A COMPREHENSIVE STUDY ON DATA EXTRACTION IN SINA WEIBO

Xiao Cui[1] and Hao Shi[2]

College of Engineering and Science, Victoria University, Melbourne, Australia

ABSTRACT

With the rapid growth of users in social networking services, data is generated in thousands of terabytes every day. Practical frameworks for data extraction from social networking sites have not been well investigated yet. In this paper, a methodology for data extraction with respect to Sina Weibo is discussed. In order to design a proper method for data extraction, the properties of complex networks and the challenges when extracting data from complex networks are discussed first. Then, the reason for choosing Sina Weibo as the data source is given. After that, the methods for data gathering are introduced and the techniques for data sampling and data clean-up are discussed. Over 1 million users and hundreds of millions of social relations between them were extracted from Sina Weibo using the methods proposed in this paper.

KEYWORDS

Social Network Analysis, Data Extraction,Sina Weibo, Data Mining

1. INTRODUCTION

The use of social networking services has exploded in the past decades, that has created data on a scale never seen before in human history. People leave numerous digital footprints in Facebook, Twitter, LinkedIn, and so on. These digital footprints paint a picture of what happens in the real world [1]. Analyzing these digital footprints is called social network analysis, which has drawn extensive attention from all walks of life[2-5]. However, mostof these studies focus on data analysis rather than data extraction. As data extraction is an indispensable step in social network analysis, studying how to extract data from social networks efficiently and effectively is not a waste of time. This paper aims to provide a comprehensive study of data extraction with respect to Sina Weibo, one of the most influential social networking sites in China [6].Through the demonstration based on real data, this paper aims to help other researchers better design the tools for data extraction so as to deal with the challenges of extracting data from complex networks.

A practical framework for extracting data from Sina Weibo is presented in this paper.Before that, the properties of complex networks are illustrated and the challenges of extracting data from complex networks are discussed in this paper. Then, the reason for choosing Sina Weibo as the data source is given. After that, a framework for data extraction is proposed by three parts: data gathering, data sampling and data clean-up.A conclusion is drawn in the end, upon the experimental results obtained from Sina Weibo.

DOI: 10.5121/ijaia.2016.7404

2. BACKGROUND

This section consists of two parts. The first part discusses the properties of complex networks. The second part illustrates the challenges when extracting data from complex networks. Recognizing them helps researchers make wiser decisions ondata extraction from complex networks like social networks.

2.1. PROPERTIES OF COMPLEX NETWORKS

Social networks are often huge, such as dozens of millions of users and the interactions between these users are even more complex. No matter what kinds of social networks they are, they do have some things in common that are seldom seen in other kinds of networks. The most well-known common properties include power law distribution, the small-world effect and strong community structure.

The degrees of vertices in social networks often follow power law distributions or long tail distributions[7]. To be more specific, vertices with lower degrees are more frequent than vertices with higher degrees. Figure 1 indicates that power law distributions exist in Flickr, LiveJournal, Orkut, and YouTube[8].

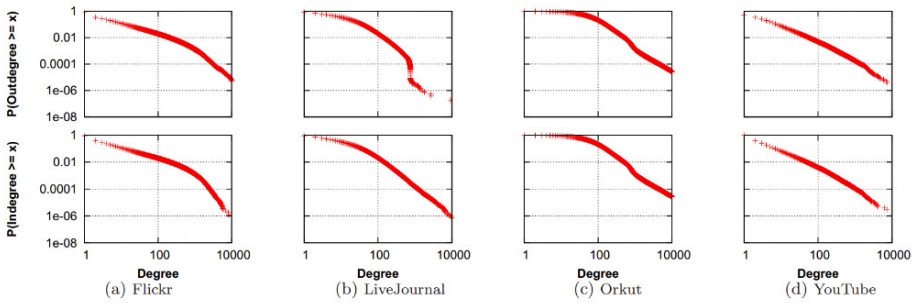


Figure 1. The proportion of in-degree and out-degree for each type of social networks. Calculation based on complementary cumulative distribution functions[8].

