2012 IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications

Design and Implementation of Facebook Crawler Based on Interaction Simulation

Zhefeng Xiao
School of Computer Science
National University of Defense Technology
Changsha, Hunan 410073-P.R.China
e-mail:zhefeng.xiao@gmail.com

Bo Liu
School of computer Science
National University of Defense Technology
Changsha, Hunan 410073-P.R.China
e-mail: boliu615@yahoo.com.cn

Huaping Hu
School of Computer Science
National University of Defense Technology
Changsha, Hunan 410073-P.R.China
e-mail: howardnudt@yahoo.com.cn

Tian Zhang
School of computer Science
National University of Defense Technology
Changsha, Hunan 410073-P.R.China
e-mail:jonathan2012@yahoo.cn

Abstract*—The extensive use of Online Social Networks (OSNs) has attracted such wide attention of academia that OSNs have become a hot topic. Based on the interaction simulation, we have designed and implemented a Facebook crawler, which can obtain the complete friend list of a Facebook user and overcome the drawback in [7] that the crawler can get at most 400 friends. We make an analysis and visualization of the crawled dataset, and find that 36.5% of Facebook users have changed the default privacy setting, compared to 26.6% and 16% in [7, 20] respectively, implying that the awareness of privacy protection of Facebook users has been greatly improved.

Keywords-Facebook; social networks; crawler; analysis; privacy

I. INTRODUCTION

From the first Online Social Network (OSN) site SixDegrees.com [1] emerging in 1997 to the largest OSN site Facebook [2], OSNs have been developing rapidly and had more and more impact on our daily life. There have been 1.438 billion users [17] who are using OSNs by the end of November 2011. Since OSNs have gone beyond search engine in the residence time for the first time [3], there will be more and more people who begin to use OSNs. Therefore, OSNs has attracted extensive attention from research community, including sociology, psychology, computer science, and so on [6, 7, 8, 9, 13, 16]. The bulk of OSN research has focused on impression management, friendship performance, networks and network structure, online/offline connections, and privacy issues [1].

Before researchers conduct the study on OSNs, first of all, they have to obtain sufficient data. There are two ways to fulfill the goal. One is to use the existing research dataset [6, 15] and the other is to use active detection tools, such as crawlers [7, 8, 9, 10].

As the largest OSN site, Facebook has had a total of 845 million monthly active users by the end of December 2011. And they post over 20 million topics and upload 250 million photos each day [4]. At present, Facebook has become the main object of studies on OSNs [11], and at the same time it has the most complex privacy policy, which has been updated frequently. Thus, we choose it as the object of our study.

We summarize the main contributions of this paper as follows:

First, we design and implement a crawler based on the interaction simulation, which is no longer limited to at most 400 friends in [7] and can crawl the complete friend list of a Facebook user;

Second, we make an analysis and visualization of the dataset crawled from Facebook. We find that 36.5% of Facebook users have changed the default privacy settings, compared to 26.6% and 16% in [7, 20] respectively, implying that the awareness of privacy protection of Facebook users has been greatly improved.

The remainder of this paper is organized as follows: The section 2 briefly introduces related work to our study. The section 3 describes the detailed implementation of the crawler algorithm in this paper. The section 4 presents the process of data collection, the analysis and visualization of dataset crawled from Facebook. And the section 5 summarizes our work and discusses a blueprint for future work.

II. RELATED WORK

Bonneau et al. [5] took a survey of several approaches for obtaining large amounts of personal data from Facebook, including public listings, false profiles, profile compromise, phishing attacks, malicious applications and the Facebook Query Language. The research dataset in [6, 13] was mined through public listing in [5]. As the privacy policy and access control mechanisms in Facebook has been continuously enhanced, obtaining user data from Facebook has become more difficult than ever. At present, the method of public listings in [5] has failed. If we want to continue to obtain data from OSN sites, the other two methods can still work - crawlers and API. But for Facebook, due to its strict access restrictions to use API, unauthorized developers are

978-0-7695-4745-9/12 $26.00 © 2012 IEEE
DOI 10.1109/TrustCom.2012.121

unable to get the friend list of users' friends. Because Facebook Query Language is one of the Facebook API, so the Facebook Query Language in [5] has also failed. But through crawlers, the user can still obtain a lot of user data for scientific research. Therefore, we choose crawlers as the main approach to obtain user data from Facebook. The goal of this paper is to study the network structure and privacy issues of OSNs.

