University of Szeged
Doctoral School of Computer Science

Network Based Data Oriented Methods for Application Driven Problems

Summary of the PhD Dissertation

by

László Hajdu

Supervisor:

Dr. Miklós Krész

Szeged
2021

1 Introduction

Networks are amazing. If you think about it, some of them can be found in almost every single aspect of our life from sociological, financial and biological processes to the human body. Even considering entities that are not necessarily connected to each other in a natural sense, can be connected based on real life properties, creating a whole new aspect to express knowledge. A network as a structure implies not only interesting and complex mathematical questions, but the possibility to extract hidden and additional information from real life data. The data that is one of the most valuable resources of this century. The different activities of the society and the underlying processes produces a huge amount of data, which can be available for us due to the technological knowledge and tools we have nowadays. Nevertheless, the data without the contained knowledge does not represent value, thus the main focus in the last decade is to generate or extract information and knowledge from the data. Consequently, data analytics and science, as well as data-driven methodologies have become leading research fields both in scientific and industrial areas.

In this dissertation, the author introduces efficient algorithms to solve application oriented optimization and data analysis tasks built on network science based models. The main idea is to connect these problems along graph based approaches, from virus modelling on an existing system through understanding the spreading mechanism of an infection/influence and maximize or minimize the effect, to financial applications, such as fraud detection or cost optimization in a case of employee rostering.

2 Basic Definitions and Notations

Most of the algorithms and methods in the dissertation are based on undirected or directed networks (graphs). To define an *undirected (directed) network*, let $G(V, E)$ be an ordered pair where V is the set of *nodes* or *vertices*, and E is the set of *edges* that are unordered (ordered) pairs of nodes. The vertices or nodes in a graph, so the elements of set V can be connected by edges. If $u, v \in V$ and e is an edge that connects u and v nodes, e can be denoted by $e = (u, v)$, $e_{u,v}$ or $e(u, v)$, and u and v are *neighbours* or *adjacents*. Let d_v or $deg(v)$ called *degree* be the number of the edges that are connected to the node v. In the case of directed networks, the order of u and v with the edge $e(u, v)$ reflects the direction of the edge. In this case, the edge $e(u, v)$ is *outgoing* or *out edge* of node u and *incoming* or *in edge* of the node v. Moreover, if G is a directed graph, then we distinguish for each vertex v the outdegree $d_{out}(v)$ and indegree $d_{in}(v)$ by a straightforward way. The neighbours of a node in a directed network can be divided into two groups according to the direction of the corresponding directed edge. The *out neighbours* are connected to node v by an edge that points from v, while the *in neighbours* are connected to v by an edge pointing towards the v.

Furthermore, to understand the content of this thesis book it is important to define the idea of *complete graphs*. For a simple network G (i.e. no loops and multiple edges are allowed) with $|V| = n$, the minimal number of edges is 0 in an *empty graph* while the maximal size of the E is $n(n-1)/2$ in a *complete graph* which is mostly denoted by K_n. The *clique* in a graph G is a subgraph that forms a complete graph. Consequently, *k-clique* is a clique that has exactly k number of nodes. In a graph G, the *maximum clique* is the largest clique, which contains the largest number of nodes, and the *maximal clique* is the clique that cannot be extended or enlarged. In directed network, let d_{v_c} be the restricted out-degree of a node v in clique c which means the out-degree of a given node inside the clique. Based on the definition the directed clique contains all directed edges from v_1 to v_2 where $d_{v1_c} > d_{v2_c}$ and no directed loops as well as every node in a clique has a different restricted out-degree. Regarding to our definitions it is also important to discuss the definition of the closed walk. A *walk* on a network is a finite or infinite alternating sequence of edges that are connecting sequence of nodes, so $e_{v_i,v_1}, e_{v_1,v_2}...e_{v_{j-1},v_j}$ is a walk that goes from v_i to v_j, however, it is called *trail* if the edges are distinct. The walk is *closed*, if the v_i equals to v_j.

