STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LIX**, Special Issue 1, 2014
10th Joint Conference on Mathematics and Computer Science, Cluj-Napoca, May 21-25, 2014

DIRECT MARKETING OPTIMIZATION USING CLIENT RELATIONAL GRAPHS

LAJOS GYŐRFFY, ATTILA BALATON, AND ANDRÁS CSERNENSZKY

Abstract. In the present paper we give an introduction to some problems which occur at a bank and can be represented by graphs. We will demonstrate network building possibilities on specific data and apply them to improve prediction and/or replace the present methods of the sector. We give a detailed analysis of the corporate transaction graph and a retail client relational graph based on data of the OTP Bank. Our main result is the optimization of the response rates of Direct Marketing (DM) campaigns using the relational network (built on the known attributes such as common telephone number, same family name, etc.). According to our new approach - in contrast to the traditional banking methods - we use neither clients' personal data itself nor account behaviours, but the structure of networks to find a better segmentation. Networks can also give us forecasting models: we restrict to sending DM offers only to certain clients who met with some graph theoretical requirements. This approach raised the DM offer acceptance rate by 1.5-2 times of the average of previous DM campaigns.

1. Introduction

Studying networks is one of the most developing area in science. In medical sciences it comes up when we are talking about infection propagation models, or about the role of connections between cells. In sociology we analyse the social networks or interactions between groups of people. The importance of information which is hidden in the graph, requires such types of analyses. The same is true in business life, especially at banks. For the used terminology see the Appendix. Knowing the structure of banking networks and relationships between entities can give us many information which one could not get by traditional approaches.

Received by the editors: May 1, 2014.

2010 *Mathematics Subject Classification.* 91D30, 05C82.

1998 *CR Categories and Descriptors.* J.4 [**Social and behavioural sciences**]: Subtopic – *Economics*; E.1 [**Data structures**]: Subtopic – *Graphs and networks*.

Key words and phrases. Direct Marketing optimization, client relational graph, collaboration graph, banking methods, banking graphs.

Unfortunately, there are lots of difficulties when one works with banking data. First of all, it is not simple either to measure or to evaluate results. One cannot measure anything, but only what is available. The quality of data is probably poor, not all items are available and a large part of data is confidential. Due to these facts the main tool of data mining problems is rather pattern recognition than stating and proving theorems.

In the literature, there are approaches to use graph methods which are to a certain extent similar to our method, but those do not deal with the Direct Marketing optimization in banks. Currently banks use random methods, regression models and decision trees based purely on demographical and behavioural data. Still, there are some problems in which banks use graph based methods in practice. For example the spread of credit default, fraud detection or churn prevention can be tackled efficiently via graphs and networks.

In case of credit default, infection models are applied on a directed network drawn from the transaction data. The idea is that the neighbours of a client who fell into credit default might share the difficulties through the transaction edges. The generalizations of Independent Cascade model [12, 13, 15] is applicable to estimate the spreading of defaults; for detailed information see Bóta et al. [4, 6, 8].

Detecting possible churn one may take into account the communities of a client graph, too. A community means a set of points among which there are more edges than expected, see in the Appendix. If a community has more inactive customers than the average, then the bank is advised to take preventive actions.

The case of fraud detection is similar to that of churn. Members of a fraudulent community are suspected also to be fraudulent.

In the previous methods the difficulty is to find the communities. There are several available algorithms to find those; we used the Sixtep software for that. For the problems relating to communities see Bartalos and Pluhár [1], Bóta et al. [5], Griechisch and Pluhár [14].

To solve the Direct Marketing campaign optimization problem, we tried to incorporate all of these approaches. While the direct application of those did not yield convincing results, the process was useful to learn the structure of the problem. The clients who have neighbours in the relational graph give better responses to the DM offers than the isolated ones, which is the main experimental result of our research.

2. Graphs in a bank

First, we have to answer some questions about the building blocks of our model. How can we construct graphs in a bank? What can be the vertices and

the edges? A natural approach is that nodes represent some kind of clients. They can be retail or corporate clients; we might include the municipalities or not; the SME sector or a part of these. For the edges we can have many choices. Here we considered only two of those, a corporate transaction graph and a retail client relational graph. In the following subsections we will describe those in more detail.

