International Scientific Conference & International Workshop
Present Day Trends of Innovations 2012
28th – 29th May 2012
Łomża, Poland

DATA ANALYSIS IN PUBLIC SOCIAL NETWORKS

Lubos Takac [1] – Michal Zabovsky [2]

[1] University of Zilina, Faculty of Management Science and Informatics, Univerzitna 8215/1, Zilina, Slovak Republic , lubos.takac@fri.uniza.sk
[2] University of Zilina, Faculty of Management Science and Informatics, Univerzitna 8215/1, Zilina, Slovak Republic , michal.zabovsky@fri.uniza.sk

ABSTRACT

Public social networks affect significant number of people with different professional and personal background. Presented paper deals with data analysis and in addition with safety of information managed by online social networks. We will show methods for data analysis in social networks based on its scale-free characteristics. Experimental results will be discussed for the biggest social network in Slovakia which is popular for more than 10 years.

Key words: social networks [1], big data [2], scale-free networks[3], data security [4], graph theory [5]

INTRODUCTION

Social networks are phenomena of these days. Increasing number of users and personal character of data leads to the problem with security and unwilling loss of privacy. The advantages of using social networks are in free and fast communication with friends usually in a form of objects such as tweets, pictures, videos and texts. Another feature is creation of networks – friends, colleagues, family members. The problem comes with the possibility to trace activities, relations and communication together with limited possibility to manage critical activities such as delete of data. The information managed by different kinds of social networks is very interesting mostly for its potential to form overall personal profile.

The aggregation of information from public profiles is very useful for specific purposes. The problem with representation comes with complex structure of social networks. The graph theory allows us to use relatively good algorithms for analysis and data processing together with very good characteristics for specific types of networks. In this paper we will define an algorithm for getting data from semi-public social networks such as Facebook, MySpace and others. Analytical processing will use methods from graph theory and will help us with formulation of basic characteristics of created network. For the experimental evaluation we will use social network Pokec[1] which is in comparison to e.g. Facebook smaller in scale but has been provided for more than 10 years and connects over 1.6 million users. It offers

[1] http://www.pokec.sk, December 2011

chat, fast e-mail, e-mail, picture and video sharing services. Pokec is the most popular social network in Slovakia and is very popular in Czech Republic as well.

GETTING SOCIAL NETWORK DATA

For analyses of social networks some public, encrypted data sets are available. The encryption is necessary for the security reasons, thus names or usernames are replaced by artificial text produced by hash function. Identification of relations is possible together with some other characteristics such as age, but the association with real person is not possible. Pokec stores profile and contact information in a way that most of the profiles are public. The reason is that default settings in Pokec for contacts and profile information are predefined as a public.

The most interesting assumption comes with the fact that if constructed network is scale-free, then we can find relatively short path between every person in created graph. This can be addressed as a special data security problem. By using knowledge on scale-free networks and based on fact that the most of the contact and profile information comes publicly available, we created web robot for getting data from Pokec social network.

The algorithm for crawling social data is defined by following steps:

1. Insert user into queue. The user is identified by the nick name.
2. Take the first nick from the queue. If the queue is empty, algorithm ends.
3. Take a profile by using Pokec's URL together with the nick added at the end of URL. Get all not processed contacts from the profile and put them to the queue. Continue with step 2.

The web robot is in principle nick crawler and constructs social graph where nicks form vertices and friendship is represented by an edge. Web robot also getting data from profiles which are publicly available and stores them for future processing. The process of harvesting data was initially run locally, since the time consuming character, features for parallel processing were added. After then ten robots get all data in two days. Constructed network includes 1.6 million users (vertices) and more than 40 million relations (edges). More than 66 percent of users have their contacts published!

SCALE-FREE CHARACTER OF SOCIAL NETWORK

Scale-free network (SFN) is a network whose degrees distribution of nodes follows power law, at least asymptotically as defined by expression

$$P(k) \sim ck^{-\gamma} \quad (1)$$

where P(k) is a probability that degree of node is k, c is normalization constant, and γ is parameter whose value is typically in the range $2 < \gamma < 3$.

In the SFN majority of nodes has a low degree but some nodes (called hubs) have enormously high node degree (see Figure 1). Hubs keep network stable and resistant

to damage [1]. For example if we cancel randomly some edges or nodes in random network and in SFN, random network tends to split in sub graphs while SFN not. SFN is then resistant to random attacks. Only direct attack to hubs can split such network type. In other words, the core of the network is stable.

