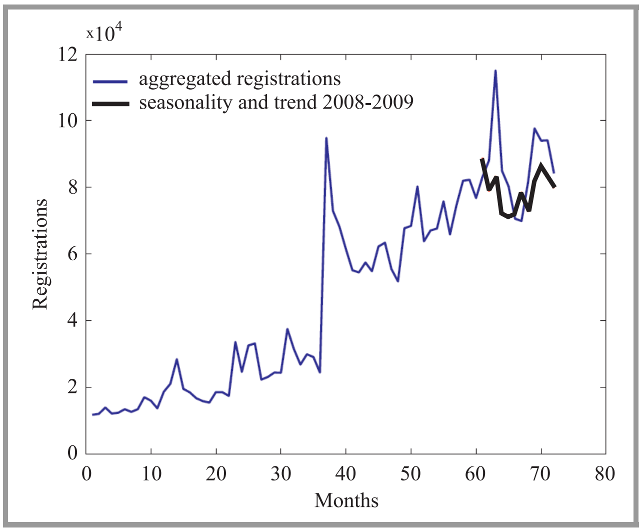


Fig. 8. Registrations forecasted using seasonality and exponential trend for 2008–2009 period.

5.2. Residuals Analysis

In order to analyze results of fitting a trend and a seasonality, residuals were analyzed. The graphs in Fig. 9 present quality of fit to learning data and values of residuals. Although the model output follows the general shape of registration process the values of residuals remain significant and, as can be seen in the lower graph in Fig. 9(a), some correlation between values of the modeled process and residuals may be found.

It must be noted that correlation (if it exists) is relatively weak – grouping of points in the lower left side of the plot is not very clear. The presence of correlation suggests that autoregression could be applied to improve the model. To assess the structure of the model an autocorrelation and a partial correlation functions were computed for a process – see Fig. 10. Both ACF and PACF plots decay relatively fast with only first coefficient being significant. Such a shape suggests correlation with the process lagged one interval (month) back, and application of AR(1) model. Values of coefficients for further (10, 11 and 12) intervals remains

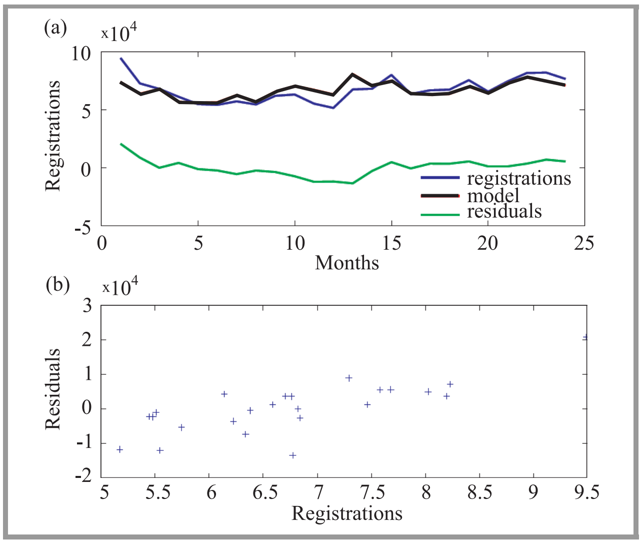


Fig. 9. Quality of fit and residuals for the model using seasonality and trend fitted to 2008–2009 period: (a) quality of fit; (b) values of residuals are plotted against process values.

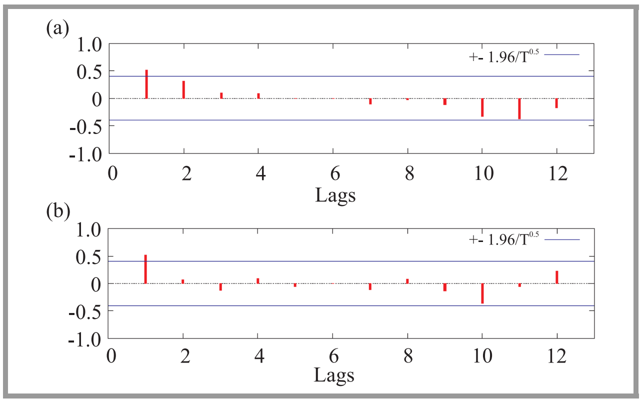


Fig. 10. Autocorrelation – ACF (a) and partial autocorrelation – PACF (b) for residuals of the model using seasonality and trend fitted to 2008–2009 period.

close to significant, which may be caused by some, even weaker correlation, however intervals of 10 or 11 months seem not to be justified by any known property of the process.

5.3. Regressive Modeling

Following analysis in Subsection 5.2, we decided to try to improve the model by applying autoregression to residuals. We started with first order model to begin with the simplest formula and eventually augment it with higher lags after assessing the results. As the model was fitted to the data with trend removed, we neglected intercept and identified only one coefficient. Shorter (2008–2009) data set was used for identification of seasonality and trend, and for computing residuals according to analysis in Subsection 5.1. The resulting AR(1) model proved to be significant, predicted values are shown in Fig. 11. The improvement attained is marginal and visible only in the beginning of the predicted process, however this is implied by the nature of AR(1) model and small values of ACF and PACF coefficients.

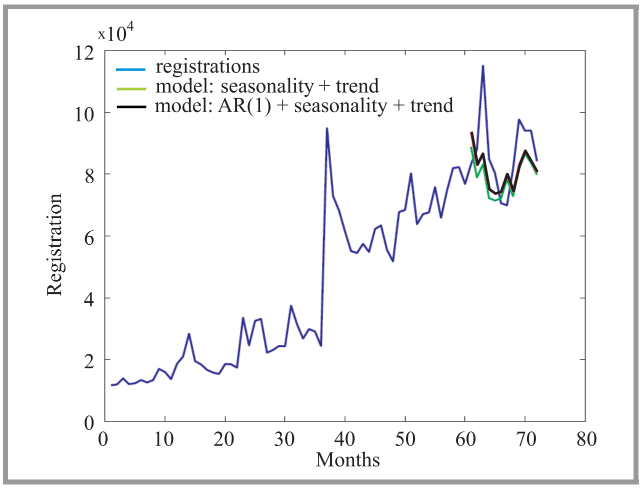


Fig. 11. Prediction by the model augmented with AR(1) versus pure seasonality with trend and data.

To assess the resolving value of the model ACF and PACF of its residuals were computed (see Fig. 12). The analysis of residuals show similarly to earlier results (see Fig. 10), relatively high value of ACF and PACF coefficients for 12th interval, however coefficients for shorter intervals are smaller than in the case of seasonality and trend modeling. Concluding: autocorrelations show that AR(1) model improves model fit with respect to shorter lags, however modeling longer dependencies may be beneficial, especial as the 12th interval has some interpretation in the nature of the analyzed process (yearly correlations caused by yearly rate of payments).

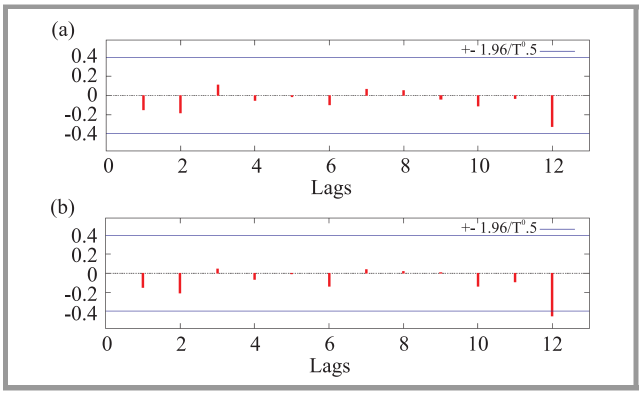


Fig. 12. Autocorrelation – ACF (a) and partial autocorrelation – PACF (b) for residuals of the model using seasonality and trend fitted to 2008–2009 period augmented with AR(1).

To check this hypothesis, AR(12) model consisting of three coefficients: for lags 1, 12 and intercept was fitted. The remaining lags (2-11) were skipped to avoid solving a poorly conditioned problem. The resulting model has significant coefficients, however not to a degree like in the AR(1) case. To assess the fit to the learning data set, Akaike information criterion (AIC) was computed. The application of AIC is

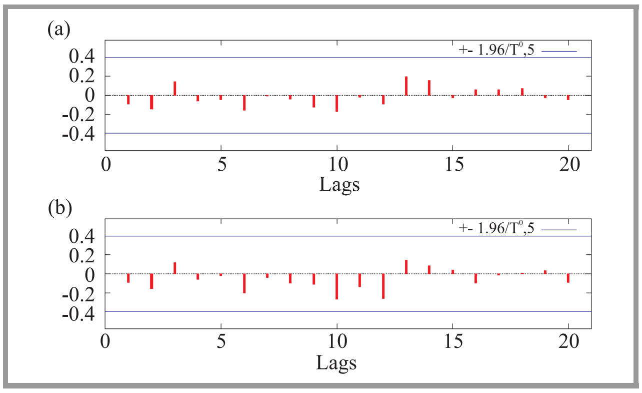


Fig. 13. Autocorrelation – ACF (a) and partial autocorrelation – PACF (b) for residuals of the model using seasonality and trend fitted to 2008–2009 period augmented with AR(12).

reasonable here, as it not only provides measure of fit to the learning data, but also provides correction for complexity of the model. For AR(12) it is a bit better than in case of the AR(1) model (476.5 vs. 481.7), also ACF and PACF (see Fig. 13) show some reduction of coefficients for higher lags. These findings may be contradicted by assessing the quality of prediction – the mean square error for AR(12) model is visibly higher (4165.9 vs. 3361.5). So although the model seems to better reflect the character of learning data its ability of prediction is lower.

To check the possibility of finding better model, we identified and verified a number of models – we tried to test how introduction of longer lags may influence quality of fit and prediction, we also tested effects of using longer period to calculate seasonality (i.e., using again 2005–2009 instead of 2008–2009). To summarize the results we computed two indexes: AIC, and mean square error of prediction to assess the possibility of practical use. The results are presented in Table 1. They show that although it is possible to attain better fit to learning data by application of higher order AR model, it does not improve the quality of prediction. Also, as suggested by preliminary analysis using longer period to identify seasonality is ineffective – seasonal changes tend to evolve similarly to trends, however two years period allows to build relatively effective model.

Table 1
Comparison of long range models

Model variant	AR lags	Intercept	AIC	Prediction error
Trend period: 2008–2009, seasonality period: 2008–2009				
Trend+seasonality	–	–	–	3689.9
AR(1)	1	–	481.7	3361.5
AR(12)	1, 12	–	476.5	4165.9
AR(12) – 2nd variant	12	–	485.5	4396.4
AR(12) – 3rd variant	1, 12	+	476.9	3797.4
Trend period: 2008–2009, seasonality period: 2005–2009				
Trend+seasonality	–	–	–	3928.1
AR(1)	1	–	496.2	3584.8
AR(11)	1, 11	–	494.4	4140.3
AR(11) – 2nd variant	1, 11	+	496.2	3804.3
AR(10)	1, 10	–	493	4055.6
AR(10) – 2nd variant	1, 10	+	494	4016.8
Trend period: 2005–2009, seasonality period: 2008–2009				
Trend+seasonality	–	–	–	10852.0
Trend period: 2005–2009, seasonality period: 2005–2009				
Trend+seasonality	–	–	–	11098.0

6. Short Range Modeling

Although long range model presented in Section 5 is usually sufficient for making strategic decisions, there are situations when more precise, shorter range predictions are necessary. An example may be assessing resources needed for proper operation of registration databases or planning the advertisement campaign – sometimes even a date of publishing advertisements or billboards may be important. To achieve this goal a completely new model with resolution of days must be built, thankfully the prediction horizon may be reduced, 4 weeks being usually enough. The advantage of a short horizon is that much more data is available. In consequence, models can be better verified. We prepared 18 learning data sets of length 12 weeks selected from period 2008–2010, each of them accompanied by 4 subsequent weeks used for validation. Later, to check properties of models, we shortened learning sets to 4 weeks with validation sets unchanged. Such a construction of data sets allowed to tune 18 models independently and compute mean errors for comparison.

6.1. Model Construction

The model was constructed following the pattern used for long range model (see Subsection 5.1). The most important is seasonality, computed as average number of registrations in subsequent days of a week. Figure 14 shows weekly pattern generated this way, compared with original values of the process. The regularity of the data results in relatively good fit even for such a simple model. The explanation of weekly changes is easier when noted that lower sales

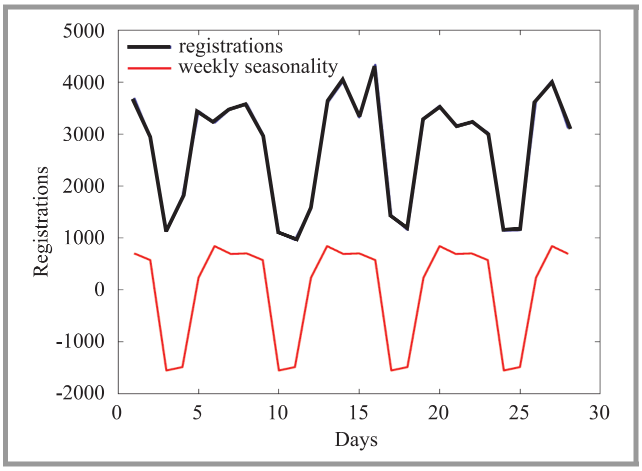


Fig. 14. Weekly seasonality versus registration process.

occurs in weekends. The reason for this may be twofold: first, weeks are scheduled for work – people usually tend to rest during weekends, second (and in fact resulting from the first), bank transfers can not be done on weekends. Payments are only possible by means of other services like, e.g., PayPal or a credit card.

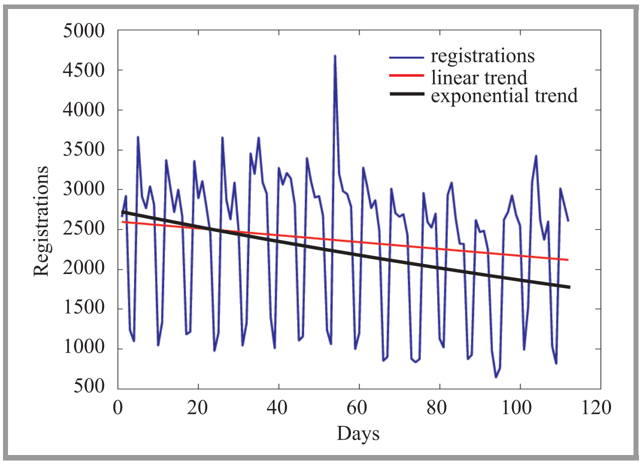


Fig. 15. Fitting trend: learning period of 12 weeks, last 4 weeks is a prediction.

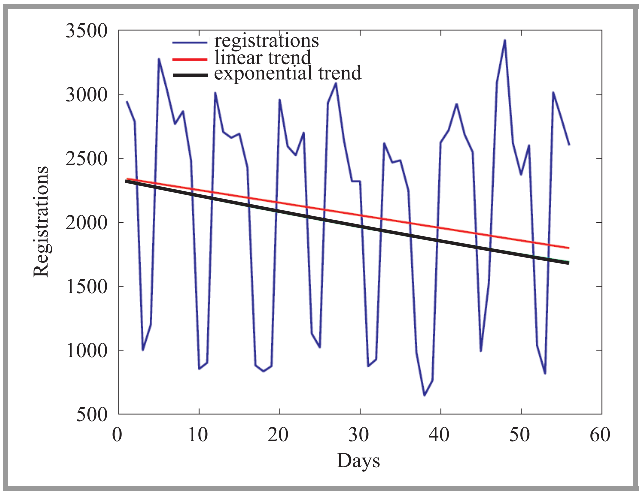


Fig. 16. Fitting trend: learning period of 4 weeks, last 4 weeks is a prediction.

Fitting a trend in a short time horizon is slightly different task than in a timespan of several years. Changes are not so pronounced. For this reason, we tried to use not only previously selected exponential trend, but also a linear one. Another question is a selection of appropriate learning period – there is a danger of unnecessarily introducing long range fluctuations, which are beyond resolution of a short range model. We tried to fit both trends to initially selected learning period (12 weeks) and shortened data set (4 weeks). The results are presented in Figs. 15 and 16 respectively. Observation of graphs allows to find out that longer learning period results in better, a bit damped, estimation. Also, the linear trend performs better, giving more stable prediction.

6.2. Model Validation

Combining seasonality and trend into a single model results in predictions presented in Fig. 17 for learning period of 12 weeks and 4 weeks (Fig. 18). Parts (a) figures show prediction compared to observed reservations while (b) two ACF and PACF plots respectively, in both cases the prediction is calculated for 25-03-2008 to 21-04-2008 being typical period for all of 18 analyzed samples.

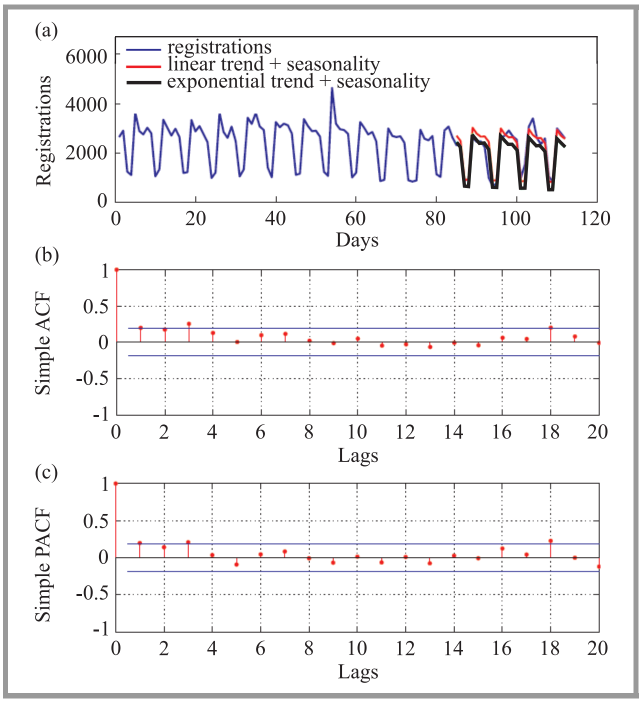


Fig. 17. Prediction for 4 weeks using seasonality and trend, learning period of 12 weeks: (a) prediction itself; (b) ACF, and (c) PACF of residuals for linear trend and seasonality.

Results are surprisingly good, especially in case of 12 week learning period and linear trend. Of course, it is impossible to predict some rapid, individual changes like e.g. in the second part of prediction, however, the fact, that all coefficients in the residuals ACF (see Fig. 17(b)) are reduced, proves the quality of proposed model. Such a shape of autocorrelation suggests that application of autoregressive models to improve prediction would be nearly impossible – and it was indeed the result of our trials. On the other hand, the PACF graph of the model tuned to shorter period of data (see Fig. 18(c)) shows some interesting properties – although coefficients for most of lags are highly reduced, the lag 14 coefficient is significant, suggesting some dependence on the span of two weeks. This hypothesis seems to be understandable – the presence of such a cycle may be somehow explained (e.g., investors may observe market in one week and then take decisions). However, building 14th order autoregressive model to encompass this is hardly feasible (and it proved to be), especially when confronted with results of modeling using 12 weeks of learning data, when this problem is overcome by averaging over longer period.

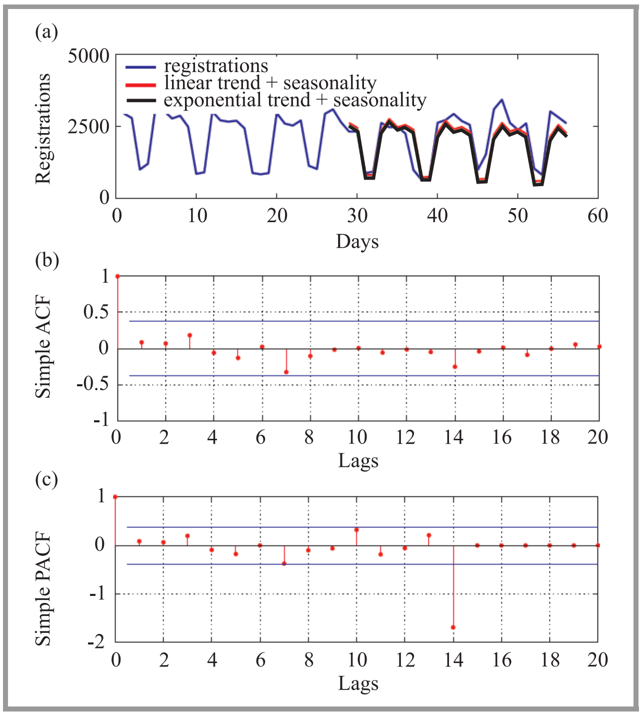


Fig. 18. Prediction for 4 weeks using seasonality and trend, learning period of 4 weeks: (a) prediction itself; (b) ACF, and (c) PACF of residuals for linear trend and seasonality.

Table 2
Comparison of short range models

Learning period	Trend	Error of 18 predictions
1 month	linear	166.12
1 month	exponential	178.45
3 months	linear	134.59
3 months	exponential	139.15

To summarize: as results for the model constructed of seasonality and linear trend tuned to longer period of data was sufficient to describe most of short range properties of registration process, and attain precision of approx. 15%, we refrained from further refinement. The results in the form of mean square error of 18 cases for all analyzed variants are presented in Table 2.

7. Composite Modeling

Encouraged by promising results acquired with long and short range models, we decided to try to construct a model, which while having long range (possibly one year) capability will allow prediction with high resolution – possibly of one day like the short range model. Such a model can be useful for making some decisions based on precise forecast of registrations, it can also provide some important information on the nature of the analyzed process. The possibility of building such a model is mostly grounded by the fact of relatively high regularity of weekly cycles what was shown in Section 6.

7.1. Model Construction

The core of the model is monthly registration sums computed by means of the long range model. The best version of the model i.e. with calculation of seasonality and trend using two years data and AR(1) model was used. Monthly sums are interpolated linearly over subsequent days of a month, as it was shown that the linear trend performs better in the short range model. Obtained this way, monthly trend is then modified with weekly seasonality calculated in similar way, as for the short range model but independently for subsequent months. This way, different shape of weekly cycle (mostly amplitude) is taken into account. During initial evaluation we found out that the amplitude of weekly cycles changes in subsequent years – typically it grows, when number of registrations grows. This phenomenon can not be modeled by summation of a trend and seasonality – to encompass it we introduced a multiplicative factor – amplitude growth rate.

7.2. Model Validation

The same, as in the previous experiments learning data consisted of daily registrations in years 2008–2009, while data from 2010 was used for validation. Five variants of the model were compared, they differed in the way of calculation of the following components:

- weekly seasonality: for the whole period or one year selected,
- amplitude growth rate: none, monthly or annual.

The reason for shortening data period used for weekly seasonality computation was the occurrence of the above mentioned changes in the amplitude of cycles. Two variants of amplitude growth rate were calculated to identify its nature: eventually it can be stated that the growth of amplitude may be seen as long range process correlated with general (yearly) trend.

Validation showed that all five models behave surprisingly well, describing most of significant properties of the data. The most important is the ability to follow general trend and

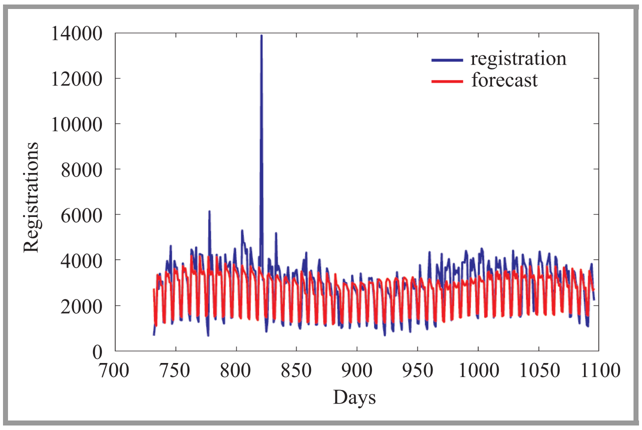


Fig. 19. Prediction for year 2010 using composite model with annual amplitude growth rate.

to model seasonal variations of weekly amplitude. Modeling of the last property is to some extent improved by introducing multiplicative component – annual rate of growth – in the most successful model (see Fig. 19).

The performance of all models is summarized in Table 3. Although the results are very good, it must be noted that still some periods when customers behave differently than usual (e.g., rise in the beginning of autumn 2010), and anomalies cannot be predicted. To analyze performance better ACF and PACF of residuals were computed (see Fig. 20). The results are difficult to interpret and probably need the further analyses. What can be stated now is that not all coefficients of ACF in the range of 1 to 50 days are sufficiently reduced, which may suggest presence of some unmodeled dependencies. Also, the PACF graph does not decay smoothly – there are some lags of length between 180–240 days, which have significant coefficients. The 6 month (approx. 180 days) lag may be to some extent attributed to two periods of higher sales observed in every year while longer may result from irregularities caused by external factors.

Table 3
Comparison of composite models

Weekly seasonality	Growth rate	Mean square error
2008–2009	none	820.95
2008	none	841.11
2009	none	831.72
2008–2009	monthly	836.42
2008–2009	yearly	816.39

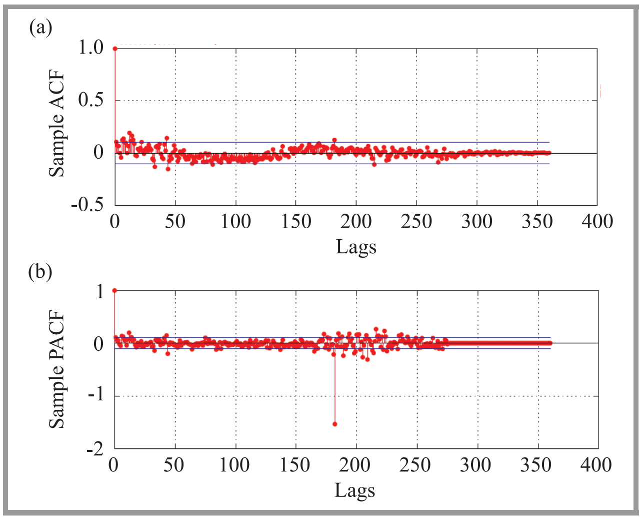


Fig. 20. Autocorrelation (a) and partial autocorrelation (b) for residuals of the composite model with annual amplitude growth rate.

Another question implied by this analysis is the presence of long range dependence in the registration process. The autocorrelations (also these computed for long range model) can not answer this question clearly – first of all, the number of samples is relatively small. Nevertheless, we tried to estimate Hurst coefficient for the process by fitting fractional Brown motion process. We ended with Hurst coefficient of 0.8 and relatively poor fit. Our supposition is that the long range dependence in the registration process is possible, however, it is likely that it is implied by other socio-economic variables, e.g., economic cycles to name most obvious one, which in turn are known to be long range dependent.

8. Conclusions

We have analyzed data and proposed models for various time scales. The most important outcome of these analyses is in our opinion identification of periodic nature of registration process. The periodicism has two scales – shorter, connected with weekly cycle and longer, visible as two periods of lower sales during the year. Another important part of the process is a trend, which in long range may be best modeled by exponential curve. These components were used to build models proposed, which proved to be precise enough for planning marketing strategies or sizing hardware.

There are also factors we do not cover in our models – mostly connected with external variables, which influence registrations. We roughly identified two such variables: one is general socio-economic situation and the second are prices. Both of them are difficult to comprehend, especially in the case of prices it is difficult to observe strategies of all dealers selling domains. However, we plan to analyze the influence of external factors deeper and to input them into the model, possibly in an aggregated form using indexes and statistics. We hope to solve the problem of nonstationarity this way and eventual long range dependence, which we could observe in the data.

References

[1] M. Mueller, “Toward an economics of the domain name system”, in *Handbook of Telecommunications Economics: Technology Evolution and the Internet*, M. E. Cave, S. K. Majumdar, and I. Vogelsang Eds., vol. 2, North-Holland, 2005, pp. 443–487.

[2] M. Jindra, “The market for Internet domain names”, in *Proc. 16th ITS Regional Conf.*, Porto, Portugal, 2005 [Online]. Available: http://userpage.fu-berlin.de/∼jmueller/its/conf/porto05/papers/Jindra.pdf, accessed 12.08.2011.

[3] “The Secondary Market For Domain Names”, *OECD Report DSTI/ICCP/TISP(2005)9/FINAL*, 2005.

[4] M. Zook, “Study of the Factors Behind the Demand for Country Code Domain Names”, ZookNIC Internet Intelligence [Online]. Available: http://www.aptld.org/pdf/APTLD-Bangkok-07-Report-FINAL.pdf, 2007, accessed: 12.08.2011.