2.1. **Transaction graph.** The most natural links between clients are the transactions, i. e. money transfers. An edge goes from the first client to the second, if there was at least one payment sent in that direction in the given period. Furthermore, weights can be associated to edges, for example the number of transactions, the sent amount, the sent relative amount, etc.

2.1.1. *Transaction graph at the OTP Bank.* In the above described way we have built up a transaction network out of the OTP Bank Corporate transaction database. In this graph vertices represent corporate clients, while edges represent connections between clients if the transactions between them are considered to be relevant. Since the transactions themselves are directed, the network is directed as well. The decision about which connections should be included in the graph is based on experiments. In a filtering process we used the following three criteria:

- Frequency of transactions: the average number of transactions in a month.
- Amount of transactions: the average transacted amount in a month.
- Relative amount: average incoming amount from one company divided by the total income from all of the partners.

If a connection did not satisfy a specific set of criteria, then this connection is not considered as an edge of the graph. Since the business partners of a company may change dynamically, we have built the network on a one year period: from November 2012 to October 2013. The basic statistics of the resulting transaction graph as follows. It has approximately 68,000 vertices, and 106,000 edges.

The largest component consists of 63,000 points, while the size of the second largest is only 14. There are also 1750 components with only two elements. Some nodes has large number of links and a bit less than the half of the nodes are connected to one node. The average degree is about 3.09, while the maximal degree is 2018. There are three nodes with more than 1000 connections and 52 with more than 100. The number of triangles is 9970 and the clustering coefficient is 0.00165, which measures the local density of the edges.

This means that - in contrast to the *social graphs* - if (A, B) and (B, C) edges are present, then the conditional probability of the existence of (A, C) grows less than in case of a relational graphs.

According to the data, the degree sequence of the transaction graph might a power-law distribution ($p(x) = Kx^{-\alpha}$) with $\alpha = 2.457$, hence it might be a scale-free (small world) network. For the computation we used the method advised by Newman [17].

2.2. **Social graphs.** Beside the transaction graph, another possibility to build graphs is to use personal data of clients.

One of the most well known social graphs in mathematics are the collaboration graphs. This is a graph which models some social network where the vertices represent participants of that network (usually individual people) and in which two distinct participants are joined by an edge if there is a collaborative relationship between them of a particular kind. The two most well-studied collaboration graphs are the Erdős collaboration graph and the Hollywood graph, see Appendix and [19].

2.2.1. *Client relational graphs in banks.* To construct collaboration-like graphs in a bank, we follow the collaboration graph models. In the Erdős collaboration graph two authors are joined, if there is a paper they co-authored. Similarly in the Hollywood graph two actors are neighbours if there is a movie in which both of them played.

At a bank there are several attributes assigned to every client. E. g. the clients mother's name, phone number, employer address, etc. We join two clients if there is an attribute in which their values coincide. Knowing these attributes, one can build a graph by each attribute, therefore many graphs may arise. Then one might take the arbitrary union of those graphs to construct a network. We applied that building method at the OTP Bank.

2.2.2. *Building a retail client relational graph at the OTP Bank.* We have constructed the relational network from the OTP Bank Retail client database out of the 1st of January 2008 to the 11th of April 2013. Our nodes are the retail customers, who have had at least one credit-application (personal loan, trade loan, mortgage loan or credit card) in the given period. We had approximately 1.8 million applicants during that time period. The edges are defined by the common attributes. At first we used 17 properties to connect vertices. Some of those are for example the applicant's address, family name, mother's name; same phone number, employer address or tax ID, etc. We also used *derived variables* such as same street (at the same settlement), or same address combined with the family name.

So we join two clients if they live in the same house or in the same street. Another edge is drawn if they work for the same company or they have common family name. It means if there are seven people with a common attribute, we get a 7-clique by this method. After taking the union of these 17 graphs, we get cliques which may have common parts, edges or just vertices. All in all, the relational graph is the union of a large number of cliques.

