OPT-*i*
An International Conference on
Engineering and Applied Sciences Optimization
M. Papadrakakis, M.G. Karlaftis, N.D. Lagaros (eds.)
Kos Island, Greece, 4-6, June 2014

AN EFFICIENT VARIABLE NEIGHBOURHOOD SEARCH HEURISTIC FOR THE INFLUENTIAL SELECTION PROBLEM

Evren Güney[1], İrem Düzdar[1], Volkan Çakır[1], Abdullah Özdemir[1], Özge Şahin[1]

[1]Istanbul Arel University
Tepekent, Istanbul
e-mail: evrenguney, iremduzdar, volkancakir, abdullahozdemir, ozgesahin@arel.edu.tr

Keywords: influential selection, combinatorial optimization, variable neighbourhood search

Abstract. *A online social network consists of individuals or entities which are tied by a certain type of interdependency such as collaboration, friendship or acquaintance. Due to the wide usage of internet, many companies actively use this new media in their marketing campaigns. Given an online social network, a person that influences his/her connections is called an influential. Since companies have limited budgets for advertising, they have to correctly select the influentials, which will forward their message to their connections over various online social networks and hope that a cascade will be triggered so that a maximum number of individuals are reached. In most of the previous studies the social networks are modeled as stochastic fields where the cascading process is represented by linear threshold or independent cascade models. It is shown that determining the top-k influential nodes within the network under this setting is NP-hard and a greedy based heuristic provides a provable approximation guarantee. Following these initial efforts many researchers worked on the improvement of the efficiency of the greedy algorithm without focusing on the solution quality. In this study, the problem of selecting the best k-influentials in social netrowk is studied from the solution quality perspective and using a Variable Neighbourhood Serach(VNS) heuristic optimal solutions are sought. Experimental analysis on certain real-life data is carried out to determine the performance of the techniques proposed.*

1 Introduction

Online social networks come into prominence with the fast growth of their number of members and frequency of usage. This fundamental change in communication of individuals raised the importance of so called "word-of-mouth", where individuals trust and act according to the recommendations of their friends instead of the advertisements they see [1]. The information spread in online social networks is very similar to those on other social netwoks, hence information, ideas or trust can be transferred among the members [2]. It is of great importance to understand the mechanisms that govern the spreading process, which is called as network diffusion or cascade by researchers [3]. This spreading processes are usually triggered by some special individuals known as "influentials" who can easily affect the decisions of people around them. Their influence spreads like a virus, affects everyone in their reach and through friends of friends they can influence individuals beyond their reach [4].

The problem of identifying the most influential individuals on an online social network (as well as social networks in general), so that the number of influenced people is maximized is called the "Influential Selection Problem (ISP)". Domingos and Richardson [5, 6] are the first reseachers on this field by studying the calculation of the network value of the individuals as an algorithmic problem. Kempe et al. [7] use the term "Influence Maximization" for the same concept and apply popular mathematical diffusion models to formulate it as a discrete optimization problem. They show ISP is NP-hard and a greedy approach can provide good approximations due to the fact that the influence spread function under the proposed diffusion models is monotone and submodular.

Although the greedy algorithm is easy to apply and successful in identifying the influentials within a network, the stochastic nature of the problem brings out excessive computational burden as the network size increases. The objective function of the influential selection problem, which is the expected number of influenced nodes in the network can not be calculated exactly, but estimated through Monte Carlo simulations. Online social networks usually consist of millions of members (nodes) and connections (edges), causing a scalability issue while estimating the influence measure. Upon this fact, many researchers propose new methods and improvements on the greedy algorithm to solve the scalability and performance issue. Leskovec et al. [8] develop a cost-effective lazy forward algorithm which prevents the calculation of influence function for all nodes. Chen et al. propose degree discount heuristics [9] which is much faster than the greedy algorithm. Following that they develop a maximum arborescence model to efficiently solve the influential selection problem for large scale social networks [10]. Sheldon et al. [11] and Kumar et al. [12] work on the underlying discrete stochastic optimizaiton problem. In the former study, a sample average approximation technique is applied, whereas in the latter a Lagrangean Relaxation scheme is proposed to tackle the problem. Xu et al. [13] use semi-definite programming techniques to determine the group of influentials, contrary to the previous reserachers who identify them one by one. Finally Kim et al., propose a scalable and parallelizable approximation algorithm called Independent Path Algorithm (IPA) and claim that IPA outperforms all the previous methods in terms of computation speed [14].

