The use of infection models in accounting and crediting

András Csernenszky[1] *–Gyula Kovács*[2] *- Miklós Krész*[3] *– András Pluhár*[4] *– Tamás Tóth*[5]

Recently one of the main directions of data mining is the study and use of network data. Our research is concentrated on the network data about the links of big and medium size of companies that can be deduced from the bank transactions. The main goals are to develop models to predict churn and bankruptcy. We implement the Domingos-Richardson cascade model, and for the parametrization and evaluation we use the database of the OTP Bank. The results suggest that the developed system is capable of supporting a wide range of applications of network problems such as churn, bankruptcy, campaign management, information diffusion etc.

Keywords: Graph mining, Influence models, Consumer value

1. Introduction

Earlier results on network processes suggest that it is worthwhile to examine the transactions between corporate clients. We believed that if something happens with a company's supplier or purchaser, it obviously has some effect on its business partners. By using a transaction data between corporate clients, we modeled the spreading of Basel II default events on a bank's corporate portfolio. To simulate this process, we used the DR cascade model with appropriate parameters. Our results show that the bankruptcy forecasting can be greatly improved by this method, provided with a careful parameterization. The research is based on the OTP's corporate transaction database and the actual computations were done by the commercial graph-mining software of Sixtep Ltd.

[1] András Csernenszky, OTP Bank Hungary
[2] Gyula Kovács, Sixtep Ltd
[3] Miklós Krész, University of Szeged
[4] András Pluhár, University of Szeged
[5] Tamás Tóth, OTP Bank Hungary

2. Network research in the Corporate Banking

There is a lot of area for network modeling in the operations of a Corporate Bank. For the market development these are the acquisition, estimating customer value, forecasting attrition (performing at the OTP), product development and sales support and campaign optimization. There are also some applications for the credit risk department such as forecasting arrears, using networks for segmenting risk-groups, money laundering investigations, mapping the client's total customer-supplier relationship for business purposes. Here we concentrate on risk decisions especially the bankruptcy forecasting (till 2009 September performing at the OTP) – this is the main topic of this paper.

3. The Independent Cascade model

First we have to understand the role of modern network theory in Economics. The theory of graph is a well developed subject with plenty of beautiful theoretical results, applications and algorithms (Bollobás 1998). It was noted only recently, that some very large, but important graphs, the so-called *Small World graphs*, have characteristics that have not been explored in the classical studies, see Albert - Barabási (2002), or Newman (2003) for comprehensive introduction.

These graphs may arise by mapping the links among people or companies that already indicates their significance in the investigation of epidemics, spread of information or economical troubles and so on (Boguna et al. 2003), (Diekmann - Heesterbeek 2000). However, these works are based on SIS or SIR models; in which repeated infections and recoveries are both possible. In our case, predicting default, no recovery is possible, and a node might get infected *without* outer influence with an apriori probability, depending on the properties of the node itself.

So let us introduce our main tool, a model that tries capturing the process when such an effect propagates on a network. It was invented by Domingos – Richardson (2001). Originally it was proposed to support marketing decisions and determining client values. Nevertheless, Kempe et al. (2003, 2005) showed it is equivalent to a model given by Granovetter (1978), which shows these models can grasp a great variety of phenomena.

However, it is not quite obvious that it might be readily used in solving finance decisions. First of all, one has to define and built up a network (graph) with weighted edges that estimate the probability of one node infecting another. We shall give some of the details of that work later. The other problem is the arising computational issues. In order to get the approximated default probabilities, and consequently the expected value of default, one has to run a large number of Monte Carlo simulations on enormous size of network. As there had not been such applications available on the market, we decided to develop such code.

