A Tale of Three Social Networks

User Activity Comparisons across Facebook, Twitter, and Foursquare

Despite recent efforts to characterize online social network (OSN) structures and activities, user behavior across different OSNs has received little attention. Yet such information could provide insight into issues relating to personal privacy protection. For instance, many Foursquare users reveal their Facebook and Twitter accounts to the public. The authors' in-depth measurement study examines users' network activities and privacy settings across Facebook, Twitter, and Foursquare. Results show that user activities are highly correlated among these three OSNs, which causes information leakage for a large fraction of Foursquare users.

Pinghui Wang and Wenbo He
McGill University, Canada

Junzhou Zhao
Xi'an Jiaotong University, China

Today, many people simultaneously use multiple online social networks (OSNs) for different services, such as keeping in touch with old friends on Facebook and retrieving popular news via Twitter. This motivates OSNs to interact and collaborate with each other, as when Twitter lets its users automatically post tweets to their Facebook profiles. Characterizing interactions among different OSNs is important and helpful for understanding people's network activities. Formally, we can model these multiple OSNs as a *multiplex network*.[1] While some have studied these networks' structures in terms of community detection,[1] evolution modeling,[2] and diffusion process analysis,[3] others employ common users as a bridge to facilitate knowledge transfer across different OSNs for applications such as link prediction and interest recommendation.[4,5]

Interactions among OSNs can also greatly increase the risk of personal information leakage. For example, many people release their Facebook and Twitter account information on their profiles for other OSNs, such as Foursquare, Pinterest, and Quora. Thus, attackers can obtain users' personal information by combining the data from multiple OSNs. Even when users don't list their account associations among different OSNs and carefully manage their privacy settings for each one, we find that an adversary could infer their account associations and sensitive information based on information revealed by their friends.

Previous work has studied Facebook, Twitter, and Foursquare separately

 Published by the IEEE Computer Society 1089-7801/14/$31.00 © 2014 IEEE

when it comes to applications such as network statistic measurement[6–9] and privacy analysis.[10–12] To raise awareness of the privacy issue among OSNs, we conducted a measurement study of users' network activities and privacy settings across Facebook, Twitter, and Foursquare. Based on our correlation analysis, we observed several new information leakage issues among these OSNs. Moreover, neither users themselves (by managing their privacy settings consistently across OSNs, for example) nor a single OSN service provider can achieve protection against the information leakage. We offer suggestions for both providers and users.

Data Collection

To collect datasets for our study, we first randomly sampled a fraction of Foursquare users and collected their profiles. Then, we collected the Facebook and Twitter profiles of sampled users who revealed links to those profiles on their Foursquare profiles. To characterize user activity dynamics, we collected sampled users' profiles once a day for two months.

Foursquare is a location-based OSN that motivates registered users to check in at certain venues (coffee shops, restaurants, shopping malls, and so on) by offering virtual or real-world rewards. Users can explore their venues of interest and leave tips and comments on the network. It has attracted considerable interest since its development in March 2009. As of September 2012, Foursquare had more than 25 million users (https://foursquare.com/about/). Because of its large size and rapid growth, our limited resources, and having only a short amount of time, we could not retrieve all network properties (for instance, users' privacy settings and friends). To address this problem, we sampled a fraction of Foursquare users at random.

Foursquare assigns each user a numeric ID with no more than eight digits. We can visit the profile of a user with ID d at https://foursquare.com/user/d. We uniformly sampled users by randomly generating numeric IDs between the minimum and maximum ID values. Because such values aren't assigned sequentially, a randomly generated ID might not correspond to a valid user, and we could waste considerable computational effort generating random IDs when the ID space is sparsely populated. Thus, we first evaluated how user IDs are dispersed over the ID space. On 23 December 2012, we queried 20,000 IDs sampled from the interval $[1, 10^8]$ at random. We observed that 99.9 percent of user ID values are smaller than 4.4×10^7. To increase the hit-to-miss ratio of sampling a validated user, we restricted the user ID values to the interval $[0, 4.4 \times 10^7]$. At each step, we queried an ID selected from $[0, 4.4 \times 10^7]$ at random. We repeated this step until we sampled 2×10^5 users. Finally, we collected Foursquare profiles of these randomly sampled users.

Among these users, 56.9 percent and 32.6 percent make their Facebook and Twitter accounts publicly available, respectively, and 16.6 percent reveal both accounts (*common users*). We collected the Facebook, Twitter, and Foursquare profiles of sampled common users once per day from January to March 2013; the measurement results we describe next are based on these samples.

Correlation of User Activities

We present our measurement results for studying the correlation of activities among Foursquare, Facebook, and Twitter users on two different levels: network and content.

Network-Level Analysis

For a user u, we denote by $f_{FS}(u)$ the number of his or her friends on Foursquare, $f_{FB}(u)$ the number of friends u has on Facebook, $f^o_{TW}(u)$ the number of friends u follows on Twitter, and $f^i_{TW}(u)$ the number of Twitter followers u has. We study the Spearman's rank correlation coefficients (SRCCs) for f_{FS}, f_{FB}, f^o_{TW}, and f^i_{TW}. SRCC is one of the most common metrics for measuring the dependence between two quantities. The SRCC of two variables ranges from –1.0 and 1.0 and assesses how well the relationship between the variables can be modeled using a monotonic function. A value of 1.0 implies that each variable is a perfect monotonically increasing function of the other, whereas a value of –1.0 implies that a perfect monotonically decreasing function exists between the two variables.