[5] P. Salvador and A. Nogueira, “Analysis of the Internet domain names re-registration market”, *Procedia Comp. Science*, p. 325–335, 2011.

[6] K. Lasota and K. Kozakiewicz, “Analysis of the similarities in malicious DNS domain names”, *Commun. Computer Inform. Sci.*, vol. 187, pp. 1–6, Springer, 2011.

[7] J. Wu, X. Li, X. Wang, and B. Yan, “DNS usage mining and its two applications”, in *Proc. Sixth Int. Conf. Digit. Inform. Manag. ICDIM 2011*, Melbourne, Australia, 2011, pp. 54–60.

[8] T. Lindenthal, “Valuable Words: Pricing Internet Domains”, Social Science Research Network, July 14, 2011 [Online]. Available: http://ssrn.com/abstract=1885465, accessed 09.01.2012.

[9] A. Tajirian, "Statistical Models for Market Approach to Domain Name Valuation", 2010 [Online]. Available: http://www.domainmart.com/news/Statistical_Models_for_Market_Approach_to_Domain_Name_Value.pdf, accessed 09.01.2012.

[10] H. Simon, *Price Management*. North-Holland, 1989.

[11] G. E. P. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control*. Holden-Day, 1970.

[12] S. Makridakis, S. C. Wheelwrighth, and R. J. Hyndman, *Forecasting: Methods and Applications*. Willey, 1998.

[13] J. Haywood and J. Randal, "Modelling seasonality and multiple structural breaks. Did 9/11 affect visitor arrivals to NZ?", *Research report 08/10*, University of Wellington, New Zealand 2008 [Online]. Available: http://msor.victoria.ac.nz/twiki/pub/Main/ResearchReportSeries/mscs08-10.pdf, accessed 12.08.2011.

[14] M. K. Francke, "Repeat Sales Index for Real Estate Prices: a Structural Time Series Approach", *Technical Working paper 2009-0*, Ortec Finance Research Center, 2009.

[15] D. A. Pierce, "A survey of recent developments in seasonal adjustment", *The American Statistician*, vol. 34, no. 3, 1980.

[16] Ch. I. Plosser, "A time series analysis of seasonality in econometric models", in *Seasonal Analysis of Economic Time Series*, A. Zellner, Ed. NBER, 1979, pp. 365–410.

[17] B. Sheharyar and D. Geltner. "Estimating real estate price movements for high frequency tradable indexes in a scarce data environment", *The Journal of Real Estate Finance and Economics*, 2010 [Online]. Available: http://www.springerlink.com/content/uxl087168u781086/, accesed: 09.08.2011.

[18] L. Ferrara and D. Guagan, "Fractional seasonality: Models and Application to Economic Activity in the Euro Area", *Working paper*, Eurostat 2006 [Online]. Available: http://epp.eurostat.ec.europa.eu, accessed 12.08.2011.

[19] R. B. Cleveland, W. S. Cleveland, J. E. McRae, and I. Terpenning, "STL: a seasonal trend decomposition procedure based on loes", *J. Official Statistics*, vol. 6, no. 1, pp. 3–73, 1990.

[20] M. Yogo, "Measuring Business Cycles: Wavelet analysis of economic time series", *Economics Letters*, vol. 100, no. 2, 2008, [Online]. Available: http://ssrn.com/abstract=1120482, accessed 10.08.2011.

Piotr Arabas received his Ph.D. in Computer Science from the Warsaw University of Technology, Poland, in 2004. Currently, he is assistant professor at Institute of Control and Computation Engineering at the Warsaw University of Technology. Since 2002 with Research and Academic Computer Network (NASK). His research area focuses on modeling computer networks, predictive control and hierarchical systems.

E-mail: parabas@ia.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

E-mail: Piotr.Arabas@nask.pl
Research Academic Computer Network (NASK)
Wąwozowa st 18
02-796 Warsaw, Poland

Przemysław Jaskóła received his M.Sc. in Computer Science from the Warsaw University of Technology, Poland, in 1999. Currently, he is a Ph.D. student in the Institute of Control and Computation Engineering at the Warsaw University of Technology. Since 2005 with Research and Academic Computer Network (NASK). His research area focuses on hierarchical optimization and computer networks.

E-mail: Przemyslaw.Jaskola@nask.pl
Research Academic Computer Network (NASK)
Wąwozowa st 18
02-796 Warsaw, Poland

Mariusz Kamola received his Ph.D. in Computer Science from the Warsaw University of Technology, Poland, in 2004. Currently, he is assistant professor at Institute of Control and Computation Engineering at the Warsaw University of Technology. Since 2002 with Research and Academic Computer Network (NASK). His research area focuses on economics of computer networks and large scale systems.

E-mail: mkamola@ia.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska 15/19
00-665 Warsaw, Poland

E-mail: Mariusz.Kamola@nask.pl
Research Academic Computer Network (NASK)
Wąwozowa st 18
02-796 Warsaw, Poland

Michał Karpowicz received his Ph.D. in Computer Science from the Warsaw University of Technology (WUT), Poland, in 2010. Currently, he is an Assistant Professor at Research and Academic Computer Network (NASK). His research interests focus on game theory, network control and optimization.

E-mail: Michal.Karpowicz@nask.pl
Research Academic Computer Network (NASK)
Wąwozowa st 18
02-796 Warsaw, Poland

Paper

RT^T_+ – Time Validity Constraints in RT^T Language

Anna Felkner and Adam Kozakiewicz

Research and Academic Computer Network (NASK), Warsaw, Poland

Abstract—Most of the traditional access control models, like mandatory, discretionary and role based access control make authorization decisions based on the identity, or the role of the requester, who must be known to the resource owner. Thus, they may be suitable for centralized systems but not for decentralized environments, where the requester and service provider or resource owner are often unknown to each other. To overcome the shortcomings of traditional access control models, trust management models have been presented. The topic of this paper is three different semantics (set-theoretic, operational, and logic- programming) of RT^T, language from the family of role-based trust management languages (RT). RT is used for representing security policies and credentials in decentralized, distributed access control systems. A credential provides information about the privileges of users and the security policies issued by one or more trusted authorities. The set-theoretic semantics maps roles to a set of sets of entity names. Members of such a set must cooperate in order to satisfy the role. In the case of logic-programming semantics, the credentials are translated into a logic program. In the operational semantics the credentials can be established using a simple set of inference rules. It turns out to be fundamental mainly in large- scale distributed systems, where users have only partial view of their execution context. The core part of this paper is the introduction of time validity constraints to show how that can make RT^T language more realistic. The new language, named RT^T_+ takes time validity constraints into account. The semantics for RT^T_+ language will also be shown. Inference system will be introduced not just for specific moment but also for time intervals. It will evaluate maximal time validity, when it is possible to derive the credential from the set of available credentials. The soundness and completeness of the inference systems with the time validity constraints with respect to the set-theoretic semantics of RT^T_+ will be proven.

Keywords*—*access control, inference system with time constraints, logic-programming semantics, role-based trust management, set-theoretic semantics.

1. Introduction

Guaranteeing that confidential data and services offered by a computer system are not made available to unauthorized users is an increasingly significant and challenging issue, which must be solved by reliable software technologies that are used for building high-integrity applications. The data, whether in electronic, paper or other form must be properly protected. The traditional solution to this problem are access control techniques, by which users are identified, and granted or denied access to a system, data and other resources, depending on their individual or group identity. This approach fits well into closed, centralized environments, in which the identity of users is known in advance.

Role-based access control (RBAC) model [1], [2] is the most flexible type of access control policy. It uses a user role to control of which users have access to particular resources. Access rights are grouped by the role name and access to resources is restricted to the users who are assigned to appropriate roles. This type of access control works well in a large-scale centralized system and is often used in enterprise environments. Quite the new challenges arise in decentralized and open systems, where the identity of users is not known in advance and the set of users can change. For example, consider a bookstore, in which students who are returning customers are eligible to get discount. However, when a person comes to the bookstore and she says that she is Mary Smith, then her identity itself will not help in deciding whether she is eligible for a discount or not. What can help in this particular situation are two credentials stating that she is a student (she has a student card) and that she owns a bookstore card. The identity of a user itself does not help in making decisions about their rights. What is needed to make such decisions is information about the privileges assigned to the user by other authorities, as well as trust information about the authority itself.

The term of *trust management* was introduced in 1996 by Blaze *et al*. in [3], who defined it as a unified approach to specify and interpret security policies, credentials and trust relationships. In trust management system an entity's privilege is based on its attributes instead of its identities. An entity's attributes are demonstrated through digitally signed credentials issued by multiple principals. A *credential* is an attestation of qualification, competence or authority issued to an individual by a third party. Examples of credentials in real life include identification documents, driver's licenses, membership cards, keys, etc. A credential in a computer system can be a digitally signed document. Such a concept of trust management has evolved since that time to a much broader context of assessing the reliability and developing trustworthiness for other systems and individuals [4]. In this paper, however, we will use the term trust management only in a meaning restricted to the field of access control.

The potential and flexibility of trust management approach stems from the possibility of *delegation*: a principal may transfer limited authority over a resource to other principals.

Such a delegation is implemented by means of an appropriate credential. This way, a set of credentials defines the access control strategy and allows deciding on who is authorized to access a resource, and who is not. RT languages combine trust management and RBAC features. To define a trust management system, a language is needed for describing entities (principals and requesters), credentials and roles, which the entities play in the system. Responding to this need, a family of role-based trust management languages has been introduced in [5]–[7]. The family consists of five languages: RT_0, RT_1, RT_2, RT^T, and RT^D, with increasing expressive power and complexity.

The core language of RT family is RT_0, described in detail in [7]. It allows describing localized authorities for roles, role hierarchies, delegation of authority over roles and role intersections. All the subsequent languages add new features to RT_0.

RT_1 introduces parameterized roles, which can represent relationships between entities.

RT_2 adds to RT_1 logical objects, which can be used to represent permissions given to entities with respect to a group of logically related objects (resources). Those extensions can help in keeping the notation concise, but do not increase the expressive power of the language, because each combination of parameters in RT_1 and each permission to a real instance of a logical object in RT_2 can be defined alternatively as a separate role in RT_0.

This paper focuses on RT^T languages, as it provides useful capabilities not found in any other languages: manifold roles to achieve both agreement of multiple principals from one set and from disjoint sets and role-product operators, which can express threshold and separation of duties policies. Similar to a role, which defines a set of principals, a manifold role defines a set of principal sets, each of which is a set of principals which cooperation satisfies the manifold role. A singleton role can be treated as a special case of a manifold role, which set of cooperating entities is a singleton set. This way, RT_0 credentials can also be expressed in RT^T language.

A threshold policy requires a specified minimum number of entities to agree on some fact, i.e., it requires agreement among k out of a set of entities that satisfy a specified condition, e.g., in a requirement that two different bank cashiers must authorize a transaction. Separation of duties policy requires a set of entities, each of which fulfills a specific role, to agree before access is granted. Both types of policies mean that some transactions cannot be completed by a single entity, because no single entity has all the access rights required to complete the transaction, that is why it is not possible to define it in RT_0.

RT^D provides mechanisms to describe delegation of role activations and selective use of role membership. This language is not covered in this paper.

A more detailed treatment of RT family can be found in [6]. The languages have a precise syntax and semantics definition. A set-theoretic semantics, which defines the meaning of a set of credentials as a function from the set of roles into the power set of entities, has been defined for RT_0 [8], [7] and we defined relational semantics, which apply also to other members of the family up to RT^T in [9].

The paper is organized as follows. Section 2 consists of the role-based trust management language syntax and description of three semantics (relational, operational and logic-programming), including example. Section 3 describes time validity in RT^T language. Section 4 shows inference system over new RT^T_+ language time constraints. An overview of the work related to RT systems and languages is given in Section 5. Final remarks are given in Conclusions.

2. The Syntax and Three Semantics of RT^T Language

Basic elements of RT languages are entities, role names, roles and credentials. *Entities* represent principals that can define roles and issue credentials, and requesters that can make requests to access resources. An entity can, e.g., be a person or program identified by a user account in a computer system or a public key. *Role names* represent permissions that can be issued by entities to other entities or groups of entities. *Roles* represent sets of entities that have particular permissions granted according to the access control policy. A role is described as a pair composed of an entity and a role name. *Credentials* define roles by appointing a new member of the role or by delegating authority to the members of other roles.

2.1. The Syntax of RT^T Language

In this paper, we use capital letters or nouns beginning with a capital letter (e.g., A, B) to denote entities and sets of entities. Role names are denoted as identifiers beginning with a small letter or just small letters (e.g., r, s). Roles take the form of an entity (the issuer of this role) followed by a role name separated by a dot (e.g., $A.r$). Credentials are statements in the language. A credential consists of a role, left arrow symbol and a valid role expression. There are six types of credentials in RT^T, which are interpreted in the following way:

$A.r \longleftarrow B$ – *simple membership*: entity B is a member of role $A.r$.

$A.r \longleftarrow B.s$ – *simple inclusion*: role $A.r$ includes (all members of) role $B.s$. This is a delegation of authority over r from A to B, because B may cause new entities to become members of the role $A.r$ by issuing credentials that define $B.s$.

$A.r \longleftarrow B.s.t$ – *linking inclusion*: role $A.r$ includes role $C.t$ for each C, which is a member of role $B.s$. This is a delegation of authority from A to all the members of the role $B.s$. The expression $B.s.t$ is called a *linked role*.

$A.r \leftarrow B.s \cap C.t$ – *intersection inclusion*: role $A.r$ includes all the entieties who are members of both roles $B.s$ and $C.t$. This is a partial delegation from A to B and C. The expression $B.s \cap C.t$ is called an *intersection role*.

$A.r \leftarrow B.s \odot C.t$ – role $A.r$ can be satisfied by a union set of one member of role $B.s$ and one member of role $C.t$. A set consisting of a single entity satisfying the intersection role $B.s \cap C.t$ is also valid.

$A.r \leftarrow B.s \otimes C.t$ – role $A.r$ includes one member of role $B.s$ and one member of role $C.t$, but those members of roles have to be different entities.

The models discussed in this paper can be, in general, very complex. Therefore, we present here only a simplified example, with the intention to illustrate the basic notions and the notation, with a focus on RT^T credentials.

Example 1 (Example of RT^T). Suppose that we need at least two out of four students to activate the subject. Using RT_0 credentials, we have to explicitly list all the students (four in this simple case) and choose two of them; this list needs to be changed each time members in the students role change. In RT^T only one credential is needed. Further, we want to have two students and one Ph.D. student, who can also (but does not have to) be a regular student. This requires just one more RT^T credential. The entire policy can be expressed as follows:

$$F.students \leftarrow F.student \otimes F.student\,, \tag{1}$$

$$F.activeSubject \leftarrow F.phdStudent \odot F.students\,. \tag{2}$$

Now, assume that the following credentials have been added:

$$F.student \leftarrow \{Alex\}\,, \tag{3}$$

$$F.student \leftarrow \{Betty\}\,, \tag{4}$$

$$F.student \leftarrow \{David\}\,, \tag{5}$$

$$F.student \leftarrow \{John\}\,, \tag{6}$$

$$F.phdStudent \leftarrow \{John\}\,, \tag{7}$$

$$F.phdStudent \leftarrow \{Emily\}\,. \tag{8}$$

Then one can conclude that, according to the policy, any pair of students from the set $\{Alex, Betty, David, John\}$ is sufficient to fulfill the role *F.students*, but to activate the subject it is required that either the pair includes *John*, or additionally *Emily* must also attend.

2.2. The Set-Theoretic Semantics of RT^T Language

The semantics of RT_0 has no potential to describe the meaning of RT^T, which supports manifold roles. Therefore, we define the meaning of a set of credentials as a relation over the set of roles and the power set of entities. Thus, we use a cartesian product of the set of roles and the power set of entities as the semantics domain of a RT language. The semantics mapping would associate a specific relation between roles and entities with each set of credentials. Such a relational approach allowed us to define a formal semantics of RT^T language presented in [9].

Example 2 (Set-theoretic semantics for Example 1). Computing consecutive relations S_i starts from an empty set, $S_0 = \phi$. According to Definition 2 from [9] only credentials 3 through 8 are mapped in S_0 into relation S_1:

$$\begin{aligned} S_1 = \{&(\{F\}, student, \{John\}), (\{F\}, student, \{Alex\}),\\ &(\{F\}, student, \{Betty\}), (\{F\}, student, \{David\}),\\ &(\{F\}, phdStudent, \{John\}),\\ &(\{F\}, phdStudent, \{Emily\})\}. \end{aligned}$$

Credential 1 adds the following instances to relation S_2:

$$\begin{aligned} S_2 = S_1 \cup \{&\\ &(\{F\}, students, \{John, Alex\}),\\ &(\{F\}, students, \{John, Betty\}),\\ &(\{F\}, students, \{John, David\}),\\ &(\{F\}, students, \{Alex, Betty\}),\\ &(\{F\}, students, \{Alex, David\}),\\ &(\{F\}, students, \{Betty, David\})\ \}. \end{aligned}$$

Credential 2 is resolved in S_3:

$$\begin{aligned} S_3 = S_2 \cup \{&\\ &(\{F\}, activeSubject, \{John, Alex\}),\\ &(\{F\}, activeSubject, \{John, Betty\}),\\ &(\{F\}, activeSubject, \{John, David\}),\\ &(\{F\}, activeSubject, \{John, Alex, Betty\}),\\ &(\{F\}, activeSubject, \{John, Alex, David\}),\\ &(\{F\}, activeSubject, \{John, Betty, David\}),\\ &(\{F\}, activeSubject, \{Emily, John, Alex\}),\\ &(\{F\}, activeSubject, \{Emily, John, Betty\}),\\ &(\{F\}, activeSubject, \{Emily, John, David\}),\\ &(\{F\}, activeSubject, \{Emily, Alex, Betty\}),\\ &(\{F\}, activeSubject, \{Emily, Alex, David\}),\\ &(\{F\}, activeSubject, \{Emily, Betty, David\})\ \}. \end{aligned}$$

The resulting relation S_3 cannot be changed using the given set of credentials, hence: $S_P = S_3$. Because the RT language considered in this example is RT^T, there is a set of sets of entities assigned to each role.

2.3. The Logic-Programming Semantics of RT^T

The second way that shows how the member sets of roles can also be calculated is to use a logic-programming semantics. The logic-programming semantics of RT_0 credentials was first introduced in [6]. A definition quoted in this subsection is a modified version of this semantics, which has been introduced in [8]. In this case the semantics is given indirectly. RT credentials are translated into a logic program and their semantics is obtained as the minimal Herbrand model of the translation. The main intention of this approach is to provide an implementation of credential resolution.

Definition 1: The logic-programming semantics of $\mathscr{P}$ is the minimal Herbrand model of $LP(\mathscr{P})$, the logic program defined as

$$LP(\mathscr{P}) = \bigcup_{c\in\mathscr{P}} lc(c)\,,$$

where function $lc(\cdot)$ translates every credential to a logic program clause as follows:

$$lc(A.r \leftarrow B) \triangleq r(A,B) : -$$

$$lc(A.r \leftarrow B.s) \triangleq r(A,\xi) : -s(B,\xi)$$

$$lc(A.r \leftarrow B.s.t) \triangleq r(A,\xi) : -s(B,\zeta),t(\zeta,\xi)$$

$$lc(A.r \leftarrow B.s \cap C.t) \triangleq r(A,\xi) : -s(B,\xi),t(C,\xi)$$

We decided to put some changes into logic-programming semantics for RT_0 and define the logic-programming semantics of RT^T.

$$\begin{aligned}
lc(A.r \leftarrow B) &= member(B,role(A,r))\\
lc(A.r \leftarrow B.s) &= member(X,role(A,r)) : -\\
&\quad member(X,role(B,s))\\
lc(A.r \leftarrow B.s.t) &= member(X,role(A,r)) : -\\
&\quad member(C,role(B,s)),\\
&\quad member(X,role(C,t))\\
lc(A.r \leftarrow B.s \cap C.t) &= member(X,role(A,r)) : -\\
&\quad member(X,role(B,s)),\\
&\quad member(X,role(C,t))\\
lc(A.r \leftarrow B.s \odot C.t) &= member(X \cup Y,role(A,r)) : -\\
&\quad member(X,role(B,s)),\\
&\quad member(Y,role(C,t))\\
lc(A.r \leftarrow B.s \otimes C.t) &= member(X \cup Y,role(A,r)) : -\\
&\quad member(X,role(B,s)),\\
&\quad member(Y,role(C,t)),X\backslash = Y
\end{aligned}$$

where $role(A,r)$ correspond to $A.r$, and $member(B,role(A,r))$ correspond to $A.r \leftarrow B$.
As in the case of the set-theoretic, we use Example 1 from Section 2 to illustrate the definition of RT semantics.

Example 3 (Logic-programming semantics for Example 1).

$$\begin{aligned}
lc(F.students \leftarrow & F.student \otimes F.student) =\\
& member(X \cup Y,role(F,students)) : -\\
& member(X,role(F,student)),\\
& member(Y,role(F,student)),X\backslash = Y
\end{aligned}$$

$$\begin{aligned}
lc(F.activeSubject \leftarrow & F.phdStudent \odot F.students) =\\
& member(X \cup Y,role(F,activeSubject)) : -\\
& member(X,role(F,phdStudent)),\\
& member(Y,role(F,students))
\end{aligned}$$

$$\begin{aligned}
lc(F.student \leftarrow Alex) &= member(Alex,role(F,student))\\
lc(F.student \leftarrow Betty) &= member(Betty,role(F,student))\\
lc(F.student \leftarrow David) &= member(David,role(F,student))\\
lc(F.student \leftarrow John) &= member(John,role(F,student))\\
lc(F.phdStudent \leftarrow John) &= member(John,role(F,phdStudent))\\
lc(F.phdStudent \leftarrow Emily) &= member(Emily,role(F,phdStudent))
\end{aligned}$$

The above rules can be easily implemented by using some prologue interpreter. Only minor syntactic changes (capital letters, etc.) are necessary.

2.4. Inference System over RT^T Credentials

RT^T credentials are used to define roles and roles are used to represent permissions. The semantics of a given set $\mathscr{P}$ of RT^T credentials defines for each role $A.r$ the set of entities, which are members of this role. The member sets of roles can also be calculated in a more convenient way by using an inference system, which defines an operational semantics of RT^T language. An inference system consists of an initial set of formulae that are considered to be true, and a set of inference rules that can be used to derive new formulae from the known ones.
Let $\mathscr{P}$ be a given set of RT^T credentials. The application of inference rules of the inference system will create new credentials, derived from credentials of the set $\mathscr{P}$. A derived credential c will be denoted using a formula $\mathscr{P} \succ c$, which should be read: credential c can be derived from a set of credentials $\mathscr{P}$.

Definition 2: The initial set of formulae of an inference system over a set $\mathscr{P}$ of RT^T credentials are all the formulae: $c \in \mathscr{P}$ for each credential c in $\mathscr{P}$. The inference rules of the system are the following:

$$\frac{c \in \mathscr{P}}{\mathscr{P} \succ c}\,, \qquad (W_1)$$

$$\frac{\mathscr{P} \succ A.r \leftarrow B.s \qquad \mathscr{P} \succ B.s \leftarrow X}{\mathscr{P} \succ A.r \leftarrow X}\,, \qquad (W_2)$$

$$\frac{\mathscr{P} \succ A.r \leftarrow B.s.t \qquad \mathscr{P} \succ B.s \leftarrow C \qquad \mathscr{P} \succ C.t \leftarrow X}{\mathscr{P} \succ A.r \leftarrow X}\,, \qquad (W_3)$$

$$\frac{\mathscr{P} \succ A.r \leftarrow B.s \cap C.t \qquad \mathscr{P} \succ B.s \leftarrow X \qquad \mathscr{P} \succ C.t \leftarrow X}{\mathscr{P} \succ A.r \leftarrow X}\,, \qquad (W_4)$$

$$\frac{\mathscr{P} \succ A.r \leftarrow B.s \odot C.t \qquad \mathscr{P} \succ B.s \leftarrow X \qquad \mathscr{P} \succ C.t \leftarrow Y}{\mathscr{P} \succ A.r \leftarrow X \cup Y}\,, \qquad (W_5)$$

$$\frac{\mathscr{P} \succ A.r \leftarrow B.s \otimes C.t \qquad \mathscr{P} \succ B.s \leftarrow X \qquad \mathscr{P} \succ C.t \leftarrow Y \qquad X \cap Y = \phi}{\mathscr{P} \succ A.r \leftarrow X \cup Y}\,. \qquad (W_6)$$

There could be a number of inference systems defined over a given language. To be useful for practical purposes an inference system must exhibit two properties. First, it should be sound, which means that the inference rules could derive only formulae that are valid with respect to the semantics of the language. Second, it should be complete, which means that each formula, which is valid according to the semantics, should be derivable in the system.

All the credentials, which can be derived in the system, either belong to set $\mathscr{P}$, rule (W_1) or are of the type: $\mathscr{P} \succ A.r \longleftarrow X$, rules ($W_2$ through W_6). To prove the soundness of the inference system, one must prove that for each new formula $\mathscr{P} \succ A.r \longleftarrow X$, the triple (A, r, X) belongs to the semantics $S_{\mathscr{P}}$ of the set $\mathscr{P}$. To prove the completeness of the inference system over a set $\mathscr{P}$ of RT^T credentials, we must prove that a formula $P \succ A.r \longleftarrow X$ can be derived by using inference rules for each element $(A, r, X) \in S_{\mathscr{P}}$. Both properties have been shown in [10], proving that the inference system provides an alternative way of presenting the semantics of RT^T.

Example 4 (Inference system for Example 1). We use the inference system to formally derive a set of entities which can cooperatively activate a subject. To make long example shorter, let us use less credentials ((1), (2), (4), (6), and (7)). Using credentials (1), (2), (4), (6), and (7) according to rule (W_1) we can infer:

$$\frac{F.students \longleftarrow F.student \otimes F.student \in \mathscr{P}}{\mathscr{P} \succ F.students \longleftarrow F.student \otimes F.student}$$

$$\frac{F.activeSubject \longleftarrow F.phdStudent \odot F.students \in \mathscr{P}}{\mathscr{P} \succ F.activeSubject \longleftarrow F.phdStudent \odot F.students}$$

$$\frac{F.student \longleftarrow \{Betty\} \in \mathscr{P}}{\mathscr{P} \succ F.student \longleftarrow \{Betty\}}$$

$$\frac{F.student \longleftarrow \{John\} \in \mathscr{P}}{\mathscr{P} \succ F.student \longleftarrow \{John\}}$$

$$\frac{F.phdStudent \longleftarrow \{John\} \in \mathscr{P}}{\mathscr{P} \succ F.phdStudent \longleftarrow \{John\}}$$

Then, using credentials (1), (6) and (4) and rule (W_6) we infer:

$$\frac{\begin{array}{c}\mathscr{P} \succ F.students \longleftarrow F.student \otimes F.student \\ \mathscr{P} \succ F.student \longleftarrow \{John\} \\ \mathscr{P} \succ F.student \longleftarrow \{Betty\} \\ \{John\} \cap \{Betty\} = \phi\end{array}}{\mathscr{P} \succ \mathbf{F.students} \longleftarrow \{\mathbf{John}, \mathbf{Betty}\}}$$

In the next step we use the newly inferred credential and additionally credentials (2) and (7) with the rule (W_5):

$$\frac{\begin{array}{c}\mathscr{P} \succ F.activeSubject \longleftarrow F.phdStudent \odot F.students \\ \mathscr{P} \succ F.phdStudent \longleftarrow \{John\} \\ \mathscr{P} \succ F.students \longleftarrow \{John, Betty\}\end{array}}{\mathscr{P} \succ \mathbf{F.activeSubject} \longleftarrow \{\mathbf{John}, \mathbf{Betty}\}},$$

showing that the set of entities $\{John, Betty\}$ is sufficient to activate the subject.

3. Time Validity in RT^T

Inference rules with time validity for RT_0 were originally introduced in a slightly different way in [8]. In this paper we will try to extend the potential of RT^T language by putting time validity constraints into this language. In this case credentials are given to entities just for some fixed period of time. It is quite natural to assume that permissions are given just for fixed period of time, not for ever. Time dependent credentials take the form: c **in** v, meaning ”the credential c is available during the time v”. Finite sets of time dependent credentials are denoted by $\mathscr{CP}$ and the new language is denoted as RT_+^T. To make notation clear we write c to denote ”c **in** $(-\infty, +\infty)$”. Time validity can be denoted as follows:

$[\tau_1, \tau_2]; [\tau_1, \tau_2); (\tau_1, \tau_2]; (\tau_1, \tau_2); (-\infty, \tau]; (-\infty, \tau);$
$[\tau, +\infty); (\tau, +\infty); (-\infty, +\infty); v_1 \cup v_2; v_1 \cap v_2; v_1 \backslash v_2$

and v_1, v_2 of any form in this list, with τ ranging over time constants.