2.1 Community Structure in Networks

The complex or real networks often have a structural peculiarity, which implies that the nodes can be grouped into sets such that the vertices inside the set have dense connection structure. This property is called *community structure*. However, in the literature there are lot of different definitions to describe the communities, they are mostly interpreted as a dense subgraph, where the nodes are more connected to each other than in the other parts of the network. The communities can have two different types, based on the number of the groups a node can belong to. In a case when a vertex can be the part of only one community, it is called *non-overlapping community* or *cluster*. On the other hand, if a node can belong to multiple communities, we are talking about *overlapping communities*.

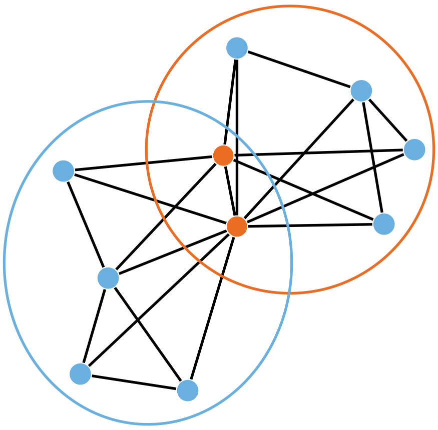


Figure 1: Example of overlapping community.

The extraction or creation of the previously defined groups or dense subgraphs from a network is called *community detection*, and it can be a computationally difficult task. The history of clustering and community detection goes back to the 70's, thus several methods can be found in the literature nowadays both for clustering [27, 25, 24, 26, 19, 10] and for overlapping community detection [11, 22, 30, 28, 3, 21]. However, in real life networks, overlapping communities are more useful due to the fact that an individual or entity can belong to multiple groups or communities. Therefore, in the dissertation we deal with overlapping communities. A good overview on different community detection methods can be found in [9].

2.2 Diffusion Models

Networks and their structure can support and effectively express the modelling of different real life processes. The origin of diffusion modelling comes from epidemiological use cases, where processes can express the spread of different viruses or diseases. However, the process can come from many other areas, for example next to medical processes, we can model the spread of information, bankcruptcy, churn, or as it is discussed in the dissertation, even the spread of behavior or mood between different individuals.

Regarding diffusion models, the first point which is needed to be shortly discussed, is the topic of compartmental models. These models discretize the mathematical representation of infectious diseases in a way, that the population is grouped into different compartments based on their state. However, it is important to note, that classical diffusion models were not used on a network, but the spread was defined based on differential equations. If we consider the compartmental models, the first important model that has to be discussed is the SIR model, which was proposed by Kermack and McKendrick in 1927 [17]. In the SIR, the individuals from the population can be in three different states. The state S represents the susceptible group, which means that if an individual has infectious contact and it is susceptible, it will be transitioned into the infectious compartment. The I denotes the group of infectious individuals that can infect people from the susceptible compartment in a case of a contact. The R means removed, so they were in state I, but are healed and immune or deceased. A simplified version of this concept, the SI model, which is dealing only with the S and I compartments is applied for the results of Section 3.1 . A good overview of general and network based epidemiological models can be found in [8, 18]

The another model that is important regarding the content of this dissertation is the Independent Cascade (IC). The mathematical formulation of the Independent Cascade was defined by Kempe, Kleinberg and Tardos [6]. In this model, the nodes can be in *active* or *inactive* state, where active means that the node is infected, while inactive means the opposite. The nodes can only go from passive to active state and the infection can be realized by the edge probabilities. To define the Independent Cascade let $G(V, E)$ be a network where for $\forall(v, u) \in E$, there is a probability $p(v, u)$ that represents the connection strength and $0 < p(v, u) \leq 1$. The model is iterative and starts with an initially infected node set A_0. To understand the process, let us define the A_i as set of the nodes that have become infected in the $i-1$-th iteration by the set A_{i-1}. The model terminates if the set of the newly infected nodes is empty. The expected number of the infected nodes in a case of the given A_0 set is denoted by $\sigma(A_0)$. Concerning the Independent Cascade model, an optimization problem called *infection maximization* [7] can be defined, where the objective is to find the initial infected set A_0 with a predefined size such that the expected number of infected nodes $\sigma(A_0)$ is maximized. The computation of the $\sigma(A)$ is #P-hard problem [5] but with simulation any precision can be reached.