Another characteristics of complex networks is the small-world effect (or six degrees of separation). Half century ago, scientists [9] had already investigated the average path length between people in Nebraska and Boston and people in Massachusetts. The results showed that anyone is just six relationships away from anyone else on Earth. The small-world effect exists in social networking sites as well. The average path length on Facebook was 4.7 [10], that on Twitter was 4.12 [11]. The average path length on YouTube was a little bit longer, at 5.1 [8].

Social networks also show a strong community structure [12]. This means that people within the same community tend to interact with each other more frequently. On the other hand, people from different communities barely connected to one another.

The properties mentioned above need to be taken into account carefully when extracting data from complex networks.

2.2. CHALLENGES WHEN EXTRACTING DATA FROM COMPLEX NETWORKS

Millions of people play online, learn online and even work online. People are living in an information explosion era they have never experienced before. Almost everything on Earth has a

digital footprint on the Internet. The full capability of social networks has yet to be reached. Social networks combined with their unique characteristics pose challenges that have never been met before. This subsection discusses the challenges when extracting data from complex networks.

The forms of the interactions can vary, even different forms of interactions exist between the same set of users, e.g., two users work at the same company but they do not like each other. Multiple types of entities are also involved, e.g., two users work together but are connected to each other through a cloud server. Extracting such data requires new theories and models.

People share their thoughts online in the form of comments, reviews, ratings, etc. Such meta information is useful for many applications. Collecting the intelligence from such data effectively is not a straightforward job but it is very necessary because this intelligence is very precious.

Thus, data extraction plays a significant role in social network analysis. In other words, the quality of the research in social network analysis is partially determined by how good data is.

3. SINA WEIBO

As Twitter is banned in China, Sina Weibo is considered a replacement for it. Sina Weibo has reached 56 million daily active users (i.e. who spend an average of one hour per day with the service) [13]. Sina Weibo has had a significant influence on Chinese society. Sina Weibo was used as the data source in this paper, because it is more informative than any other social networks in terms of the contents, the interaction between users and, most importantly, the verification system. Sina Weibo allows users to insert images, videos, music, long articles (more than 140 characters) and even polls without any plug-ins being required (as shown inFigure 2).

Figure 2. The contents on Sina Weibo are more than text.

The interaction between users is everywhere, e.g., users are allowed to leave comments on someone's weibo even reply to others' comments on someone else's weibo. Sina Weibo also encourages its users to participate in its identity verification program. Verified users are categorised into eleven groups (as listed in Table 1).

Table 1. Verified types

Verified Type	Domain
Agency	usually referring to welfare organizations, sports clubs, arenas, and other non-governmental organizations.
Application Software	usually used to promote the use of an application
Brand	corporate accounts, usually using Sina Weibo to promote their brand values
Campus	universities' official accounts, student associations' accounts, etc
Government	usually referring to local authorities
Hall of Fame	famous individuals from all walks of life
Media	usually referring to news agencies, television broadcasters, even self-media
Pioneer	grassroots, usually taking an active part in Sina Weibo, whose identities have been successfully verified by Sina Weibo
Website	a window on a website, usually giving an absorbing summary through Sina Weibo but linking details to its own webpages
Weibo Girl	girls who are addicted to sharing selfies with others and who use Sina Weibo as a platform to promote themselves

Sina Weibo even provides verification services for ordinary people. A 'Pioneer' badge is granted as long as the applicant's real identity is verified and the minimum requirement of being active is satisfied. Unlike 'Pioneer', more information is required so as to grant a badge of one of the rest, e.g., applying for a corporate account requires a business license, an official letter with stamp and signature of official representatives, the certificate of trademark registration, the brand letter of authorisation, etc. Because of the strict verification policy, the public has a chance to communicate with the real celebrities and real giants from all walks of life. The trustworthiness and authenticity of the contents posted by verified accounts is guaranteed. Consequently, authenticity stimulates more users to participate actively in Sina Weibo [14].