Example 5 (Time validity for Example1). In our scenario, it is quite natural to assume that *Alex*, *Betty*, *David* and *John* are students only for a fixed period of time. The same with *John* and *Emily* as Ph.D. students. Thus, credentials (3)–(8) should be generalized to:

$$F.student \longleftarrow \{Alex\} \textbf{ in } v_1, \tag{9}$$

$$F.student \longleftarrow \{Betty\} \textbf{ in } v_2, \tag{10}$$

$$F.student \longleftarrow \{David\} \textbf{ in } v_3, \tag{11}$$

$$F.student \longleftarrow \{John\} \textbf{ in } v_4, \tag{12}$$

$$F.phdStudent \longleftarrow \{John\} \textbf{ in } v_5, \tag{13}$$

$$F.phdStudent \longleftarrow \{Emily\} \textbf{ in } v_6, \tag{14}$$

stating that credentials (3)–(8) are only available during v_1, v_2, v_3, v_4, v_5, and during v_6, respectively. On the other hand, credentials (1) and (2) are always valid, as they express some time-independent facts. Now, by using (1), (2) and (9)–(14), we want to be able to derive that for example the set $\{Alex, Betty, John\}$ can cooperatively activate the subject during all of the period: $v_1 \cap v_2 \cap v_5$ or $\{Betty, John\}$ during the time $v_2 \cap v_4 \cap v_5$.

3.1. Set-Theoretic Semantics of RT_+^T

Now we can adapt our set-theoretic semantics of RT^T language to the new form of credentials. The semantics can be defined formally in the following way:

Definition 3: The semantics of a set of credentials $\mathscr{CP}$, denoted as $S_{\mathscr{CP}}$, is the smallest relation S_i, such as:

1. $S_0 = \phi$
2. $S_{i+1} = \bigcup_{(c \textbf{ in } v) \in \mathscr{CP}} f(S_i, c) \qquad$ for $i = 0, 1, \ldots$

that is closed with respect to function f, which describes the meaning of credentials in the same way as in [9].

3.2. Logic-Programming Semantics of RT^T_+

When considering the logic-programming semantics of RT^T_+, two possible scenarios must be analyzed: validation of authority at a given time instant and establishing authority for a period of time. In the first scenario, the logic-programming semantics is calculated at a precise time instant, by only considering those time-dependant credentials which are valid at that moment. In view of the fact that there will be no big changes, we will not provide a precise definition of the semantics. The second scenario is more complex, since it involves computing intersections of validity periods. Yet this case is as a future work. Feasibility of creating such semantics is underlined by the fact that development of an inference system for this case proved to be possible, as illustrated in the next section.

4. Inference System over RT^T_+ Credentials

Now, we can adapt inference system over RT^T credentials to take time validity into account. Let $\mathscr{CP}$ be a given set of RT^T_+ credentials. The application of inference rules of the inference system will create new credentials, derived from credentials of the set $\mathscr{CP}$. A derived credential c valid in time τ will be denoted using a formula $\mathscr{CP} \succ_\tau c$, which should be read: credential c can be derived from a set of credentials $\mathscr{CP}$ during the time τ.

Definition 4: The initial set of formulae of an inference system over a set $\mathscr{CP}$ of RT^T_+ credentials are all in the form: c **in** $v \ \in \mathscr{CP}$ for each credential c valid in time v in $\mathscr{CP}$. The inference rules of the system are the following:

$$\frac{c \textbf{ in } v \in \mathscr{CP} \qquad \tau \in v}{\mathscr{CP} \succ_\tau c}, \tag{CW_1}$$

$$\frac{\mathscr{CP} \succ_\tau A.r \leftarrow B.s \quad \mathscr{CP} \succ_\tau B.s \leftarrow X}{\mathscr{CP} \succ_\tau A.r \leftarrow X}, \tag{CW_2}$$

$$\frac{\begin{array}{c}\mathscr{CP} \succ_\tau A.r \leftarrow B.s.t \qquad \mathscr{CP} \succ_\tau B.s \leftarrow C \\ \mathscr{CP} \succ_\tau C.t \leftarrow X\end{array}}{\mathscr{CP} \succ_\tau A.r \leftarrow X}, \tag{CW_3}$$

$$\frac{\begin{array}{c}\mathscr{CP} \succ_\tau A.r \leftarrow B.s \cap C.t \qquad \mathscr{CP} \succ_\tau B.s \leftarrow X \\ \mathscr{CP} \succ_\tau C.t \leftarrow X\end{array}}{\mathscr{CP} \succ_\tau A.r \leftarrow X}, \tag{CW_4}$$

$$\frac{\begin{array}{c}\mathscr{CP} \succ_\tau A.r \leftarrow B.s \odot C.t \qquad \mathscr{CP} \succ_\tau B.s \leftarrow X \\ \mathscr{CP} \succ_\tau C.t \leftarrow Y\end{array}}{\mathscr{CP} \succ_\tau A.r \leftarrow X \cup Y}, \tag{CW_5}$$

$$\frac{\begin{array}{cc}\mathscr{CP} \succ_\tau A.r \leftarrow B.s \otimes C.t & \mathscr{CP} \succ_\tau B.s \leftarrow X \\ \mathscr{CP} \succ_\tau C.t \leftarrow Y & X \cap Y = \phi\end{array}}{\mathscr{CP} \succ_\tau A.r \leftarrow X \cup Y}. \tag{CW_6}$$

All the credentials, which can be derived in the system, either belong to set $\mathscr{CP}$, rule (CW_1) or are of the type: $\mathscr{CP} \succ_\tau A.r \leftarrow X$, rules ($CW_2$ through CW_6). This new inference system mainly extends the inference rules from previous section, by replacing rules (W_i) with (CW_i) and considering only valid time-dependent credentials from $\mathscr{CP}$.

To prove the soundness of the inference system we must prove that for each new formula $\mathscr{CP} \succ_\tau A.r \leftarrow X$, the triple (A,r,X) belongs to the semantics $S_{\mathscr{CP}}$ of the set $\mathscr{CP}$. Let us first note that all the formulae $\mathscr{CP} \succ_\tau A.r \leftarrow X$, such as $A.r \leftarrow X \in \mathscr{CP}$ are sound. This is proven in Lemma 1.

Lemma 1: If $A.r \leftarrow X \in \mathscr{CP}$ then $(A,r,X) \in S_{\mathscr{CP}}$.

Proof: The relation $S_{\mathscr{CP}}$, which defines the semantics of $\mathscr{CP}$, is a limit of a monotonically increasing sequence of sets $S_0, S_1 \ldots$ such that $S_0 = \phi$. According to Definition 3: $f(S_0, A.r \leftarrow X) = (A,r,X)$. Hence, $(A,r,X) \in S_1$ and because $S_1 \subseteq S_{\mathscr{CP}}$ then $(A,r,X) \in S_{CP}$. ■

To prove the soundness of the inference system over $\mathscr{CP}$, we must prove the soundness of each formula $\mathscr{CP}_\tau \succ A.r \leftarrow X$, which can be derived from the set $\mathscr{CP}$. This is proven in Theorem 1.

Theorem 1 (soundness): If $\mathscr{CP} \succ A.r \leftarrow X$ then $(A,r,X) \in S_{\mathscr{CP}}$.

Proof: Like the proof of Theorem 1 in [10], but relying on the above Lemma 1 instead of Lemma 1 from [10]. ■

To prove the completeness of the inference system over a set $\mathscr{CP}$ of RT^T_+ credentials, we must prove that a formula $\mathscr{CP} \succ A.r \leftarrow X$ can be derived by using inference rules for each element $(A,r,X) \in S_{\mathscr{CP}}$. This is proven in Theorem 2.

Theorem 2 (completeness): If $(A,r,X) \in S_{\mathscr{CP}}$ then $\mathscr{CP} \succ A.r \leftarrow X$.

Proof: Like the proof of Theorem 2 in [10], but relying on the above Lemma 1 instead of Lemma 1 from [10]. ■

4.1. Inferring Time Validity of Credentials

This inference system evaluates maximal time validity, when it is possible to derive the credential c from $\mathscr{CP}$. It enhances formula $\mathscr{CP} \succ_\tau c$ to $\mathscr{CP} \succ\!\succ_v c$, specifying that at any time $\tau \in v$ in which $\mathscr{CP}$ has a semantics, it is possible to infer the credential c from $\mathscr{CP}$. To make notation clear we write $\succ\!\succ$ to denote $\succ\!\succ_{(-\infty,+\infty)}$. The inference rules of the system are the following:

$$\frac{c \textbf{ in } v \in \mathscr{CP}}{\mathscr{CP} \succ\!\succ_v c}, \tag{CWP_1}$$

$$\frac{\mathscr{CP} \succ\!\succ_{v_1} A.r \leftarrow B.s \quad \mathscr{CP} \succ\!\succ_{v_2} B.s \leftarrow X}{\mathscr{CP} \succ\!\succ_{v_1 \cap v_2} A.r \leftarrow X}, \tag{CWP_2}$$

$$\frac{\begin{array}{c}\mathscr{CP} \succ\!\succ_{v_1} A.r \leftarrow B.s.t \qquad \mathscr{CP} \succ\!\succ_{v_2} B.s \leftarrow C \\ \mathscr{CP} \succ\!\succ_{v_3} C.t \leftarrow X\end{array}}{\mathscr{CP} \succ\!\succ_{v_1 \cap v_2 \cap v_3} A.r \leftarrow X}, \tag{CWP_3}$$

$$\frac{\begin{array}{c}\mathscr{CP} \succ\!\succ_{v_1} A.r \leftarrow B.s \cap C.t \qquad \mathscr{CP} \succ\!\succ_{v_2} B.s \leftarrow X \\ \mathscr{CP} \succ\!\succ_{v_3} C.t \leftarrow X\end{array}}{\mathscr{CP} \succ\!\succ_{v_1 \cap v_2 \cap v_3} A.r \leftarrow X}, \tag{CWP_4}$$

$$\frac{\begin{array}{c}\mathscr{CP} \succ\!\succ_{v_1} A.r \leftarrow B.s \odot C.t \qquad \mathscr{CP} \succ\!\succ_{v_2} B.s \leftarrow X \\ \mathscr{CP} \succ\!\succ_{v_3} C.t \leftarrow Y\end{array}}{\mathscr{CP} \succ\!\succ_{v_1 \cap v_2 \cap v_3} A.r \leftarrow X \cup Y}, \tag{CWP_5}$$

$$\frac{\begin{array}{cc}\mathscr{C}\mathscr{P} \succ\succ_{v_1} A.r \leftarrow B.s \otimes C.t & \mathscr{C}\mathscr{P} \succ\succ_{v_2} B.s \leftarrow X \\ \mathscr{C}\mathscr{P} \succ\succ_{v_3} C.t \leftarrow Y & X \cap Y = \phi\end{array}}{\mathscr{C}\mathscr{P} \succ\succ_{v_1 \cap v_2 \cap v_3} A.r \leftarrow X \cup Y}, \qquad (CWP_6)$$

$$\frac{\mathscr{C}\mathscr{P} \succ\succ_{v_1} c \quad \mathscr{C}\mathscr{P} \succ\succ_{v_2} c}{\mathscr{C}\mathscr{P} \succ\succ_{v_1 \cup v_2} c}. \qquad (CWP_7)$$

The key rule is (CWP_1). It claims that $\mathscr{C}\mathscr{P}$ can be used whenever it is valid. Rules (CWP_2) - (CWP_6) simply claim that an inference rule can be used only when all its premises are true and that the validity of the resulting credentials is the intersection of validity periods of all the premises. Finally, the rule (CWP_7) claims that if a credential c can be inferred both with validity v_1 and with validity v_2, then c can be inferred with validity $v_1 \cup v_2$. $\mathscr{C}\mathscr{P} \succ\succ_v$ generalises $\mathscr{C}\mathscr{P} \succ_\tau$. They are both equivalent whenever $v = [\tau, \tau]$. Because several possible ways may exist to infer a certain c from $\mathscr{C}\mathscr{P}$, all providing a different period of validity, the rule (CWP_7) can be used several times to broaden c's validity.

Definition 5 (maximal inference): An inference terminating in $\mathscr{C}\mathscr{P} \succ\succ_v c$ is called maximal if and only if:

1) there exists no $v' \supset v$ such that $\mathscr{C}\mathscr{P} \succ\succ_{v'} c$, and

2) every sub-inference terminating in $\mathscr{C}\mathscr{P} \succ\succ_{v''} c'$, for $c' \neq c$, which does not use c in its premises, is maximal.

The first condition ensures that the rule (CWP_7) has been used as much as possible to infer the validity of c. The second condition ensures that this property is propagated through the whole inference tree. Maximal inferences guarantee that v in (CWP_1) is the maximal time validity for $A.r \leftarrow X$.
For these inferences we can prove soundness and completeness of $\mathscr{C}\mathscr{P} \succ\succ_v$ by means of Theorem 3, which proof relies on the following Lemma.

Lemma 2: $\mathscr{C}\mathscr{P} \succ_\tau c$ implies that there exists a v containing τ such that $\mathscr{C}\mathscr{P} \succ\succ_v\ c$.

Proof: It suffices to replicate inference for $\mathscr{C}\mathscr{P} \succ_\tau c$, replacing every appearance of rule (CW_i) with (CWP_i), and v will be the intersection of the validity of all the credentials $\mathscr{C}\mathscr{P}$ used in the inference and will be at least $[\tau, \tau]$. ■

Theorem 3 ([soundness and completeness for maximal inferences): Let $\mathscr{C}\mathscr{P} \succ\succ_v c$ be a maximal inference and set $\mathscr{C}\mathscr{P}$ of RT_+^T credentials be defined. Then $\mathscr{C}\mathscr{P} \succ_\tau c$ if and only if $\tau \in v$.

Proof: By induction on the depth of $\mathscr{C}\mathscr{P} \succ\succ_v c$. For the base case, $\mathscr{C}\mathscr{P}$ must contain a credential c **in** v. If $\tau \in v$ we can trivialy conclude thanks to (CW_1). By induction, $\mathscr{C}\mathscr{P} \succ_\tau A.r \leftarrow X$ if and only if $\tau \in v$. And vice versa, assuming by contradiction that there is a $\tau' \notin v$ such that $\mathscr{C}\mathscr{P} \succ_{\tau'} c$; but then the inference leading to $\mathscr{C}\mathscr{P} \succ\succ_v c$ would not be maximal, because Lemma 2 would contradict the assumption.
For the inductive step, we prove by case analysis on the last rule used. Analysis of (CWP_i) for $i = 2...6$ is trivial, as it adapts the reasoning from proof in [10] in the same way as done above for the base case. The most difficult cases are when using rule (CWP_7). If $\mathscr{C}\mathscr{P} \succ\succ_v c$ terminates with an appearance of (CWP_7), then $v = v_1 \cup v_2$. This case is particular, because formulae $\mathscr{C}\mathscr{P} \succ\succ_{v_1} c$ and $\mathscr{C}\mathscr{P} \succ\succ_{v_2} c$ are not maximal. Let $\mathscr{C}\mathscr{P} \succ_\tau c$. By Lemma 2, there exists a v' containing τ such that $\mathscr{C}\mathscr{P} \succ\succ_{v'} c$. Now, it is that $v' \subseteq v$, otherwise $\mathscr{C}\mathscr{P} \succ\succ_v c$ would not be maximal. And vice versa, let $\tau \in v$ and let $\mathscr{C}\mathscr{P} \succ\succ_{v'} c$ be the deepest sub-inference of $\mathscr{C}\mathscr{P} \succ\succ_v c$, which premises do not require c (hence, $\mathscr{C}\mathscr{P} \succ\succ_{v'} c$ has been obtained by using (CWP_i), for $i \neq 7$) and such that $\tau \in v'$. By definition of the rules of inference system (inferring time validity), each of these premises has a time validity containing τ_i; since these premises have been obtained by maximal inferences, by induction we can replace $\succ\succ \ldots$ with $\succ_\tau$. Now, we have to use (CW_i) and conclude. ■

Example 6 (Time validity in inference system for Example 1). Let us get back to our example and to make long example shorter, let us use less credentials: (1), (2), (10), (12), and (13). According to rule (CWP_1) we can infer:

$$\frac{F.students \leftarrow F.student \otimes F.student \in \mathscr{C}\mathscr{P}}{\mathscr{C}\mathscr{P} \succ\succ F.students \leftarrow F.student \otimes F.student}$$

$$\frac{F.activeSubject \leftarrow F.phdStudent \odot F.students \in \mathscr{C}\mathscr{P}}{\mathscr{C}\mathscr{P} \succ\succ F.activeSubject \leftarrow F.phdStudent \odot F.students}$$

$$\frac{F.student \leftarrow \{Betty\}\ \mathbf{in}\ v_2 \in \mathscr{C}\mathscr{P}}{\mathscr{C}\mathscr{P} \succ\succ_{v_2} F.student \leftarrow \{Betty\}}$$

$$\frac{F.student \leftarrow \{John\}\ \mathbf{in}\ v_4 \in \mathscr{C}\mathscr{P}}{\mathscr{C}\mathscr{P} \succ\succ_{v_4} F.student \leftarrow \{John\}}$$

$$\frac{F.phdStudent \leftarrow \{John\}\ \mathbf{in}\ v_5 \in \mathscr{C}\mathscr{P}}{\mathscr{C}\mathscr{P} \succ\succ_{v_5} F.phdStudent \leftarrow \{John\}}$$

When we want to check if two different students can cooperate, from credentials (1), (10), (12) and rule (CWP_6) we infer:

$$\frac{\begin{array}{c}\mathscr{C}\mathscr{P} \succ\succ F.students \leftarrow F.student \otimes F.student \\ \mathscr{C}\mathscr{P} \succ\succ_{v_2} F.student \leftarrow \{Betty\} \\ \mathscr{C}\mathscr{P} \succ\succ_{v_4} F.student \leftarrow \{John\} \\ \{Betty\} \cap \{John\} = \phi\end{array}}{\mathscr{C}\mathscr{P} \succ\succ_{v_2 \cap v_4} \mathbf{F.students} \leftarrow \{\mathbf{Betty}, \mathbf{John}\}}$$

In the next step we use it and additionally credentials (2), (13) and rule (CWP_5):

$$\frac{\begin{array}{c}\mathscr{C}\mathscr{P} \succ\succ F.activeSubject \leftarrow F.phdStudent \odot F.students \\ \mathscr{C}\mathscr{P} \succ\succ_{v_5} F.phdStudent \leftarrow \{John\} \\ \mathscr{C}\mathscr{P} \succ\succ_{v_2 \cap v_4} F.students \leftarrow \{Betty, John\}\end{array}}{\mathscr{C}\mathscr{P} \succ\succ_{v_2 \cap v_4 \cap v_5} \mathbf{F.activeSubject} \leftarrow \{\mathbf{Betty}, \mathbf{John}\}}$$

showing that the set of entities that can cooperatively activate a subject is: $\{Betty, John\}$ during the time: $v_2 \cap v_4 \cap v_5$.

5. Related Work

Traditional access control systems usually rely on RBAC model [1], [2], which groups the access rights by the role name and limits the access to a resource to those users, who are assigned to a particular role.

The term trust management was first applied in the context of distributed access control in [3]. The first trust management system described in the literature was PolicyMaker [11], which defined a special assertion language capable of expressing policy statements, which were locally trusted, and credentials, which had to be signed using a private key. The next generation of trust management languages were KeyNote [12], which was an enhanced version of PolicyMaker, SPKI/SDSI [13] and a few other languages [14]. All those languages allowed assigning privileges to entities and used credentials to delegate permissions from its issuer to its subject. What was missing in those languages was the possibility of delegation based on attributes of the entities and not on their identity.

Trust management, introduced in [3], has evolved since that time to a much broader context of assessing the reliability and developing trustworthiness for other systems and individuals [4]. In this paper, however, we used the term trust management only in a meaning restricted to the field of access control.

The meaning of roles in RT captures the notion of groups of users in many systems and has been borrowed from RBAC approach. The core language of RT family is RT_0, described in detail in [7]. It allows describing localized authorities for roles, role hierarchies, delegation of authority over roles and role intersections. All the subsequent languages add new features to RT_0. A more detailed overview of the RT family framework can be found in [5], [6], [15].

Time-dependant credentials were introduced in [8] but just for RT_0 language. Because RT^T language is more complex, powerful and it allows to express security policies more suited to real needs, we decided to develop extensions to this specific language, which has not been done before.

6. Conclusions

This paper deals with modeling of trust management systems in decentralized and distributed environments. The modelling framework is the RT^T language from a family of role-based trust management. Three types of semantics for a set of RT^T credentials have been introduced in the paper. A set-theoretic semantics of RT^T has been defined as a relation over a set of roles and a power set (set of sets) of entities. All the members of a set of entities related to a role must cooperate in order to satisfy the role. In the case of logic-programming semantics, RT credentials are translated into a logic program. This way, our definitions cover the full potential of RT^T, which supports the notion of manifold roles and it is able to express structure of threshold and separation-of-duties policies. Using RT^T one can define credentials stating that an action is allowed if it gets approved by members of more than one role. This enables defining complex trust management models in a real environment. An operational semantics of RT^T is defined as a inference system, in which credentials can be established from an initial set of credentials using a simple set of inference rules. The core part of the paper is a formal definition of a sound and complete inference system, in which credentials can be derived from an initial set of credentials using a set of inference rules. The semantics is given by the set of resulting credentials of the type $A.r \longleftarrow X$, which explicitly show a mapping between roles and sets of entities. Using RT^T_+ one can define credentials, which state that an action is allowed if it gets approval from members of more than one role. This improves the possibility of defining complex trust management models in a real environment. The goal of this paper is the introduction of time validity constraints to show how that can make RT^T language more realistic. The properties of soundness and completeness of the inference system with respect to the semantics of RT^T_+ are proven. Inference systems presented in this paper are simple, but well-founded theoretically. It turns out to be fundamental mainly in large-scale distributed systems, where users have only partial view of their execution context.

Acknowledgements

This work is partially funded by the Polish Ministry of Science and Higher Education as part of project “Secure workstation for special applications”, grant number OR00014011.

References

[1] D. F. Ferraiolo, R. S. Sandhu, S. I. Gavrila, D. R. Kuhn, and R. Chandramouli, “Proposed NIST standard for role-based access control”, *ACM Trans. Inf. Syst. Secur.*, no. 3, pp. 224–274, 2001.

[2] R. S. Sandhu, E. J. Coyne, H. L. Feinstein, and C. E. Youman, “Role-based access control models”, *IEEE Computer*, no. 2, pp. 38–47, 1996.

[3] M. Blaze, J. Feigenbaum, and J. Lacy, “Decentralized trust management”, in *Proc. 17th IEEE Symp. Secur. Privacy*, Oakland, CA, USA, 1996, pp. 164–173.

[4] W. M. Grudzewski, I. K. Hejduk, A. Sankowska, and M. Wańtuchowicz, *Trust Management in Virtual Work Environments: A Human Factors Perspective*. CRC Press Taylor & Francis Group, 2008.

[5] N. Li and J. Mitchell, “RT: a role-based trust-management framework”, in *Proc. 3rd DARPA Inform. Surviv. Conf. Exp.*, IEEE Computer Society Press, Oakland, CA, USA, 2003, pp. 201–212.

[6] N. Li, J. Mitchell, and W. Winsborough, “Design of a role-based trust-management framework”, in *Proc. IEEE Symp. Secur. Privacy*, IEEE Computer Society Press, Oakland, CA, USA, 2002, pp. 114–130.

[7] N. Li, W. Winsborough, and J. Mitchell, “Distributed credential chain discovery in trust management”, *J. Comput. Secur.*, no. 1, pp. 35–86, 2003.

[8] D. Gorla, M. Hennessy, and V. Sassone, “Inferring dynamic credentials for role-based trust management”, in *Proc. 8th ACM SIGPLAN Conf. Princip. Pract. Declar. Program. PPDP 2006*, Venice, Italy, 2006, pp. 213–224.

[9] A. Felkner and K. Sacha, “The semantics of role-based trust management languages”, in *Proc. CEE-SET 2009*, Kraków, Poland, 2009, pp. 195–206, (preprints).

[10] A. Felkner and K. Sacha, “Deriving RT^T credentials for role-based trust management”, *e-Inf. Softw. Engin. J.*, vol. 4, pp. 9–19, 2010.

[11] M. Blaze, J. Feigenbaum, and M. Strauss, "Compliance checking in the policymaker trust management system", in *Proc. 2nd Int. Conf. Financial Cryptography*, London, United Kingdom, 1998, pp. 254–274.

[12] M. Blaze, J. Feigenbaum, and A. D. Keromytis, "The role of trust management in distributed systems security", in *Secure Internet Programming*, J. Vitek, C. D. Jensen, Eds. Springer, 1999, pp. 185–210.

[13] D. Clarke, J.-E. Elien, C. Ellison, M. Fredette, A. Morcos, and R. L. Rivest, "Certificate chain discovery in SPKI/SDSI", *J. Comp. Secur.*, no. 9, pp. 285–322, 2001.

[14] P. Chapin, C. Skalka, and X. S. Wang, "Authorization in trust management: features and foundations", *ACM Comp. Surv.*, no. 3, pp. 1–48, 2008.

[15] M. R. Czenko, S. Etalle, D. Li, and W. H. Winsborough, "An Introduction to the Role Based Trust Management Framework RT", Tech. Rep. TR-CTIT-07-34, Centre for Telematics and Information Technology University of Twente, Enschede, 2007.

Anna Felkner graduated from the Faculty of Computer Science of Białystok University of Technology (M.Sc., 2004) and the Faculty of Electronics and Information Technology of Warsaw University of Technology (Ph.D., 2010). At present she is an Assistant Professor at Network and Information Security Methods Team in NASK Research Division. Main scientific interests concern the security of information systems, especially access control and trust management.
E-mail: anna.felkner@nask.pl
Research and Academic Computer Network (NASK)
Wąwozowa st 18
02-796 Warsaw, Poland

Adam Kozakiewicz got his M.Sc. in Information Technology and Ph.D. in Telecommunications at the Faculty of Electronics and Information Technology of Warsaw University of Technology, Poland. Currently he works at NASK as Assistant Professor and Manager of the Network and Information Security Methods Team, also as part-time Assistant Professor at the Institute of Control and Computation Engineering at the Warsaw University of Technology. His main scientific interests include security of information systems, parallel computation, optimization methods and network traffic modeling and control.
E-mail: adam.kozakiewicz@nask.pl
Research and Academic Computer Network (NASK)
Wąwozowa st 18
02-796 Warsaw, Poland

Paper

Application of Social Network Analysis to the Investigation of Interpersonal Connections

Marcin Mincer[a] and Ewa Niewiadomska-Szynkiewicz[a,b]

[a] Institute of Control and Computation Engineering, Warsaw University of Technology, Warsaw, Poland
[b] Research and Academic Computer Network (NASK), Warsaw, Poland

Abstract—Social network analysis (SNA) is an important and valuable tool for knowledge extraction from massive and unstructured data. Social network provides a powerful abstraction of the structure and dynamics of diverse kinds of interpersonal connection and interaction. In this paper, we address issues associated with the application of SNA to the investigation and analysis of social relationships of people. We provide a brief introduction to representation and analysis of social networks, SNA models and methods. The main objective is to investigate the application of SNA techniques to data mining in case of two social networks Facebook and Twitter. The presented simulations illustrate how social analysis can be used to determine the interpersonal connections, importance of actors in a given social network and detect communities of people. We then discuss strength and weakness of SNA techniques.

Keywords*—*centrality measures, communities detection, social network, social network analysis.

1. Introduction

During the last decade social networks (SN) have become extremely popular and have been attracted attention of scientists of different disciplines, such as sociology, epidemiology, economy, computer engineering, telecommunication and many others [1]–[7]. Many systems in nature and technology are examples of social networks, i.e., systems composed of a large number of highly interconnected individuals (actors), whose structure is irregular, complex and dynamically evolving in time. Communication networks, such as the Internet and the World Wide Web, are examples of SN.

