RETRIEVING DATA FROM PUBLIC SOCIAL NETWORKS BASED ON SCALE-FREE CHARACTERISTIC

Ľuboš TAKÁČ
Department of Informatics, Faculty of Management Science and Informatics,
University of Žilina, Univerzitná 8215/1, 010 26 Žilina, Slovak Republic, e-mail: lubos.takac@fri.uniza.sk

ABSTRACT

This paper deals with retrieving data and data analysis of online social networks. We show an easy and universal approach to retrieve data and relations from any just partially public online social network. We explain the assumptions under which this approach works and we verify it on a practical example of online social network Pokec which has been popular in Slovakia for more than 10 years. We show some statistics and analysis of this online social network and the way how can be collected data abused. We recommended in this paper some operations and procedures for providers and for users to prevent or reduce this kind of security problems.

Keywords: *online social networks, scale-free networks, large data analysis, large database, web crawling, security of social networks*

1. INTRODUCTION

Online social networks (OSNs) are nowadays very popular. There are only a few people which are not connected to at least one of them. Others, including us and most of people are connected in. We are using OSN because of their advantages: free and fast communication with friends and colleges; sharing our news, photos, videos and opinions; finding out new friends with similar interests; etc. Some people are using this kind of communication also when they are unable to talk with someone about something face to face or when someone would like to present something "anonymously". By the way, on the internet is almost everything traceable and that is the reason why we are writing anonymously with quotation marks. There are also some disadvantages of OSNs which are people not aware of: loss of privacy; all communication and opinions can be stored after deleting and not only by provider of OSN but sometimes by anyone; all data about you and your friends can be abused. There is nothing new when we see in TV or newspaper Facebook communication of a criminal or a celebrity. Or when you go to a job interview and your potential employer knows about you more than is written in CV.

We show in detail algorithm how to obtain data from any just partially public OSN such as Facebook, MySpace etc. For the testing and analyzing purposes we choose Slovakian OSN Pokec (because of smaller scale) which has been provided for more than 10 years in Slovak and Czech Republic and connect over 1.6 million people[1]. It offers chat, fast mail, mail, sharing photos and videos. Pokec is the most popular social network in Slovakia and also is very popular in Czech Republic and nothing has changed on that after the Facebook has come.

2. GETTING SOCIAL NETWORK DATA

For analyzing social network data there are some public data sets which are encrypted because of security. For example nick names or user names are changed with hash function. You can identify which objects are in relation and other characteristics (average age, average contact count etc.) but you cannot identify or map object to concrete person. You can analyze data globally without private information about persons. Thus prepared data can be published by provider for analytical purposes. We get data in another way.

As skilled Pokec users we notice that most of profiles data and also user contacts are public. Maybe the reason is that the default setting in Pokec for contacts and profile information is public for all or maybe just people like to publish them. By using knowledge about scale-free networks [section 3] and the fact that most contacts and profile data are public we tried to get these data by web crawling. There exist some web crawlers (for example Nutch, Heritrix, etc.) but it can be difficult to adjust if it is even possible for this special case. So we created an own simply web robot for this case. Before crawling we assume at most 2 million users in Pokec so we decided to get whole network not only random sample.

If Pokec network of user's relation is scale-free then almost all people are in one graph component and from almost every person exists short path to each other. We will verify the assumption that Pokec contacts network is scale-free in next section after collecting data.

Algorithm of our web robot which crawl social data looks as follows:

1. Insert user into queue, for example my nick name.
2. Take first nick out from the queue and process it[2], if the queue is empty, **go to 4** else **go to 3**.
3. Get all contacts relations from nick and put all of them, which were not processed yet, in a queue. This step is also responsible for storing nicks and network relations, **go to 2**.
4. END

[1] Based on our analysis, see section 4.1.

[2] You can easy get user profile and friendships by adding nick name after OSN Pokec web page URL.