3 Main Topics of the Dissertation

The main parts of the dissertation can be grouped into 5 different topics that are the following:

- Community Detection and Infection Modelling on Public Transportation Networks.
- Infection and Communities.
- Uplift Network Model for Targeted Interventions.
- Temporal Network Analytics for Fraud Detection in the Banking Sector.
- Network Based Crew Rostering.

The results of the research connected to these topics will be discussed briefly in this section of the thesis book.

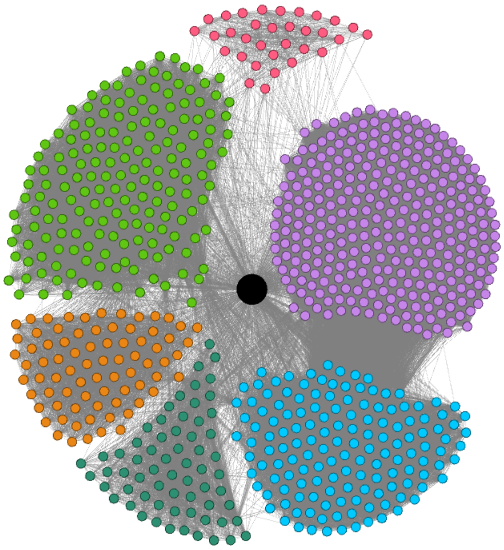


Figure 2: Contact network.

3.1 Community Detection and Infection Modelling on Public Transportation Networks

We introduced a new methodology which aims to examine the nature of the community structure on public transportation networks, as well as to analyze existing systems from epidemiological point of view. The methodology creates two additional networks from the so called contact network (Figure 2.) that describes the original connections between the passengers. The definition of the different networks used in this chapter can be seen in Table 1.

Table 1: Network definitions used in this chapter

	Contact network	Transfer network	Community network
Nodes	Passengers	Atomic passenger groups	Passengers traveling on at least two vehicle trips
Edges	Passengers are connected if they are physically present on the same vehicle trip at the same time (undirected)	Atomic passenger groups are connected if they share a passenger (directed)	Passengers are connected if they are traveling together on at least two vehicle trips (undirected)
Attributes	Contact duration and start time	Number of transfer passengers between atomic groups	Connection strength measuring the similarity of the travel patterns between passengers

Our methodology creates the transfer network in three steps. In the first step, we partitioned our original G(V,E) contact network into subgraphs along vehicle trips. In the second and third step of the creation, we detected maximal cliques using the Bron-Kerbosch algorithm [1] and built the transfer network F (see Table 1.). The resulting network provides a good representation of the system usage, so at the end of this stage we detected the frequent vehicle trip combinations and the passenger group movements. These information can be very useful for the public transportation company, in order to monitor the habits and the behaviour of the individuals on the vehicle trips.

In the second part of the corresponding chapter, we proposed a novel weighted network structure called community network denoted by H, which is using a novel link based metric called connection strength. The connection strength defines the edge weights in the community network and takes into account the number of transfers a pair of passengers makes together, but also penalizes the usual travel patterns between them, since these patterns do not indicate strong hidden connection. We also validated our method with real world scenarios from the public transportation system of Twin Cities Minneapolis.

3.1.1 Epidemic Spreading Risk Application

One application of identifying the communities within the transit network is infrastructure security. Understanding passenger communities enables more efficient and accurate tracking of infectious disease spread on a public transportation system, since one of the main challenges in modeling epidemic spreading is accurately mapping the relationships between individuals traveling on the same vehicle. In order to simulate an epidemic outbreak on the contact network, we used the well-known discrete compartmental susceptible-infected (SI) model. Similarly to the procedure in [2], we randomly selected 100 passengers from the network to be initially infected. Due to the probabilistic nature of the simulation model, we ran the SI infection model $k = 10000$ times to quantify the likelihood of each node being in an infectious state at the end of the fifth time step. The output of our method was the ranking of the vehicle trips, where infection is most likely to appear.

3.2 Infection and Communities

We examined the connection between community detection and diffusion models. Our main objective was to introduce a novel community-based infection maximization method, which also can be used as a benchmarking system in order to rank different community detection methods. In the first part of the chapter after we discussed the Independent Cascade Model and the Greedy method to maximize the infection in a network, we defined a new community-based reduction techniques. Let $G(V*) \subset G(V)$ be a reduced node set, where the greedy algorithm chooses from $G(V*) \setminus A_0$ in every iteration. Let $f(v) : v \to Z$ be a function that assigns an integer to every node. The nodes are ordered based on their $f(v)$ value, and the nodes with the highest $f(v)$ scores can be included in the set $G(V*)$.