The studies on OSN crawlers include [7, 8, 9]. [8] proposed two new unbiased crawl strategies: Metropolis-Hasting random walk (MHRW) and a re-weighted random walk (RWRW). [9] introduced a design of a social network parallel crawler. But [8, 9] did not present the implementation and technical challenge of the crawler in detail. [7] described the detailed implementation of a social network crawler. It used the breadth-first search and uniform sampling as the crawling strategies to run the crawler on Facebook, and then compared the two strategies. However, the crawler in [7] has a drawback: As Facebook limited each friend webpage to return at most 400 friends at that time, the crawler could only get at most 400 friends of a user.

Nowadays, Facebook has changed its privacy policy and further strengthened its access control policy against crawlers, which limits each friend webpage to return at most 60 friends, no longer 400. In this paper, based on interactive simulation, we have designed and implemented the crawler algorithm, which can obtain all the friend list of a Facebook user.

III. ALGORITHM

Based on the interaction simulation, we design and implement a multi-threaded crawler using Java. The crawler can obtain the complete friend list of users, and overcome the drawback in [7] that the crawler can get only the 400 friends of users. The detailed algorithm is shown in Algorithm 1. The crawler uses the strategy of breadth-first search and works as follows: First, the crawler automatically logs in with a real Facebook user credential, and then visits the seed profile and its friend webpage. Second, the crawler extracts the friend lists from the webpage through regular expression and put them in the queue to be visited. Finally, the crawler visits the profile in the queue to be visited in turn by FIFO (First In and First Out), and cycle the process. Due to the access control mechanism in Facebook, each friend webpage returns at most 60 friends. Therefore, in order to visit the complete friend list, the crawler must simulate the normal interaction behavior of a user with Facebook. Through analyzing the traffic between the browser and Facebook captured by HttpFox, a plugin for Firefox, we find that the interaction of the Ajax code embedded in the user's friend page and Facebook is always going as shown in Figure 1. By simulating the interactive behavior of the Ajax code with Facebook, we can extract the complete friend list, which is done through multiple extractions and get a list of at most 60 friends each time. The experiment result in section 4.2 proves the algorithm correctness.

In this paper, the crawler focuses on obtaining the social graph, i.e., the friend links. In order to protect user privacy, it only collects the user ID and do not collect user name, photos, other personal information or interaction information among users.

ALGORITHM 1: THE PSEUDO CODE OF CRAWLER ALGORITHM

```
1. Open the Facebook webpage;
2. Log in with a real user credential;
3. Visit the seed profile and its friend webpage;
4. Step 1:
5. get the friendNum; {extract the total number of
   friends of user from the friend webpage;}
6. If friendNum % 60 != 0
   timesofExtract = (friendNum - friendNum %
   60 + 60) / 60;
7. For (i = 2; i < timesofExtract; i++)
8.  Create a URI; {directing to a friend webpage
   including friend between 60*(i-1) and 60*i
   }
9.  Put the URI into the queue to be visited;
10. Go to step 3;
11. Step 2:
12. Visit the URI; {directing to a friend webpage
   including at most 60 friends in the queue to be
   visited}
13. Extract the friend list through regular
   expression;
14. Put the profile in the queue to be visited;
15. Step 3:
16. Visit the next item in the queue to be visited;
17. If the item is a profile
18.  Visit the profile and his friend webpage;
19.  Go to step 1 until the crawl ends;
20. If the item is a URI {directing to a friend list
   webpage including at most 60 friends}
21.  Go to step 2 until the crawl ends;
```

IV. DATA ACQUISITION AND ANALYSIS

A. Data Acquisition

To validate the algorithm, this paper used a notebook with Intel i5 CPU, 6G RAM, and broadband Internet access at a rate of 2Mb/s, and had been running the crawler from October 1, 2011 to November 1, 2011. We got a dataset of a total of 262,526 users. After data collection, we encrypted the user ID with MD5 function in order to protect the user privacy.