Most of the new research focuses improving the scalability and solution time performance of the influential selection problem. In this study, the solution quality is the primary focus. The influential selection problem is proven to be NP-hard [7]. Even calculating the objective function is shown to be #P-hard [10]. Thus, improving the solution quality is a challenging issue. Given a feasible solution found by any of the available methods in the literature, our aim is to improve the solution quality. It is recently shown by Zhang et al. [15] that social networks

display a community structure. There are many small-sized communities in social networks which are connected to each other through a few number of common individuals. Building upon this observation, a neighbourhood search within the communities to identify better influentials is a viable approach. For this purpose, we applied a popular local search heuristic, variable neighbourhood search (VNS) metaheuristic, on the influence maximization problem.

The rest of the paper is organized as follows. In the next section, we further detail the influential selection problem (ISP) and the information diffusion processes. In Section 3, we explain our solution approach and analyze the details of VNS heuristic In Section 4, we report experimental results. Finally, the last section concludes the paper.

2 Influential Selection Problem

The ISP is formulated on a directed and weighted network $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{W})$. Here $\mathcal{V}$ the set of nodes, which are the individuals or members of the social network with a size of $|\mathcal{V}|$=V. $\mathcal{E}$ represents the set of edges or arcs in the social network that corresponds to any kind of a connection scheme in the online social network, i.e. being friends, followers, common-link sharers. Finally, $\mathcal{W}$ represents the weight set, i.e. each link $(u, v) \in \mathcal{E}$ from node u to v has a corresponding weight $w_{uv} \in \mathcal{W}$. These weights show the level of influence of individual u on v. The number of edges in the network is $|\mathcal{E}|$=$|\mathcal{W}|$=E. The undirected network version of the same problem can be considered as well. For instance two individuals being friends can be represented as a single edge (u, v) in an undirected network, where as it requires two edges (u, v) and (v, u) to represent the friendship relation in a directed network. The edge weights can be equal if the influence relation is assumed to be symmetric or they can be unequal when this relation is unsymmetric, which is a more realistic scenario.

2.1 Diffusion Models

Information diffusion or propogation mechanisms are well-studied topics under the sociology discipline and two general models are generally accepted in the analysis of online social networks. The first model is the linear threshold model and states that a node will be influenced if the fraction of its influenced neighbours are larger than a certain threshold [16]. In other words, a node v is influenced by each neighbour u by a weight w_{uv} such that those weights satisfy $\sum_{u \in N(v)} w_{uv} \leq 1$. The linear threshold model initialy determines a threshold θ_v for every node uniformly at random from the interval $[0, 1]$. θ_v is a node specific property and shows how easily a node can be influenced by its neighbours. Given a random choice of thresholds and an initial set of active nodes $\mathcal{V}_a$, the diffusion process runs deterministically in discrete steps. These steps may represent a small time interval such as taking an action in social media after seeing a friend's action. In step t all nodes that are active in step $t-1$ remain active, and any node v for which the total weight of its active nieghbours is at least θ_v are also activated; i.e. $\sum_{u \in N(v)} w_{uv} \geq \theta_v$. This process continues until no more nodes can be activated.

The second diffusion model considered is the independent cascade model [17], where a node u that is influenced or activated in step t has only one chance to influence any neighbour v successfully with a probability w_{uv}. These probabilities are independent of each other and if node u is unsuccessful, it has no chance to influence node v any more. When a node u activates its neighbour v in step t, then in step $t+1$ node v tries to activate all of its inactive neighbours. The process again starts with an initial active set of nodes $\mathcal{V}_a$ and runs until no more activations are possible.

There are many different variants of these diffusion models available in the literature, each

one representing a different real life phenomenon, but in both diffusion models there are certain common basics. Firstly, a node u is said to be active or influenced if it does the action of interest or accepts the idea that is shared to it. Whenever a node is active, it can not deactivate and stays as an active node all through the process. Secondly, when a node becomes active, it tries to influence all of its neighbours without any discrimination.

2.2 Mathematical Models

For both diffusion models the influence of a set of nodes $\mathcal{V}$ is the expected number of active nodes at the end of the process and is shown by $\sigma(\mathcal{V})$. The influence maximization problem tries to identify the initial active set $\mathcal{V}_a$, such that when diffusion process starts with this initial active nodes, the expected influence $\sigma(\mathcal{V}_a)$ is maximized. There are different versions of the problem, where the initial active set may have a fixed size k; or there may be a fixed budget B and a cost c_v associated with initially activating a node v. In this study, the fixed size version of the problem is considered and the mathematical formulation of the ISP is as follows.