As we discussed earlier, communities are dense, connected subgraphs, where the nodes have stronger connection with each other than with the other parts of the network. If the subgraphs are dense enough, infection or influence can spread among the nodes more easily. Nodes that are connecting different communities in multiple dense subgraphs have a special position in the network, which can be used for influence maximization. Let $f_c(v) : v \to Z$ be a function for the community value, that assigns to each node the number of communities it belongs to. The idea is to reward nodes that are in central position between the dense subnetworks. Two reduction based methods were introduced in the dissertation. The steps of the first method are the following:

1. Detection of overlapping communities in the network.

2. The method computes $f_c(v) : v \to Z$ for each node .

3. Creating the $G(V*)$ reduced search space using the $f_c(v)$ values.
4. Infection maximization with the greedy method using $G(V*)$.

We also introduced a simplified heuristic, where we skip the fourth step of the previously defined method.

3.2.1 Evaluation of the Method on Undirected Networks

First we tested our method on 1080 undirected networks generated by Andrea Lancichinetti and Santo Fortunato [20]. To compute the community value and also benchmark existing methods, we used eight different undirected community detection methods from the literature. The Table 2. shows how the different algorithms were working on the networks compared to the original greedy heuristic.

Table 2: Number of times methods provided the best results. Three variants of the algorithm are shown: the unmodified heuristic with the selection set reduced to 20%, to 10% and the simplified heuristic.

	20% selection set	10% selection set	Simplified heuristic
Greedy(Kempe)	207	243	335
CPM	71	61	2
COPRA	10	1	0
GCE	0	0	0
Infomap	732	769	737
MOSES	2	0	0
OSLOM	2	2	6
SBM inference	27	0	0
SLPA	29	4	0

3.2.2 Further Analysis of the Concept: Directed Hub Percolation Method

After we tested the basic idea on different undirected community detection methods, our objective was to examine the methodology on directed networks using two directed community detection methods, where we can extract additional structural information about the nodes. We introduced the Directed Hub Percolation Method (DHPM), which is an extension of an existing undirected method [3]. The structure of the method is similar to the original one, except we added a parameter to the end of the algorithm as well as changed the undirected elements to directed. Using the DHPM, we defined a new value in directed networks called the hub value. Just as in a case of the community value let $f(v)$ $f : v \to Z$ be a function, which assigns a number to each node. Let $f_{hv}(v)$ be a function that assigns the hub value h_v to the nodes of the network indicating how many directed cliques contain the node. We compared our DHPM method with the Directed Clique Percolation [29] and the original Greedy method on artificially generated and real networks.

3.3 Uplift Network Model for Targeted Interventions

We have developed a new Uplift Network Model, which is able to optimize targeted interventions on a network, in order to reduce global negative effects. The model was discussed as a part of a interdisciplinary research, where the objective was to improve the mental well-being of individuals in organizations and companies with targeted psychological interventions. The model itself is an extension of the Generalized Independent Cascade Model [4]. To define the Uplift Network Model formally, let $G = (V, E)$ be an undirected or directed network, where $\forall(v, u) \in E$ edge has a $p(v, u)$ probability where $0 < p(v, u) \leq 1$ and $\forall v \in V$ node has an a priori $v^{apriori}$ and an uplift v^{uplift} probability, where both of them are between 0 and 1, i.e. $0 \leq v^{apriori} \leq 1$ and $0 \leq v^{uplift} \leq 1$. The $v^{apriori}$ probabilities are used, when the node did not get intervention, while the v^{uplift} means the probability of the node after intervention. The objective of the optimization is to search for a set of nodes where the change of the probabilities from $v^{apriori}$ to v^{uplift} maximizes the difference from a reference simulation, where there was no intervention, so it minimizes the global effect.

3.3.1 Psyhological Use Case

The model was tested on a social network, which was created based on real world data from 14 Norwegian nursery homes. The mental level of the employees and the results during the use case was measured using the WHO-5 questionnaire [31]. We also compared the effect of targeted interventions compared to a random interventional strategy. The results of the psychological use case can be seen in Table 3.