Mentioned algorithm is trivial and it works because of OSN Pokec has weak security and small size. If we would like to retrieve social data from for example facebook, linkedin or badoo, etc., we have to use sophisticated algorithm. These OSNs are very large (hundreds of millions users) and we cannot retrieve all data but only sample of it. This can be done by changing algorithm as follows:

1. Insert user into queue, for example my nick name.
2. Take **random** nick out from the queue and process it, **if we crawled more than N nicks**; **go to 4** else **go to 3**.
3. Get all contacts relations from nick and put all of them, which were not processed yet, in a queue. This step is also responsible for storing nicks and network relations, **go to 2**.
4. END

Another problem can be that some OSNs web sites are restricted by number of requests per minute or request count limit from unique IP address. These security measures are purposed against web robots. Although it is possible to bypass it by using more robots at several IP addresses what requires additional hardware capacity or requesting only after specified time what requires additional time. Both of these methods, however, would not help us if the OSN is secured by captcha. In that case it will be very complicate to get the data.

These web robots are something like nick crawlers and they obtain contacts network with topology or social graph where the nicks are vertices and friendships are edges. Web robot also obtains another data which are in profiles published for all. On Fig. 1 we can see data layer of web robot program and database storage for obtained data. We used MySQL database and Java for programing it.

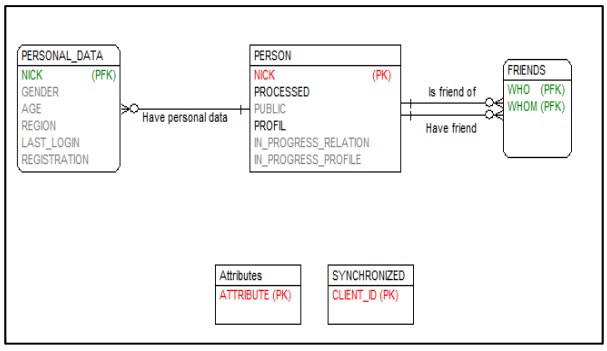


Fig. 1 The figure shows relational database model. Data layer of web robot program and also storage for obtained data from OSN "Pokec".

After running the trivial web robot (first algorithm) first time, it processed roughly hundred thousand nicks per day. We did some changes to run it parallel. We created lock system for crawling and also for profile processing. Every web robot can do either nick crawling or profile parsing. Always when a web robot takes either unprocessed nicks or unprocessed profiles they are marked with flags (on Fig. 1 Table PERSON, columns IN_PROGRESS_RELATION and IN_PROGRESS _PROFILE) that it could not be taken by other web robots. This allows running robots parallel also on more computers. Crawling speed depends on web robots count, internet connection speed and server of OSN speed. We use one PC, HP server Intel® Xeon® with 2 CPUs 2.53GHz, 8 GB RAM with 64 bit operating system Windows Server 2008 R2 Enterprise, with 25 web robots for crawling and 25 for parsing. With fast broadband internet connection we get all data in two days. We get all the network data and there was not any security restriction.

We get social network data with over than 1.6 million users and more than 40 million relations between them. More than 66 percent of users have their contacts (friendships) published. For more data statistics see section 4.

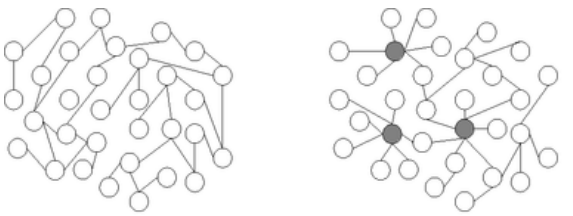
3. SCALE-FREE CHARACTERISTICS OF OSN POKEC

Scale-free network (SFN) is a network whose degrees distribution of nodes follows power law, at least asymptotically as follows (**formula** 1).

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

where: $P(k)$ – probability that degree of node is k, c – normalization constant, γ – parameter.

In SFNs most of nodes has a low degree but some nodes (called hubs) has enormously high node degree (See Fig. 2). These 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 subgraphs while SFN not. SFN is resistant to random attack. Only direct attack to hubs can split this network. Maybe we lost some edges and nodes but the core of the network is stable.



(a) Random network (b) Scale-free network

Fig. 2 The difference between (Erdös-Rényi) random and scale-free network [3] is in nodes degree distribution. In scale-free networks there are some nodes called hubs which has exponentially more edges than the others nodes.

Because of this we are convinced that if we get a social network via our web robots where 66% of users have public contacts we get roughly all of nodes. It is the same like random attack to scale-free network when we canceled edges from 34% of nodes. It is very unlikely that this network splits into subgraphs and the core of network is broken. And because of this we know that we get almost all data and users from the network.