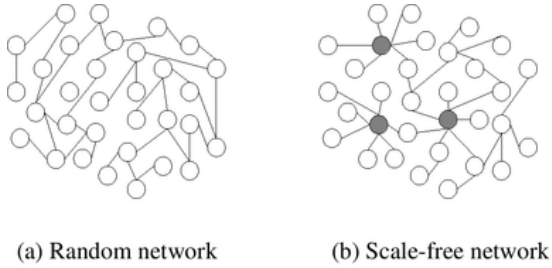

Figure 1 The difference between random (Erdös-Rényi) and scale-free network [5] is in nodes degree distribution. In scale-free networks there are some nodes called hubs which have exponentially more edges than the others nodes.

Based on previous definition, we suppose that if we create social network by web robot where 66% of users have public contacts we can recognize it as a significantly good sample. It is similar to random cancelation of edges in scale-free network when we canceled edges from 34% of nodes. It is very unlikely that this network splits into sub graphs and the core of network is broken.

Now we have to show that the network we get is scale-free. After analyzing nodes degrees of obtained network we have seen very nice power law distribution in figure 2. Hence Pokec network can be considered as scale-free.

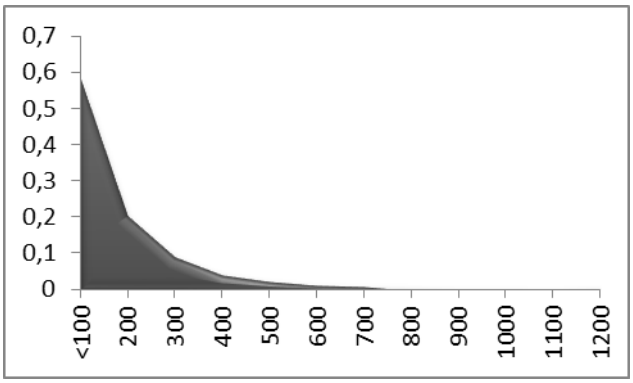

Figure 2 The figure shows nodes degree distribution of network Pokec obtained by a web robot. The chart clearly shows power law distribution which is specific for the scale-free networks. On the x-axis are abundances and on the y-axis are probabilities.

Another typical characteristic of SFN is very short distance between nodes. According to Cohen and Havlin, the diameter or the mean distance between nodes in typical SFN with $2 < \gamma < 3$ can be bounded by interval [2]

$$<\ln \ln N, \ln N / \ln \ln N> \quad (2)$$

where N is number of nodes in network.

During the experimental evaluation we made a test where we calculated the shortest distance between several randomly selected nodes. We could not calculate average distance exactly because of complexity of the task. Table 1 shows results of the experiment on 100 random nodes. The network was composed by 1 637 068 vertices and 46 171 896 edges.

Table 1 Table shows output from experiment of 100 random selected pairs.

	DISTANCE	TIME (min)
MIN	3.000000000	0.758416667
MAX	6.000000000	589.603600000
AVG	4.670588235	119.338999400

The average shortest distance calculated in experiment lies in appropriate interval for this network. Calculated distance is:

$$\ln \ln N < \textbf{diameter} < \ln N / \ln \ln N$$
$$2.66 < 4.67 < 5.38$$
$$N = 1\,637\,068$$

This shows that calculated average shortest distance between nodes from the experiment lies in the interval specified for typical scale-free network with respect to N.

DATA ANALYSIS

We did some analysis on obtained data from users profiles. At first we have to note that it is necessary to take into account that not all users fill they profile truthfully and some people should have more than one profile. But we believe that most of people fill their profiles seriously so the resulting statistics are not very distorted.

Table 2 shows basic facts such as overall user count, gender representation, number and percentage of users which published their friendships contacts and age. Figure 3 shows overall structure of users based on distribution by age.

Table 2 Basic facts about Pokec social network.