Table 3: Comparison of the WHO-5 percentage score mean increase per person

Number and percent of interventions	Mean increase in score after random administrations	Mean increase in score after targeted administrations	Difference between targeted and random
10 (3.6%)	0.74	0.83	0.09
20 (7.2%)	1.43	1.61	0.18
50 (18.0%)	3.55	3.94	0.38
100 (36.0%)	6.85	7.64	0.79
200 (71.9%)	13.58	14.64	1.06

Despite the fact that the mood can spread only at a low level between the individuals, our method was able to increase the total score of the questionnaire with only targeted strategy.

3.4 Temporal Network Analytics for Fraud Detection in the Banking Sector

We have introduced a method, which is able to detect special patterns in temporal networks. Since the inspiration has come from a bank and the defined pattern is highly connected to this area, the algorithm was introduced through a financial use case, where the objective was to detect suspicious patterns connected to money cycle transfers. Let a transfer cycle be a closed walk where:

- The repetition of the nodes along a cycle is allowed, while for the edges it's not allowed

- Let $t_e^0, t_e^1 ... t_e^{n-1}, t_e^n$ be the sequence of the timestamps along a valid cycle, where every $t_e^i \leq t_e^{i+1}, i = 0...n-1$ so the timestamps are in ascending order.
- Let a_e^0 be the amount of the first transaction and a_e^i be the amount of the further transactions with $i = 1...n$. Then $a_e^0 \cdot (1-\alpha) \leq a_e^i \leq a_e^0 \cdot (1+\alpha)$, meaning that the difference between the amount of the first transaction and other transactions is bounded by the previously given real value $0 \leq \alpha \leq 1$. The α value is a parameter of the method and is defined by the specific user requirements.

The figure 3. shows the transactions of 5 different clients. Based on the cycle definition, along the 1-2-3-1 closed walk the timestamps are in ascending order and it is a cycle if the $\alpha \geq 0.1$ nevertheless for example the 5-2-3-4-5 is not because of timestamps and amounts.

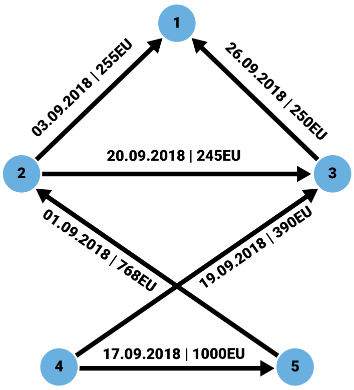


Figure 3: An example of a graph which contains a transactions cycle.

First the algorithm orders the edges based on the timestamps to have the ordered list of the starting transactions. In the second step of the algorithm, we used a modified Depth First Search algorithm to collect the cycles. Each cycle is identified by its starting transaction, so after each cycle is found from the given transaction, the algorithm set the state of the edge to visited, so that the further cycles cannot use the given transaction.

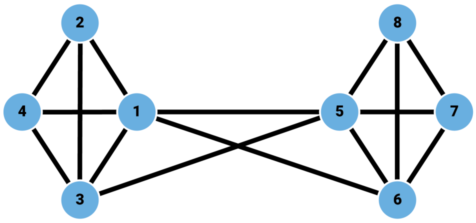
3.4.1 Real-world Use Case on Bank Transaction Network

We tested the algorithm on the real anonymized data from 2016 that contains the transactions of a Hungarian Bank. Our method is part of a fraud detection system which has been working at the bank since 2018. Based on the feedback of the bank, the introduced algorithm detects 2-4 suspected fraud activities in every month, as well as it detected some fraud-related client base in the last years.

3.5 Network Based Crew Rostering

We introduced a novel network based heuristic for the crew rostering problem. To define the problem let C be the set of workers or employees at a company and S be the set of shifts that has to be carried out. The aim is to assign the crew members to the shifts with minimal cost. Consequently, let $f = S \rightarrow C$ be an assignment, where the shifts are covered by the workers and exactly one worker is assigned to each shift. The basic employment cost is based on workers' expected worktime defined by their contract. So in the optimal case, every worker will work according to their expected worktime. However, in case of working over the expected worktime, employers have to provide extra salary for this overtime. Therefore, the $cost = \alpha * overtime + \beta * employment\ cost$, where α and β are pre-defined weights. The objective is to find a solution that minimizes the previously defined formula, and meets with different regulations that are tipically defined in national laws.