B. Data Analysis

In this section, we make visualization and analysis on the dataset crawled from Facebook.

First, we use the open source software, NodeXL [19], to make visualization on a subset of the collected dataset, as

shown in Figure 2, including a visualization of 5000 users, 10000 users, 20000 users and 30000 users subgraph.

Second, we make an analysis on the dataset. Formally, a social network is an undirected graph G = <V, E>, where V is the set of nodes (users) and E is the set of edges (links). General metrics on OSN structure include degree, clustering coefficient, diameter and so on. In this paper, we use the functions of SNAP (Stanford Network Analysis Platform) [14] for the data analysis, and the analysis results is shown in Table 1.

TABLE 1: THE RESULT OF DATA ANALYSIS

N. of unique visited user	262,526
N. of unique dis. neighbors	2,172,432
N. of unique edges	2,646,681
Avg. diameter	4.85
Minimal degree	1
Maxmum degree	4,918
Avg. degree	2.59
Avg. PageRank	0.99
Avg. clustering coefficient	0.04
crawling period	01/10/11-01/11/11

According to the Facebook statistics [4], each user in Facebook has more than 130 friends on average. However, the result in Table 1 is only 2.59. Moreover, the average clustering coefficient in Table 1 is 0.04, but the result in [8] is 0.16. Thus, we can conclude that the dataset crawled from Facebook is biased. The main reason is that breadth-first search used in this paper is a biased crawling strategy [18], which always skews towards the users with larger degree.

C. *Privacy Setting*

we find that only 166,802 users' friend list can be seen by public in the total 262,526 users crawled, which shows that 36.5% of the Facebook users have changed the default privacy setting and set the user's friend list invisible to strangers. In comparison to 26.6% and 16%, the results in [7, 20] respectively, the figure has increased by up to 20.5%, which implies that the Facebook user's awareness of privacy protection has been greatly improved.

D. *Summary*

In the 166,802 users whose friend lists are visible to the public, we find that 13,776 users have more than 60 friends, accounting for 8.3%; 1,728 users have more than 130 friends, accounting for 1.0%. It proves two facts: First, our crawler can get the complete friend list; second, due to the biased crawl strategy and the crawl time limitation, most of the user's friend list is still in the queue to be visited and not crawled completely.

V. CONCLUSION

In conclusion, based on the interaction simulation, we design and implement a crawler, which can obtain the complete friend list of Facebook users and overcome the drawback in [7] that the crawler can only obtain at most 400 friends of a user. We make an analysis and visualization of the dataset crawled from Facebook, and find that 36.5% of Facebook users have changed the default privacy settings, set the friends list invisible to strangers, compared to 26.6% and 16% in [7, 20] respectively, implying that the Facebook user's awareness of privacy protection has been greatly improved.

The future researches include: First, study distributed crawlers to improve the crawling performance; second, use unbiased crawling strategy to obtain unbiased dataset.

ACKNOWLEDGMENT

This work is supported by the National Nature Science Foundation of China (Grant No. 90818028).

REFERENCES

[1] D. M. Boyd and N. B. Ellison, "Social Network Sites: Definition, History, and Scholarship," J. Comp.-Mediated Commun., vol. 13, no. 1, Oct. 2007, pp.210−30.