We give a sketch of the Independent Cascade model (Domingos-Richardson 2001). For a network *G* and probabilities assigned to the edges of *G*, and a set of *active (infected)* vertices *A*, the initial infection will affect the other vertices in the future. The time is modeled by discrete steps. In each steps the vertices that got the infection in the previous step might infect healthy vertices connected to them with probabilities assigned to the connecting edges. (The infection by different vertices is considered to be independent.) The process goes until there are newly infected vertices, otherwise halts. In a pseudo code:

1. Infected dataset = Active dataset A
2. The new infected vertices are these ones which are infected by the edges where one of the vertices is "active." The probability that a healthy vertex v stays healthy at the given step is the product $\prod_{v \in A} q_{uv}$, where u is a neighbor of v infected in the previous step. Here $q_{uv} = (1 - p_{uv})$ and p_{uv} is the infection parameter of the edge uv.
3. Vertices infected in the previous period = Active dataset;
4. If there is no new infection then STOP, else back to 2.

Note, that we have generalized the original Independent Cascade model such that not only an initial infected set can be given, but a probability distribution describing the a priori infection probability of the vertices.

4. Parameter estimation

Dataset:

There are several possibilities to compute the parameters within the model in order to get the best estimation for the bankruptcy probability of a corporate client. Our first task is to define the transaction dataset, in which lots of hard question have to be answered. Which is better to analyze all corporate clients with all of its transactions or just the debtors? Can we build up a meaningful weighted network drawing information out of our data warehouse? Certainly the vertices should be the clients, edges (some of the) transactions and the weights have to correspond to the transfers on those edges. Which length of time period should we monitor looking for transaction, and upon which assumptions should we declare there is an edge between the vertices representing two clients? How old basel2 default events have effects on a company's business partners, how long does it hold, and how big is it?

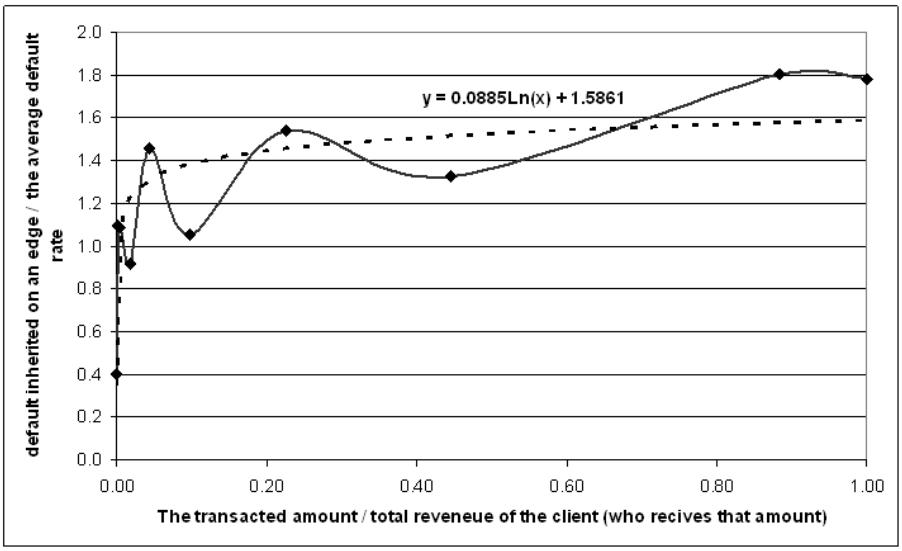
Our research clearly showed the followings. It was better for us to use just the debtors network because its homogeneity (it both means that the credit portfolio is divers from the account holders portfolio, and only just here has the default definition any sense). The other parameters were more flexible, but after some trials we

have reached the consensus that the business relation-graph should be based on at least 6 month, but not more than one year transactions data. Furthermore at most 3 months basel2 default event observation needed for the optimal result. In practice it means that to get a fresh report for November 2009, then the edges of the transaction graph should be based on the data from February 2009 to July 2009, and the default events are to be observed from August 2009 to October 2009. For the parameter estimations we used data from the previous year to re-measure its effect on a one year time-interval.[6] In the finally examined database there were 21 696 corporate client with 34 388 connections.