Figure 4: Example of the conflict graph corresponding to two consecutive days.



Our solution defines the optimization problem on a conflict graph where the nodes are the shifts and they are connected by an edge if they can not be carried out by the same person. An example network can be seen in Figure 4. The introduced method is a two step graph coloring algorithm, so it colors the previously defined network using the set of employees as colors. The steps of the method are the following:

1. Initial rostering.

a. Estimate the number of workers.

b. Generate days-off patterns.

c. Build a conflict graph.

d. Color the graph.

2. Tabu Search to improve the solution.

3.5.1 Evaluation of the Method and a Real-world Use Case

During the testing, we analyzed the behaviour of our method both on artificially generated and real-life input data. The real data comes from the Public Transportation Company of Szeged. The results of the method have been compared to the results of an integer programming model on small input sizes. As a conclusion it can be said, that our method were able to outperform the IP model regarding to the running time, as well as it reached the theoretical lower bound (best possible solution) in many cases producing a fast and efficient solution for our original problem.

4 Summary of the Author's Contribution

The following section summarizes the main results of the dissertation. Table 4. shows the relation between the thesis points and main chapters of the thesis, as well as the corresponding publications.

I. We introduced a method to analyze the vulnerability of a public transportation system, using the passenger co-traveling network of the actual city. In the first part of the chapter, we presented a new community definition for networks that expresses the passenger connections on public transportation and examined the usage of the system giving frequent trip changes, helping the public transportation company to improve their service. After the passenger community detection and system analysis, we examined the different scenarios in a case of an epidemic and detected the critical and risky vehicle trips in order to minimize the spread on the public transportation network in the city. We introduced and tested our method with the public transportation network of Minneapolis. Our publication connected to this topic can be found in [14].

II. An interesting concept was introduced that is trying to connect the infection models and the community detection. The main idea of this thesis point is based on a fact that an infection on a network spreads easier inside a community, than in the other parts of the network. We introduced a new methodology where on one hand, the results of different community detection methods are able to improve the efficiency of the infection maximization problem, on the other hand, the infection maximization can provide a reliable benchmark system in order to rank the different community detection methods. After the concept was introduced, we tested the methods from both sides on real and randomly generated networks, and gave comparison on existing community detection algorithms. The methodology is studied in [12, 13]

III. We introduced an infection maximization model for targeted intervention which is based on the Generalized Independent Cascade [4], applying a use case from psychology. In the first part of this chapter, we focused on the new infection maximization model and on a method to minimize the network effect in a case of an existing influence or infection. The second part of the chapter examined one of the possible use cases, where the objective is to improve mental well-being in organizations with targeted psychological interventions. The model was tested on a social network that was created based on employee data of Norwegian nursing homes. Our research connected to this chapter can be seen in [23].

IV. A new method was introduced to analyse temporal networks in order to detect fraud in the banking environment. The motivation of this research comes from the financial sector, where the detection of special network motifs can reveal fraud like activity in the system. The main objective of this part was to give a general view on financial fraud detection and introduce a new cycle detection method, that is able to detect special money transfer cycles in a transaction network. The method was tested on real life transaction data from 2016 and it has been using in a Hungarian bank since 2018. The corresponding paper to this research can be found in [15].

V. We solved the crew rostering problem with a heuristic using a network to represent the search space of the optimization problem. The goal of our method is to minimize the cost of the employment and to create a solution that meets with the given regulations. The two step method was tested on real life problems from a public transportation company, as well as on randomly generated inputs. In the end of the chapter, we compared the results of our method with the results of a mathematical model. The results are studied in [16].

Table 4: Correspondence between thesis points, publications and chapters.

	[14]	[12]	[13]	[23]	[15]	[16]
I.	●					
II.		●	●			
III.				●		
IV.					●	
V.						●
Chapter	2.	3.	3.	4.	5.	6.

Selected references

[1] C. Bron and J. Kerbosch. Algorithm 457: Finding all cliques of an undirected graph. *Commun. ACM*, 16(9):575–577, Sept. 1973.