4. DATA GATHERING

Having a complete dataset is a nearly impossible mission for three reasons. Frist, there are more than 500 million users registered on Sina Weibo. Second, privacy is a serious matter for Sina Weibo. Third, only limited access is provided for third-party developers. Instead, this paper collects a partial dataset from Sina Weibo. The following things must be clearly defined before extracting data from social networks: social relations, search algorithm, and seeds. After that, the architecture used to extract data from social networks is discussed, with respect to the REST API provided by Sina Weibo.

4.1. SOCIAL RELATIONS

Four types of social connections are defined. Given two users $v_{1,1}$ and $v_{1,2}$ (as shown in Figure 3), there are:

- neither $v_{1,1}$ nor $v_{1,2}$ follow each other
- $v_{1,2}$ follows $v_{1,1}$ but $v_{1,1}$ does not follow back
- $v_{1,1}$follows $v_{1,2}$ but $v_{1,2}$ does not follow back
- $v_{1,1}$ and $v_{1,2}$ follow each other

Type 4 is the only symmetric relation. In this case, $v_{1,1}$ is a friend of $v_{1,2}$ and $v_{1,2}$ is also a friend of $v_{1,1}$. Type 4 is considered a bilateral friendship. In this paper, bilateral friendships were used to extract data.

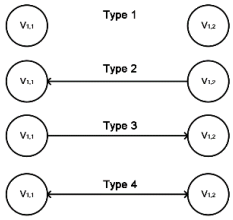
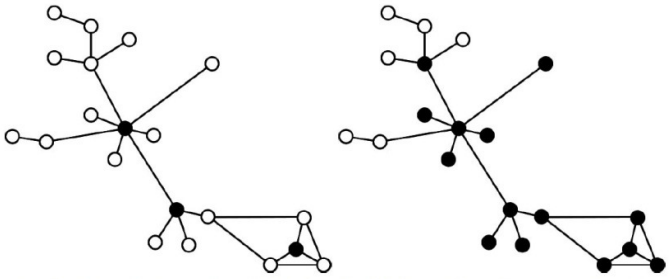


Figure 3. Four types of social relations in Sina Weibo

4.2. SEARCH ALGORITHM

Breath-first search (BFS) is a well-known graph traversal algorithm that has been widely used as a crawling strategy to extract data from social networks [8, 15-19]. A BFS program was written and used to extract data from Sina Weibo. The traversal starts from a set of vertices (i.e. seeds) and continues by visiting the vertices adjacent to the last vertices (as shown in Figure 4).



(a) Starting from vertices in black (b) Visiting the adjacent vertices and colouring them black

Figure 4. An efficient traversal where 11 vertices were visited after the first iteration

It is worth mentioning that selecting proper seeds is important. For example, using the seeds with low degree (i.e. having few connections) (as shown in Figure 5) as the starting points can lead to an inefficient traversal (e.g. only 6 vertices were visited after the first iteration), compared to the seeds with high degree (e.g. 11 vertices were visited after the first iteration). The next section discusses how to choose proper seeds for data extraction from social networks.

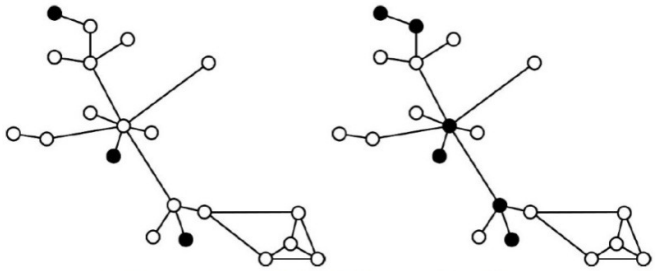


Figure 5. An inefficient traversal where 6 vertices were visited after the first iteration

4.3. CHOICE OF SEEDS

Table 2 lists 6 accounts used in this paperas the starting points for data extraction. As the screen names include Chinese characters, for convenience, their domain names on Sina Weibo were used instead.