Obviously, it is not fruitful to join two or more customers if they live in the largest street of Budapest, they have a common but very familiar name or they work for a big company where there are lots of employees. As a rule of thumb we do not join more than 10 people with the same attribute. With this restriction we reach our goal, namely not to connect all employees of the OTP Bank or all Smiths, but to join the co-workers of a small company or the ones who have a common but rare name in Hungary, which means that they are possibly relatives.

2.2.3. *The properties of the graph.* Our final graph has approximately 1.3 million non-isolated nodes from the 1.8 million clients. The number of edges are between two and three millions.

We also analysed the components, the degrees and the clustering coefficient as in the case of the transaction graph. The largest component is a giant one, consisting of one million points, while the size of the second largest is 55. There are about 80,000 components with only two elements. Here the maximum degree is 30 and there are 350,000 nodes with one degree. So, in contrast to the transaction graph, this degree sequence does not seem to have power law degree distribution. The average degree is about 4.07. There are more than three million triangles, and the clustering coefficient $C = 0.2060$.

In summary, we got two quite different networks (in size, structure, degree distribution, clustering coefficient, etc.). However, they have similar properties (average degree, a giant components, some isolated pairs), too.

To obtain these results, we used the Sixtep software [18] which beside providing the information of the graph parameters (components, structure, visualisation) can create communities and clusters, and can handle infection models which we use in the next section.

In the next section we present the Direct Marketing optimization problem and a solution for it which is based on the retail client relational graph and gave better response rates than the traditional methods used at the bank.

3. Direct Marketing optimization

One of the greatest challenge in a DM campaign is to find a (usually fixed size) target group with the greatest expected response rates. Here we could detect the effect of the network, namely those clients who are better embedded

into the networks of the bank (the vertices have high degrees), have a higher response rate. We stress that only the topology of network is considered, not the previous DM responses or other behavioural data. Merely, the role of the client-vertices in the network determines the response rate.

Before we go deeper to the possible methods, we describe the mechanism of a Direct Marketing campaign.

3.1. **DM campaigns.** When a bank decides to start a Direct Marketing campaign, they select a fixed number group of their clients (e. g. 10000), and send them direct mails. The question is, what would be the best target? We analysed personal loan and credit card campaigns. In both cases the bank sends a mail to the selected clients with an offer, which is valid only for the targeted people. The credit card DM envelope contains a plastic card, and instructions how to activate it. The offers have an expiration (45 or more days). Since a campaign is very expensive, it should produce the best possible acceptance rate. We do not have to consider the cost of a mail, since the number of offers is fixed. At the OTP Bank there are already methods for target selection. These use demographical and behavioural variables to make predictions for the response rate of every client. These do not use graphs directly, that is why we decided to introduce the network based methods.

3.2. **Evaluation of the campaigns.** Since we constructed our retail client relational graph from the data drawn from the 1st of January 2008 until the 11th of April 2013, we used the DM offer responses from the 12th of April 2013 to the end of November 2013 to avoid the intersection of the time periods. We combined the personal loan and credit card DM campaigns, and we got more than 350,000 clients who received a DM offer. This proved to be big enough to draw conclusions. Part of these clients accepted the offer, and claimed the personal loan or the credit card. Let this ratio be X. That is,

$$X = \frac{\text{clients who accepted the offer in the evaluation time period}}{\text{clients who got an offer in the evaluation time period}}.$$

We define the DM-acceptance value of an offered client with one if she accepted the offer and zero otherwise.

3.3. **Optimizing DM campaigns on the relational graph.** We tried two different approaches for the optimization, namely using infection models and communities. However, these approaches alone did not yield satisfactory results. Still, we discuss them shortly because they played a role to find our final method.

3.3.1. *First approach.* At first, we tried to apply a variant of the generalized Domingos-Richardson infection model on the retail client relational graph. The reason for this was that it worked well on the transaction graph, where the infection was the credit default ("bankruptcy") of a client. Unfortunately, the method here did not yield good results.[1] See again Domingos and Richardson [12] and Bóta et al. [4, 6, 8].