[2] A. Bóta, L. Gardner, and A. Khani. Identifying critical components of a public transit system for outbreak control. *Networks and Spatial Economics*, 08 2017.

[3] A. Bóta and M. Krész. A high resolution clique-based overlapping community detection algorithm for small-world networks. *Informatica*, 39:177–187, 01 2015.

[4] A. Bóta, M. Krész, and A. Pluhár. Approximations of the generalized cascade model. *Acta Cybernetica*, 21:37–51, 01 2013.

[5] W. Chen, C. Wang, and Y. Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1029–1038, 09 2010.

[6] K. David, K. Jon, and T. Éva. Maximizing the spread of influence through a social network. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 137-146, 07 2003.

[7] K. David, K. Jon, and T. Éva. Influential nodes in a diffusion model for social networks. *Lecture Notes in Computer Science*, 3580:1127–1138, 07 2005.

[8] O. Diekmann and J. Heesterbeek. Mathematical epidemiology of infectious diseases: Model building, analysis and interpretation. *Wiley Series in Mathematical and Computational Biology, Chichester, Wiley*, 01 2000.

[9] S. Fortunato. Community detection in graphs. *Physics Reports*, 486, 06 2009.

[10] M. Girvan and M. Newman. Community structure in social and biological networks. *proc natl acad sci*, 99:7821–7826, 11 2001.

[11] S. Gregory. Finding overlapping communities in networks by label propagation. *New journal of Physics*, 12(10):102018, 2010.

[12] L. Hajdu, A. Bóta, and M. Krész. Community based influence maximization in the independent cascade model. *2018 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 237–243, 2018.

[13] L. Hajdu, A. Bóta, and M. Krész. Evaluating the role of community detection in improving influence maximization heuristics. *(Under Review)*, 2021.

[14] L. Hajdu, A. Bóta, M. Krész, A. Khani, and L. M. Gardner. Discovering the hidden community structure of public transportation networks. *Networks and Spatial Economics*, 20:209–231, 2020.

[15] L. Hajdu and M. Krész. Temporal network analytics for fraud detection in the banking sector. *ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium*, pages 145–157, 2020.

[16] L. Hajdu, A. Tóth, and M. Krész. Graph coloring based heuristic for crew rostering. *Acta Cybernetica*, (4):643–661, 2020.

[17] F. M. Ivo. W.o. kermack and a.g. mckendrick: A seminal contribution to the mathematical theory of epidemics (1927). *A Historical Introduction to Mathematical Modeling of Infectious Diseases*, pages 59–87, 01 2017.

[18] M. Keeling and K. Eames. Networks and epidemic models. *Journal of the Royal Society, Interface / the Royal Society*, 2:295–307, 10 2005.

[19] B. Kernighan and S.-D. Lin. An efficient heuristic procedure for partitioning graphs. *Bell Syst. Tech. J.*, 49:291–307, 1970.

[20] A. Lancichinetti and S. Fortunato. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 80:016118, 08 2009.

[21] A. Lancichinetti, S. Fortunato, and J. Kertész. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11, 03 2008.

[22] C. Lee, F. Reid, and A. M. abd N. Hurley. Detecting highly overlapping community structure by greedy clique expansion. *arXiv:1002.1827*, 2010.

[23] D. Lipovac, L. Hajdu, S. Wie, and A. Q. Nyrud. Improving mental wellbeing in organizations with targeted psychosocial interventions. *Business Systems Research*, 11:86–98, 2020.

[24] U. Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17:395–416, 01 2004.

[25] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proc. Fifth Berkeley Symp. on Math. Statist. and Prob., Vol. 1*, 1967.

[26] F. Murtagh and P. Contreras. Algorithms for hierarchical clustering: An overview, ii. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 09 2017.

[27] M. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 69:026113, 03 2004.

[28] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435:814–818, 07 2005.

[29] G. Palla, I. Farkas, P. Pollner, I. Derenyi, and T. Vicsek. Directed network modules. *New Journal of Physics*, 9, 04 2007.

[30] M. Rosvall and C. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences of the United States of America*, 105:1118–1123, 02 2008.

[31] WHO. Wellbeing measures in primary health care/the depcare project. 2020.