Table 2. Choices of seeds

Account	Description
Hejiong	A famous actor, who has many connections with other celebrities. He has more than 50 million followers
panshiyi	A business magnate, who is the chairman of SOHO China, the largest prime office real estate developer in China. He has more than 17 million followers
rmrb	An official newspaper of the government of China, a giant in mass media. It has more than 28 million followers
haroldlee	An ordinary person who works at a consulting company. He lives in Beijing.
wraithree	An ordinary person who works at a shopping center. She lives in Shanghai.
jerjj	An ordinary person who works at an IT company. He lives in Guangzhou.

Three 'Pioneer' accounts from the top three cities in China were selected. Recent research shows that people are still bounded by physical distance even though the earliest social networking services appeared decades ago [20]. Thus, picking one 'Pioneer' user from Beijing (Northern China), one 'Pioneer' user from Shanghai (Eastern China) and one 'Pioneer' user from Guangzhou (Southern China) reduced the possibility of finding existing users that have been visited before. The reason the other three accounts 'hejiong', 'panshiyi' and 'people's daily' were selected, is that they have many connections with other verified users from all walks of life, for example, 'hejiong' has more than 500 friends (i.e. bilateral friendships) who are 'VIP' accounts, including 'Hall of Fame' accounts, 'Brand' accounts, 'Campus' accounts, 'Government' accounts, etc. These friends also have many connections with others. Thus, using them as the starting points increased the possibility of finding new users from all walks of life, and eventually created a ripple effect and more efficient traversal. Once the seeds are selected, the architecture used to extract data has to be defined.

4.4. CRAWLERS FOR SOCIAL NETWORKING SERVICES

'Crawler' is a generic term for any program used to automatically discover and scan website by following links from one webpage to another [21]. As most social networking sites use Dynamic HyperText Markup Language (DHTML) (i.e. a generic term for any technologies used to create web pages that are not static web pages) to make their websites more lively, using a web crawler to retrieve information from a combination of markup tags and programming scripts is not that easy. Also, heavy web traffic is involved (e.g. in order to get the list of friends of a user, the web crawler needs to explore several pages when one page is not enough to list all the friends).

Instead, this paper suggests to use the official API to extract data from the server. Sina Weibo provides a REST API for third-part developers. The crawler designed for use in this paper was built on the official API (as shown in Figure 6). This made the implementation more efficient. For example, in order to get the list of friends of a user, the crawler only needs to send a request to the server. All the data extracted was integrated into one database.

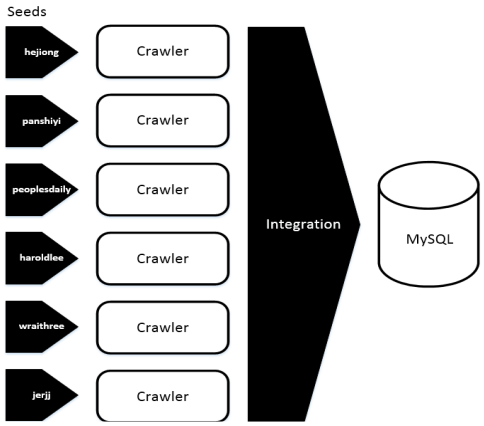


Figure 6. The architecture for data extraction

4.5. API ACCESS

OAuth authentication is required when trying to access to the REST API if using it to acquire user profiles and the social connections between users. An access token (i.e. a permit) is granted to the crawler (as shown in Figure 7) once an OAuth request is authorised by the resource owner. The crawler uses the token to access to the resources protected by default.