[2] http://www.facebook.com

[3] http://goarticles.com/article/Social-network-first-run-win-the-search-engine/5686840/

[4] http://www.facebook.com/press/info.php?statistics

[5] Joseph Bonneau , Jonathan Anderson , George Danezis, Prying Data out of a Social Network, Proceedings of the 2009 International Conference on Advances in Social Network Analysis and Mining, p.249-254, July 20-22, 2009 [doi>10.1109/ASONAM.2009.45]

[6] Alan Mislove , Massimiliano Marcon , Krishna P. Gummadi , Peter Druschel , Bobby Bhattacharjee, Measurement and analysis of online social networks, Proceedings of the 7th ACM SIGCOMM conference on Internet measurement, October 24-26, 2007, San Diego, California, USA [doi>10.1145/1298306.1298311]

[7] Catanese, S., Meo, P.D., Ferrara, E., Fiumara, G., and Provetti, A. Crawling Facebook for Social Network Analysis Purposes. WISM 2011.

[8] M. Gjoka, M. Kurant, C. Butts, and A. Markopoulou. Walking in facebook: a case study of unbiased sampling of OSNs. In Proceedings of the 29th conference on Information communications, pages 2498−2506. IEEE Press, 2010.

[9] D. Chau, S. Pandit, S. Wang, and C. Faloutsos. Parallel crawling for online social networks. In Proceedings of the 16th international conference on World Wide Web, pages 1283-1284. ACM, 2007.

[10] http://www.80legs.com

[11] Huber, M., Mulazzani, M., Weippl, E.: Social networking sites security: Quo Vadis, in Proceedings of the Second International Conference on Social Computing (SocialCom),pp.1117-1122 (2010)

[12] Tao Stein , Erdong Chen , Karan Mangla, Facebook immune system, Proceedings of the 4th Workshop on Social Network Systems, p.1-8, April 10-13, 2011, Salzburg, Austria [doi>10.1145/1989656.1989664]

[13] J. Bonneau, J. Anderson, F. Stajano, and R. Anderson, "Eight Friends Are Enough: Social Graph Approximation via Public Listings," in SNS '09: Proceeding of the 2nd ACM Workshop on Social Network Systems, 2009.

[14] http://snap.stanford.edu/snap/index.html

[15] "Uniform sampling of facebook users: Publicly available datasets." http://odysseas.calit2.uci.edu/fb/, 2009.

[16] Ellison, N., Steinfield, C., & Lampe, C. (2007). The benefits of Facebook ''friends'': Exploring the relationship between college students' use of online social networks and social capital. Journal of Computer-Mediated Communication, 12(3), article 1. Retrieved July 30, 2007 from http://jcmc.indiana.edu/vol12/issue4/ellison.html

[17] "ComScore: Google+ Grows Worldwide Users From 65 Million In October To 67 Million In November". December 22, 2011.

http://techcrunch.com/2011/12/22/googlesplus/?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed%3A+Techcrunch+%28TechCrunch%29

[18] L.Becchetti, C.Castillo, D.Donato, and A.Fazzone, "A comparison of sampling techniques for web graph characterization," in LinkKDD, 2006.

[19] http://nodexl.codeplex.com/

[20] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou, "A walk in facebook: Uniform sampling of users in online social networks," http://arxiv.org/abs/0906.0060, 2009.

application/x-javascript	http://www.facebook.com/ajax/browser/list/friends/all/?uid=[redacted]&offset=60&dual=1&__a=1
image/jpeg	http://profile.ak.fbcdn.net/hprofile-ak-ash2/174476_[redacted]_7270992_q.jpg
image/jpeg	http://profile.ak.fbcdn.net/hprofile-ak-snc4/276073_[redacted]_101252561_q.jpg
image/jpeg	http://profile.ak.fbcdn.net/hprofile-ak-snc4/174490_[redacted]_7258240_q.jpg
image/jpeg	http://profile.ak.fbcdn.net/hprofile-ak-snc4/276269_[redacted]_5763679_q.jpg
image/jpeg	http://profile.ak.fbcdn.net/hprofile-ak-snc4/27379_[redacted]_9881_q.jpg
image/jpeg	http://profile.ak.fbcdn.net/hprofile-ak-snc4/275302_[redacted]_8080851_q.jpg
image/jpeg	http://profile.ak.fbcdn.net/hprofile-ak-snc4/49132_[redacted]_4686_q.jpg