	COUNT	%
Users	1 637 068	100.00
Men	802 556	49.11
Women	831 725	50.89
Public contacts	1 088 838	66.62
Public age	1 125 734	68.88

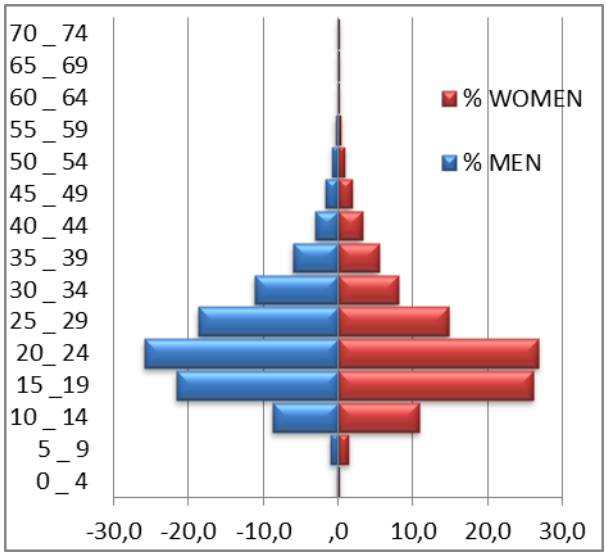

Figure 3 The figure shows age distribution of Pokec users. There are dominated group of young people between 15 and 30. Approximately 33% of women and 29% of men have their age nonpublic.

Friendships or relations between users are in Pokec oriented. It means that if user1 has friend user2, user1 need not to be a friend of user2. But most of all there are friendships on both sides. Most connected nodes (users with most friends) are called in scale-free networks hubs. Interesting fact is that hubs in Pokec are not people but commercial companies which advertise through this network.

Table 3 Average user friendship count and maximal number of friends.

	AVG	MAX
Has friends	22.181	13 840
Is friend	20.996	9 449

Pokec is holding and displaying last visit date and registration date for every profile. So there is an easy way how to get monthly visit rate or number of new registered user per month. Other profile data are divided into 58 categories such as region, physical and psychical characteristics, hobbies, job, languages skills, social and sexual preferences, smoking and alcoholic habits, etc. If a user writes anything to the profile, it is publicly available for all users. Only age and friendships can be nonpublic in Pokec. The basic recommendation is to create more powerful options for user to manage their personal data. We calculated from obtained data that the average profile is filled to 40.33 %. Based on this database detailed marketing research on user interests and preferences segmented for example by age, region, and gender can be done. The user profile also includes video and photo albums which are optional and can be secured by password but lots of them are public for all. We did not deal with these types of data.

CONCLUSION

We have shown methods for obtaining social and personal data from semipublic social network using scale-free character of such networks. Discussed current problems with security and privacy of data managed by social networks led us to formulation of recommendations in the form of simple rules for reduction its effects on

user's privacy. Experimental part of work, applied on the largest social network in Slovakia, disclosed potential problems with security together with possible ways of personal data processing by the third party. In future work we plan to repeat our test to see modification of network in time. In addition to this basic work we would like to check user's behavior changes according to security of personal information.

This work was supported by the Agency of the Slovak Ministry of Education for the Structural Funds of the EU, under project ITMS:26220120007.

REFERENCES

[1] BARABÁSI, A. L.: *Linked: How Everything Is Connected to Everything Else and What it Means for Business, Science, and Everyday Life.* 2002. ISBN 0-452-28439-2

[2] COHEN, C., HAVLIN S.: *Scale-Free Networks Are Ultra small.* Physical Review Letters, 7 Feb. 2003

[3] GAO, J., MENG, Y.: *Research on Degree Control Resistance to Frangibility Of Scale-free Networks.* Second International Conference on Communication Software and Networks, Feb. 2010

[4] TAKÁČ, Ľ.: *Data processing over very large databases.* Winter School MICT, Šachtičky Slovakia, 3-8 Jan. 2011

[5] ERDŐS, P., RÉNYI, A.: *On the evolution of random graphs*. Institute of Mathematics, Hungarian Academy of Science, 1960

[6] TAKÁČ, Ľ.: *Visualization of Large Multivariate Data Sets using Parallel Coordinates.*, Transcom, University of Žilina, 2011

[7] ALBERT, R., JEONG, BARABÁSI, A. L.: *Diameter of the World Wide Web.* Nature 401, 1999

[8] ZÁBOVSKÝ, M., ZÁBOVSKÁ, K.: *Big Data.* Proceedings of 12th International Conference System Integration 2010, 6-7 Sept. 2010

[9] MATIAŠKO, K., VAJSOVÁ, M., ZÁBOVSKÝ, M., CHOCHLÍK, M.: *Database Systems.* Edis, University of Žilina, 2008. ISBN 978-80-8070-820-7