ISP:

$$\max z = \sigma(\mathcal{V}_a) \tag{1}$$

$$\text{s.t. } |V_a| = k \tag{2}$$

$$V_a \in \mathcal{V} \tag{3}$$

In ISP, the objective function (1) maximizes the expected influence. The first constraint (2) sets the number of initially activated influentials to k and the last set of constraints (3) dictates the initial set of influentials to be a subset of the set of nodes $\mathcal{V}$.

Kempe et al.[7] prove that ISP is NP-hard, thus it is gruelling to find the optimal solution for large problems within a reasonable duration. Even calculating the objective function $\sigma(\mathcal{V}_a)$ by itself is a very complicated. Chen et al. [10] show that calculating the objective function can be reduced to the s-t connectivity problem and it is known that this problem is #P-hard. Roughly speaking, if NP-hardness is about the complexity of finding an optimal solution to the problem, then #P-hardness is about counting all possible solutions to the problem and it is considered to be as hard as or even harder than the former [18].

3 Methods to Solve ISP

The difficulty of exact calculation of $\sigma(\mathcal{V}_a)$, the influence of an initial active set, and the condition that ISP is NP-hard, so no polynomial time algorithm is available for solving them, forces the researchers to develop alternative methods. Kempe et al. show that the influence function $\sigma(\mathcal{V}_a)$ is submodular and monotone when the underlying diffusion mechanisms are linear threshold and independent cascade. Thus the natural greedy heuristic developed by Cournejols et al. [19] has an approximation guarantee of $(1 - 1/e)$ of the optimum solution, where e is the base of natural logarithm. Kempe et al. use Monte Carlo simulation to determine the expected value of the influence function, which generates an apporximation arbitrarily close to the limit of the greedy heuristic, i.e., $(1 - 1/e - \varepsilon)$.

The greedy heuristic basically determines one node v at each step, which maximizes the expected influence if node v is added to the set of active nodes $\mathcal{V}_a$. The algorithms complexity

is in the order of $O(kVAR)$, where V and A are the size of the node and arc sets as defined previously. k is the size of the initial active set and R is the number of runs of the Monte Carlo simulations to obtain a good estimate of the expected influence of the initial active set. It is reported in [8, 10] that for large networks the greedy algorithm's running time performance is poor. Hence, many researchers developed new methods or improvements on the greedy algorithm to improve the running time performance.

In this study, without ignoring the importance of solution time performance, we aim to improve the solution quality. Although greedy algorithm guarantees a solution within 63% of the optimum solution, there can still be a huge gap between the greedy solution and the optimum. Recent researchs show that many complex networks, especially social networks display a community structure [15]. In other words, there exists small sub-networks within the larger network that a group of nodes are connected to each other very densely and the number of connections to the nodes that are outside of this sub-group of nodes is comparibly small in size. Due to its simplicty, the greedy approach can not identify this structure of the network. Hence applying a neighbourhood search to identify better influentials within the communities is a viable approach. To benefit from this fact, we applied variable neighbourhood search (VNS) metaheuristic, which is one of the most well-known local search heuristics [20] on the influential selection problem.

3.1 Variable Neighbourhood Search on Influential Selection

Variable neighbourhood search (VNS) is a popular technique to solve combinatorial optimization problems. The basic idea of VNS is a systematic change of neighbourhood within a local search. In our case, we use VNS to obtain such initial active nodes that, we improve the solution quality of IMP.

The basic steps of VNS is described in Algorithm 1.

Algorithm-1 VNS Algorithm

1. Initialize. Select neighbourhood structures N_k, for
   $k = 1, ..., k_{\max}$, find an initial solution x;
2. Repeat the following until stopping condition is met:
   - 2.1. Set $k = 1$;
   - 2.2. Repeat the following steps until $k = k_{\max}$;
   - 2.2.1. Shaking: Generate a point x' at random from the k^{th} neighbourhood of x
   - 2.2.2. Local Search: Apply some local search method with x' as initial solution, denote with x'' the local optimum
   - 2.2.3. Move Or Not: If this local optimum is better than the incumbent, move there and continue the search with $N_{1,}$ otherwise set $k := k + 1$;

In our case, every VNS solution is a set of active nodes and the local optimum is the active nodes set to initialize the diffusion process with the maximum expected influence.