A social network is formally defined as a set of actors or social groups, and relationships such as: friendship, collaboration, business, political, etc. The first approach to capture the global properties of such systems is to model them as graphs which nodes represent the actors and links the relationships between them. Nevertheless, most of real world networks are characterized by the similar topological properties, such as relatively small characteristic path lengths, high clustering coefficients, degree correlations, which make them radically different from regular lattices and random graphs. Hence, in many cases the standard models from graph theory cannot be applied, and the dedicated techniques and methods have to be used.

In the beginning, the social network became a field of interest of sociology that did not use mathematical graph theory. It has appeared soon that merging experience of sociology and graph theory needed the dedicated formal social network analysis (SNA) methods.

Social network analysis is a group of graph theory based techniques that can be used to retrieve meaningful knowledge from networks formed by various actors. In the recent past, SNA techniques have been rapidly increasing their advance into a wide variety of applications and systems [2], [3], [8]–[13]. Due to powerful computers, emerging and widely adapted platforms such as Facebook, Twitter, Foursquare, nk.pl, and many others SNA has become commonly used approach to interpersonal connections analysis. Data about relationships of people are commonly available like never before, and applying analytical methods to them became a source of unique and valuable knowledge.

In literature, one can find an extensive survey of state of the art in SNA techniques and methods [1], [7], [14]–[18]. The topological and structural properties of social networks are considered. The major results and concepts in SN, with focus on the fundamental concept, i.e., scale-free and small-world properties, and current approaches to SN analytical analysis and simulation are described and discussed. Numerous books and papers present models demonstrating the main features of evolving networks, network topologies, and summarize software, currently used in the analysis of complex network systems.

The main aim of this paper is to present the application of SNA methods to retrieve meaningful information, from commonly used social media platforms. The goal of presented case studies is to show that SNA can be a strong technique to investigate the interpersonal connections. However, the application of SNA has some limitations and requirements for input data. The remainder of this paper is organized as follows. In Sections 2 and 3, we provide the introduction to SNA techniques. We focus on social networks properties and popular measures in SN. In Section 4, we describe two popular algorithms of communities detection. In Section 5, we present and discuss the results of simulation experiments. Two of them show the effectiveness of application of SNA techniques to data mining, in the case of the social networks Facebook and Twitter. The goal of these experiments was to illustrate how social analysis can be used to determine the social relationships of people. The next test concerned with cliques detection, show limitations of SNA techniques. The paper concludes in Section 6.

2. Properties of Social Networks

Various measures are used to classify a network to be the social network. They are commonly used by researchers and commercial users to analyze characteristics of social networks to be considered. The most important measures that come from the graph theory are presented below.

2.1. Basic Measures

Node degree. The simple measure for an individual actor in a network is the degree of the corresponding node. From the graph theory the degree k_i of the node i is defined as follows:

$$k_i = \sum_{j=1}^{N} a_{ij} = \sum_{j=1}^{N} a_{ji}\,, \tag{1}$$

where N denotes a number of nodes in a network, a_{ij} element of the coincidence matrix A defined as follows: $a_{ij} = 1$ if the nodes i and j are interconnected, $a_{ij} = 0$ otherwise.

Shortest path. A critical primitive in large scale graph problems is the estimation of the shortest path – a path between two network nodes in a given network, such as the sum of their weights corresponding to edges is minimized. The average shortest path for the whole network is widely used in SNs to capture characteristic features of these networks. The average shortest path is calculated as follows:

$$l = \frac{1}{N(N-1)} \sum_{i \neq j} d(i,j) \approx \frac{\ln N}{\overline{k}}\,, \tag{2}$$

where N denotes a number of nodes in a network, $d(i,j)$ the shortest path between nodes i and j, $\overline{k}$ the average degree of nodes calculated according to Eq. (1).

Clustering coefficient. Clustering coefficient G_i of the node i is defined as a fraction of existing edges between neighbors of the node i, and all edges that are possible between those neighbors. In undirected network, the maximal number of edges is computed as $\frac{k_i(k_i-1)}{2}$, where k_i denotes the degree of the i-th node. Clustering coefficient of the node i is computed as follows:

$$G_i = \frac{2|\{e_{jk}\}|}{k_i(k_i-1)}\,, \tag{3}$$

where $\{e_{jk}\}$ denotes a set of edges connecting neighbors of the node i.
We can calculate average clustering coefficient ($\overline{G}$) in a network:

$$\overline{G} = \frac{1}{N} \sum_{i=1}^{N} G_i\,, \tag{4}$$

where N denotes a number of nodes.

2.2. Properties of SN

The common properties of social networks are:

- scale-free networks,
- clusterability,
- small-world networks.

Social networks are conjectured to be *scale-free*. The typical scale-free network consists of a few nodes with high degree, and long tail of nodes with low degree. It is a common structure of most networks encountered in nature that was investigated by R. Albert and A.-L. Barabási and described in [19] and [20]. R. Albert and A.-L. Barabási observed that in case of social networks a degree distribution follows a power law:

$$P(k) \propto k^{-\alpha}\,, \tag{5}$$

where $P(k)$ denotes a probability that a degree of randomly selected node will be equal to k.
Typical social network consists of a set of communities, grouping strongly connected actors – the value of $\overline{G}$ defined in Eq. (4) is usually high (close to 1). Hence, we can say about high clusterability of SN [7].
The small-world networks were investigated by D. Watts, and described in [18]. It was proved that networks that widely occur in nature, especially communities of people are small-world networks. The typical feature of so-called *small-world* networks is that an average shortest path l defined in Eq. (2) is very small relative to the number of nodes N forming a network. It can be observed that in social networks $l \approx \ln(N)/\overline{k}$.

3. Centrality Measures

In many social network applications, the main objective of data analysis is to identify the *most important actors* in a network. We consider a network node (an actor) to be a prominent one, if it is extensively involved in relationships with other nodes that form a social network. Moreover, an importance of a node relies on the number of prominent nodes that are connected to this node. A variety of statistical parameters – centrality measures were designed to show differences in the importance of actors. They are described in details in literature [7], [17]. To calculate these measures direct and indirect, inter-node connections have to be considered. In this section we present definitions of the most noteworthy and popular measures.

3.1. Betweenness Centrality

A betweenness centrality is a very important measure, while considering flows in a network. The large betweenness value means that a given actor is connected with many other actors (directly and indirectly). The betweenness centrality for the i-th node is calculated as follows [15]:

$$Cb_i = \sum_{j}^{N}\sum_{k}^{N} \frac{g_{jik}}{g_{jk}}, \qquad i \neq k \neq j, \tag{6}$$

where g_{jik} denotes a number of shortest paths linking nodes j and k passing through the node i, g_{jk} a number of paths not including the node i.

Usually, Cb_i is normalized to values from $[0,1]$ by multiplying through $\frac{2}{(N-1)(N-2)}$, where N denotes a number of nodes.

3.2. Closeness Centrality

A view of a node centrality can be based on closeness or distance. The question is how close is a node to all other nodes in a network. This measure is very important and commonly used in the graph theory. In general, a closeness Cc_i of the node i is defined as the inverse of the sum of distances between the node i and all other nodes in a network:

$$Cc_i = \frac{1}{\sum_{j\neq i} d(i,j)}, \tag{7}$$

where $d(i,j)$ denotes the shortest path between node i and j.

This closeness measure can be viewed as a time required to spread information from a given node to all other reachable nodes in a network [15].

Another definition of closeness was proposed by M. E. J. Newman in [5]. Cc_i is defined as the average shortest path from the node i to all other reachable nodes

$$Cc_i = \frac{\sum_{j\neq i} d(i,j)}{N-1}, \tag{8}$$

where $N \geq 2$ denotes a number of nodes in a network.

3.3. Eigenvector Centrality

The eigenvector centrality measure highlights the importance of the node i within a social network. The value of this measure relies on a number of other prominent nodes that are linked to the node i. The eigenvector centrality corresponds to the network coincidence matrix A. According to formula (9), the centrality of the node i is proportional to the sum of centralities of all nodes that are connected to the i-th node.

$$Ce_i = \frac{1}{\lambda}\sum_{j=1}^{N} a_{ij} Ce_j, \tag{9}$$

where Ce_j is the eigenvector centrality of the j-th node, N is a number of nodes in a network and λ is the constant value, a_{ij} an element of the coincidence matrix A.

4. Community Detection

Social networks are usually formed by smaller subnetworks (communities). It is obvious that community consists of subset of actors (nodes) with dense inter-node connections within this subset. The links to nodes from other communities are less dense. The communities detection, which idea is to divide a network into communities is one of the most interesting and important problem in the investigation, and analysis of networks. It is a challenging task, especially when consider overlapping communities and the dynamics of networks. In social networks, the overlapping is natural, as people usually belong to many communities. Many algorithms of communities detection in complex systems have been developed and described in literature [14], [21], [16], [22]. Two common techniques, i.e., an algorithm developed by M. Girvan, M. E. J. Newman and modified by A. Clauset, and an algorithm proposed by V. D. Blondel *et al.* are described below.

4.1. Clauset&Newman Algorithm

The first algorithm of communities detection was developed by Girvan and Newman, and described in [22]. It was improved by Clauset [16]. The idea of the Clauset&Newman algorithm is to identify the edges in a network, which links different communities. This identification is based on the betweenness centrality measure Eq. (6) that is extended to the case of edges. The communities detection is performed in two phases. In the first phase, the betweenness centrality measures are calculated for all edges in the network. Next, the edge with the highest betweenness is identified and removed from the set of edges. A high value of the betweenness centrality is typical to nodes connecting two communities – many shortest paths linking nodes from different communities pass through such edge. In this way, we can split our network into subnetworks. In every iteration, a dendrogram is produced to illustrate how the network splits into communities with the successive removal of edges. The first phase stops when all edges are removed from the set. The final result is the dendrogram that demonstrates the clustering structure of a network. The algorithm switches to the second phase. In the second phase, the calculated dendrogram is analyzed, and a number of communities forming the network is estimated based on the value of a modularity coefficient Q. The modularity coefficient Q defined in Eq. (10) is calculated for all splits performed in successive iterations in the first phase, and demonstrated in the dendrogram.

$$Q = \sum_{l=1}^{M} e_{ll} - p_l^2, \qquad p_l = \sum_{m=1}^{M} e_{lm}, \tag{10}$$

where M denotes the number of groups, e_{lm} denotes the fraction of edges linking two groups l and m, e_{ll} the fraction of edges linking nodes from the same community l, p_l the fraction of edges with at least one end vertex inside the community l.

4.2. Blondel Algorithm

There are many alternative methods for communities detection. One of them was developed by V. D. Blondel *et al.*

and is described in [14]. It is a simple heuristic technique based on modularity optimization that calculates a network partition in a short computation time. The authors claim in [14] that their algorithm outperforms many other methods in terms of quality of communities detection and computation time.

The algorithm is composed of two phases that are repeated iteratively. It starts from the assumption that every node is assigned to a different communities, hence the initial number of communities is equal the number of nodes N in a network. Next, for each node i and all its neighbors j values of modularity coefficient Q (10) are detected under the assumption that the node i is moved to the community of j. The calculations are repeated for all neighbors of i. Finally, the node i is moved to the community, for which the gain of Q is the highest one. The calculations are repeated for all nodes in a network, until no further improvement can be achieved. The algorithm switches to the second phase. A new aggregated network is built. In every community detected during the first phase, all nodes from this community are aggregated into one *super-node*. The weights of the edges between super-nodes are equal to the sum of weights of edges, linking two communities corresponding to these super-nodes. Hence, a new network is formed by these super-nodes. The second phase is completed and the first phase of the algorithm is executed for the aggregated network. Then, both phases are repeated iteratively, until no further improvement in the modularity coefficient can be achieved. The result of the algorithm is the partition of the original network into communities. Moreover, the algorithm also computes division inside computed groups.

5. Numerical Experiments

Multiple experiments were performed for data acquired from widely used social networks. The goal was to verify the results of application of SNA methods to knowledge extraction from massive commonly available data. In our tests, we validated and compared two techniques for community detection, described in the previous section: Clauset&Newman and Blondel *et al.* algorithms. Four series of experiments were performed for data acquired from the social platforms. The objective of the first set of tests was to compare the performance of described grouping techniques. Next, two series of experiments were performed for data about interpersonal connections, acquired from two commonly used platforms Facebook and Twitter. Different kinds of social networks were considered. The last series, of tests was performed for data acquired from the thesixtyone.com web page. The objective was to detect cliques of malicious voters.

5.1. Comparison of Algorithms of Communities Detection

We validated the communities detection algorithms through simulation. The accuracy and performance of three algorithms were compared, two presented in Section 4 and MCL (Markov Clustering) technique described in [23]. All experiments were performed for data containing members of Karate club from San Francisco. The results of calculations, i.e., discovered communities are presented in Figs. 1–3.

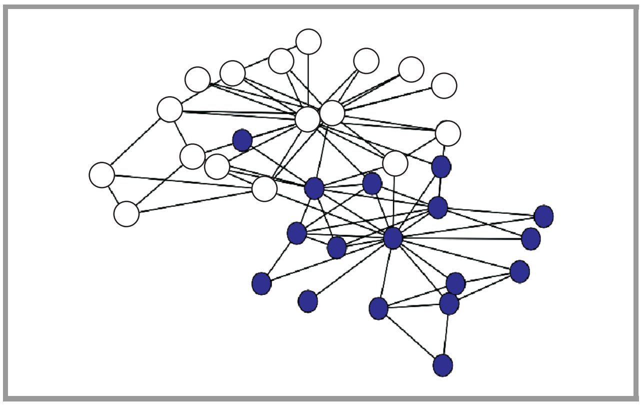


Fig. 1. Detected communities (Clauset&Newman aglorithm).

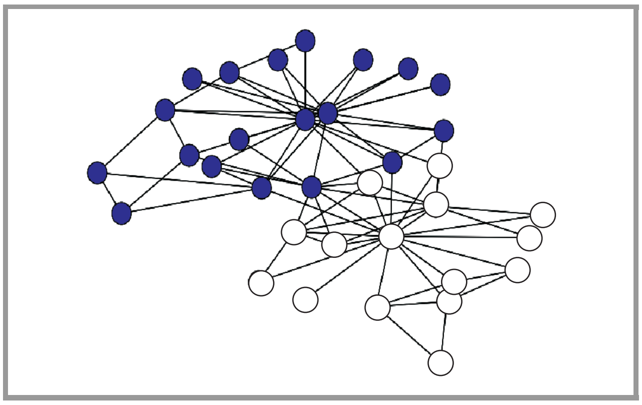


Fig. 2. Detected communities (MLC aglorithm).

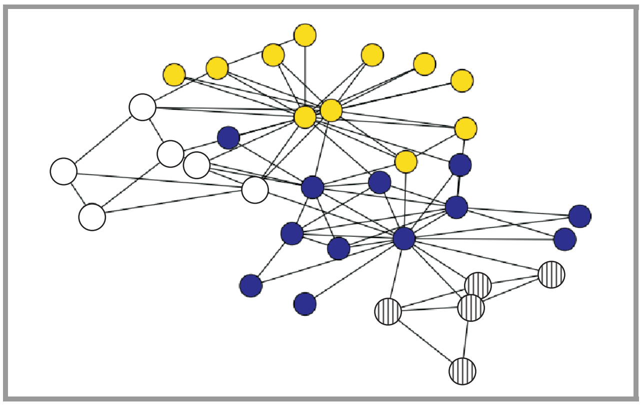


Fig. 3. Detected communities (Blondel aglorithm).

We obtained similar groupings with these three algorithms, although with some differences. Both Clauset&Newman and MLC algorithms discovered two groups, which differed only in two nodes. The disadvantage of the MLC algorithm is that it has to be tuned manually, so it is difficult to use. The Blondel algorithm identified four groups, but after dividing them into two pairs and merging members of these pairs, the results were the same as calculated using the Clauset&Newman algorithm.

Table 1
Communities detection – calculation time

Algorithm	Calculation time [s]
Griewan&Newman	5.051
MLC	4.979
Blondel	2.680

The goal of the second series of experiments was to compare the performance and efficiency of the algorithms. The network formed by 931 nodes and 73 228 edges was considered. The calculation times of communities detection using different algorithms are collected in Table 1. The results presented in this section indicate that the Blondel algorithm produced more accurate results, and it was about 2 times faster than the other methods. However, from the perspective of community detection accuracy, the suggestion is to use more than one algorithm and compare the results.

5.2. Social Network from Facebook

Facebook is a social networking service and website that connects people with other people, and share data between people. A user can create a personal profile, add other users as friends, exchange data, create and join common interest communities. The objective of our experiment was to validate the theorems formulated in SN domain on real network. We extracted a subnetwork from the Facebook database. Next, we calculated the centrality measures described in this paper, and finally divided this network into smaller communities. The test network was a special-kind network, so called ego-network. In such network, all nodes are connected to the *central node* (apart from being connected among themselves). In our case, the central node was one of the authors and the rest of the network was formed by his friends that were registered in Facebook.

We started our experiment from calculating the centrality measures of all nodes in our network. The results – values of degree, closeness, betweenness and eigenvector centrality measures are depicted in Figs. 4–7. From the experimental results, we can observe that for most nodes in the test network the calculated centrality measures are similar, low values. The results confirm theory about free-scale nature of social networks. It means that in SNs, usually only a few nodes are important and the other nodes are similarly not so much important. Moreover, it can be noticed that the correlation between centrality measures calculated for all nodes in the network is positive, i.e., in case of all nodes, a high value of one measure for a given node involves high values of other measures for this node.

Next, measures for the whole network were computed, i.e., a shortest path and a clustering coefficient. We ob-

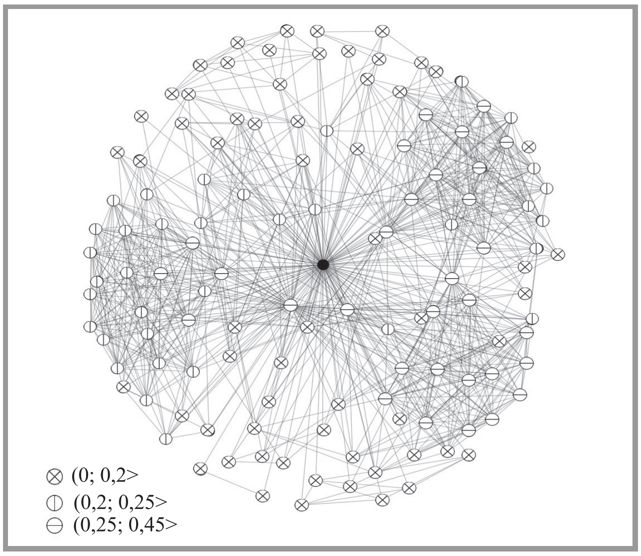


Fig. 4. Degree centrality of nodes; the Facebook network.

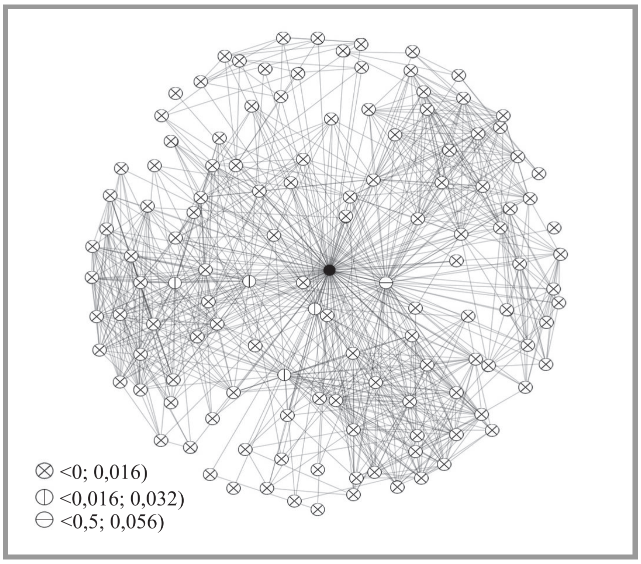


Fig. 5. Betweenness centrality of nodes; the Facebook network.

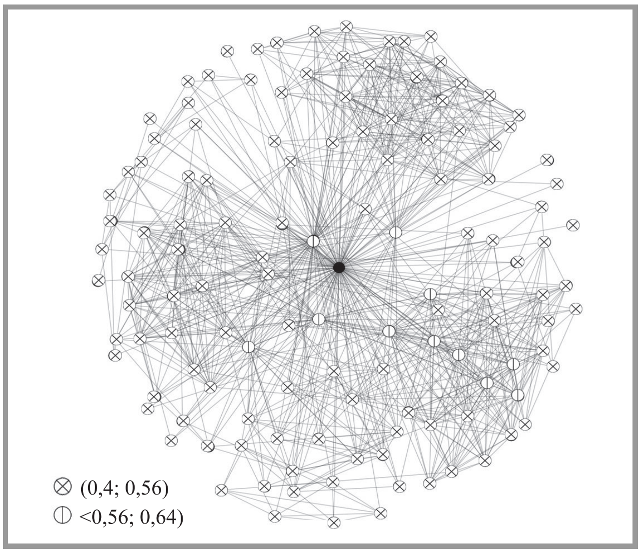


Fig. 6. Closeness centrality of nodes; the Facebook network.

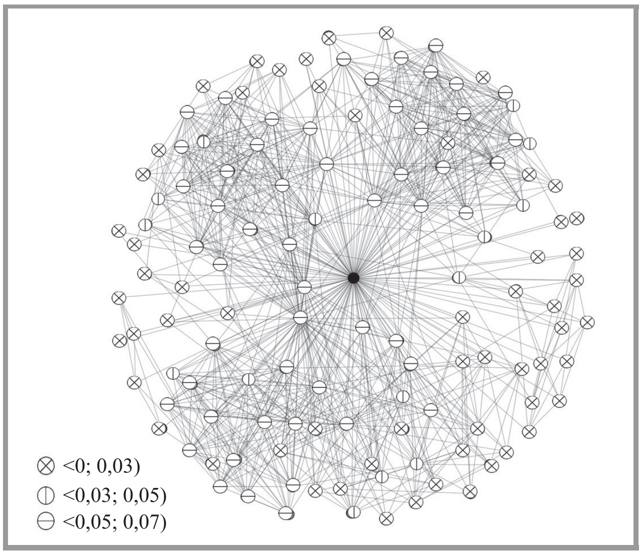


Fig. 7. Eigenvector centrality of nodes; the Facebook network.

tained the following values: the average clustering coefficient was quite high and equal to 0.7487, the average shortest path length was rather low and equal to 1.887. Such values of these measures are typical to small-world networks. Hence, the results of our experiments confirmed that our test network is a typical SN.

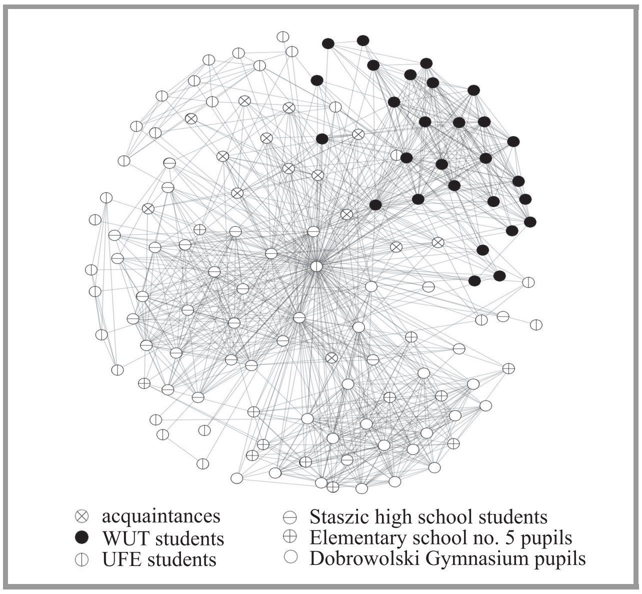

⊗ acquaintances
● WUT students
⦶ UFE students
⊖ Staszic high school students
⊕ Elementary school no. 5 pupils
○ Dobrowolski Gymnasium pupils

Fig. 8. Community detection using Blondel *et al.* algorithm; the Facebook network.

Finally, we used the Blondel *et al.* algorithm of communities detection in our test network. The results – communities extracted from the network are presented in Fig. 8. To verify the results of the experiment, we manually (based on our knowledge) detected the communities that were formed by friends of the author from different periods of his life (primary school, high school and university). After comparison of calculated and manually detected groups, we obtained the accuracy of the Blondel *et al.* algorithm equal to 68%.

5.3. Social Network from Twitter

The next series of experiments was performed for data acquired from the Twitter platform. Twitter is a social networking and microblogging service. The users of Twitter can exchange text-based posts called *tweets*. A tweet is a maximum 140 characters long but can be augmented by pictures or audio recording. The main concept of Twitter was to build a social network formed by friends and followers. Friends are people who you follow, followers are those who follow you. Hence, the person who has many followers in Twitter is recognized as an important actor in a given network. The Twitter system collects not only data about people who send tweets but also those who decide to forward these tweets to other users of Twitter. Moreover, the tweets are aggregated to speed up Twitter. Hashtag (a word included in a tweet preceded with a hash # symbol) is added to some tweets. Next, tweets with the same hashtag are aggregated into one stream. Therefore, persons who are interested in a popular topic have an easy and fast access to information concerned with this selected topic.

Similarly to the previous set of tests, we tried to extract knowledge about examined social networks. We compared two social networks formed by two different groups of users tweeting about two topics: a pop starlet Justin Bieber and July Oslo massacre on Utoya island. In the first step of our experiment we collected tweets with the hashtags #justinbierber and #oslo, and formed two groups corresponding to two hastags. Then information about senders of all collected tweets were downloaded from the Internet. Two social networks (one for each tag) were built with nodes corresponding to the senders and edges linking nodes that followed one another. The measures described in Section 3 were calculated for both networks. The computed values of an average clustering coefficient, node degree and shortest path length are presented in Table 2. From the results we can observe that people twitting about Justin form a community with higher connectivity. It seems credible because this group consists of young people – typical users of social networking platforms such as Twitter, Facebook etc. and fans of the singer. The "oslo" network was formed by loosely connected people just as a response to one event

Table 2
Twitter networks characteristics

Measure	#justinbieber	#oslo
Number of nodes	1470	519
Number of edges	7081	636
Maximal degree of node	1414	403
Avg. clustering coefficient	0.137	0.1017
Avg. node degree	9.38	2.45
Avg. shortest path length	3.06	5.95

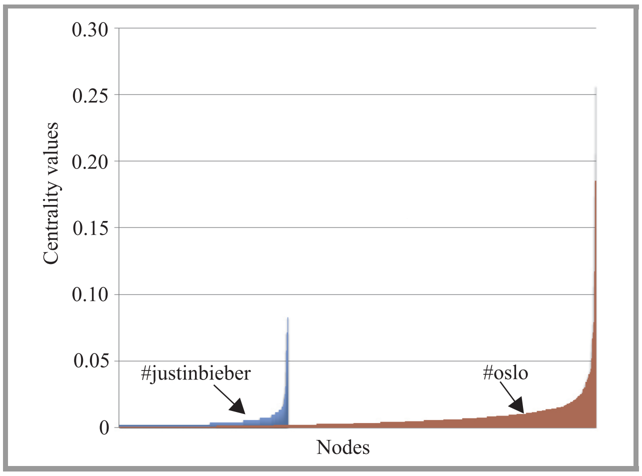


Fig. 9. Degree centrality of nodes; the Twitter network.

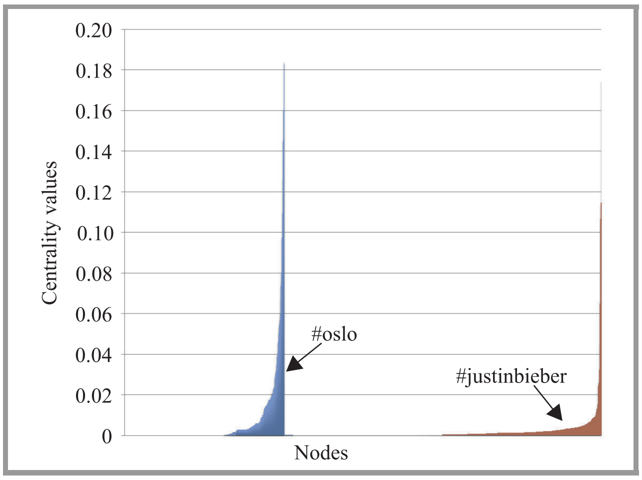


Fig. 10. Betweenness centrality of nodes; the Twitter network.