3.3.2. *Second approach.* The other method was to consider communities which has "infected" vertices. In the previous cases infection was an undesirable property: possible churn or default. Now a client is called infected by accepting the offer. We classified the communities (it contains infected nodes or not), then compared the response rates in both. Again, the rates were almost the same.

3.3.3. *Third approach.* However, we noticed a remarkable fact: the clients who were in a community had a higher response rate ($1.23X$) than the average of all clients (X). Considering the fact that only one third of the clients are in communities (out of the 350,000 people who have got a DM offer around one third was in a community) it gives useful restriction for the search space. According to our computation, had we sent DM mails only to the clients who are in a community (there are more than one third of the 1.8 million tested clients, too), the bank would get 1.23 times bigger profit on a single DM campaign.

Furthermore, the clients who are in some component of our graph (170,000 from the 350,000 DM-offered clients) has also a higher acceptance rate ($1.12X$) than the average (X). We found this phenomenon very useful. When calculating the response rates only for those points of degree more than one (130,000 clients) and then more than two (100,000 clients), we got the following ratios: The vertices with more than one neighbours in the graph have $1.19X$ response rates. For the clients, who have more than two neighbours, the average response rate is $1.23X$.

There are some possible explanations for this phenomenon. If a client is joined to another client, for example by an "employer address" edge, than they are possible co-workers, they might have shared information and decided together to accept the new credit card. The same holds for the family name

[1]The hypothesis was that the acceptance of the DM offers spreads similarly to the defaults on the other graph. We set the a priori infection values 1, if a client has already accepted a Direct marketing offer, and 0 otherwise. The edges were unweighted, so the infection spread along the edges with equal probability. Still the results were not better on the evaluating set than a random selection; measuring by the Gini-index the results was even not positive. It is either because of the different structure of the two graphs, or the difference between the credit default and the DM offer acceptance is bigger than we thought before.

or mother's surname (possible relatives), for the address (neighbours or flatmates), for the e-mail domain names (co-workers), etc. That is the more coincidences in data the more chance to accept the offer.

3.4. **Improving the graph.** After that, we made several attempts to improve response rates. Since adding new edges to the network was not efficient[2], the next idea was to reduce the number of connections and retain only the best types of edges. For doing that, we used two methods. The first was that we built 17 different graphs, one for each type of edge, and then looked at the response rates one by one.

The second idea was to run a logistic regression to measure which type of edges are the most relevant ones. Here the target variable was the DM-acceptance, and the explanatory variables were 17 dummies, one for each attribute. If a vertex was connected to an other vertex, for example by the common surname, the *surname* dummy variable got the value one, if not, it got a zero value. The same goes for the other 16 variables. Therefore every node got 17 zero-one flags, one for each types of edges. Then we used the PASW Modeller 13 software and applied the forward stepwise logistic regression method.

By the first selection method (selecting those types of edges which have the best response rates) we have found five types of edges useful: *the employer address, the employer tax ID, the address* and two other types of links.[3]

By the logistic regression method we have found eight dummy variables, which were significant at the 5% significance level. All of the above mentioned five important variables were among them. Two of the eight (*surname combined with address* and *surname*) had negative effect. This means that if two clients are joined by those type of edges, they probably accept DM offers less times than the average. One possible explanation for the last observation is that if there is one credit card or a personal loan in a family, they do not need an other.

Next we decided to use our five best variables to construct our second graph which gave us much better results than on the first with the 17 variables.

3.5. **Reduced client relational graph.** By the five types of edges we got 800,000 connected vertices and 1.5 million edges. However we have less clients in that graph, the DM acceptance rate is higher, $1.54X$. Out of the 350,000 clients in the evaluation set there are 75,000 non-isolated vertex in that graph.

[2]Our first idea was to create new edges. Since we have incorporated all possible properties what were available, we gave 100,000 and then 500,000 more edges randomly to the graph, but the efficiency dropped to $1.04X$.

[3]Because confidelity, we can not list the best variables, just as the exact value of the X ratio.