Vertices:
There are at least three approaches about what should be the initial influence distribution of the vertices. The first is to simply write 1 if the company is bankrupted in a given time period and 0 otherwise. (This corresponds to the original Independent Cascade model, where a vertex is either healthy or infected.) The other is to write the apriori default probabilities, the score values, to the vertices. For that case we generalized the model allowing fractional infection in the input. Our research confirmed that the second one yields better solution, increasing the efficiency on the reported target segment by 25 percent.
Another possible solution is to use the first method (0 and 1 to the vertices) and after multiplying the influences by the apriori probabilities. It has somehow different meaning; however it still increases the efficiency with 20 percent. In spite of its better classification we use for reporting the second method, because here the companies with higher influence rates have more (direct) bankrupted business partner and hence the interpretation and the acceptance of the results are easier.

[6] Our transaction graph comes from the data between January 2008 and Jun 2008, the default events from July 2008 to September 2008, while the measuring period is from October 2008 to September 2009.

Figure 1. Curve-fitting by a one variable function



Edges:

To get a useful model we also have to deal with the edge parameters. Here the main question is: 'How to model the probability of influence on the edges?' An obvious answer is to assign the average probability of influence on the edges. It means that for the bankrupted company we considered those companies that were directly connected to it and also bankrupted within the following twelve months, estimated the influence probability with the ratio of the bankrupted and all companies.

However, for the estimation of the infection probability there are better functions than a linear function of that ratio. Instead of the ratio, we took the ratio of the transacted amount among the clients per the total revenue of the client (the receiver) on that period. By scaling the function on the axis and the re-measured effect is on the other axis, we found that a logarithmic curve fits best to it;[7] see it on Figure 1.

To sum up the results, we can see the effects of different treats in Figure 2. For example if we write 0-1 to the vertices, but we use a one variable sigmoid function to estimate the influence of the edges we get 2.72 times more defaulted company at the top 10% percent of the portfolio than the average default rate.

[7] We tried other curve fitting methods like linear, polynomial, exponential, power fitting, etc…, but the R^2 here was the largest with 83%.

Figure 2. The classification power of various treatments

The top 10%'s default-rate per the average default-rate	Constant (as influence value) on the edges	One variable sigmoid function (as influence value) on the edges
Default/ non default flag (0,1) on the vertices	2,18	2,72
Writing score values on the vertices, and 1 if it is in default	3,82	4,25

Source: own creation

5. Results:

Using the parameterized Independent Cascade model, we found segments where the expected bankruptcy is 3-4 and even 10 to 12 times of the average. Of course there are some other standard variables that are used to assess the risk; such as the size of a given company, the importance its sector in the Hungarian economy (and even weather it is a municipality or not). Incorporating all these predictions, a monthly report is being built, which is installed to the OTP monitoring system that is used by the credit monitoring and the credit controllers department. According to the preliminary results, by the improved monitoring system the bank can significantly reduce the loss on bankruptcies.

References

Albert, R. – Barabási, A. L. 2002: Statistical mechanics of complex networks. *Reviews of Modern Physics*, **74**.

Boguna, M. - Pastor-Satorras, R. – Vespignani, A. 2003: Absence of epidemic threshold in scale-free networks with degree correlations. *Phys. Rev. Lett*, Vol. **90** Issue **2.** pp. 028701-1—4.

Bollobás, B. 1998: Modern Graph Theory. *Springer*, New York.

Diekmann, O. - Heesterbeek, J. A. P. 2000: Mathematical epidemiology of infectious diseases: Model Building, Analysis and Interpretation. *John Wiley & Sons*, New York.

Domingos, P. - Richardson, M. 2001: Mining the Network Value of Costumers.*7th Intl. Conf. on Knowledge Discovery and Data Mining*.

Granovetter, M. 1978: Threshold models of collective behavior. *American Journal of Sociology* 83/6, 1420-1443. p.

Kempe, D. - Kleinberg, J. – Tardos, E. 2003: *Maximizing the Spread of Influence though a Social Network*. Proc. 9th ACM SIGKDD Intl, Conf. on Knowledge Discovery and Data Mining.

Kempe, D. - Kleinberg, J. – Tardos, E. 2005: Influential Nodes in a Diffusion Model for Social Networks. Proc. 32nd International Colloquium on Automata, *Languages and Programming* (ICALP).

Newman, M.E.J. 2003: *The structure and function of complex networks*. Preprint cond-mat/0303516.