3.1.1 Initial Solution

We apply two initial solution strategies to start the neighbourhood search algorithm. Our first strategy is to apply the greedy heuristic to obtain a good initial solution. Our second strategy is just choosing the initial active nodes randomly until there are k initial nodes.

3.1.2 Determining a Local Minimum

VNS starts with an initial set of active nodes and at each step of VNS, we move to another feasible active nodes set. We make a local search around this new set and if we find a better solution, we update our active nodes set. To determine the local maximum we need to calculate the objective function of the corresponding problem for each feasible set. Hence, for each set we run Monte Carlo simulation runs to determine the influence function $\sigma(\mathcal{V}_a)$.

3.1.3 Neighbourhood Structure of VNS

At each step of VNS algorithm, we create candidate neighbourhood sets by using certain moves. Our moves are SWAP to keep the solution feasible, i.e. there are k influentials. The neighbourhood size parameter $N_k = 1, 2$, so in each move one or two nodes are swapped, added or dropped.

In the SWAP move we randomly drop one or two nodes and then add one or two new ones so that the total number of selected nodes is unchanged. The number of neighbours generated by each of these three moves is a parameter, which we call r. In our experiments we set $r = 3$, so that a total of three neighbours are created at each step. The stopping condition is the repetition of the main step 30 times without any improvement in the best solution.

4 Experimental Results

In this section, we report the solution quality and CPU time of our heuristic methods. We compare our heuristic results with the solutions generated by the greedy method and by the two popular centrality methods: degree-centrality and betweenness centrality. The solution found by our method is shown as z_{VNS}. Also z_G, z_D and z_B are defined as the best objective values of the greedy method, degree centrality method and betweenness centrality method, respectively. We also compare the average CPU times of these four methods. We test different scenarios for the VNS algorithm with a combination of two types of influence diffusion models : linear threshold and independent cascade and two types of initial active set strategies: random and greedy. Thus, in total there are four different scenarios to be experimented.

Our testbed consists of the available data from arXiv, which is the same source used in the experimental study of [7, 10]. In this data set each node is an author and the arcs show the co-authorship relation. There are a total of 37,154 nodes and 231,584 arcs. All experiments are carried out on Dell PowerEdge 2400 with two 64-bit, 2.66-GHz Xeon 5355 Quad Core processors and 28GB SDRam memory, operating within Windows 2008 Server environment.

In the independent cascade model the propogation probability is assumed to $p = 0.01$. For the Monte Carlo Simulations the sampling size $R = 20,000$. The size of the initial active node set ranges from $k = 1$ to $k = 50$.

The results for four scenarios are displayed in Figure 1-4. Observe that, when the initialization method is greedy VNS heuristic outperforms all the other three methods, however when a random initializaiton is applied the solution quality worsens. In the first case, the average performance improvement in the solution quality is 4.9 % over the greedy method.

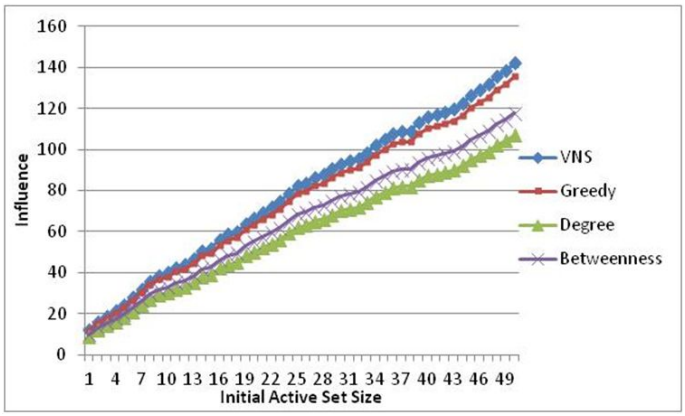

Figure 1: Influence Spread of different algorithms (IC, greedy initialization)

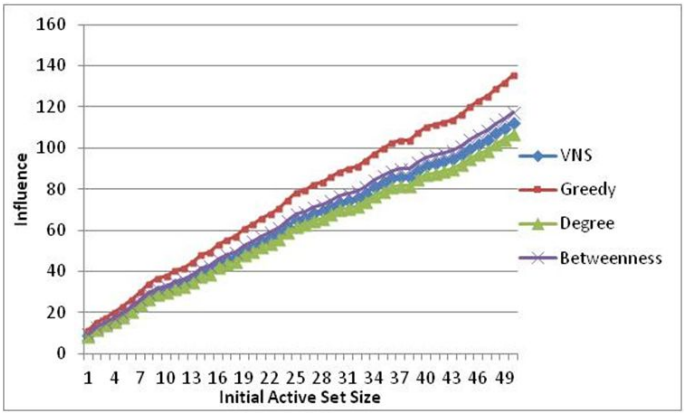

Figure 2: Influence Spread of different algorithms (IC, random initialization)