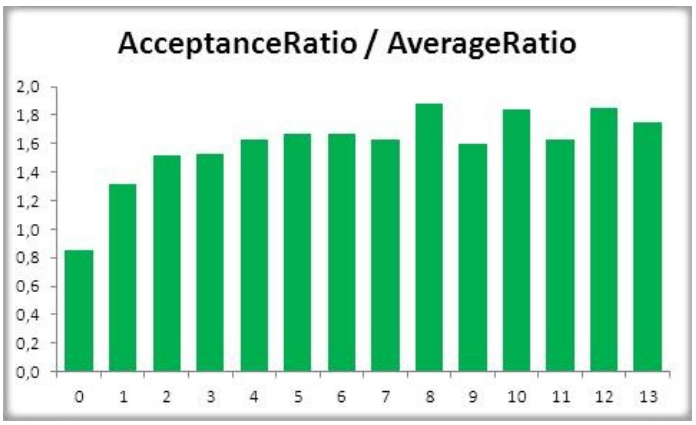

FIGURE 1. DM acceptance rates by the degrees of the reduced graph

This means that expectedly all of these 800,000 clients have 1.54 times higher response rate in average than in the set of all clients. Similarly to the bigger graph with 17 types of edges, we looked at the points with two and three minimum degree, and got by 62 and 65% better results on 56,000 and 44,000 clients than the average. On the Figure 1. we can see the acceptance rates of the 350,000 points from degree zero to thirteen. Most of the vertices have zero degree (these are the isolated nodes in the reduced graph), there are 275,000. From the other 75,000 points 20,000 has degree one, 12,000 have two, 10,000 have three, 7000 have four and there are more than 500 which have degree 13.

4. Methods and results

4.1. **A simple method for the optimization.** ¿From our previously defined three approaches we recur to the third one again: connecting clients having common attributes defines a graph. The clients, who are in that graph accept DM offers in much higher rate than the others. Even considering the clients with more than two or three neighbours gives us even better results.

The most natural method is the following: If we selected the target group from the 800,000 clients in the above defined reduced graph instead of all customers at the bank, we would get expectedly 1.54 times better response rates. If we restrict ourselves to those points having at least two or three degrees, we can get up to 1.65 times better rates.

The 1.54 and 1.65 times better responses give the bank a quiet big profit, but we could separate better these 800,000 (or even the 1.3 million) customers by a logistic regression. Furthermore, by using regression all clients get a DM score value which is also useful for the bank.

4.2. **Using logistic regression on the bigger client relational graph.** The reduced graph consists of only five types of edges, while the bigger graph have 17 types. There are two types with significant negative effect which are not in the reduced graph. That is why we chose the bigger one to run the logistic regression. We applied the same model that we used for variable selection. There the target variable was the DM-acceptance and the explanatory variables were the 17 dummies defined above and in the Appendix (surname dummy).

We partitioned the 170,000 clients (who are in the intersection of the DM offer evaluating set and the big relational graph) randomly to a 70-30% training and test set. We used again forward stepwise logistic regression method and obtained the same eight significant edge-variables as before. After getting the coefficients, it was possible to give a DM score value to the vertices of the test set (and to all 1.3 million clients in the graph).

Since we got the coefficients only on the 70% training set, then we could independently measure the results on the 30% test set which contained 51,000 clients.

We ordered the test sample (ascending) by the regression score values, and put them into 20 equal bins (5-5 percent in each). The whole test sample had an acceptance rate $1.12X$ which was the same as that was earlier on the 170,000 clients. Then we compared the real acceptance rates with the predicted ones.

4.3. **Results.** In the bin with the highest scores the DM acceptance rate was $2.37X$ in average. In the second it was $2.07X$ and in the third $1.77X$. In the 4th, 5th and 6th that was $1.63X$, $1.45X$ and $1.58X$. Between the 6th and the 7th there was a bigger difference, because in the 7th the average response rate was only $1.06X$. After the 9th we got only smaller values than X and in the last and worst bin the rate was $0.66X$.