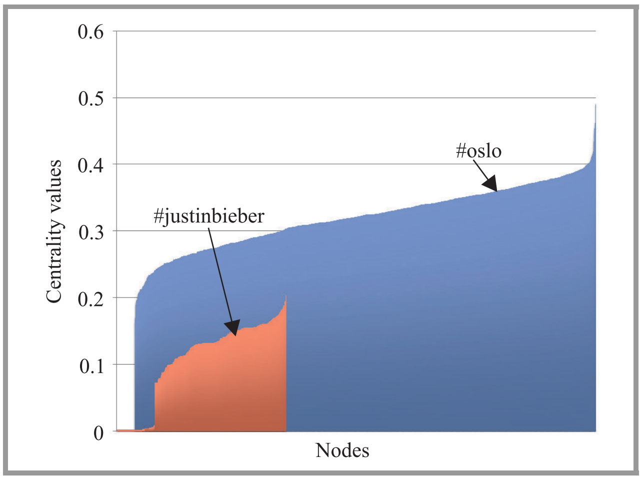


Fig. 11. Closeness centrality of nodes; the Twitter network.

that had suddenly happened. Next, we calculated the centrality measures. The results are depicted in Figs. 9–12. Finally, we checked the scale-free structure of the Twitter network. The calculated degree distribution is presented in Fig. 13, and the retweet distribution using approximately 65% of tweets in Fig. 14. Our experiments proved that the topic-based networks, as Twitter, are typical scale-free networks.

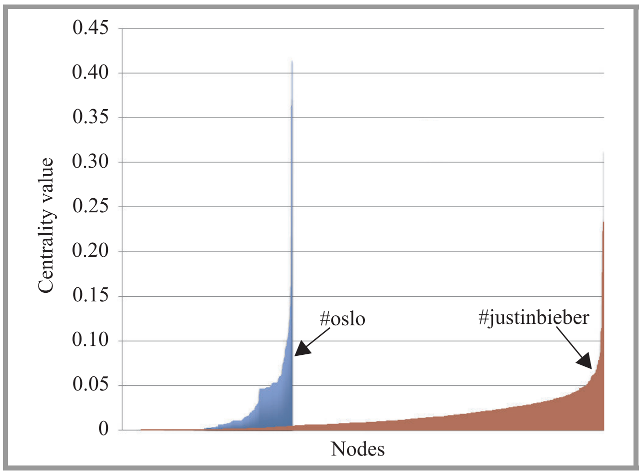


Fig. 12. Eigenvector centrality of nodes; the Twitter network.

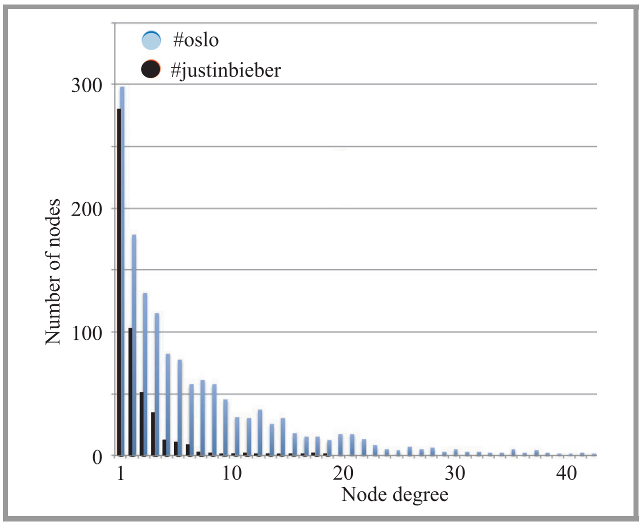


Fig. 13. Degree distribution; the Twitter network.

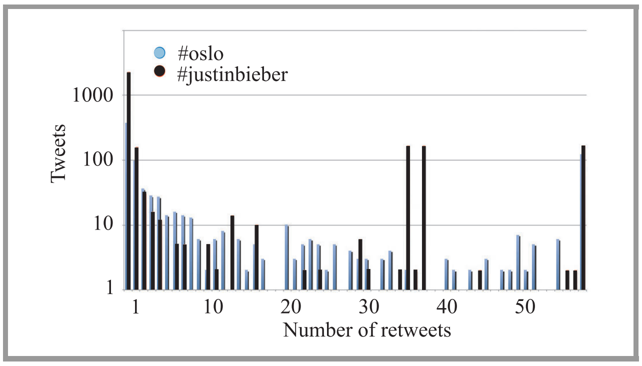


Fig. 14. Retweet distribution using approximately 65% of tweets.

The untypical result is a peak in histograms describing “oslo” network, Fig. 14. It can be caused by a variety of actors with high centrality forming this network. We can expect that participants of this network are typical leaders, i.e., news agencies or newspaper accounts, and another leaders – *information brokers* such as journalists.

5.4. Detecting Cliques of Malicious Voters

The last series of experiments was performed for data acquired from the thesixtyone.com web page. This page is owned by the record company. It presents pieces of songs done by young, rather unknown artists that try to release their first album. The company founds a recording of such an album for the group, which get the highest number of votes given by the web users. In general, democratic voting is widely adopted by many web applications. Unfortunately, in case of such a voting there is an obvious room for abuses, such as bribing the voters or using voting bots in order to get the highest number of votes. For this reason, the owner of the thesixtyone.com is interested in discovering such cliques of malicious voters.

The formulation of the problem was as follows: given the list of objects and lists of users, who voted on these objects, identify cliques of malicious voters. Every voter could vote on many objects, but on one object each voter can vote only once. The SNA techniques were employed to solve the problem. First, the social network formed by voters was generated, and then the communities detection algorithms were employed to identify groups of voters. Finally, we tried to recognize suspicious groups – cliques of malicious voters. The network was created under following assumptions.

- The network was formed by voters (network nodes). Each edge linked two voters voted on the same object.
- The weights were assigned to each edge; $weight = \frac{1}{L}$, where L denoted a number of times that connected voters voted on the same object.
- The edges with values of *weight* grater than an assumed threshold value *cut-off level* were removed from considerations (only persons who often vote on similar objects were suspected).

The Blondel algorithm was used to detect cliques. The network was divided into groups. The smallest one consisting of 106 voters (11.38% of network nodes) was recognized as a clique of malicious voters.

In order to verify the performance of the proposed method, we performed several experiments for data generated by our network simulator. The simulator applies NetworkX library. It was used to generate networks with properties similar to the thesixtyone network. Next, the list of malicious voters was generated. Using different parameters we generated networks with different properties (number of cliques, size of cliques, etc.). Multiple experiments were performed for a network formed by 500 nodes, *cut-off level*=1/3, and different input parameters. In general, the results were unsatisfactory. On average, only 5% of voters recognized as suspected persons were among real malicious voters.

The results of this experiment show the limitations of application of simple grouping techniques to social networks analysis. It is often difficult to divide actors who behave in a similar way into groups. The key issue is to define the adequate criterion or measure for the selecting procedure when strong differences between actors can not be observed. In such cases other methods of analysis applied to larger set of data should support the simple SNA techniques (see [12]). In case of our experiment we probably could reduce the number of badly classified voters considering data from not one but series of voting records.

6. Summary and Conclusion

The paper provides the short overview of social network analysis techniques. The common properties of social networks were summarized. By performing experiments for real life social networks available in two different types of popular social services Facebook and Twitter, we tried to show that SNA is a valuable tool for extracting knowledge from networks encountered in nature, especially networks formed by people. Our results confirm that both Facebook and Twitter are typical social networks, i.e., scale-free and small-world networks. It is worth mentioning that SNA techniques are based on data processing, and unfortunately, they may fail for more complex problems when network properties and available data are not enough to make a decision and solve a task.

References

[1] S. N. Dorogovtsev and J. F. F. Mendes, “Evolution of networks”, *Advances Phys.*, vol. 51, no. 4, pp. 1079–1187, 2002.

[2] W. Gruszczyński and P. Arabas, “Application of social network to improve effectiveness of classifiers in churn modelling”, in *Proc. 3rd Int. Conf. Comput. Aspects of Social Netw. CASoN’11*, Salamanca, Spain, 2011.

[3] M. Kamola, B. C. Piech, and E. Niewiadomska-Szynkiewicz, “Reconstruction of a social network graph from incomplete call detail records”, in *Proc. 3rd Int. Conf. Comput. Aspects of Social Netw. CASoN’11*, Salamanca, Spain, 2011.

[4] M. E. J. Newman, “Modularity and community structure in networks”, *Proc. Nat. Academy Sci. USA*, vol. 103, no. 23, pp. 8577–8582, 2006.

[5] M. E. J. Newman, A. L. Barabasi, and D. J. Watts, *The Structure and Dynamics of Networks*. USA: Princeton University Press, 2006.

[6] M. E. J. Newman, “Communities, modules and large-scale structure in networks”, *Nature Phys.*, vol. 8, pp. 25–31, 2011.

[7] S. Wasserman and K. Faust, *Social Network Analysis*. USA: Camridge University Press, 2009.

[8] S. Eubank, H. Guclu, V. S. A. Kumar, M. V. Marathe, A. Srinivasan, Z. Toroczkaj, and N. Wang, “Modelling disease outbreaks in realistic urban social networks”, *Nature*, vol. 429, pp. 180–183, 2004.

[9] A. D. Henry, “Belief-oriented segregation in policy networks”, *Procedia – Social Behav. Sci.*, vol. 22, pp. 14–26, 2011.

[10] B. Karrer and M. E. J. Newman, “Competing epidemics on complex networks”, *Phys. Rev.*, vol. 84, pp. 1–14, 2011.

[11] S. L. Magsino, *Applications of Social Network Analysis for Building Community Disaster Resilience*. USA: The National Academies Press, 2009.

[12] M. A. Porter, P. J. Mucha, M. E. J. Newman, and A. J. Friend, “Community structure in the united states house of representatives”, *Physica*, vol. 386, pp. 414–438, 2007.

[13] Z. Tarapata and R. Kasprzyk, Graph-based optimization method for information diffusion and attack durability in networks. *Lecture Notes Artif. Intel.*, vol. 6086, pp. 698–709, 2010.

[14] V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in lagre networks", *J. Statistical Mechanics: Theory and Experiment*, no. 10, pp. 1–12, 2008.

[15] S. P. Borgatti, "Centrality and network flow", *Social Netw.*, vol. 27, pp. 55–71, 2005.

[16] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks", *Rev. Modern Phys.*, vol. 70, pp. 66–111, 2004.

[17] A. Fronczak and P. Fronczak, *Świat sieci złożonych. Od fizyki do Internetu*. Warsaw: PWN, 2009 (in Polish).

[18] D. Watts, "Networks, dynamics and the small-world phenomenon", *The American J. of Sociol.*, vol. 105, no. 2, pp. 493–527, 1999.

[19] R. Albert and A. L. Barabási, "Statistical mechanics of complex networks", *Rev. Modern Phys.*, vol. 47, pp. 47–97, 2002.

[20] A. L. Barabási and R. Albert, "Emergence of scaling in random networks", *Science*, vol. 286, pp. 509–512, 1999.

[21] S. Boccaletti, M. Ivanchenko, V. Latora, A. Pluchino, and A. Rapisarda, "Detecting complex network modularity by dynamical clustering", *Phys. Rev.*, vol. 75, no. 4, pp. 1–4, 2007.

[22] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks", *Proc. Nat. Academy Sci. USA*, vol. 99, no. 12 pp. 7821–7826, 2002.

[23] A. J. Enright, S. Van Dongen, and C. A. Ouzounis, "An efficient algorithm for large-scale detection of protein families", *Nucleic Acid Res.*, vol. 30, no. 7, pp. 1575–1584, 2002.

Marcin Mincer received his B.Sc. in Computer Science from the Warsaw University of Technology, Poland, in 2011. Currently he is a M.Sc. student in the Institute of Control and Computation Engineering at the Warsaw University of Technology. Since 2011 he is involved in the ECONET project of EU. His research area focuses on social network analysis applied on emerging Internet social media platforms.

E-mail: M.Mincer@stud.elka.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Ewa Niewiadomska-Szynkiewicz, D.Sc. (2005), Ph.D. (1995), M.Eng., Professor of Control and Information Engineering at the Warsaw University of Technology, head of the Complex Systems Group. She is also the Director for Research of Research and Academic Computer Network (NASK). She is the author and co-author of three books and over 120 journal and conference papers. Her research interests focus on complex systems modeling and control, computer simulation, global optimization, parallel computation, computer networks and ad hoc networks. She was involved in a number of research projects including EU projects, coordinated the Groups activities, managed organization of a number of national-level and international conferences.

E-mail: ens@ia.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

E-mail: ewan@nask.pl
Research and Academic Computer Network (NASK)
Wąwozowa st 18
02-796 Warsaw, Poland

Paper

Effective Design of the Simulated Annealing Algorithm for the Flowshop Problem with Minimum Makespan Criterion

Jarosław Hurkała and Adam Hurkała

Institute of Control and Computation Engineering, Warsaw University of Technology, Warsaw, Poland

Abstract—In this paper we address the n-job, m-machine flowshop scheduling problem with minimum completion time (makespan) as the performance criterion. We describe an efficient design of the Simulated Annealing algorithm for solving approximately this NP-hard problem. The main difficulty in implementing the algorithm is no apparent analogy for the temperature as a parameter in the flowshop combinatorial problem. Moreover, the quality of solutions is dependent on the choice of cooling scheme, initial temperature, number of iterations, and the temperature decrease rate at each step as the annealing proceeds. We propose how to choose the values of all the aforementioned parameters, as well as the Boltzmann factor for the Metropolis scheme. Three perturbation techniques are tested and their impact on the solutions quality is analyzed. We also compare a heuristic and randomly generated solutions as initial seeds to the annealing optimization process. Computational experiments indicate that the proposed design provides very good results – the quality of solutions of the Simulated Annealing algorithm is favorably compared with two different heuristics.

***Keywords*—flowshop, heuristics, makespan, simulated annealing.**

1. Introduction

The flowshop problem has been studied by many researches because of the educational character and many real-life applications. Various optimization techniques with different assumptions have been used to solve this problem. The regular flowshop problem consists of a group of m machines and a set of n jobs to be processed on these machines. Each job is processed one at a time and only once on each machine (preemption is not allowed). A job cannot be processed simultaneously on more than one machine. The same processing order of jobs applies to each of the m machines.

In this paper, we consider the classical flowshop-sequencing problem with minimum completion time (makespan) and assume infinite buffer at any machine in the processing sequence, so that jobs may form queues and wait between the machines without blocking them. The makespan criterion can be defined as a completion time, at which all jobs complete processing or equivalently as a maximum completion time of jobs. The flowshop scheduling problem with the makespan criterion is indicated by $n/m/F/C_{max}$ and the aim is to find the order of jobs that minimizes the makespan.

The n-job m-machine flowshop problem belongs to the class of NP-hard problems [1]. Because the search space grows exponentially as the number of jobs increases, obtaining the optimal solutions for large-size problems with exact methods in reasonable time is impossible. As a consequence, many researchers have been developing various heuristics for this problem. These include constructive heuristics [2], [3], metaheuristics like Simulated Annealing [4]–[6] and Tabu Search [7]–[10], evolutionary algorithms, such as Genetic Algorithm [11], [12], and other neighbor search approaches [13].

In this paper, we comprehensively describe the design of the Simulated Annealing algorithm for the purpose of effectively solving the flowshop problem. In Section 2, we present the objective function and propose an alternative approach for calculating the makespan. In Sections 3 and 4, we explain in details the design of the Simulated Annealing algorithm for the flowshop problem and briefly describe two constructive heuristics that we compare the outcomes with. The results of the experiments are shown in Section 5. In Section 6 some concluding remarks are presented.

2. Problem Definition

The classical flowshop problem, we focus on in this paper, can be defined as follows, using the notation by Nowicki, Smutnicki [10], Grabowski, Pempera [8] and Wodecki, Bożejko [6]. We consider a set of n jobs $J = \{1, 2, \ldots, n\}$, and a set o m machines $M = \{1, 2, \ldots, m\}$. Job $j \in J$, consists of a sequence of operations $O_{j1}, O_{j2}, \ldots, O_{jm}$, where operation O_{jk} corresponds to the processing of job j on machine k and takes p_{jk} time. The goal is to find the sequence of jobs that minimizes the completion time of all jobs.

Let $\pi = (\pi(1), \pi(2), \ldots, \pi(n))$ be a permutation of jobs and Π be the set of all permutations. Each permutation $\pi \in \Pi$ defines a processing order of jobs on each machine. We want to find a permutation $\pi^* \in \Pi$ such that:

$$C_{max}(\pi^*) = \min_{\pi} C_{max}(\pi), \tag{1}$$

where $C_{max}\ (\pi)$ is the makespan of the processing order given by π, and can be found by the following recursive formula:

$$C_{jk}(\pi) = \max\left\{C_{\pi(j-1)k},\ C_{\pi(j)k-1}\right\} + p_{\pi(j)k}, \tag{2}$$

$$C_{max}(\pi) = C_{nm}(\pi), \tag{3}$$

where $\pi(0) = 0,\ C_{0k} = 0,\ k = 1,2,\ldots,m,\ C_{j0} = 0,\ j = 1,2,\ldots,n$.

It is also well known in the literature that the makespan associated with permutation π can be found by:

$$C_{max}(\pi) = \max_{1 \le t_1 \le \cdots \le t_{m-1} \le n} \left(\sum_{j=1}^{t_1} p_{\pi(j)1} + \cdots + \sum_{j=t_{m-1}}^{n} p_{\pi(j)m} \right). \tag{4}$$

This optimization problem can be considered as finding the longest (critical) path from node (1, 1) to (m, n) in a grid graph. Each path in such graph is composed of horizontal and vertical sub-paths.

In this paper, we propose another approach of calculating the makespan. Instead of using the recursive formula or solving the longest path problem, we have created an algorithm that simulates the flowshop, hence finding the maximum completion time of jobs. For the overview of the algorithm see Algorithm 1.

Algorithm 1 Flowshop simulation

```
$C_{max} \leftarrow 0,\ \delta t \leftarrow 0,\ q_1 \leftarrow \pi$
repeat
  $C_{max} \leftarrow C_{max} + \delta t$
  $\delta t \leftarrow \infty$
  for $k$ from $m$ downto 1 do
    if $r_k > 0$ then
      $r_k \leftarrow r_k - \delta t$
      if $r_k = 0$ then
        if $k+1 \le m$ then
          $add(q_{k+1}, c_k)$
        end if
        $c_k \leftarrow \emptyset,\ \delta t \leftarrow 0$
      else
        $\delta t \leftarrow \min\{\delta t, r_k\}$
      end if
    else if $q_k \neq \emptyset$ then
      $c_k \leftarrow removeFirst(q_k)$
      $r_k \leftarrow p_{c_k k}$
      $\delta t \leftarrow \min\{\delta t, r_k\}$
    end if
  end for
until $\delta t = \infty$
return $C_{max}$
```

The design of the algorithm is fairly simple. Let $c_k^i \in J \cup \{\emptyset\}$ be the job processed on machine k in iteration i, r_k^i be the remaining time of processing this job on machine k in iteration i, and $q_k^i \subset J \cup \{\emptyset\}$ be the job queue at machine k in iteration i. The makespan can be calculated from the following formula:

$$C_{max}(\pi) = \sum_i \delta t^i, \tag{5}$$

where δt^i is the time step by which we increase the makespan in iteration i:

$$\delta t^i = \min_{k=1,\ 2,\ldots,m} r_k^i. \tag{6}$$

The remaining time of processing on machine k in iteration $i+1$ is calculated as follows:

$$r_k^{i+1} = \begin{cases} r_k^i - \delta t^i & \text{if } r_k^i > 0 \\ p_{c_k^{i+1}k} & \text{if } r_k^i = 0 \wedge q_k^i \neq \{\emptyset\} \\ 0 & \text{otherwise} \end{cases}, \tag{7}$$

where $\delta t^0 = 0, r_k^0 = 0,\ c_k^0 = \{\emptyset\}, k = 1,2,\ldots,m$.

The job processed on machine k in iteration $i+1$ is found by:

$$c_k^{i+1} = \begin{cases} c_k^i & \text{if } r_k^i > 0 \wedge r_k^i - \delta t^i > 0 \\ q_k^i(1) & \text{if } r_k^i = 0 \wedge q_k^i \neq \{\emptyset\} \\ \{\emptyset\} & \text{otherwise} \end{cases}, \tag{8}$$

where $q_k^i(1)$ is the first element in the job queue, $k = 1,2,\ldots,m$.

The queue at machine k in iteration $i+1$ is calculated as follows:

$$q_k^{i+1} = \begin{cases} q_k^i \cup \left\{c_{k-1}^i\right\} & \text{if } r_{k-1}^i > 0 \wedge r_{k-1}^i - \delta t^i = 0 \\ q_k^i - \{q_k^i(1)\} & \text{if } r_k^i = 0 \wedge q_k^i \neq \{\emptyset\} \\ q_k^i & \text{otherwise} \end{cases}, \tag{9}$$

where $q_1^0 = \pi, q_{k+1}^0 = \{\emptyset\}, k = 1,2,\ldots,m-1$.

3. Simulated Annealing

The Simulated Annealing (SA) was first introduced by Kirkpatrick [14], while Černý [15] pointed out the analogy between the annealing process of solids and solving combinatorial problems. Researchers have been studying the application of the SA algorithm in various fields of optimization problems, but more importantly, it was shown that SA can be applied to sequencing problems [16].

The process of Simulated Annealing can be described as follows. First, an initial solution must be specified as a starting point. Then, repeatedly, a candidate solution is randomly chosen from the neighborhood of the current solution. If the newly generated solution is better than the current one, it is accepted and becomes the new current solution. Otherwise, it still has a chance to be accepted with, so called, acceptance probability. This probability is determined by the difference between objective function of the current and the candidate solution, and depends on a control parameter, called temperature, taken from the thermodynamics. After a number of iterations the temperature is decreased and the process continues as described above. The annealing process is stopped either after a maximum number of iterations or when a minimum temperature is reached. The best solution that is found during the process is considered a final. For the algorithm overview see Algorithm 2.

Algorithm 2 Simulated Annealing

```
Require: Initial schedule $\pi_0$
$\pi^* \leftarrow \pi_0$
for $i$ from 1 to $N$ do
    for $t$ from 1 to $N_{temp}$ do
        $\pi \leftarrow perturbate(\pi_0)$
        $\delta \leftarrow C_{max}(\pi) - C_{max}(\pi_0)$
        if $\delta < 0$ or $e^{-\delta/k\tau} > random(0,1)$ then
            $\pi_0 \leftarrow \pi$
        end if
        if $C_{max}(\pi) < C_{max}(\pi^*)$ then
            $\pi^* \leftarrow \pi$
        end if
    end for
    $\tau \leftarrow \tau * \alpha$
end for
return $\pi^*$
```

In order to solve the flowshop problem with the SA algorithm, the annealing process needs to be adapted to this particular problem and values of several parameters must be determined.

The main step of the SA is the procedure of generating a candidate solution from the neighborhood of the current one, which is often called a perturbation scheme or transition operation. Although there are many ways to accomplish this task, we have examined the three most popular techniques:

- Interchanging two adjacent jobs.
- Interchanging two jobs.
- Moving a single job.

The key element of SA is to define the temperature decrease schedule, also called the cooling scheme. The main issue at this point is to determine values for the following parameters:

- initial temperature,
- function of temperature decrease in consecutive iterations,
- the number of iterations at each temperature (Metropolis equilibrium),
- minimum temperature at which the algorithm terminates or alternatively the maximum number of iterations as the stopping criterion.

The cooling process is usually simulated by decreasing the temperature by a factor, called the reduce factor. Let τ be the temperature and α be the reduce factor. Then the annealing scheme can be represented as the following recursive function:

$$\tau^{i+1} = \alpha * \tau^i, \tag{10}$$

where i is the number of current iteration in which the cooling schedule takes place.

Another building block of SA that has to be customized is the acceptance probability function, which determines whether to accept or reject candidate solution that is worse than the current one. The most widely used function is:

$$p(\delta, \tau) = e^{-\delta/k\tau}, \tag{11}$$

where δ is the difference between the objective function of the candidate (π) and the current solution (π_0):

$$\delta = C_{max}(\pi) - C_{max}(\pi_0), \tag{12}$$

and k is the Boltzmann constant found by:

$$k = \frac{\delta^0}{\log \frac{p^0}{\tau^0}}, \tag{13}$$

where δ^0 is an estimated minimal difference between objective function of two solutions, p^0 is the initial value of the acceptance probability and τ^0 is the initial temperature. Notice that we use decimal logarithm rather than natural, which is most widely seen in the literature. Moreover, rather than average, we use estimation of the minimal difference between solutions.

After thorough analysis of the SA application for the flowshop problem, we have arrived at the following initial values of all the aforementioned parameters that should be used to achieve the best results and make the most of the Simulated Annealing algorithm – see Table 1.

Table 1
Initial values of Simulated Annealing parameters

Param.	Description	Value
α	Reduce factor	$1-\frac{7}{N}$
τ^0	Initial temperature	0.99
δ^0	Estimated minimal difference between solutions	1
p^0	Initial value of acceptance probability	1
k	Boltzmann constant	$1/\log\left(\frac{1}{0.99}\right)$
N_{temp}	Number of iterations at each temperature	10
N	Number of SA iterations	1000000

4. Heuristic Algorithms

4.1. CDS Algorithm

The flowshop problem with two machines and the makespan criterion ($n/2/F/C_{max}$) can be solved by applying the famous Johnson's optimal rule, saying that job i precedes job j in an optimal sequence if:

$$\min\{p_{i1}, p_{j2}\} \leq \min\{p_{i2}, p_{j1}\}. \tag{14}$$

The Johnson's algorithm implementing this rule can be described in the following four steps:

1. Let $U = \{j : p_{j1} < p_{j2}\}$ and $V = \{j : p_{j1} \geq p_{j2}\}$.
2. Sort U in non-descending order by p_{j1}.
3. Sort V in non-ascending order by p_{j2}.
4. Set $\pi^* = U \cup V$ is the optimal job sequence.

Many researchers have tried to extend the rule for larger problems with more machines. An algorithm named CDS was proposed in [2] for the flowshop problem with makespan performance criterion, that effectively solves instances with any number of machines.

The algorithm is based on a heuristic application of the Johnson's rule to a two-machine sub-problem, obtained by merging machines to artificial machine centers.

The CDS algorithm creates $m-1$ two-machine sub-problems:

$$p^*_{j1} = \sum_{k=1}^{i} p_{jk}, \tag{15}$$

$$p^*_{j2} = \sum_{k=i+1}^{m} p_{jk}, \tag{16}$$

where i is the number of sub-problem, $j = 1, 2, \ldots, n$. Each sub-problem is solved with the Johnson's algorithm and one of the obtained $m-1$ sequences with the lowest makespan becomes the final solution of the main m-machine problem.

4.2. NEH Algorithm

Another approach for solving the flowshop problem is to construct the schedule by adding one job at a time to the sequence of jobs instead of calculating the makespan for the entire sequence of jobs at once. An excellent example of such algorithm is the NEH heuristic proposed in [3], which is considered the best constructive heuristic for the makespan flowshop problem.

This heuristic method is based on the assumption that in the process of constructing the schedule a job with higher value of total processing time on all machines should have higher priority and be taken into consideration before other jobs.

The algorithm consists of the following four steps:

1. Sort jobs in non-ascending order of total processing time on all machines.
2. Take the first of the remaining (unscheduled) jobs.
3. Find a position of the job in the partial sequence that minimizes makespan of the extended by this job partial sequence.
4. If there are more unscheduled jobs, go to Step 2.

At each iteration there are k possible places, at which a job can be inserted, where k is the iteration number. At the last iteration, the best partial sequence extended by the remaining job is the final schedule and the solution of the makespan problem.

5. Results

The design of the Simulated Annealing algorithm has been tested on a subset of the collection of flowshop problems developed by Taillard [17]. We have selected following different problem sizes: $n = \{20, 50, 100\} \times m = \{5, 10, 20\}$, and chosen first 4 instances of each of the 9 problem classes, which gave us a total of 36 instances. Each instance was solved 20 times and the best result was taken as final.

The difference between the algorithms was calculated by the following formula:

$$\eta(x,y) = \frac{x-y}{y}. \tag{17}$$

For the small-size problems (instances with 20 jobs or 5 machines) the SA algorithm is beyond compare. For large-size problems (50×10, 50×20, 100×10, 100×20) it outperforms the CDS algorithm on average by 13% and is better than the NEH algorithm by more than 4% (see Tables 2–4 for more detailed results).