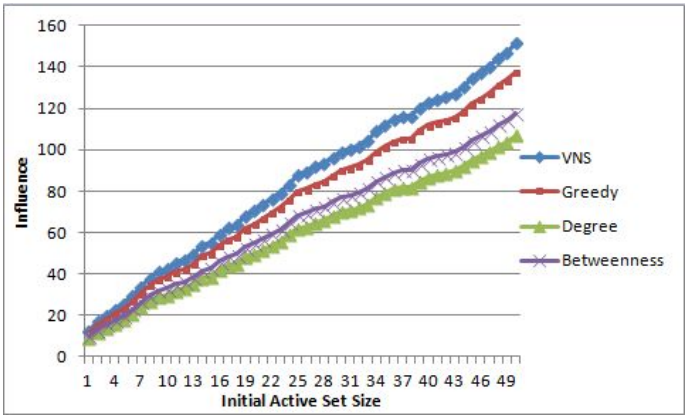

Figure 3: Influence Spread of different algorithms (LT, greedy initialization)

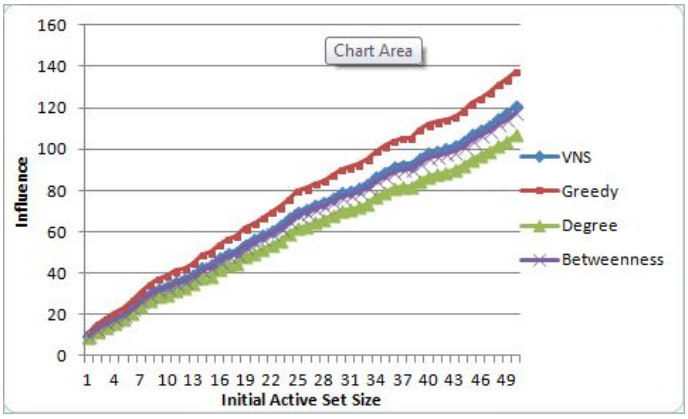

Figure 4: Influence Spread of different algorithms (LT, random initialization)

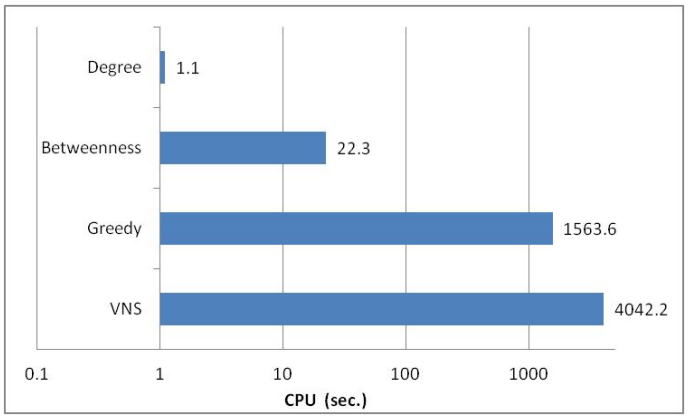

Figure 5: Comparison of average running time of different algorithms

Our experiments corrseponding to the other scenarios are continuing and will be added to the camera-ready version of this paper. We also compared the CPU performances of our heuristic method with other three methods. Since the degree heuristics are very simple they have the smallest computational durations. Our method is slower than the others due to extra computations, which is expected. As the network size increases the percent difference in terms of duration between our method and the greedy method diminishes. The results are displayed in Figure 5.

5 Conclusions

In this work, the influential selection problem on complex social networks is considered. We proposed a Variable Neighbourhood Search heuristic to improve the solution quality, which has been neglected by researchers until now. We tested two popular diffusion models, which are the linear threshold method and the independent cascade method. Experimental results indicate that VNS algorithm using the greedy method in the initialization phase can improve the solution quality with extra computational burden. As a direction for the future study more efficient greedy methods available in the literature can be incorporated to the solution scheme. Also different versions of the diffusion methods can be experimented. Finally, some other metaheuristics can be applied to improve the solution performance.