Then we analysed the best k percent of the test sample. Following the best 5% with the $2.37X$ rates, the highest 10% had $2.22X$. The best 15% had $2.07X$, the best 20% had $1.96X$ and the highest 25% had $1.86X$. The best half of the test sample had $1.5X$ and as we mentioned above the whole test sample had an acceptance rate $1.12X$.

Counting the number of clients whose targeting resulted in more than $2X$ response rates we got the following. We had 170,000 analysed clients from the DM evaluation out of the 1.3 million nodes in the 17-variables-graph. The 15% of the 51,000 had more than twice larger acceptance probability than X. This result suggests that we can also select the 15% of the 1.3 million clients with this method and this results in the acceptance rate $2X$, too. There are a

bit less than 200,000 customers which is a very large part of the OTP's clients and which could give a huge profit to the bank.

4.4. **Future works at the bank.** The next step of this study is to test the method in a new set of customers. However, to introduce a new model in a bank is a huge project, mainly depending on the management. There is some chance that in a few months the bank will send out the DM mails according to our scores, and after the expirations of the offers we would see how the model works in real stress.

5. Summary

At first, we built a corporation transaction and a retail client relational graph out of the database of the OTP Bank. After the analysis of the networks, we focused on the Direct Marketing campaign optimization problem. We found a graph building and modelling process which results two times better DM offer response rates than the bank had earlier.

We emphasize once more that what differentiate these models from the previous ones. In our method we used neither the concrete demographical attributes of the clients (age, exact address or employers) nor the previous DM offer acceptances or any other behavioural data. We used only the structure of our created network and the degree of the vertices. This new approach of using networks can give more efficient DM campaigns and better results.

Acknowledgement

The publication is supported by the European Union and co-funded by the European Social Fund. Project title: ?Telemedicine-focused research activities on the field of Mathematics, Informatics and Medical sciences? Project number: TÁMOP-4.2.2.A-11/1/KONV-2012-0073

References

[1] I. Bartalos and A. Pluhár, *Közösségek és szerepük a kisvilág gráfokban.* Alkalmazott Matematikai Lapok **29** (2012) pp. 55–68.

[2] V. Batagelj and A. Mrvar, *Some analyses of Erdős collaboration graph.* Social Networks, vol. 22 (2000), no. 2, pp. 173–186.

[3] B. Bollobás: *Modern Graph Theory.* Springer-Verlag (2002)

[4] A. Bóta, A. Csernenszky, L. Győrffy, Gy. Kovács, M. Krész, A. Pluhár *Applications of the Inverse Infection Problem on bank transaction networks.* **Submitted** (2014).

[5] A. Bóta, L. Csizmadia and A. Pluhár, *Community detection and its use in Real Graphs.* Proceedings of the 2010 Mini-Conference on Applied Theoretical Computer Science - MATCOS 10 (2010), pp. 95–99.

[6] A. Bóta, M. Krész and A. Pluhár *Approximations of the Generalized Cascade Model.* Acta Cybernetica **21** (2013) pp. 37–51.

[7] A. Bóta, M. Krész and A. Pluhár, *Dynamic Communities and their Detection.*, Acta Cybernetica Volume **20** (2011) pp. 35–52.
[8] A. Bóta, M. Krész and A. Pluhár, *Systematic learning of edge probabilities in the Domingos-Richardson model.* Int. J. Complex Systems in Science Volume **1(2)** (2011) pp. 115–118.
[9] Z. Brys, B. Buda, A. Pluhár, *Hálózatkutatás a medicinában és határterületein.* Lege Artis Med. 2012 Jul, 22(6-7): pp. 445–449.
[10] Chaomei Chen, C. Chen. *Mapping Scientific Frontiers: The Quest for Knowledge Visualization.* Springer-Verlag, New York, 2003. pp. 94.
[11] A. Csernenszky, Gy. Kovács, M. Krész, A. Pluhár and T. Tóth, *The use of infection models in accounting and crediting.* Proc. of the Challenges for Analysis of the Economy, the Business, and Social Progress International Scientific Conference, Szeged, November 19–21, 2009.
[12] P. Domingos and M. Richardson, *Proc. 7th Intl. Conf. on Knowledge Discovery and Data Mining*, pp. 57–66. (2001).
[13] M. Granovetter, *American J. of Sociology* **83**, pp. 1420–1443. (1978).
[14] E. Griechisch and A. Pluhár, *Community Detection by using the Extended Modularity.* Acta Cybernetica Volume **20** (2011) pp. 69–85.
[15] D. Kempe, J. Kleinberg, and É. Tardos, *Proc. 9th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, (2003).
[16] M. Krész and A. Pluhár, *Economic Network Analysis.* In Encyclopedia of Social Network Analysis and Mining, Springer 2013.
[17] M. E. J. Newman, *Power laws, Pareto distributions and Zipf's law* Contemporary Physics, **46** (2005), pp. 323?-351 .
[18] *The Sixtep software* www.sixtep.hu.
[19] *Wikipedia: Collabration graph* http://en.wikipedia.org/wiki/Collaboration_graph.