Table 2
Average results of the CDS algorithm

$\eta(CDS,T)$ [%]		m = 5	10	20	Avg
n	20	7.48	14.96	11.73	11.39
	50	6.65	14.78	15.92	12.45
	100	5.22	10.21	14.44	9.96
				Total average:	11.27

Table 3
Average results of the NEH algorithm

$\eta(NEH,T)$ [%]		m = 5	10	20	Avg
n	20	2.69	4.75	3.19	3.54
	50	0.62	5.04	6.59	4.09
	100	0.76	2.11	5.36	2.74
				Total average:	3.46

Table 4
Average results of the SA algorithm

$\eta(SA,T)$ [%]		m = 5	10	20	Avg
n	20	0.00	0.00	0.00	0.00
	50	0.00	0.57	0.83	0.47
	100	0.02	0.16	0.99	0.39
				Total average:	0.29

While the SA algorithm is superior in terms of solution quality, it requires more time to compute the results than heuristic algorithms like CDS or NEH. Nevertheless, on a 3.1 GHz CPU it has found all the solutions in a reasonable time – ranging from less than 30 s for 20×5 instances to about 6 minutes for 100×20 instances (see Table 5).

Table 5
Average solution times of the SA algorithm

$SA_{Time}[s]$		m = 5	10	20
n	20	28	65	128
	50	43	105	208
	100	72	178	356

In order of brevity, we present the results obtained only with ‘move a single job’ permutation scheme, since it generally finds better solutions than the other two presented techniques. This can be explained by the fact that this particular scheme generates the largest neighborhood of the current solution and it requires only $O(n)$ operations to move from the current to any permutation (transition path length). The main reason this technique outperforms the other two, is that, it changes position not only of a pair jobs, but can also change position of every job in the entire sequence in just one execution. See Table 6 for comparison of permutation schemes.

Table 6
Comparison of the perturbation schemes

Permutation scheme	Neighbor-hood size	Transition path length	Number of positions changed
Interchanging two adjacent jobs	$n-1$	$O(n^2)$	2
Interchanging two jobs	$\frac{n(n-1)}{2}$	$O(n)$	2
Moving a single job	$(n-1)^2$	$O(n)$	$O(n)$

The Simulated Annealing algorithm has found 21 optimal solutions: 11 for 5-machine instances, 6 for 10-machine instances and 4 for 20-machine instances. The best result of the NEH algorithm is the solution of instance #12 – only 0.18% worse than optimal, while for the CDS algorithm it is 0.66% (instance #2). On the other hand, in the worst case of the SA algorithm, the makespan of instance #17 is only 1.14% higher than optimal, while for NEH it is 7.88% (instance #31) and for CDS it is 19.59% (instance #14). See Table 7 for detailed results of all the flowshop problem instances.
We have tested three types of starting conditions of the SA optimization process:

- Initial permutation is chosen at random.
- Solution generated by the CDS algorithm is taken as the initial permutation.
- Solution generated by the NEH algorithm is taken as the initial permutation.

As the SA algorithm finds solutions equal or better than both the CDS and the NEH algorithms approximately after half of the cycle or earlier, the initial permutation has little impact on the final solution. This property, however, makes the proposed design of SA algorithm self-sufficient.

6. Conclusions

We have presented an effective design of the Simulated Annealing algorithm for the flowshop problem with minimum makespan criterion and have shown that it outperforms both the CDS and NEH heuristics in terms of solution quality. Even though SA is not the fastest heuristic

Table 7

Detailed results of the Taillard flowshop problem instances

#	$n \times m$	Results				$\eta(SA,ET)$	$\eta(NEH,ET)$	$\eta(CDS,ET)$	$\eta(NEH,SA)$	$\eta(CDS,SA)$
		Taillard	SA	NEH	CDS	[%]	[%]	[%]	[%]	[%]
1	20×5	1278	1278	1286	1334	0.00	0.63	4.38	0.63	4.38
2	20×5	1359	1359	1365	1368	0.00	0.44	0.66	0.44	0.66
3	20×5	1081	1081	1159	1253	0.00	7.22	15.91	7.22	15.91
4	20×5	1293	1293	1325	1409	0.00	2.47	8.97	2.47	8.97
5	50×5	2724	2724	2733	2934	0.00	0.33	7.71	0.33	7.71
6	50×5	2834	2834	2843	3020	0.00	0.32	6.56	0.32	6.56
7	50×5	2621	2621	2640	2856	0.00	0.72	8.97	0.72	8.97
8	50×5	2751	2751	2782	2843	0.00	1.13	3.34	1.13	3.34
9	100×5	5493	5493	5519	5901	0.00	0.47	7.43	0.47	7.43
10	100×5	5268	5268	5348	5466	0.00	1.52	3.76	1.52	3.76
11	100×5	5175	5175	5219	5378	0.00	0.85	3.92	0.85	3.92
12	100×5	5014	5018	5023	5303	0.08	0.18	5.76	0.10	5.68
13	20×10	1582	1582	1680	1771	0.00	6.19	11.95	6.19	11.95
14	20×10	1659	1659	1729	1984	0.00	4.22	19.59	4.22	19.59
15	20×10	1496	1496	1557	1735	0.00	4.08	15.98	4.08	15.98
16	20×10	1377	1377	1439	1547	0.00	4.50	12.35	4.50	12.35
17	50×10	2991	3025	3135	3386	1.14	4.81	13.21	3.64	11.93
18	50×10	2867	2887	3032	3306	0.70	5.76	15.31	5.02	14.51
19	50×10	2839	2852	2986	3243	0.46	5.18	14.23	4.70	13.71
20	50×10	3063	3063	3198	3565	0.00	4.41	16.39	4.41	16.39
21	100×10	5770	5770	5846	6255	0.00	1.32	8.41	1.32	8.41
22	100×10	5349	5352	5453	6004	0.06	1.94	12.25	1.89	12.18
23	100×10	5676	5679	5824	6155	0.05	2.61	8.44	2.55	8.38
24	100×10	5781	5812	5929	6461	0.54	2.56	11.76	2.01	11.17
25	20×20	2297	2297	2410	2587	0.00	4.92	12.63	4.92	12.63
26	20×20	2099	2099	2150	2351	0.00	2.43	12.01	2.43	12.01
27	20×20	2326	2326	2411	2565	0.00	3.65	10.28	3.65	10.28
28	20×20	2223	2223	2262	2490	0.00	1.75	12.01	1.75	12.01
29	50×20	3850	3893	4082	4424	1.12	6.03	14.91	4.85	13.64
30	50×20	3704	3722	3921	4260	0.49	5.86	15.01	5.35	14.45
31	50×20	3640	3666	3927	4204	0.71	7.88	15.49	7.12	14.68
32	50×20	3723	3760	3969	4403	0.99	6.61	18.26	5.56	17.10
33	100×20	6202	6271	6541	7263	1.11	5.47	17.11	4.31	15.82
34	100×20	6183	6239	6523	7064	0.91	5.50	14.25	4.55	13.22
35	100×20	6271	6338	6639	7193	1.07	5.87	14.70	4.75	13.49
36	100×20	6269	6323	6557	7002	0.86	4.59	11.69	3.70	10.74

algorithm, the computation time on modern computers is acceptable. Furthermore, the design proposed in this paper is similar to the general design and can be easily adapted and used to solve other combinatorial problems (by just changing the value of estimated minimal difference between solutions).

References

[1] M. R. Garey, "The complexity of flowshop and jobshop scheduling", *Math. Oper. Res.*, vol. 1, no. 2, pp. 117–129, 1976.

[2] H. G. Campbell, R. A. Dudek and M. L. Smith, "A heuristic algorithm of the *n*-job, *m*-machine sequencing problem", *Manag. Sci.*, vol. 16, pp. 630–637, 1970.

[3] M. Nawaz, E. Enscore Jr, and I. Ham, "A heuristic algorithm for the m-machine, n-job flowshop sequencing problem", *OMEGA Int. J. Manag. Sci.*, vol. 11, pp. 91–95, 1983.

[4] F. A. Ogbu and D. K. Smith, "The application of the simulated annealing algorithm to the solution of the $n/m/Cmax$ flowshop problem", *Comput. Oper. Res.*, vol. 17, no. 3, pp. 243–253, 1990.

[5] J. Hurkała and A. Hurkała, "Effective design of the simulated annealing algorithm for the flowshop problem with minimum makespan criterion", in *9th Int. Conf. Decision Support Telecomm. Inform. Society DSTIS 2011* , Warsaw, Poland, 2011.

[6] M. Wodecki and W. Bożejko, "Solving the flow shop problem by parallel simulated annealing", *LNCS*, vol. 2328, pp. 597–600, 2006.

[7] E. Taillard, "Some efficient heuristic methods for flow shop sequencing", *Eur. J. Oper. Res.*, vol. 47, pp. 65–74, 1990.

[8] J. Grabowski and J. Pempera, "New block properties for the permutation flow-shop problem with application in TS", *J. Oper. Res. Soc.*, vol. 52, pp. 210–220, 2001.

[9] E. Nowicki, "The permutation flow shop with buffers: a tabu search approach", *Eur. J. Oper. Res.*, vol. 116, pp. 205–219, 1999.

[10] E. Nowicki and C. Smutnicki, "A fast tabu search algorithm for the permutation flowshop problem", *Eur. J. Oper. Res.*, vol. 91, pp. 160–175, 1996.

[11] C. R. Reeves, "A genetic algorithm for flowshop sequencing". *Comput. Oper. Res.*, vol. 22, pp. 5–13, 1995.

[12] C. R. Reeves and T. Yamada, "Genetic algorithms, path relinking, and the flowshop sequencing problem", *Evol. Comput.*, vol. 6, no. 1, pp. 230–234, 1998.

[13] S. R. Hejazi and S. Saghafian, "Flowshop-scheduling problems with makespan criterion: a review", *Int. J. Prod. Res.*, vol. 43, no. 14, pp. 2895–2929, 2005.

[14] S. Kirkpatrick, C. D. Gellat and M. P. Vecchi, "Optimization by simulated annealing", *Science*, vol. 220, pp. 671–680, 1983.

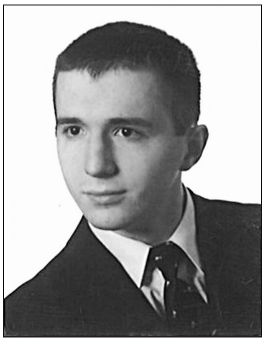
[15] V. Černý, "Thermodynamical approach to travelling salesman problem: An efficient simulation algorithm". *J. Optim. Theory Appl.*, vol. 45, pp. 41–51, 1985.

[16] C. Koulamas, S. R. Antony, and R. Jaen, "A survey of simulated annealing applications to operations research problems", *Omega*, vol. 22, no. 1, pp. 41–56, 1994.

[17] E. Taillard, "Benchmarks for basic scheduling problems", *Eur. J. Oper. Res.*, vol. 64, pp. 278–285, 1993.

Jarosław Hurkała received his M.Sc. degree in Computer Science with honors from the Warsaw University of Technology, Poland, in 2010. Currently, he is a Ph.D. student in the Institute of Control and Computation Engineering at the Warsaw University of Technology. His research area focuses on scheduling problems, heuristic algorithms, fairness and multicriteria optimization.

E-mail: j.hurkala@elka.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland



Adam Hurkała received his M.Sc. degree in Computer Science with honors from the Warsaw University of Technology, Poland, in 2010. Currently, he is a Ph.D. student in the Institute of Control and Computation Engineering at the Warsaw University of Technology. His research area focuses on information security and cryptography.

E-mail: a.hurkala@elka.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Diffusion in Networks

Rafał Kasprzyk

Faculty of Cybernetics, Military University of Technology, Warsaw, Poland

Abstract—In this paper a concept of method and its application examining a dynamic of diffusion processes in networks is considered. Presented method was used as a core framework for system CARE (Creative Application to Remedy Epidemics).

Keywords—complex networks, diffusion, probabilistic finite-state machine.

1. Introduction

Diffusion is a process, by which information, viruses, gossips and any other behaviors spread over networks [1]–[5], in particular, over social networks.

The standard approach is a simplified assumption that behaviors (information, viruses, gossips) spread in the environment, which is modeled, using very simple construction of *Regular Graphs* like GRID-based graph or similar, very rarely *Random Graphs*. Standard approaches do not explain the real dynamic of diffusion in real-world networks, in particular:

- why even slightly infectious behavior (e.g., contagious diseases) can spread over a network for a long time;
- how to choose nodes to maximize or minimize diffusion range (e.g., how to choose individuals to vaccinate, in order to minimize the epidemic's range);
- what is the mechanism of arising secondary behaviors centres.

The drawbacks of the standards diffusion models is that they do not take into account an underling real-world networks topology. Who (or what) is connected to whom (what), seems to be a fundament question. Apparently, networks derived from data on real life cases (most often: networks growing spontaneously) are neither *Regular Graphs* nor *Random* ones. As it turned out, real networks, which have been intensively studied recently have some interesting features. These features, which origins are nowadays discovered, modeled [6]–[11] and examined [12]–[15] significantly affect dynamics of the diffusion processes within real-world networks. Three very interesting models of real-world networks which have been introduced recently, e.g., *Random Graphs*, *Small World* and *Scale Free*, will be described later in this paper.

We have to also remember that all kinds of behavior spreading over the network have their unique properties, and we should be able to model them. The notion of a state machine seems to be useful in this modeling situation. Using probabilistic finite-state machines [16], [17] we can model a spreading of vast variety of behaviors. For example, we are able to build models of diseases with any states (e.g., *susceptible*, *infected*, *carrier*, *immunized*, *dead*, etc.), and probabilities of transitions from one state to another, resulting from social interactions (contacts). Again, the underling contacts (social network topology) seem to have a huge impact on the dynamic of diffusion processes, what has been already mentioned.

2. Definitions and Notations

Let's define network as follows:

$$Net(t) = \Big\langle G(t) = \langle V(t), E(t)\rangle, \{f_i(v,t)\}_{\substack{i\in\{1,\ldots,NF\},\\ v\in V(t)}}, \\ \{h_j(e,t)\}_{\substack{j\in\{1,\ldots,NH\}\\ e\in E(t)}} \Big\rangle,$$

where:

$G(t) = \langle V(t), E(t)\rangle$ – simple dynamic graph, $V(t), E(t)$ – sets of graph's vertices and edges, $E(t) \subset \big\{\{v, v'\} : v, v' \in V(t)\big\}$ (the dynamic [18] means that $V(t)$ and $E(t)$ can change over time);

$f_i : V(t) \to Val_i$ – the i-th function describe on the graph's vertices, $i = 1, \ldots NF$, (NF – number of vertex's functions), Val_i – is a set of f_i values;

$f_j : E(t) \to Val_j$ – the j-th function describe on the graph's edges, $j = 1, \ldots NH$, (NH – number of edge's functions), Val_j – is a set of h_j values.

We assume that values of function's ($f_i(\cdot)$ and $h_j(\cdot)$ can also change over time.

In this paper we were particularly interested in relationship between the structure of real-world networks and the dynamic of any behaviors on them. Due to this fact, we focused on the characteristics of the graph $G(t)$, while functions on the graph's vertices (nodes) and edges (links) were omitted.

Simple dynamic graphs are very often represented by a matrix $A(t)$, called adjacency matrix, which is a $V(t) \times V(t)$ symmetric matrix. The element $a_{ij}(t)$ of adjacency matrix equals 1 if there is an edge between vertices i and j, and 0 otherwise.

The first-neighborhood of a vertex v_i denote as $\Gamma_i^1(t)$ is defined as set of vertices immediately connected with v_i, i.e.,

$$\Gamma_i^1(t) = \big\{v_j \in V(t) : \{v_i, v_j\} \in E(t)\big\}.$$

The degree $k_i(t)$ of a vertex v_i is the number of vertices in the first-neighborhood of a vertex v_i, i.e.,

$$k_i(t) = \left|\Gamma_i^1(t)\right|.$$

The path starting in vertex v_i and ending in vertex v_j is a sequence of $\langle v_0, v_1, \ldots, v_{k-1}, v_k \rangle$, where $\{v_{i-1}, v_i\} \in E(t) \forall\, i = 1, \ldots, k$. The length of a path is defined as the number of links in it. The shortest path length starting in vertex v_i and ending in vertex v_j is denoted as $d_{ij}(t)$.

Now we can define diameter D as the longest shortest path, i.e.,

$$D(t) = \max_{v_i, v_j \in V(t)} \left\{ d_{ij}(t) \right\}.$$

Let's denote the number of existing edges between the first-neighborhood of a vertex v_i as $N_i(t)$, i.e.,

$$N_i(t) = \left| \{v_l, v_k\} : v_l, v_k \in \Gamma_i^1(t) \wedge \{v_l, v_k\} \in E(t) \right|.$$

Now, we can define a very important concept, called as the local clustering coefficient C_i for a vertex v_i, which is then be given by the proportion of $N_i(t)$ and divided by the number of edges that could possible exist between first-neighborhood of a vertex v_i (every neighbor of v_i is connected to every other neighbor of v_i). Formally:

$$C_i(t) = \begin{cases} \dfrac{2N_i(t)}{k_i(t)\big(k_i(t)-1\big)}, & \left|\Gamma_i^1(t)\right| > 1 \\ 0, & \left|\Gamma_i^1(t)\right| \leq 1. \end{cases}$$

The clustering coefficient C for the whole network is define as the average of C_i overall $v_i \in V$, i.e.,

$$C(t) = \frac{1}{|V(t)|} \sum_{v_i \in V(t)} C_i(t).$$

The degree distribution $P(k,t)$ of a network is defined as the fraction of nodes in the network with degree k. Formally:

$$P(k,t) = \frac{|V_k(t)|}{|V(t)|},$$

where: $|V_k(t)|$ is the number of nodes with degree k; $|V(t)|$ is the total number of nodes.

2.1. Models of Real-World Networks

Most of the real-world networks are found to have: small average path length, relatively small diameter, high clustering coefficient, and degree distributions that approximately follow a power law, i.e., $P(k,t) \sim k^{-\gamma}$, where γ is a constant. These features, which origins are nowadays discovered indeed affect dynamic of the diffusion processes within networks. Understanding the balance of order and chaos in real-world networks is one of the goals of the current research on so called complex networks.

Identifying and measuring properties of a real-world networks is a first step towards understanding their topology. The next step is to develop a mathematical model, which typically takes a form of an algorithm for generating networks with the same statistical properties.

For a long time real networks without visible or known rule of organization were described using Erdös and Rényi model of *Random Graphs* [8], [9]. Assuming equal probability and independent random connections made between any pair of vertices in initially not connected graph, they proposed a model suffering rather unrealistic topology. Their model has now only a limited usage for modeling real-world network.

Not long ago Watts and Strogatz proposed *Small World* model [11] of real-world networks as a result of simple observation that real networks have topology somewhere between regular and random one. They began with *Regular Graph*, such as a *Ring*, and then "rewire" some of the edges to introduce randomness. If all edges are rewired a *Random Graph* appears. The idea of this method was depicted in Fig. 1.

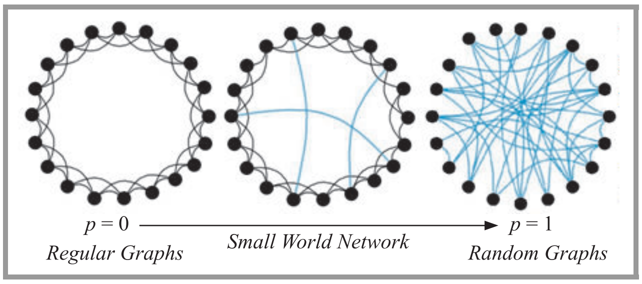


Fig. 1. The idea of *Small World* network model.

The process of rewiring affects not only the average path length but also clustering coefficient. Both of them decrease as probability of rewiring increases. The interesting property of this procedure is that for a wide range of rewiring probabilities the average path length is already low, while clustering coefficient remains high. This correlation is typical for real-world networks.

Barabási and Albert introduced yet another model [6] of real-world networks so called *Scala Free* network as a result of two main assumptions: constant growth and preferential attachment. They showed why the distribution of nodes degree is described by a power law. The process of network generation is quite simple. The network grows gradually, and when a new node is added, it creates links (edges) to the existing nodes with probability proportional to their connectivity. In consequence nodes with very high degree appears (so called *hubs* or *super-spreaders*), which are very important for communication in networks.

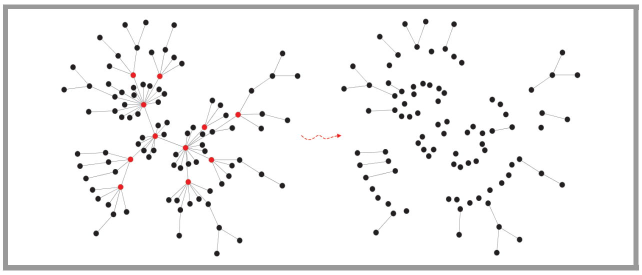


Fig. 2. The role of hubs in *Scale Free* network.

There are many modification of this basic procedure for generating networks. Now it is considered that *Scale Free* models of real-world networks are the best ones (Fig. 2).

2.2. Measures of Nodes Importance

In Fig. 3, there is an example of real social network. Nodes represent individuals and link social interactions.

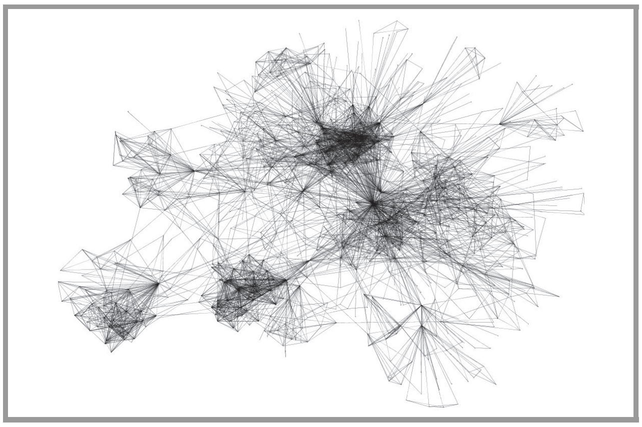


Fig. 3. An example of real social network.

The most basic and frequently asked question is how to identify the most important nodes. The answer can help maximize or, on the other hand, minimize diffusion dynamic of any behaviors within networks. We decided to use the so called centrality measures to assess nodes importance. No single measure of centre is suited for the application. Sever noteworthy measures are: degree centrality, radius centrality, closeness centrality, betweenness centrality, eigenvector centrality. Thanks to these measures we can show, for example, how to disintegrate the network with minimum number of steps and in consequence minimize diffusion area, in particular how to optimize vaccination strategies [19].

Degree centrality. The degree centrality (Fig. 4) gives the highest score of influence to the vertex with the largest number of first-neighbors. It is traditionally defined analogous to the degree of a vertex, normalized over the maximum number of neighbors this vertex could have:

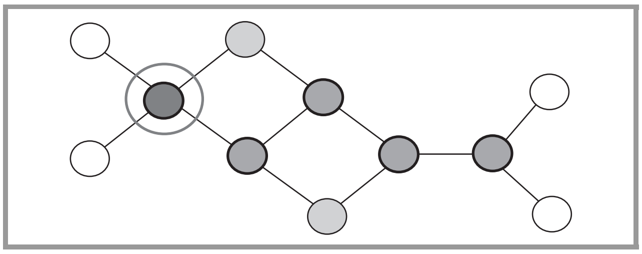


Fig. 4. Importance of nodes according degree centrality.

$$dc_i(t) = \frac{k_i(t)}{|V(t)|-1}.$$

Radius centrality. It chooses the vertex with the smallest value of the longest shortest path starting in each vertex (Fig. 5). So, if we need to find the most influential node for the most remote nodes, it is quite natural and easy to use this measure:

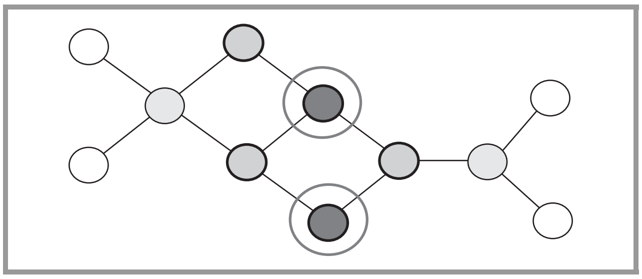


Fig. 5. Importance of nodes according radius centrality.

$$rc_i(t) = \frac{1}{\max\limits_{v_j \in V(t)} d_{ij}(t)}.$$

Closeness centrality. The closeness centrality (Fig. 6) focuses on the idea of communications between different vertices and the vertex, which is “closer” to all vertices and gets the highest score:

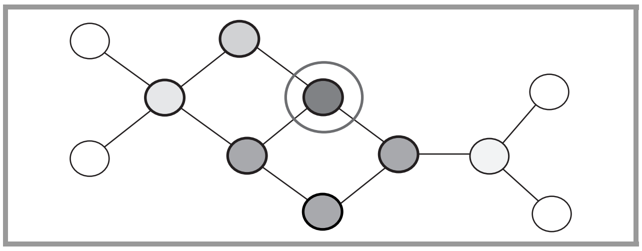


Fig. 6. Importance of nodes according closeness centrality.

$$cc_i(t) = \frac{|v(t)|-1}{\sum\limits_{v_j \in V(t)} d_{ij}(t)}.$$

Betweenness centrality. It can be defined as the percent of the shortest paths connecting two vertices that pass through the considered vertex (Fig. 7). If $p_{l,i,k}(t)$ is the set of all shortest paths between vertices v_l and v_k passing through vertex v_i and $p_{l,k}(t)$ is the set of all shortest paths between vertices v_l and v_k then:

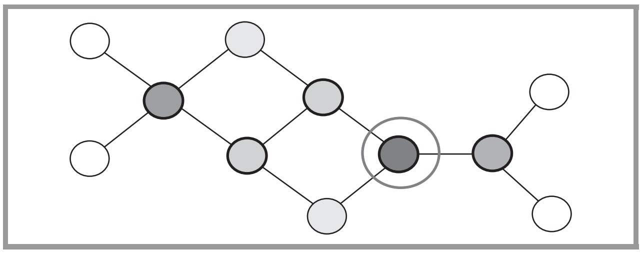


Fig. 7. Importance of nodes according betweeness centrality.

$$bc_i(t) = \frac{\sum\limits_{l<k} \frac{p_{l,i,k}(t)}{p_{l,k}(t)}}{\big(|V(t)|-2\big)\big(|V(t)|-1\big)}.$$

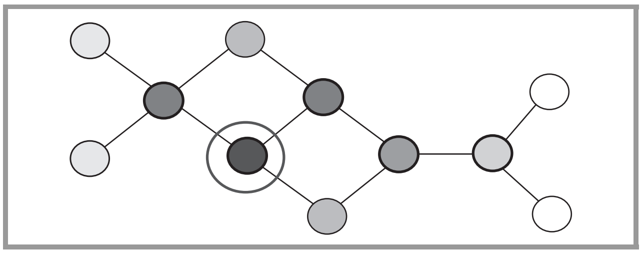


Fig. 8. Importance of nodes according eigenvector centrality.

Eigenvector centrality. While degree centrality gives a simple count of the number of connection, a vertex has eigenvector centrality acknowledges that not all connections are equal (Fig. 8). If we denote the centrality of vertex v_i by $ec_i(t)$ then we can allow for this effect by making $ec_i(t)$ proportional to the centralities of the v_i's first-neighbors,

$$ec_i(t) = \frac{1}{\lambda} \sum_{j=1}^{|V(t)|} a_{ij}(t) ec_j(t) \, .$$

Using matrix notation, we have as follows:

$$\overrightarrow{ec(t)} = \frac{1}{\lambda} A(t) \overrightarrow{ec(t)} \, .$$

So we have $A(t)\overrightarrow{ec(t)} - \lambda I \overrightarrow{ec(t)} = 0$ and the λ value we can calculate using $\det\big(A(t) - \lambda I\big) = 0$. Hence, $\overrightarrow{ec(t)}$ is an eigenvector of adjacency matrix with the largest value of eigenvalue λ.