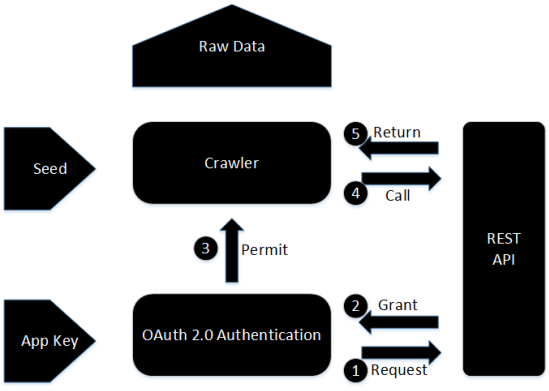


Figure 7. The mechanism of how to access to the REST API

The API only allows the third-party application to make a limited number of calls per hour (as listed in Table 3). It is worth mentioning that a maximum of 150 calls are allowed per hour, regardless of what API function is being called. Because of the rate limits, it is a time consuming task to acquire millions of user profiles and the social connections between them. This is why multiple crawlers were deployed at one time, with each of them having its own App Key (i.e. a string used to identify the application when making requests to the API). In an ideal case, 900 calls can be made per hour when adopting 6 crawlers at the same time. In the next section, the data structure is discussed.

Table 3. An example of rate limits on Sina Weibo

API function	**Number of calls per hour**	**Number of calls per day**
statuses/update	15	50
statuses/repost	15	50
friendships/create	15	50
users/show	150	N/A
statuses/count	150	N/A
friendships/friends/bilateral	150	N/A

4.6. Data Structure

The data structure determines the information to be collected. A user profile was defined as consisting of the following attributes (as listed in Table 4). Because little data is in the necessary format, extra computation and calls were required to make the data 'userful'. For example, in order to calculate the number of comments a user has received so far, the crawler had to retrieve the number of comments of each weibo the user posted on Sina Weibo and then add them up.

Table 4. User profile

Attribute	**Description**	**Example**
uid	A unique number (8 digits) assigned to a user profile	12144623
screen name	The name a user chooses to use for communicating with others online	Jerry Xu XuXu
province	The province where a user lives	Hebei
city	The city where a user lives	Guangzhou
gender	Male or female	Male
followers	The number of followers a user has	256
followees	The number of followees a user has	320
friends	The number of riends a user has	125
weibo	The number of weibo a user has posted on Sina Weibo	1200
comments	The number of comments a user has received so far	1156
reposts	The number of times a user has been retweeted	16
likes	The number of times a user has been liked	632
verified type	The type of verification a user belongs to	Pioneer
wei_age	The number of years a user has used Sina Weibo	4

The data structure illustrated in Table 5 explains how the social connections are stored. The 'uid' on the left hand side (LHS) always follows the 'uid' on the right hand side (RHS). The social connections that were stored are those between users whose profile data had been stored.

Table 5. The data structure for social relations

LHS	RHS
1065517411	2651153623
2864652252	2104908771
1889636460	2105665795

Figure 8 shows an example to explain it. Users whose profile data had been stored (i.e., $v_{1\ldots6}$) are colored in black and users whose profile data had not been stored (i.e., $v_{1\ldots6}$) are colored in white, then the social connections $\langle v_1, v_4\rangle$, $\langle v_4, v_1\rangle$, $\langle v_2, v_4\rangle$, $\langle v_3, v_4\rangle$, $\langle v_5, v_4\rangle$ and $\langle v_6, v_4\rangle$are retrieved, and $\langle v_6, v_7\rangle$, $\langle v_6, v_9\rangle$, $\langle v_7, v_9\rangle$, $\langle v_9, v_7\rangle$and $\langle v_9, v_8\rangle$are ignored.

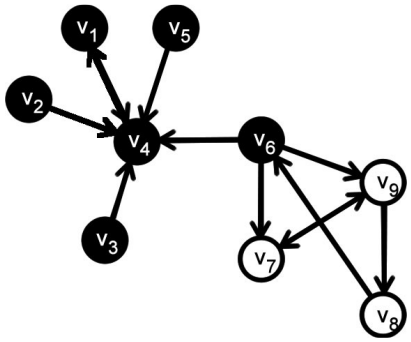


Figure 8. Retrieving social connections between the selected users. Users colored in black indicate that the profile data had been stored. Users colored in white indicate that the profile data had not been stored.