(a)

image/jpeg	http://profile.ak.fbcdn.net/hprofile-ak-snc4/274247_[redacted]_6497068_n.jpg
image/jpeg	http://profile.ak.fbcdn.net/hprofile-ak-snc4/203304_[redacted]_394877_s.jpg
application/x-javascript	http://www.facebook.com/ajax/browser/list/friends/all/?uid=[redacted]&offset=120&dual=1&__a=1
image/jpeg	http://profile.ak.fbcdn.net/hprofile-ak-snc4/49852_[redacted]_8323_q.jpg
image/jpeg	http://profile.ak.fbcdn.net/hprofile-ak-snc4/275601_[redacted]_409240610_q.jpg
image/jpeg	http://profile.ak.fbcdn.net/hprofile-ak-snc4/275884_[redacted]_4278780_q.jpg
image/jpeg	http://profile.ak.fbcdn.net/hprofile-ak-snc4/41655_[redacted]_68_q.jpg

(b)

Figure 1. the interaction traffic captured by HttpFox: (a) get the friend list between 60 and 120; (b) get the friend list between 120 and 180

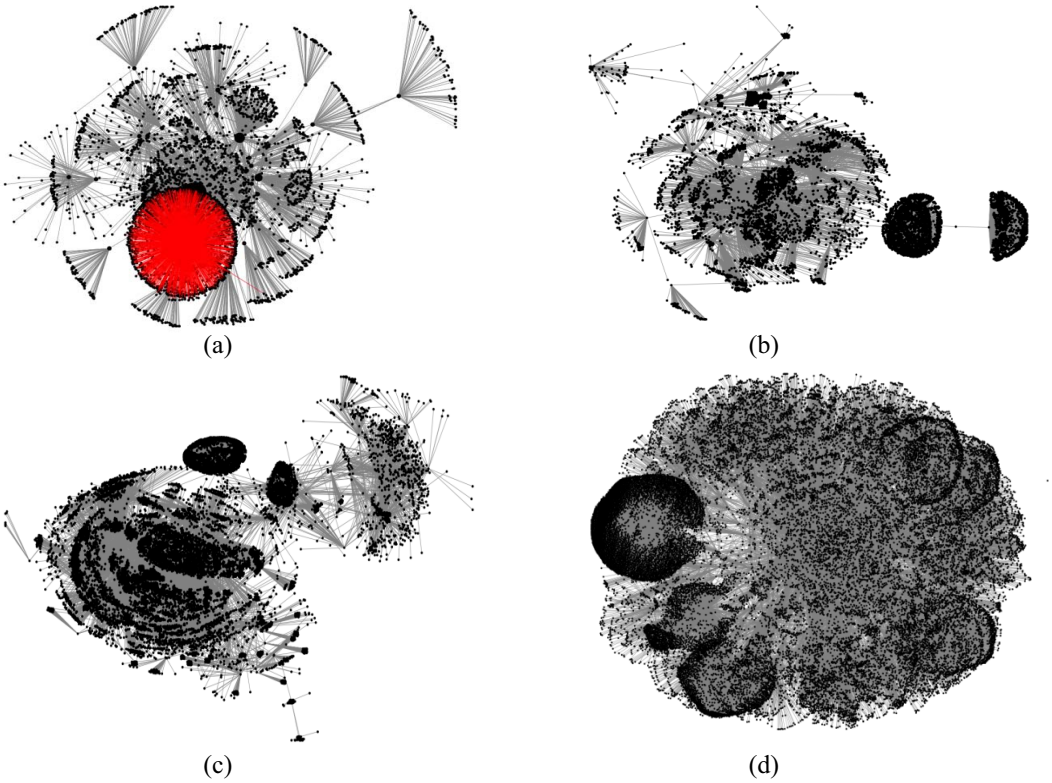

Figure 2. visualization: (a) 5000 users subgraph; (b) 10,000 users subgraph; (c) 20,000 users subgraph; (d) 30000 users subgraph