Now we have to show that the network we get is scale-free. For simplicity we transformed obtained network graph which is directed into undirected. So an edge is

between nodes if there exist at least one relation. After analyzing nodes degrees of obtained network we see very nice power law distribution in figure 3 so network Pokec can be considered as scale-free. For the non-scale-free network we cannot get the core of the network using this approach but maybe only some small part. In that case data statistics in next section would be biased and would not correspond to the fact.

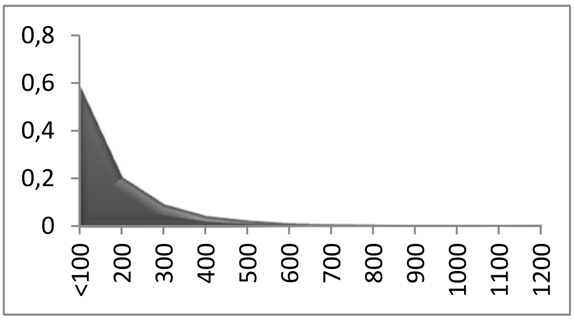


Fig. 3 The figure shows nodes degree distribution of network Pokec obtained by a web robot. The chart clearly shows power law distribution which is specific for the scale-free networks. On the x-axis are abundances and on the y-axis are probabilities.

Another typical characteristic of SFNs is very short distance between nodes [4]. According Cohen and Havlin analysis [2] the diameter or the mean distance between nodes in typical SFN with $2 < \gamma < 3$ can be bounded by interval:

$$< lnlnN,\ lnN/\ lnlnN > \qquad (2)$$

where: N - is nodes count in network.

We did 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. See on table 2 graph size and on table 1 average time for calculating the distance between two nodes. So we choose randomly 100 pairs of nodes and calculate they average shortest distance (See Tab. 1).

Table 1 Table shows output from experiment of 100 random selected pairs and they minimum, maximum and average shortest distance and calculation time.

	Distance	Time (min)
MIN	3	0,758416667
MAX	6	589,6036
AVG	**4,670588235**	119,3389994

Table 2 Table shows graphs parameters of the experiment.

Graph	Count
Nodes (N)	1 637 068
Edges	46 171 896

The average shortest distance calculated in experiment lies in appropriate interval for this network (See formula 2.).

$$lnlnN < \boldsymbol{diameter} < lnN/lnlnN$$
$$2{,}66 < \mathbf{4.67} < 5{,}38,$$
$$N = 1\,637\,068$$

Formula 2 Formula 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***.

Most of networks which are self-organized for example internet, social relations, collaboration network, protein - protein interaction, airline networks etc. are scale-free. Social network is dynamic structure which is changing over time. New nodes are coming into existence and old unused nodes are ceasing to exist as users are registering and unregistering. Edges are changing as users are changing relations between them.

4. DATA ANALYSIS

We did some data analysis of 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.

4.1. People and their privacy

Table 3 Table shows user count, gender representation, number and percentage of users which published their friendships contacts and age.

	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

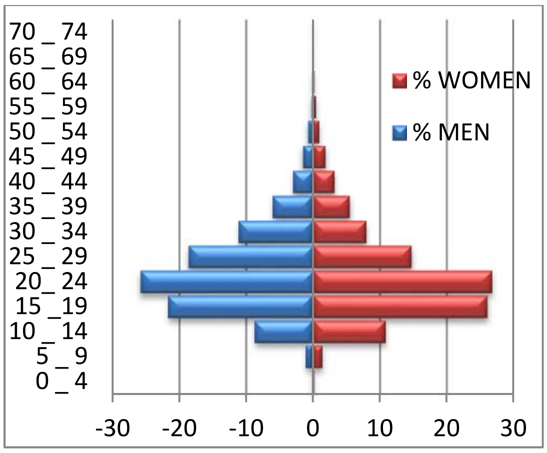


Fig. 4 The figure shows age distribution of Pokec users. There are dominated mainly young people aged 15-30 years. Approximately 33 % of women and 29 % of men have their age nonpublic.

4.2. Friendships

Friendships or relations between users are in Pokec oriented. It means that if user1 has friend user2, user1

need not to be a friend for user2. But most of all there are friendships on both sides.