Appendix: Definitions and notations

Graph or network. A graph $G = (V, E)$ consists of two finite sets V and E. The elements of V are called the vertices and the elements of E the edges of G. Each edge is a pair of vertices. If x and y are vertices and there is an edge x, y, then we say x and y are **neighbours**. The **degree of the x vertex** is the number of edges which contains x. By replacing the set E with a set of ordered pairs of vertices, we obtain a **directed graph**. A **weighted graph** associates a label (weight) with every edge in the graph. See more precisely Bollobás [3]. Sometimes we use network instead of graph, nodes instead of vertices and links or connections instead of edges.

Community. A community is a group of vertices that are densely connected to each other and sparsely connected to other vertices in the graph. See Bartalos and Pluhár [1], Bóta et al. [5], Griechisch and Pluhár [14].

Clustering coefficient. The C clustering coefficient measures the local density of a graph. Formally:

$$C = \frac{\text{3 times the number of triangles}}{\text{the number of two length paths}}.$$

Erdős graph. Collaboration graph of mathematicians, where the vertices of two scientist are joined by an edge whenever the scientists co-authored a paper together. See Batagelj and Mrvar [2].

The Hollywood graph. Collaboration graph of movie actors, also known as the Hollywood graph or co-stardom network, where the vertices of two movie actors are joined by an edge whenever the players appeared in a movie together. See Chen and Chen [10].

Client relational graph. The vertices of that graph are clients. There is an edge between two clients, if they have at least one common attribute. The 17 attributes that we examined were e. g. surname, mother's name, surname and mother's name, address, employer address, employer tax ID, phone number, same settlement and street, same e-mail domain, e-mail address, same name and mother's name, same family name and address, etc. The graph at the OTP Bank consisted of 1.3 million vertices and more than two million edges.

Reduced client relational graph. After examining the relevance of the 17 types of edges, we built a smaller graph from the five most relevant types of edges. It had 800,000 vertices and 1.5 million edges.

Transaction graph. The vertices of that graph are clients (e. g. corporate clients). The directed edges are the transactions from the payer to the payee. The possible weights can be the sent amount or the frequency of transactions.

DM-acceptance. Consider a customer who got a DM offer from the bank. The value of the DM-acceptance is one if she accepted and zero if not.

Surname dummy variable. Consider a vertex of the client relational graph. If it was connected to an other vertex by common surname, the *surname* dummy variable got the value one, if not, it got a zero value. The same went for the other mentioned 16 variables.

Response rate (acceptance rate) of a DM campaign.

$$X = \frac{\text{clients who accepted the offer in the given time period}}{\text{clients who got an offer in the given time period}}.$$

University of Szeged, Bolyai Institute, Aradi Vértanuk tere 1., 6720 Szeged, Hungary
E-mail address: lgyorffy@math.u-szeged.hu

Eötvös Loránd University, Faculty of Informatics, Pázmány Péter sétány 1/C, 1117 Budapest, Hungary
E-mail address: balcsi4@inf.elte.hu

OTP Bank Plc., Risk Analyses and Regulation Directorate, Babér utca 7., 1131 Budapest, Hungary
E-mail address: Csernenszkya@otpbank.hu