2.3. Model of Diffusion

All in all, who is connected to whom seems to be crucial for diffusion in networks, but all kinds of behaviors have their unique properties. In consequence, we defined the model of diffusion in network as a vector, with three elements:

$$Diff(t) = \langle Net(t), PSM_{x=1,2,\ldots,N}, Gen(v,t) \rangle \, ,$$

where:

$Net(t)$ – network model of system constitutes diffusion environment;

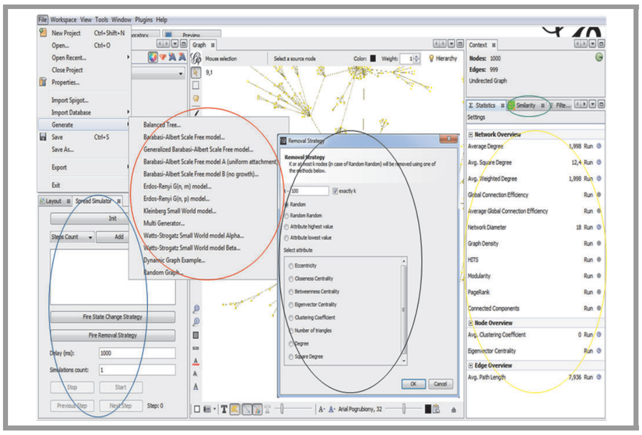
PSM_x – probabilistic finite-state machine model of considered behavior (information, virus, gossip and so on);

$Gen : V(t) \rightarrow \mathrm{SIG}$ – specific function for simulation needs (generator of signals), which assigns for each vertex in each simulation step a set of signals as a result of vertices' first-neighborhood and theirs states. These signals are received and processed by *PSM* on each vertex.

Thus, both concepts, i.e., probabilistic state machine models and real-world networks topology are highly pertaining to the presented idea subject and objectives. The aim is to uncover the diffusion mechanisms hidden in the structure of networks.

3. Simulation Environment

Our simulation environment is based on well known *Gephi* platform [20] for interactive visualization and networks exploration. The simulation environment has been implemented as a set of plugins. This kind of extensions is feasible thanks to the *Gephi* architecture based on MVC (*Model-View-Controller*) and *Service Locator* patterns. MVC pattern isolates algorithms and data from GUI (Fig. 9), permitting independent development, testing and maintenance of each one. *Service Locator* is an implementation of the IoC (*Inversion of Control*) pattern. It is a technique that allows removing dependencies from the code.



Weight: 1

9,1

Delay (ms): 1000

Simulations count: 1

Step: 0

Removal Strategy
K or at most k nodes (in case of Random Random) will be removed using one of the methods below.

k: 100 ☑ exactly k

● Random
○ Random Random
○ Attribute highest value
○ Attribute lowest value

Select attribute
○ Eccentricity
○ Closeness Centrality
○ Betweenness Centrality
○ Eigenvector Centrality
○ Clustering Coefficient
○ Number of triangles
○ Degree
○ Square Degree

Nodes: 1000
Edges: 999
Undirected Graph

Network Overview
Average Degree 1,998 Run
Avg. Square Degree 12,4 Run
Avg. Weighted Degree 1,998 Run
Global Connection Efficiency Run
Average Global Connection Efficiency Run
Network Diameter 18 Run
Graph Density Run
HITS Run
Modularity Run
PageRank Run
Connected Components Run

Node Overview
Avg. Clustering Coefficient 0 Run
Eigenvector Centrality Run

Edge Overview
Avg. Path Length 7,936 Run

Fig. 9. GUI of simulation environment.

We added to *Gephi* new functionalities, such as: complex networks generators, scenarios for centrality measures utilization in simulation of diffusion, and finally the ability to simulate diffusion of any behaviors in any networks.

Gephi architecture allows us to develop the code according to SOLID principles (*Single responsibility*, *Open-closed*, *Liskov substitution*, *Interface segregation*, *Dependency inversion*) that is five basic principles of object-oriented programming and design. It makes the code very extensible and scalable.

4. Simple Case Study

Let us now analyze a very simple case study of the diffusion process from the field of epidemiology. One of the most extensively studied epidemic models is SIS (*Susceptible-Infected-Susceptible*). In each time step, the susceptible individuals are infected by each infected neighbors with probability *beta* and the recovering rate of infected individuals to susceptible ones is *alfa*. Parameter *lambda* is known in literature as speed of spreading or virulence of the disease and is define as:

$$lambda = beta / alfa \, .$$

Figure 10 representing PSM_1 diagram of SIS model of a disease prepared in our simulation environment with $lambda = 0.5/0.1 = 5$.

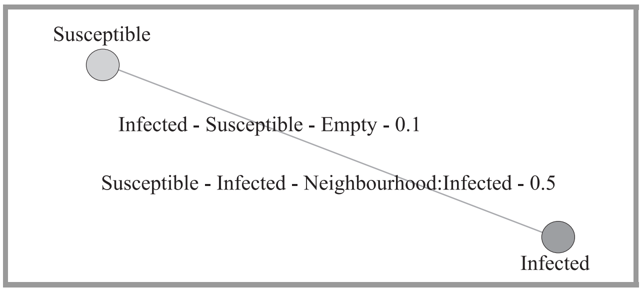

Susceptible
Infected - Susceptible - Empty - 0.1
Susceptible - Infected - Neighbourhood:Infected - 0.5
Infected

Fig. 10. SIS model of a disease.

The central question then becomes: how network topology may affect diffusion process. We focus on the SIS model of a disease spreading in networks with different topology. We use three networks: *Scale Free* (SF), *Random Graphs* (RG) and *Regular Graphs* that is exactly GRID-base one (very popular graph used in cellular automata). All networks consist of 10 000 nodes and about 20 000 edges. Average degree of nodes are similar and close to 4.

At time 0 small number of nodes (1%) is chosen randomly and infected. Then, the simulation of diffusion process is started. Each simulation was repeated 1000 times. Dynamic of disease diffusion in different networks as a function of *lambda* is presented in Figs. 11–15.

We can see that if *lambda* is high (e.g., $lambda = 5$), topology of networks have small impact on diffusion dynamic. According to Fig. 11, the number of infected individuals rose sharply and flattened out at a very high level (about 90%).

When *lambda* parameter decreases diffusion dynamic are more and more dependent on network topology. For $lambda = 0.5$ (Fig. 12) diffusion dynamic in GRID-based graph is significantly different from diffusion in *Scale Free* and *Random Graphs*. First of all, the number of infected

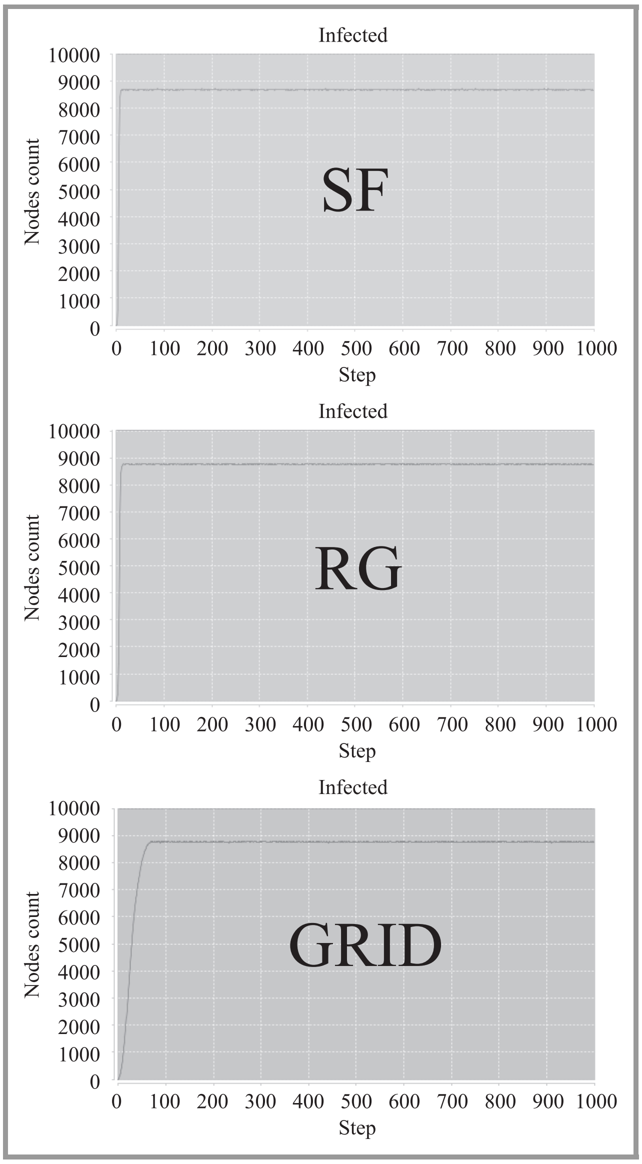


Fig. 11. SIS model of a disease with $lambda = 5$ in networks with different topology.

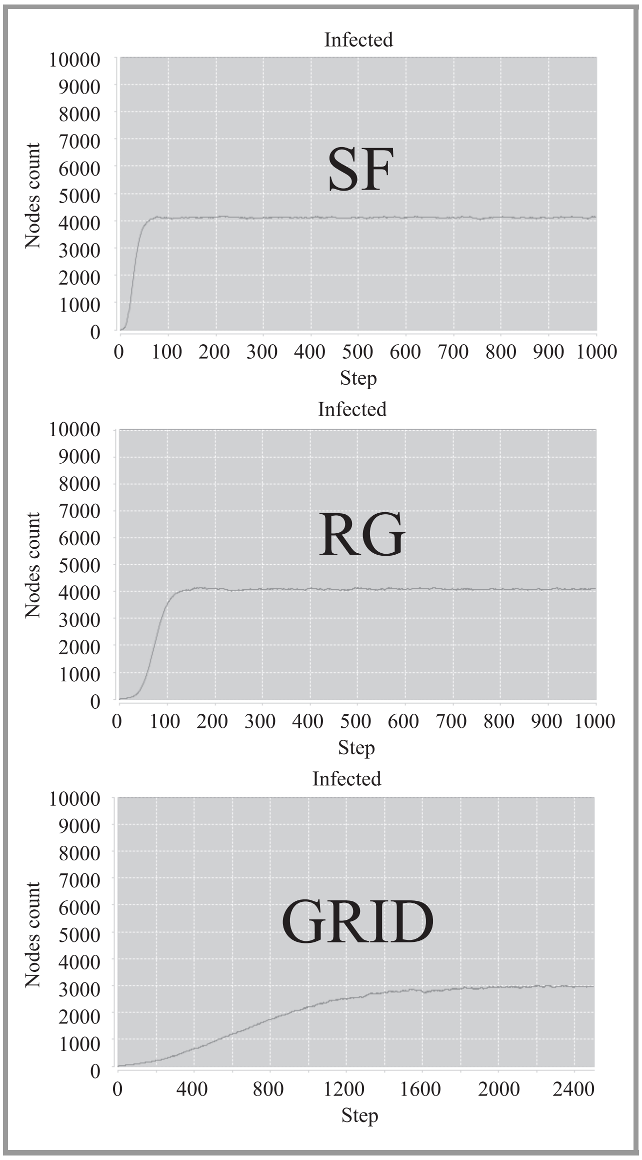


Fig. 12. SIS model of a disease with $lambda = 0.5$ in networks with different topology.

individuals rose slower, secondly flattened out at a lower level (about 30% by contrast with 40% for *Scale Free* and *Random Graphs*).

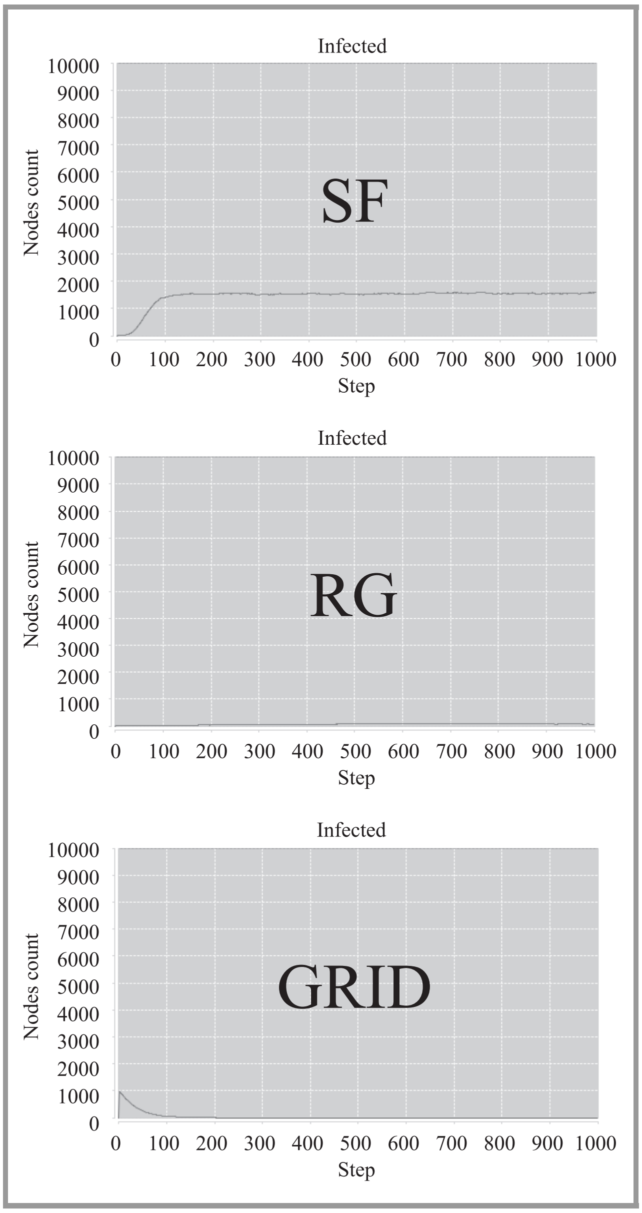


Fig. 13. SIS model of a disease with $lambda = 0.25$ in networks with different topology.

It turn out that for *lambda* = 0.25 (Fig. 13) the virus of infection disease disappear from population modeled as GRID-base graph (even though 10% individuals were infected at start time).

For *lambda* = 0.2 (Fig. 14) the virus of infection diseases also disappear from population modeled as *Random Graphs* (even though 10% individuals were infected at start time).

For *lambda* = 0.15 (Fig. 15) the virus is able to spread only in *Scale Free* network. It is an answer to the question: Why even slightly contagious diseases can plague human population over a long time without being epidemic. Not long ago it was also analytically proved that in *Scale Free* network there is no epidemic threshold for *lambda* value [5].

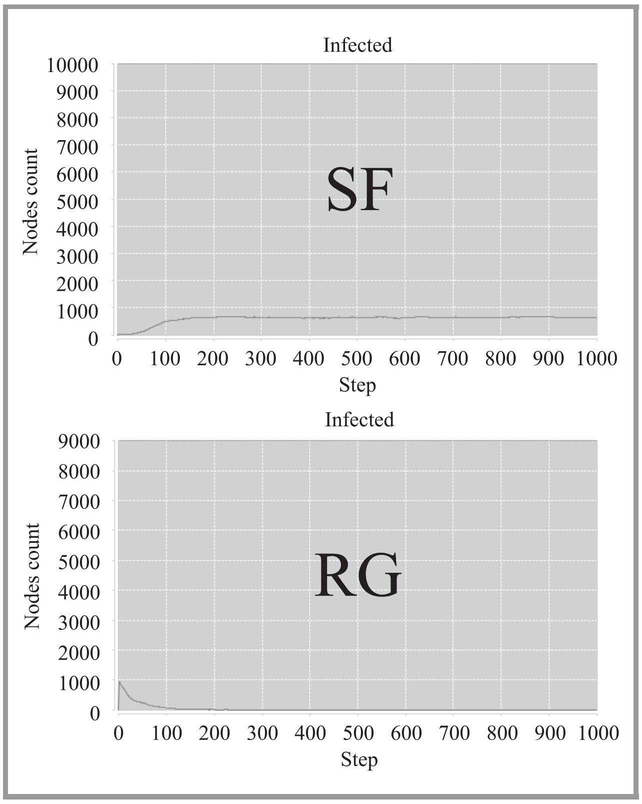


Fig. 14. SIS model of a disease with $lambda = 0.2$ in networks with different topology.

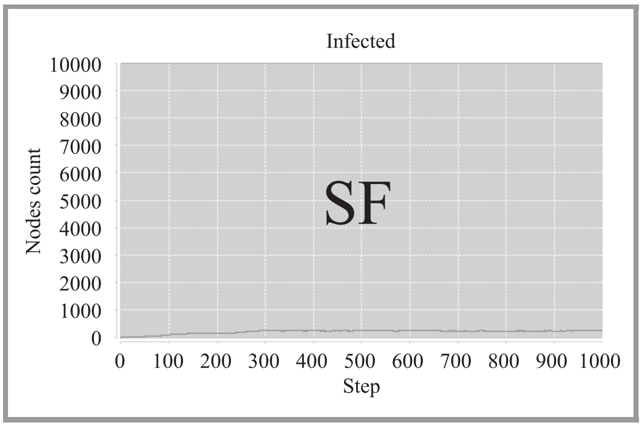
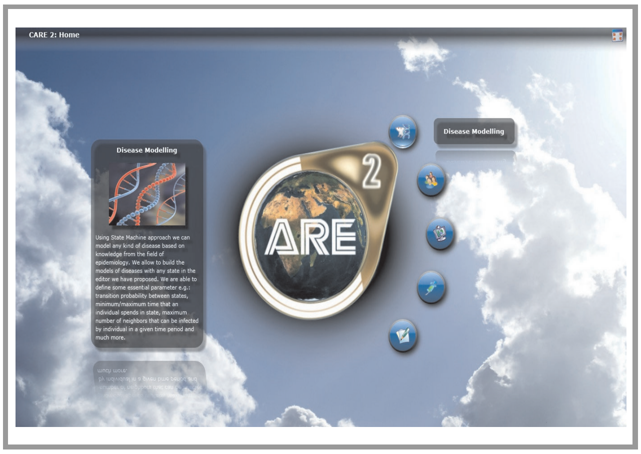


Fig. 15. SIS model of a disease with $lambda = 0.15$ in network with different topology.

5. System CARE

As practical utilization of our research system called CARE (*Creative Application to Remedy Epidemics*) was developed [21]–[23]. CARE is *Decision Support System*, which help decision makers to fight with epidemic. CARE contains five modules: *Disease Modeling*, *Social Network Modeling*, *Simulation*, *Vaccination* and *Questionnaires*.

In the *Disease Modeling* module, using probabilistic finite-state machine approach, we can model any kind of disease based on knowledge from the field of epidemiology. We allow to build the models of diseases with any states and transitions in the editor we have proposed.



Using State Machine approach we can model any kind of disease based on knowledge from the field of epidemiology. We allow to build the models of diseases with any state in the editor we have proposed. We are able to define some essential parameter e.g.: transition probability between states, minimum/maximum time that an individual spends in state, maximum number of neighbors that can be infected by individual in a given time period and much more.

Fig. 16. CARE user interface.

In *Social Network Modeling* module we can model and generate social networks using complex network theory. Using proposed generators we obtain synthetic networks but with the same statistical properties as real-world social networks. The algorithms generate networks that are *Regular Graphs*, *Random Graphs*, *Small World* networks, *Scale Free* networks or modifications thereof.

Using *Simulation* module we can visualize and simulate how the epidemic will spread in a given population. The system proposes two ways of information visualization. The first way is called "*Layout*" and helps user to manipulate networks and to set up some parameters of simulation. The alternative way is "*Geo-contextual*" one which allows to visualize networks on the world map. The system estimates the expected outcomes of different simulation scenarios and generate detailed reports. The user can assess the results and the effectiveness of the chosen vaccination strategy.

Based on the centrality measures *Vaccination* module helps the user to identify so called "*super-spreaders*" and to come up with the most efficient vaccination strategy [19]. The identification and then vaccination or isolation of the most important individuals of a given network helps decision makers to reduce the consequence of epidemics, or even stop them early in the game.

The crucial step in fighting against a disease is to get information about the social network subject to that disease. *Questionnaires* module helps building special polls based on sociological knowledge to help discover network topology. Polls designed in this way are deployed on mobile devices to gather data about social interaction.

6. Conclusion

In this paper we presented the model of diffusion in networks and the simulation environment based on *Gephi* platform. We would like to admit that we are a little bit closer to understand diffusion in networks. The solutions presented in the paper have practical implementation as a system to fight with infection diseases called CARE. Now CARE is a subsystem of monitoring, early warning and forecasting system SARNA, which was build at MUT and was put into practice in the Government Safety Centre in Poland [24]. It is worth to mentioned that CARE has its counterpart to fight with malwares in the Internet called VIRUS [25].

Acknowledgements

This work was partially supported by the research project GD-651/2011/WCY of Cybernetic Faculty at Military University of Technology.

References

[1] Godin S.: Unleashing the Ideavirus, Hyperion, New York, 2001.

[2] J. Leskovec, L. Adamic, and B. A. Huberman, "The dynamics of viral marketing", *ACM Trans. Web*, vol. 1, no. 1, article 5, 2007.

[3] A. L. Lloyd and R. M. May, "How viruses spread among computers and people", *Science*, vol. 292, no. 5520, pp. 1316–1317, 2001.

[4] D. López-Pintado, "Diffusion in complex social networks", *GAMES and Economic Behavior*, vol. 62, no. 2, pp. 573–590, 2008.

[5] R. Pastor-Satorras and A. Vespignani, "Epidemic spreading in scale-free networks", *PRL*, vol. 86, no. 14, pp. 3200–3203, 2001.

[6] A. L. Barabási and R. Albert, "Emergency of scaling in random networks", *Science*, vol. 286, pp. 509–512, 1999.

[7] A. L. Barabási and R. Albert, "Topology of evolving networks: local events and universality", *PRL*, vol. 85, no. 24, pp. 5234–5237, 2000.

[8] P. Erdös and A. Rényi, "On random graphs", *Publicationes Mathem.*, vol. 6, pp. 290–297, 1959.

[9] P. Erdös and A. Rényi, "On the evolution of random graphs", Publications of the Mathematical Institute of the Hungarian Academy of Sciences 5, pp. 17–61, 1959.

[10] M. E. J. Newman, "Models of the small world: A review", *J. Stat. Phys.*, vol. 101, pp. 819–841, 2000.

[11] D. J. Watts and S. H. Strogatz, "Collective dynamics of "small-world" networks", *Nature*, vol. 393, pp. 440–442, 1998.

[12] A. L. Barabási and R. Albert, "Statistical mechanics of complex networks", *Rev. Modern Phys.*, vol. 74, pp. 47–97, 2002.

[13] M. E. J. Newman, "The structure and function of complex networks", *SIMA Rev.*, vol. 45, no. 2, pp. 167–256, 2003.

[14] S. H. Strogatz, "Exploring complex networks", *Nature*, vol. 410, pp. 268–276, 2001.

[15] X. Wang and G. Chen, "Complex networks: Small-world, scale-free and beyond", *IEEE Circ. Sys. Mag.*, vol. 3, no. 1, pp. 6–20, 2003.

[16] A. Sokolova and E. P. de Vink, *Probabilistic Automata: System Types, Parallel Composition and Comparison*, LNCS 2925, Heidelberg: Springer, 2004, pp. 1–43.

[17] E. Vidal, F. Thollard, C. de la Higuera, F. Casacuberta, and R. C. Carrasco, "Probabilistic Finite-State Machines – Part I", *IEEE Trans. Pattern Anal. Machine Intel.*, vol. 27, no. 7, pp. 1013–1025, 2005.

[18] F. Harary and G. Gupta, “Dynamic Graph Models”, *Mathl. Comput. Modelling*, vol. 25, no. 7, pp. 79–87, 1997.

[19] R. Kasprzyk, “The vaccination against epidemic spreeding in complex networks”, *Biuletyn ISI*, no. 3(1/2009), pp. 39–43, 2009.

[20] M. Bastian, S. Heymann, and M. Jacomy, “Gephi: an open source software for exploring and manipulating networks”, in *Proc. Int. AAAI Conf. Weblogs Social Media*, San Jose, CA, USA, 2009.

[21] R. Kasprzyk, A. Najgebauer, and D. Pierzchała, “Modelling and simulation of infection disease in social networks”, *Comput. Collective Intel.*, LNAI 6922, pp. 388–398, 2011.

[22] R. Kasprzyk, D. Pierzchała, and A. Najgebauer, “Creative application to remedy epidemics”, in *Risk Analysis VII & Brownfields V*, C. Brebia, Ed. WIT Press, 2010, pp. 545–562.

[23] R. Kasprzyk, B. Lipiński, K. Wilkos, M. Wilkos, and C. Bartosiak, “CARE – creative application to remedy epidemics”, *Biuletyn ISI*, no. 3(1/2009), pp. 45–52, 2009.

[24] A. Najgebauer, D. Pierzchała, and R. Kasprzyk, “A distributed multilevel system for monitoring and silumation of epidemics”, in *Risk Analysis VII & Brownfields V*, C. A. Brebbia, Ed. WIT Press, 2010, pp. 583–596.

[25] R. Kasprzyk, “Symulator rozprzestrzeniania się złośliwego oprogramowania w sieciach komputerowych”, *Symulacja w badaniach i rozwoju*, vol. 1, no. 2, pp. 139–150, 2010 (in Polish).

Rafał Kasprzyk was commissioned as 2Lt. in 2004 and one year later received his M.Sc. Eng. degree after individual studying in DSS (Decision Support System) at Cybernetics Faculty as the 1st place graduate from Military University of Technology in 2005. He was promoted to the rank of Lt. and Capt. in 2008 and 2010 respectively. In 2012 he received Ph.D. degree in the field of computer science. He has worked as a lecturer at Cybernetics Faculty since 2005. He has participated in many scientific projects connected with combat simulation and crisis management. His main interest are graph and network theory, decision support systems, computer simulation, homeland security and cyber-security.

E-mail: rkasprzyk@wat.edu.pl
Faculty of Cybernetics
Military University of Technology
Gen. S. Kaliskiego st 2
00-908 Warsaw, Poland

Paper

Compact Broadband Rat-Race Coupler in Multilayer Technology Designed with the Use of Artificial Right- and Left-Handed Transmission Lines

Kamil Staszek, Jacek Kołodziej, Krzysztof Wincza, and Sławomir Gruszczyński

Department of Electronics, AGH University of Science and Technology, Kraków, Poland

Abstract—The paper presents a compact broadband rat-race coupler for the first time designed and realized in a multilayer microstrip technology. To achieve both broad operational bandwidth and a compact size the 270° transmission line of a conventional rat-race, coupler has been replaced by a –90° left-handed transmission line realized with the use of a quasi-lumped element technique. Moreover, to achieve better compactness of the resulting coupler, all 90° right-handed transmission lines have been realized with the use of the same technique. It has been also proved that simple LC approximation of a left-handed transmission line can be successfully used for the design. Moreover, it has been shown that when appropriately chosen, the multilayer dielectric structure allows for realization of structures designed with the use of this simple approximation, for both right-handed and left-handed transmission lines, without loosing too much of a performance.

Keywords—broadband rat-race couplers, left-handed transmission lines, quasi-lumped elements.

1. Introduction

Rat-race couplers are well-known components often used in microwave circuits. They allow for in-phase and out-of-phase power division with the fourth port being isolated, and in a classic approach they are composed of three 90° and one 270° transmission lines. The two major drawbacks of conventional rat-race couplers are their relatively large size and narrow bandwidth. The problem of size reduction is usually approached with the use of lumped-element technique [1], [2], whereas, different techniques have been developed over the years to broader the operational bandwidth [3]–[7]. One of the possible approaches that allows for achieving broadband operation of rat-race couplers involves the application of an ideal 180 transformer, which is difficult to realize [3]. Another technique allowing for significant bandwidth increase is based on the substitution of a 270° transmission line by a –90° left-handed transmission line [8]–[11]. Such an approach also allows for significant size reduction. In [8], the first trials of such devices are shown, in which left-handed transmission lines have been realized with the use of SMD components. Another example of such a rat-race coupler is shown in [9], where right-handed and left-handed lines have been implemented with the use of complementary split rings resonators. In [10], a compact broadband rat-race coupler has been considered, in which, to achieve smaller size, both right- and left-handed transmission lines have been realized as artificial lines, with the use of metal-air-metal capacitors and short-open circuited stubs. It was shown in [10] that such an approach is suitable for high frequency and low frequency applications, however, only full wave simulations are presented.

In this paper, we present for the first time the design and realization of a compact broadband rat-race coupler in a microstrip multilayer technology. The presented coupler has been designed with the use of both right- and left-handed artificial transmission lines. Right-handed transmission lines have been modeled with the use of a cascade connection of several RH cells and the appropriate number of cells has been selected, so that the effect of substituting the 90° transmission lines by the artificial lines on the overall coupler's performance can be neglected. Similarly, the –90° left-handed transmission line has been modeled with the use of LH cells, and the optimum number of cells has been chosen to ensure from one hand, the good performance of the coupler, and on the other hand, the feasibility of further circuit implementation in a microstrip multilayer technology. Both theoretical and experimental results are presented in this paper.