Acknowledgments: This research has been supported by TÜBİTAK (The Scientific and Tech-

nological Research Council of Turkey) under the Grant no: 1507-7130101

REFERENCES

[1] Heidemann, J., Klier, M., Probst, F., 2012, Online Social Networks: A Survey of a global Phenomenon, Computer Networks, Vol.56, 3866-3878 (2012)

[2] Newman, M.E.J., The Structure and Function of Complex Networks, SIAM Rev. 45, 167-256 (2003)

[3] Rogers, E.M., Diffusion of Innovations, Free Press, NY (2003)

[4] Brown, D., Hayes, N., Influencer Marketing: Who Really Influences Your Customers?, Butterworth-Heinemann Publication (2008)

[5] Domingos ,P. and Richardson, M., Mining the Network Value of Customers, Proc. of 7th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining, 57-66 (2001)

[6] Richardson M., and Domingos, P., Mining Knowledge Sharing Sites for Viral Marketing, Proc. of 8th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining, 61-70 (2002)

[7] Kempe, D., Kleinberg, J.M. and Tardos, E., Maximizing the Spread of Influence Through a Social Network. Proc. of 9th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining, 137-146, (2003)

[8] Leskovec, J., Krause, K., Guestrin, C., Faloustos, C. VanBriesen, J., Glance, N.S., Cost-effective Outbreak Detection in Networks, Proc. of 13th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining, 420,429 (2007)

[9] Chen, W., Wang, Y., Yang, S., Efficient Influence Maximization in Social Networks ,Proc. of 15th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining, 199-208 (2009)

[10] Chen, W., Wang, Y., Yang, S., Scalable Influence Maximization for Prevalent Viral Marketing in Large-scale Social Networks,Proc. of 16th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining, 1029-1038 (2010)

[11] Sheldon, D., Dilkina,B., Elmachtoub, A., Finseth, R. Sabharwal, A., Conrad, J., Shmoys, D., Allen, W., Amundsen, O., Vaughan, B., Maximizing the Spread of Cascades Using Network Design, Proc. of the 26th Conf. on.Uncertainity in Artificial Intelligence, 517-526 (2010).

[12] Kumar, A., Wu, X., Zilberstein, S., Lagrangian Relaxation Techniques for Scalable Spatial Conservation Planning, Proc. of the Twenty-Sixth Conference on Artificial Intelligence, 309-315 (2012)

[13] Xu, K., Guo, X., Li, J., Lau, R.Y.K., Liao, S.S.Y. Discovering Target Groups in Social Networking Sites: An Effective Method for Maximizing Joint Influential Power, Electronic Commerce Research and Applications, 318-334 (2012)

[14] J. Kim, S.K. Kim, H.Yu, Scalable and Parallelizable Processing of Influence Maximization for Large-Scale Social Networks, Proc. ICDE Conference, 266-277 (2013).

[15] Zhang, X., Zhu, J., Wang, Q., Zhao, H., Identifying Influential Nodes in Complex Networks With Community Structure, Knowledge-Based Systems, 74-84 (2013)

[16] Granovetter, M., Threshold Models of Collective Behaviour, The American Journal of Sociology, 83(6), 1420-1443 (1978)

[17] Goldenberg, J., Libai, B., Muller, E., Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth, Marketing Letters, 211-223 (2001)

[18] Bodlaender, H.L., Wolle, T., A Note on the Complexity of Network Reliability Problems, Technical Report, Inst.Information and Computing Sciences, Utrecht Univ., TR., UU-CS-2004-01, (2004)

[19] Cornuejols, G., Fisher, M., Nemhauser, G., Location of Bank Accounts to Optimize Float, Management Science, 23 (1977)

[20] Hansen, P. Mladenovic, N., Variable Neighbourhood Search, Computers& Operations Research, vol.24, 1097-1100 (1997)

[21] J.T. Oden, T. Belytschko, I. Babuska, T.J.R. Hughes, Research directions in computational mechanics. *Computer Methods in Applied Mechanics and Engineering*, **192**, 913–922, 2003.

[22] J.H. Argyris, M. Papadrakakis, L. Karapitta, Elastoplastic analysis of shells with the triangular element TRIC. M. Papadrakakis, A. Samartin, E. Oñate eds. *4th International Colloquium on Computation of Shell and Spatial Structures (IASS-IACM 2000)*, Chania, Crete, Greece, June 4-7, 2000.