5. DATA SAMPLING

Data sampling is necessary, as otherwise the data of 1 million users would have been too big to be handled within a reasonable time. Considering a group of 1 million vertices, the number of edges reaches 500 billion for a complete graph (i.e., every pair of vertices is connected by an edge) and 500 million for a random graph [22] where all pairs of vertices are connected with probability 0.001. However, estimating the number of social connections based on the number of users in a social network is much more complicated than estimating the number of edges based on the number of vertices in a random graph.

Although the probability that two users have a social connection between them is very hard to measure, an understanding of the volume of the data for 1 million users is still possible.Manipulating a graph of such size is a big challenge considering the computing hardware available. Even though many tasks such as community detection require only linear processing time with respect to the number of edges, they are still time consuming, considering the graph has hundreds of millions of edges. The storage requirements are even more demanding [23]. Thus, how to sample data must be carefully considered.

Simple random sampling is not able to reflect the markup of the population because of the randomness of the selection. Stratified sampling is suggested, where samples were drawn based on the distribution of different verified types of users in Sina Weibo.

6. DATA CLEAN-UP

Social networks have the properties of being large scale but low density. As listed in Table 6, a graph of Face book with 721.1 million of users only had 68.7 billion edges, where the density of the graph was 2.641×10^{-7} [24]. A graph of Twitter [11] had a density of 8.45×10^{-7}, which is also very low. As has been well-recognised, data sparsity complicates analytic computation because it makes the problem much noisier [25].

Table 6. Social networks of sparse graph

Social network	Vertices	Edges	Density
Facebook	721.1 million	68.7 billion	0.0000002641
Twitter	41.7 million	1.47 billion	0.000000845

In this research, two methods were used to improve the density of the data. First, bilateral friendships were used to expand the search of users. It eliminated zombie accounts (i.e., fake or artificial accounts, most time, used for spamming) because zombie accounts are rarely followed back by the other users [26]. Second, a better connected graph such as the largest connected graph was used as the input of the experiments. The largest connected graph provided a graph in which any two users were connected to each other directly or indirectly. Isolated users were removed because they were not connected in any way.

7. CONCLUSION

This paper provides a novel method to extract data from Sina Weibo. Considering the complexity of Sina Weibo, the methods proposed in this paper is very efficient. This paper also provides potential solutions for graph search, seed choosing, and multithread crawling. Through the demonstration based on Sina Weibo, this paper helps other researchers better design their tools for data extraction with respect to other social networking sites such as Face book, Twitter and LinkedIn.This paper also provides a valuable data set for further investigation. Such data set includes 1,192,972 users and 181,575,370 social connections.

REFERENCES

[1] Alef, D. (2010). Mark Zuckerberg: The face behind Facebook and social networking. Titans of Fortune Publishing.

[2] Beutel, A., Akoglu, L., and Faloutsos, C. (2015). Fraud detection through graph-based user behavior modeling. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. ACM pp.1696-1697.

[3] Culotta, A., Kumar, N. R., and Cutler, J. (2015). Predicting the Demographics of Twitter Users from Website Traffic Data. InAAAI. pp. 72-78.

[4] Byrd, K., Mansurov, A., and Baysal, O. (2016). Mining Twitter data for influenza detection and surveillance. In Proceedings of the International Workshop on Software Engineering in Healthcare Systems. ACM. pp. 43-49.

[5] Crossley, N., Bellotti, E., Edwards, G., Everett, M. G., Koskinen, J., andTranmer, M. (2015).Social network analysis for ego-nets: Social network analysis for actor-centred networks. Sage.

[6] Alexa (2010). Alexa traffic rank for weibo.com. http://www.alexa.com/siteinfo/www.weibo.com

[7] Ahn, Y. Y., Han, S., Kwak, H., Moon, S., and Jeong, H. (2007). Analysis of topological characteristics of huge online social networking services. InProceedings of the 16th international conference on World Wide Web. pp. 835-844

[8] Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P. and Bhattacharjee, B. (2007). Measurement and analysis of online social networks. In: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement. ACM, pp. 29-42.