2. Theoretical Analysis

The concept of a compact broadband rat-race coupler is explained in Fig. 1. The coupler utilizes a well-known idea of substituting 270° right-handed transmission line by a –90° left-handed transmission line. Both right- and left-handed lines have been divided in n equal unit cells. The right-handed unit cell consists of a series inductor L_{RH} and a shunt capacitor C_{RH}, whereas, the left-handed unit cell consists of shunt inductor L_{LH}, and a series capacitor C_{LH}. The values of the lumped elements shown in Fig. 1 would, therefore, depend on the characteristic impedance of the transmission line Z, its electrical length Θ, operating frequency f and the number of applied unit cells n for both

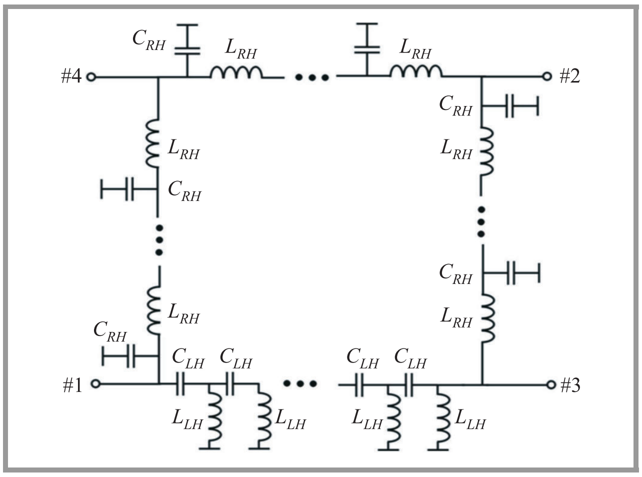


Fig. 1. A schematic diagram of a compact broadband rat-race coupler utilizing artificial right and left-handed transmission lines.

right- and left-handed artificial lines. From the transmission line theory of both right- and left-handed transmission lines we can derive simple formulas for all values of lumped elements shown in Fig. 1, which are

$$L_{RH} = \frac{\Theta Z}{2\pi f n}, \tag{1}$$

$$C_{RH} = \frac{\Theta}{2\pi f n Z}, \tag{2}$$

$$L_{LH} = \frac{Zn}{2\pi f |\Theta|}, \tag{3}$$

$$C_{LH} = \frac{n}{2\pi f Z |\Theta|}, \tag{4}$$

where: Θ – electrical length of the transmission line section in radians, Z – characteristic impedance of the transmission line section, f – frequency at which the electrical length is specified, and n – number of the applied unit cells for transmission-line section approximation.

It should be noted that for the case of a left-handed transmission line, the electrical length is negative (equals $\Theta = -\pi/4$ in case of a rat-race coupler), whereas the elements of its equivalent circuit are positive. It is also known that the higher number of unit cells n is taken, the better approximation of the ideal transmission line section by its artificial equivalent for both right- and left-handed transmission lines. However, to achieve smaller size of the resulting coupler low value of n is preferable. On the other hand, by analyzing formulas (1)–(4) one can see that in case of a right-handed artificial transmission line the higher n, the lower values of the resulting inductors L_{RH} and capacitors C_{RH}. Whereas, in case of a left-handed transmission line this dependence is opposite. Therefore, in the case of a left-handed artificial line, the number of unit cells n is a trade-off between the quality of the transmission line approximation and the feasibility of further physical realization. The dependence of the rat-race parameters versus number of cells n (for both right-handed and left-handed lines) have been investigated and it turned out that for $n = 8$ ($\Theta = 11.25°$ of each unit cell) the achieved parameters are almost identical as the parameters of the coupler modeled by the ideal transmission lines. To simplify further realization of the left-handed unit cells, the number of cells have been reduced to $n = 6$, which gave lower values of capacitors C_{LH} and inductors L_{LH}. Figure 2 presents calculated frequency characteristics of the coupler, in which 90° right-handed transmission line sections have been modeled by 8 LC unit cells, whereas, the –90° left-handed transmission line section has been modeled by 6 CL unit cells. It seems that good performance have been obtained. The coupler's bandwidth equals 55% for RL $>$ 20 dB and 106% for $I >$ 20 dB, and the coupler features broadband differential phase.

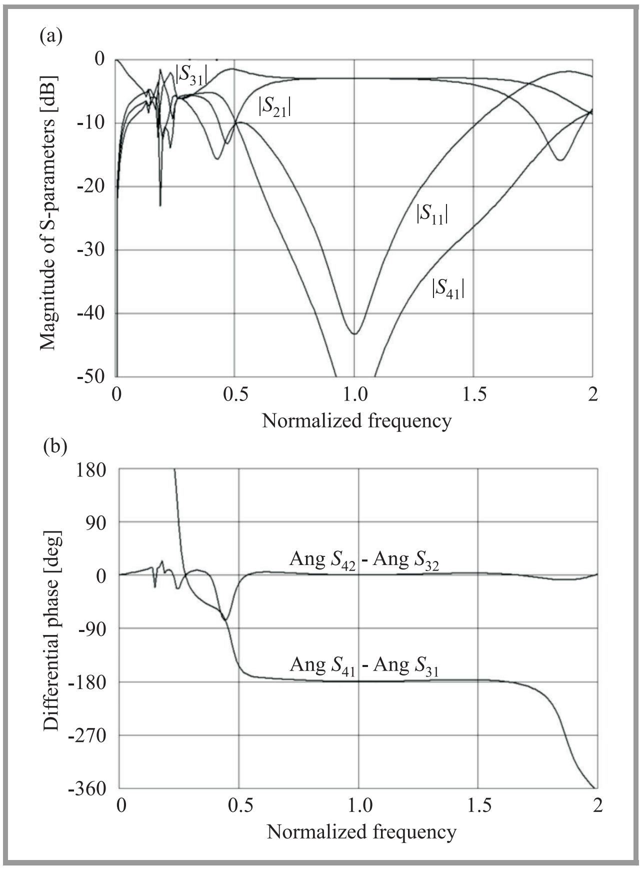


Fig. 2. Calculated amplitude (a) and differential phase (b) characteristics of a compact rat-race coupler in which 90° right-handed transmission lines have been divided into 8 cells and the –90° left-handed transmission line has been divided into 6 section. Results of circuit simulations.

3. Experimental Results

The presented approach for compact broadband rat-race realization has been experimentally verified. For the purpose of simultaneous realization of large series capacitors and shunt inductors a multilayer structure presented in Fig. 3 has been chosen. In such a structure, in which a thin dielectric layer is placed on a thick dielectric layer, it is possible to realize large series capacitors and large inductors having reduced spurious shunt capacitance. The coupler has been designed for the center frequency of 0.5 GHz, for which the following values of lumped elements have been found $L_{RH} = 4.42$ nH and $C_{RH} = 0.884$ pF (for $Z = 70.7\ \Omega$, $n = 8$, $\Theta = \pi/4$) for the right-handed transmission line and $L_{LH} = 86$ nH and $C_{LH} = 17.2$ pF (for $Z = 70.7\ \Omega$, $n = 6$, $\Theta = -\pi/4$) for the left-handed transmission line. The detailed layout of the developed rat-race coupler is shown in Fig. 4. To ensure symmetry of the right-handed transmission lines, the series inductors LRH have been divided into two equal inductors, between which shunt capacitors CRH are inserted. To minimize metallization pads required for CRH realization, the bottom pads of the capacitors have been grounded. The CLH capacitors have been realized on two sides of the thin laminate, whereas, LLH inductors have been realized as planar square spiral inductors grounded in the center, as it is presented in Fig. 4.

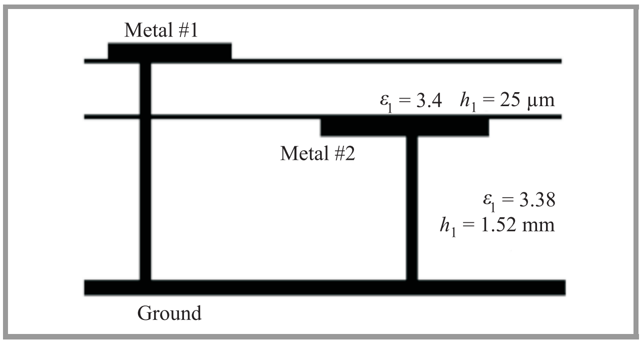


Fig. 3. Cross-sectional view of the dielectric structure used for the design of a compact broadband rat-race coupler.

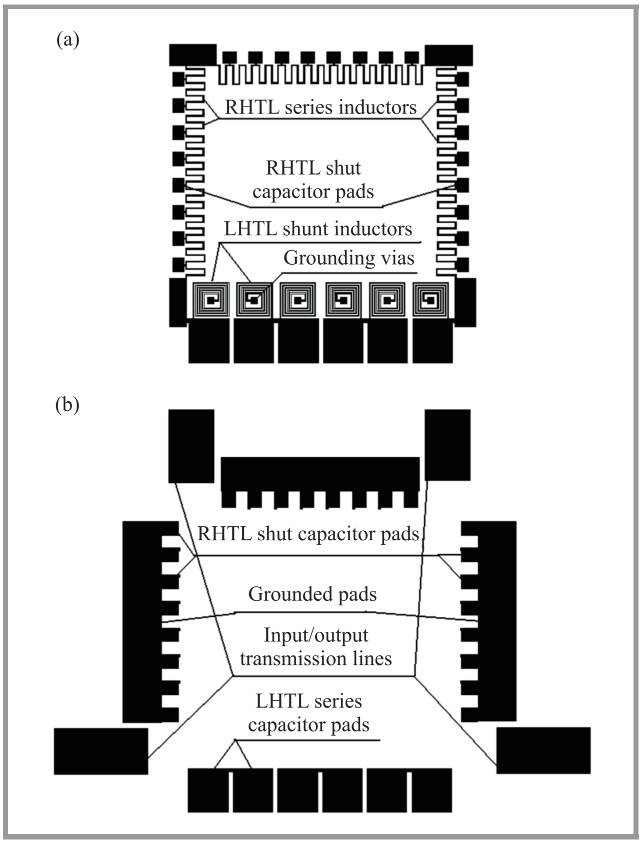


Fig. 4. Layout of the developed compact broadband rat-race coupler (not to scale): (a) upper layer; (b) lower layer.

The designed coupler has been analyzed electromagnetically and the obtained results are plotted in Fig. 5. As it is shown, the designed coupler features slightly narrower bandwidth in terms of return losses (*BW* = 44%). However, the isolation, transmission and coupling, and also differential phase characteristics are better within broader bandwidth than the bandwidth of the coupler modeled by simple LC and CL unit cells. This is due to the well-known effect of spurious shunt capacitances, and series inductances of the left-handed transmission line that affects its phase characteristic, resulting in composite right/left-handed transmission lines (CRLH TL). It is important to underline

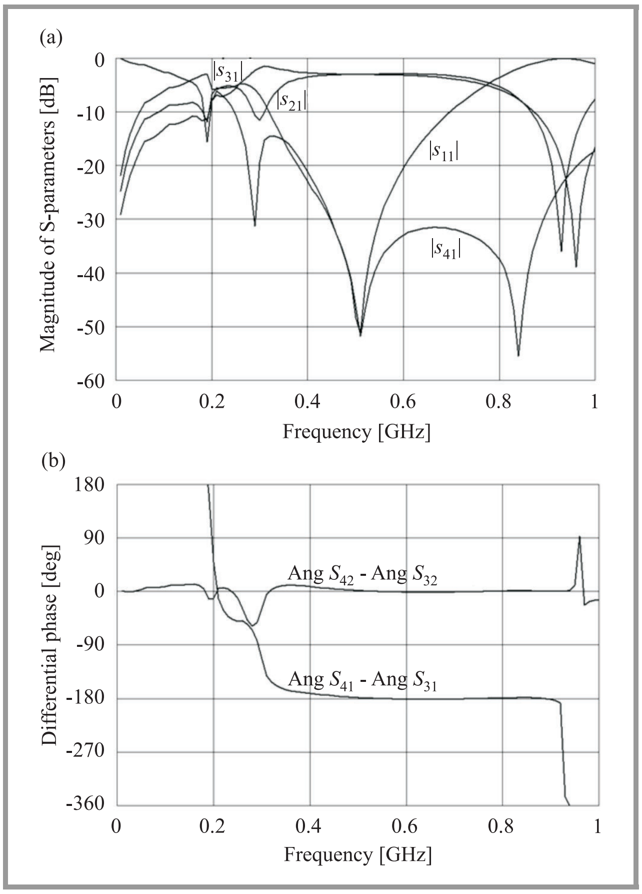


Fig. 5. Calculated amplitude (a) and differential phase (b) characteristics of the developed compact rat-race coupler shown in Fig. 4. Results of electromagnetic calculations.

that CRLH TLs can be used for maximizing bandwidth, in terms of differential phase response, as it was shown in [8]. However, the obtained results prove that the presented simplified approach is also suitable for broadband compact rat-race realization, especially in a multilayer microstrip technique. The achieved size of the developed rat-race coupler in comparison with the classic one is shown in Fig. 6. In this case a significant size reduction has been achieved.

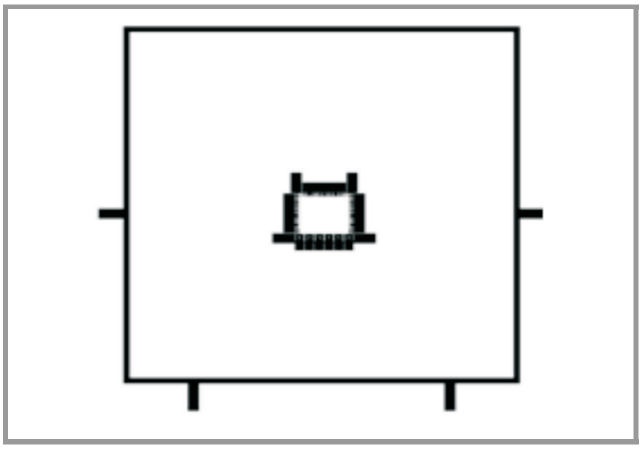


Fig. 6. Size comparison between a classic and the developed rat-race couplers.

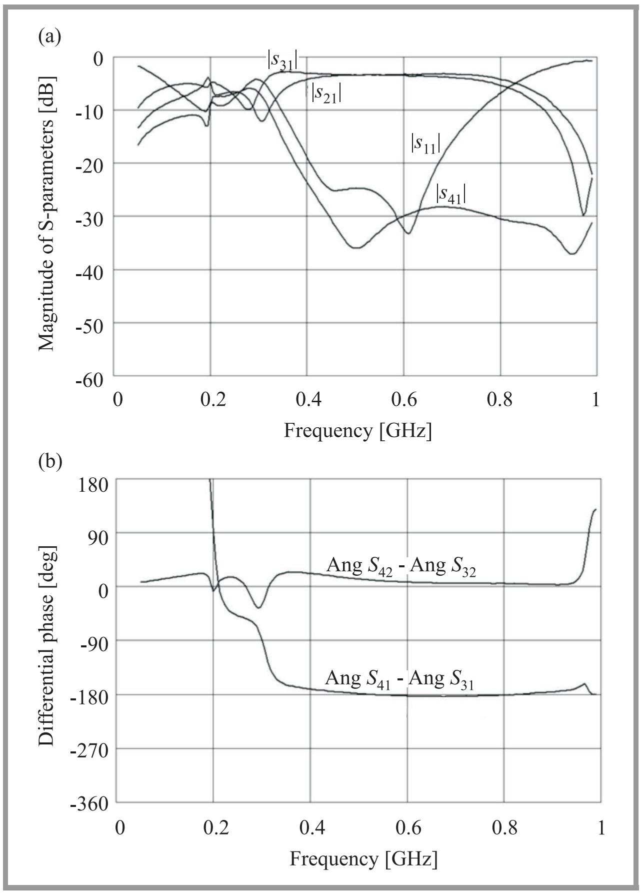


Fig. 7. Measured frequency characteristics of the manufactured compact rat-race coupler: (a) amplitude characteristics; (b) differential phase characteristics.

The designed coupler has been manufactured and the measured results are shown in Fig. 7. A good agreement has been achieved between the calculated and measured characteristics in terms of both amplitude and differential phase. The measured return losses and isolation are better than 20 dB in slightly broader bandwidth than the calculated one. The achieved differential phase ripple does not exceed 10 and the overall insertion loss equals 0.4 dB.

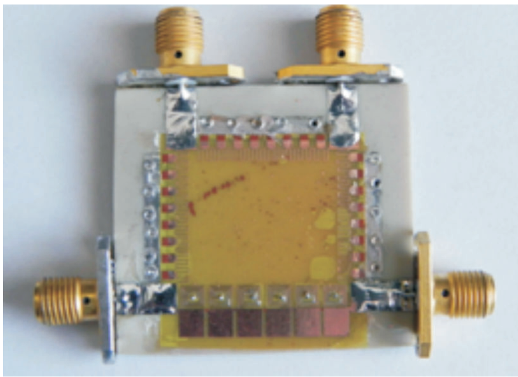


Fig. 8. Photograph of the manufactured compact broadband rat-race coupler.

Figure 8 presents the photograph of the compact broadband rat-race coupler developed in a multilayer microstrip technology.

4. Conclusions

The design of a compact broadband rat-race coupler in a multilayer microstrip technique, and utilizing artificial right- and left-handed transmission lines has been presented for the first time. A simple approach has been proposed, in which both right- and left-handed transmission lines are represented, respectively, by LC and CL lumped-element networks. Simple formulas for the lumped elements constituting both right- and left-handed transmission lines have been given for the arbitrary number of unit cells of each line. Moreover, it has been shown that such a simplified approach gives a good performance of the resulting coupler, and also that such couplers are easily realized in a microstrip multilayer structure. This is due to the fact that a microstrip structure, in which a thin dielectric layer is placed over a thick one, is suitable for large series capacitors and shunt inductors realization. Although, the formulas for lumped-elements of an artificial CRLH transmission line that gives the maximum bandwidth of the coupler are known [8], the obtained results prove that the presented simplified approach can be also used for the design of compact broadband rat-race couplers in a microstrip multilayer technique.

Acknowledgment

This work has been founded by The Polish National Science Centre under grant no. UMO-2011/01/D/ST7/00789.

References

[1] R. W. Vogel, "Analysis and design of lumped- and lumped-distributed-element directional couplers for MIC and MMIC applications", *IEEE Trans. Microwave Theory Tech.*, vol. 40, pp. 253–262, 1992.

[2] T. Hirota, A. Minakawa, and M. Muraguchi, "Reduced-size branch-line and rat-race hybrids for uniplanar MMIC's", *IEEE Trans. Microwave Theory Tech.*, vol. 38, pp. 270–275, 1990.

[3] Ch. Y. Chang, Ch. Ch. Yang, "A novel broad-band chebyshev-response rat-race coupler", *IEEE Trans. Microwave Theory Tech.*, vol. 47, no. 4, pp. 455–462, 1999.

[4] S. Gruszczynski, K. Wincza, "On the design of a broadband rat-race coupler in broadside line technique", in *Proc. 13th Conf. Microwave Technique COMITE 2005*, Prague, Czech Republic, 2005.

[5] S. March, "A wide-band stripline hybrid ring", *IEEE Trans. Microwave Theory Tech.*, vol. MTT-16, pp. 361–361, 1968.

[6] J. L. B. Walker, "Improvements to the design of the 0-180° rat race coupler and its application to the design of balanced mixers with high LO to RF isolation", in *Proc. MTT-S Int. Microwave Symp. Digest*, Denver, USA, 1997, vol. 2, pp. 747–750.

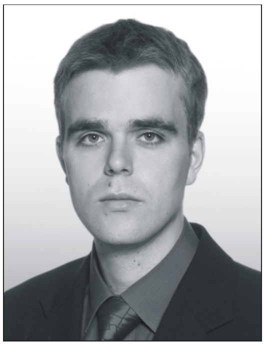
[7] S. Rehnmark, "Wide-band balanced line microwave hybrids", *IEEE Trans. Microwave Theory Tech.*, vol. MTT-25, no. 10, pp. 825–830, 1977.

[8] C. Caloz, T. Itoh, *Electromagnetic metamaterials: Transmission Line Theory and Microwave Applications*. New York: Wiley, 2006.

[9] H. Okabe, C. Caloz, T. Itoh, "A compact enhanced bandwidth hybrid ring using and artificial lumped element left-handed transmission-line section", *IEEE Trans. Microwave Theory and Tech.*, vol. 52, no. 3, pp. 798–804, 2004.

[10] G. Siso, M. Gil, J. Bonache, and F. Martin, "Applications of resonant-type metamaterial transmission lines to the design of enhanced bandwidth components with compact dimensions", *Microwave Opt. Technol. Lett.*, vol. 50, pp. 127–134, 2008.

[11] G. Monti and L. Tarricone, "Compact broadband monolithic 3-dB coupler by using artificial transmission lines", *Microwave Opt. Technol. Lett.*, vol. 50, pp. 2662–2667, 2008.



Kamil Staszek received his M.Sc. degree in Electronics Engineering from AGH University of Science and Technology, Kraków, Poland in 2011. Currently he is working toward Ph.D. degree at the AGH University in the field of Microwave Engineering. He has co-authored 5 scientific conference and journal papers. His scientific interest includes design of passive microwave network components and multiport S-parameter measurement techniques.

E-mail: kstaszek@agh.edu.pl
Department of Electronics
AGH University of Science and Technology
Mickiewicza av. 30
30-059 Kraków

Jacek Kołodziej was born in Stąporków, Poland, on Feb. 6, 1974. He received the M.Sc. degree in Electronics and Telecommunication with specialization of Electronics Equipment Designing in 1999 and Ph.D. honors degree in Electronics in 2007, both from the AGH University of Science and Technology Kraków, Poland. Since 2008 he has been the adjunct professor lecturing: Information Technology, Machine to Machine Communication and Software Engineering for Embedded System. His scientific interests include advanced electronic circuits in telecommunication, wireless sensor networks, analog/digital converters, sigma-delta modulators, adaptive non-uniform sampling delta modulators, software engineering, testing and reliability. He is co-author of 30 conference and journal papers including Electronics and Telecommunications Quarterly, WSEAS Transactions on Circuits and Systems, WSEAS Transactions on Communications. He is IEEE member, Student Branch Chancellor and WSEAS reviewer. For his organizing, scientific and didactic activities he received 5 AGH Rector's Awards for best academic teachers and twice the award of the Mayor of Kraków. He has been participating in several R&D Polish and European projects being personally responsible for system management and communication procedures.

E-mail: jacek.kolodziej@agh.edu.pl
Department of Electronics
AGH University of Science and Technology
Mickiewicza av. 30
30-059 Kraków

Krzysztof Wincza was born in Wałbrzych, Poland, on May 27, 1979. He received the M.Sc. degree and the Ph.D. degree in Electronics And Electrical Engineering from the Wrocław University of Technology, Poland, in 2003 and 2007, respectively. In 2007, he joined the Institute of Telecommunications, Teleinformatics and Acoustics, Wrocław University of Technology. In 2009, he joined the Faculty of Electronics at AGH University of Science and Technology becoming an Assistant Professor. Dr. Wincza was the recipient of The Youth Award presented at the 10th National Symposium of Radio Sciences (URSI) and the Young Scientist Grant awarded by the Foundation for Polish Science in 2001 and 2008, respectively. He has co-authored 41 scientific papers.

E-mail: krzysztof.wincza@agh.edu.pl
Department of Electronics
AGH University of Science and Technology
Mickiewicza av. 30
30-059 Kraków

Sławomir Gruszczyński was born in Wrocław, Poland, on December 14, 1976. He received the M.Sc. degree and the Ph.D. degree in Electronics and Electrical Engineering from the Wrocław University of Technology, Poland, in 2001 and 2006, respectively. Since 2001 to 2006 he has been working for Telecommunications Research Institute, Wrocław Division. From 2005 to 2009, he worked at the Institute of Telecommunications, Teleinformatics and Acoustics, Wrocław University of Technology. In 2009, he joined the Faculty of Electronics at AGH University of Science and Technology. He has coauthored 45 scientific papers. He is a member of the IEEE, and a member of Young Scientists' Academy at Polish Academy of Sciences (PAN) and Committee of Electronics and Telecommunications at Polish Academy of Sciences (PAN).

E-mail: slawomir.gruszczynski@agh.edu.pl
AGH University of Science and Technology
Mickiewicza av. 30
30-059 Kraków

Information for Authors

Journal of Telecommunications and Information Technology **(JTIT)** is published quarterly. It comprises original contributions, dealing with a wide range of topics related to telecommunications and information technology. **All papers are subject to peer review**. Topics presented in the JTIT report primary and/or experimental research results, which advance the base of scientific and technological knowledge about telecommunications and information technology.

JTIT is dedicated to publishing research results which advance the level of current research or add to the understanding of problems related to modulation and signal design, wireless communications, optical communications and photonic systems, voice communications devices, image and signal processing, transmission systems, network architecture, coding and communication theory, as well as information technology.

Suitable research-related papers should hold the potential to advance the technological base of telecommunications and information technology. Tutorial and review papers are published only by invitation.

Manuscript. TEX and LATEX are preferable, standard Microsoft Word format (.doc) is acceptable. The author's JTIT LATEX style file is available:
http://www.nit.eu/for-authors

Papers published should contain up to 10 printed pages in LATEX author's style (Word processor one printed page corresponds approximately to 6000 characters).

The manuscript should include an abstract about 150–200 words long and the relevant keywords. The abstract should contain statement of the problem, assumptions and methodology, results and conclusion or discussion on the importance of the results. Abstracts must not include mathematical expressions or bibliographic references.

Keywords should not repeat the title of the manuscript. About four keywords or phrases in alphabetical order should be used, separated by commas.

The original files accompanied with pdf file should be submitted by e-mail: redakcja@itl.waw.pl

Figures, tables and photographs. Original figures should be submitted. Drawings in Corel Draw and PostScript formats are preferred. Figure captions should be placed below the figures and can not be included as a part of the figure. Each figure should be submitted as a separated graphic file, in .cdr, .eps, .ps, .png or .tif format. Tables and figures should be numbered consecutively with Arabic numerals.

Each photograph with minimum 300 dpi resolution should be delivered in electronic formats (TIFF, JPG or PNG) as a separated file.

References. All references should be marked in the text by Arabic numerals in square brackets and listed at the end of the paper in order of their appearance in the text, including exclusively publications cited inside. Samples of correct formats for various types of references are presented below:

[1] Y. Namihira, "Relationship between nonlinear effective area and mode field diameter for dispersion shifted fibres", *Electron. Lett.*, vol. 30, no. 3, pp. 262–264, 1994.

[2] C. Kittel, *Introduction to Solid State Physics*. New York: Wiley, 1986.

[3] S. Demri and E. Orłowska, "Informational representability: Abstract models versus concrete models", in *Fuzzy Sets, Logics and Knowledge-Based Reasoning*, D. Dubois and H. Prade, Eds. Dordrecht: Kluwer, 1999, pp. 301–314.

Biographies and photographs of authors. A brief professional author's biography of up to 200 words and a photo of each author should be included with the manuscript.

Galley proofs. Authors should return proofs as a list of corrections as soon as possible. In other cases, the article will be proof-read against manuscript by the editor and printed without the author's corrections. Remarks to the errata should be provided within one week after receiving the offprint.

Copyright. Manuscript submitted to JTIT should not be published or simultaneously submitted for publication elsewhere. By submitting a manuscript, the author(s) agree to automatically transfer the copyright for their article to the publisher, if and when the article is accepted for publication. The copyright comprises the exclusive rights to reproduce and distribute the article, including reprints and all translation rights. No part of the present JTIT should not be reproduced in any form nor transmitted or translated into a machine language without prior written consent of the publisher.
For copyright form see: http://www.nit.eu/for-authors

A copy of the JTIT is provided to each author of paper published.

Journal of Telecommunications and Information Technology has entered into an electronic licencing relationship with EBSCO Publishing, the world's most prolific aggregator of full text journals, magazines and other sources. The text of *Journal of Telecommunications and Information Technology* can be found on EBSCO Publishing's databases. For more information on EBSCO Publishing, please visit www.epnet.com.

(Contents Continued from Front Cover)

INSTYTUT ŁĄCZNOŚCI
PANSTWOWY INSTYTUT BADAWCZY

Editorial Office

National Institute
of Telecommunications
Szachowa st 1
04-894 Warsaw, Poland

tel. +48 22 512 81 83
fax: +48 22 512 84 00
e-mail: redakcja@itl.waw.pl
http://www.nit.eu