Table 4 Table shows average user friendships count and maximum friendship count in both directions.

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

Most connected nodes (users with most friendships) called in scale-free networks hubs (as we mentioned in section 3). Hubs in Pokec are not usually people but commercial companies which advertise through this network.

4.3. Visit rate

There are published for all in each profile user last visit date and registration date. So there is an easy way how to get monthly visit rate or number of new registered user per month. Also a provider of Pokec declares about one million unique users per month and we have confirmed this assertion.

Table 5 Table shows number of different users visit per month and number of new registered users per month. Monthly visit rate is very high because it is about 70% of all users in network.

	COUNT per month
Different user visits	1 116 971
New registrations	24 829

4.4. Other profiles data

Other profiles 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 something to the profile, it is public for all. Only age and friendships can be nonpublic in OSN Pokec. We calculated from obtained data that the average profile is filled to 40.33 %. Base on this database a detailed marketing research on user interests, preferences etc. segmented for example by age, region, gender etc. 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.

5. SOCIAL NETWORK SECURITY

OSNs are very good invention but all of the inventions have their weaknesses. We think that the main weakness of OSNs is loss of privacy. We do not think about abusing private information by provider. We assume that the providers of OSNs are trustworthy. The problem is that everyone with public profile losses his or her privacy depending on the information that are published about him or her.

Imagine that you are walking around town and you meet stranger who asks you about your age, job, hobbies, your children etc. We think that no one normal would have answered these questions to a stranger but it is the same as the public profile on the social network. These data can be easily abused. Leakage of personal information, especially one's identity, may invite malicious attacks from the real world and cyberspace, such as stalking, reputation slander, personalized spamming, and phishing [5].

For example in our obtained data from Pokec each user has an email address nick-name@pokec.sk. Using direct mailing we can propagate our opinions, services or products according users age, gender, preferences or interests. It is possible to reach about million people that way (see Table 5 in previous section, visits per month). That would be abusing for marketing and promotional purposes. There is also a lot of other possibilities on how to abuse personal information.

The question is how to avoid these problems. For a user of social networks we recommend to make friendships private. It means that either no one or only our friends can see our contacts. Make profile less personal and make it available only for our friends or for a person chosen by us. It is also important if we do not want to be traceable to make friendships with only persons who comply with these rules. In case of compliance with these rules there is no possibility to obtain social and personal data as easy as we showed.

For providers of OSNs we recommended to make all data (friendships, profile, photo and video albums, etc.) private by default and only if a user wants to publish something he sets it as public for friends or public for all. Actually it is contrariwise and therefore people often do not consider that their data are accessible for all. We think that such simple change would be very helpful.

With using this approach it would increase the number of users with private contacts and then it would be problem to get the whole social network as is showed in section 2.

Many edges (friendships) in network would be private and the network would be broken without these edges into many subnets that would be difficult to find and it would be untraceable.

6. CONCLUSIONS

In this article we have demonstrated on a practical example how to obtain social and personal data from any just partially public OSN using scale-free characteristics of these networks. We get network relation graph from OSN, we show that it is scale-free and based on this we convinced that we get almost whole network. We have shown some statistics and we have indicated possible ways of using / abusing of collected data of concrete OSN Pokec which has been very popular in Slovakia and which has about one and half million active users. We have discussed about current OSNs security and privacy problems and we recommended simple rules for users and for providers to avoid or reduce them.

In the future work we would like to deal with security of OSNs and all personal data which it contains and try to solve current state or try to design concept in such a way that will be no as easy as now to get a lot of personal data from it. We thing that there are two possible approaches to do it: one is the approach of people which are using OSNs and the second is the security design of OSNs by creators or providers.

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

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

[5] ENISA, Security Issues and Recommendations for Online Social Networks, Position Paper, Nov. 2007.

[6] http://www.pokec.sk, Dec. 2011

Received April 25, 2012, accepted October 3, 2012

BIOGRAPHY

Ľuboš Takáč was born on 16.01.1986. In 2010 he graduated (MSc) with distinction at the Department of Transport Networks of the Faculty of Management Science and Informatics at University of Žilina. He is now post-graduate student at the Department of Informatics at the same faculty. His PhD thesis title is "Data processing over very large databases". His scientific research and interest is focusing on large databases, data mining, online social networks and data visualization.