[9] Travers, J. and Milgram, S. (1969). An experimental study of the small world problem. Sociometry, pp. 425-443.
[10] Ugander, J., Karrer, B., Backstrom, L. and Marlow, C. (2011). The anatomy of the Facebook social graph. ArXiv preprint arXiv:1111.4503.
[11] Kwak, H., Lee, C., Park, H. and Moon, S. (2010). What is Twitter, a social network or a news media? In: Proceedings of the 19th international conference on world wide web. ACM, pp. 591-600.
[12] Girvan, M., and Newman, M. E. (2002). Community structure in social and biological networks. In Proceedings of the national academy of sciences, 99(12), pp. 7821-7826.
[13] Clark, D., Crandall, R. and Mei, Y. (2013). 4th Annual China 2.0 Conference Underscores Business Innovation, Social Impact and U.S-China Links
[14] Chen, J. and She, J. (2012). An analysis of verifications in microblogging social networks – Sina Weibo. In: Distributed Computing Systems Workshops (ICD-CSW), 2012 32nd International Conference on. IEEE, pp. 147-154.
[15] Catanese, S., De Meo, P., Ferrara, E. and Fiumara, G. (2010). Analyzing the Facebook friendship graph. arXiv preprint arXiv:1011.5168.
[16] Chau, D. H., Pandit, S., Wang, S. and Faloutsos, C. (2007). Parallel crawling for online social networks. In: Proceedings of the 16th international conference on world wide web. ACM, pp. 1283-1284.
[17] Gjoka, M., Kurant, M., Butts, C. T. and Markopoulou, A. (2010). Walking in Facebook: A case study of unbiased sampling of osns. In: INFOCOM, 2010 Proceedings IEEE. IEEE, pp. 1-9.
[18] Wilson, C., Boe, B., Sala, A., Puttaswamy, K. P. and Zhao, B. Y. (2009). User interactions in social networks and their implications. In: Proceedings of the 4th ACM European conference on computer systems. ACM, pp. 205-218.
[19] Ye, S., Lang, J. and Wu, F. (2010). Crawling online social graphs. In: Web Conference (APWEB), 2010 12th International Asia-Pacific. IEEE, pp. 236-242.
[20] Backstrom, L., Sun, E. and Marlow, C. (2010). Find me if you can: improving geographical prediction with social and spatial proximity. In: Proceedings of the 19th international conference on world wide web. ACM, pp. 61-70.
[21] Rosenfeld, J. M. (2002). Spiders and crawlers and bots, oh my: The economic efficiency and public policy of online contracts that restrict data collection. Stan. Tech. L. Rev., 2002, pp. 3-4.
[22] Erdos, P. and Renyi, A. (1959). On random graphs i. Publ. Math. Debrecen, 6, pp. 290-297.
[23] Fortunato, S. (2010). Community detection in graphs.Physics reports,486(3), pp. 75-174.
[24] Backstrom, L., Boldi, P., Rosa, M., Ugander, J., andVigna, S. (2012). Four degrees of separation. InProceedings of the 4th Annual ACM Web Science Conference. pp. 33-42.
[25] Newman, M. E. and Girvan, M. (2004). Finding and evaluating community structure in networks. Physical review E, 69(2), p. 026113.
[26] Jiang, M., Cui, P., Beutel, A., Faloutsos, C., and Yang, S. (2014). Detecting suspicious following behavior in multimillion-node social networks. InProceedings of the 23rd International Conference on World Wide Web. pp. 305-306.

Authors

Xiao Cui received the Ph.D. degree from Victoria University, Melbourne, Australia, in the area of social network analysis, in 2016. His research interests include artificial intelligence, social network analysis, machine learning, data mining and web development.

Hao Shi is an Associate Professor in College of Engineering and Science at Victoria University, Australia. She completed her PhD in the area of Computer Engineering at University of Wollongong and obtained her Bachelor of Engineering degree at Shanghai Jiao Tong University, China. She has been actively engaged in R&D and external consultancy activities. Her research interests include p2p Network, Location-Based Services, Web Services, Computer/Robotics Vision, Visual Communications, Internet and Multimedia Technologies.