We conducted statistical analysis based on two user groups: highly connected and moderately connected. A user is highly connected if his or her f_{FS}, f_{FB}, f^o_{TW}, and f^i_{TW} are all larger than those same average values for all sampled users. Otherwise, the user is moderately connected. Note that 13 percent of the users we sampled are highly connected. Table 1 shows

Table 1. Spearman's rank correlation coefficients (SRCCs) for user network structure.

	Highly connected users			Moderately connected users		
SRCC	f_{FS}	f_{FB}	f^{o}_{TW}	f_{FS}	f_{FB}	f^{o}_{TW}
f_{FB}	0.49			0.48		
f^{o}_{TW}	0.34	0.32		0.32	0.31	
f^{i}_{TW}	0.44	0.38	0.83	0.41	0.36	0.82

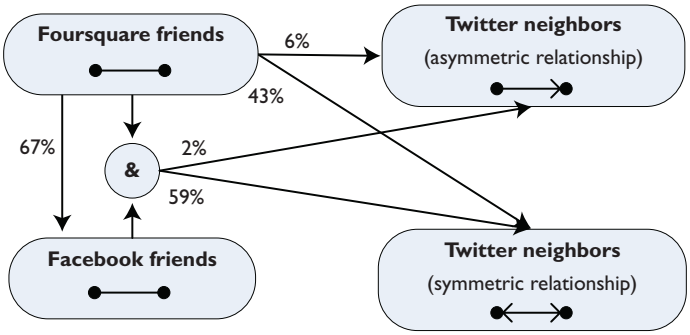

Figure 1. User relationship correlations across Foursquare, Facebook, and Twitter. All three online social networks are highly correlated and overlap with each other.

SRCCs for f_{FS}, f_{FB}, f^{o}_{TW}, and f^{i}_{TW}. We can see that these variables are positively correlated; their SRCCs are all larger than 0.3. SRCCs for the highly connected user group are slightly larger than those for the moderately connected users. This tells us that users with many friends on Foursquare also tend to have a large number of friends on Facebook and followers on Twitter.

Next, we answer the following questions: How likely are friends on Foursquare to be friends on Facebook? How likely are friends on Foursquare to follow each other on Twitter? Three relationship types are possible between two registered Twitter users:

- asymmetric — u follows v, or v follows u, with no reciprocity;
- symmetric — u and v follow each other; or
- none — u and v have no connection.

For two users who are friends on Foursquare and reveal their Facebook accounts, we can easily determine whether they're friends on Facebook when one makes his or her Facebook friend list publicly available. Similarly, we can determine two Foursquare friends' relationship on Twitter. Figure 1 shows our measurement results. We obtained these results based on a large number of sampled users who revealed their Facebook and Twitter accounts on Foursquare; they reflect the correlations among users' relationships across Foursquare, Facebook, and Twitter, although they might exhibit biases introduced by sampling disadvantages, such as self-selection. Note that removing and bounding the error is difficult because we don't have access to Facebook, Twitter, or Foursquare's entire network. From the figure, we can see that 67 percent of friend pairs on Foursquare are also friends on Facebook; 43 percent of friend pairs on Foursquare follow each other on Twitter. This appears much more frequently than asymmetric relationships. Moreover, 59 percent of user pairs who are friends on both Facebook and Foursquare follow each other on Twitter. This indicates that Twitter users are more likely to follow each other when they are friends on both Facebook and Foursquare. These findings show that all three OSNs are highly correlated, and overlap with each other.

Content-Level Analysis

For Foursquare, $c_{FS}(u)$ denotes the number of check-ins for user u, $p_{FS}(u)$ the number of photos u has uploaded to Foursquare, and $l_{FS}(u)$ the number of u's favorite Foursquare pages. Similarly, $p_{FB}(u)$ is the number of photos u has uploaded to Facebook, $l_{FB}(u)$ is the number of u's favorite Facebook pages, and $t_{TW}(u)$ is the number of tweets u posted on Twitter. Table 2 shows the SRCCs of these users' content attributes. We can see that they're positively correlated with each other. The SRCC of c_{FS} and p_{FB} is 0.28, and the SRCC of c_{FS} and t_{TW} is 0.40. This tells us that users who more frequently check in on Foursquare tend to post more photos on Facebook and tweets on Twitter. The SRCC of p_{FS} and p_{FB} is 0.25, and that of l_{FS} and l_{FB} is 0.36, which indicates that users uploading more photos to Facebook usually post more photos on Foursquare, and those with more favorite webpages on Facebook tend to expose more favorite webpages on Foursquare. These findings indicate that users' activities are well correlated among different OSNs.

We next study the correlations between the dynamics of user activities on the three OSNs. Let $\Delta c^{(m)}_{FS}(u)$ and $\Delta c^{(s)}_{FS}(u)$ denote the average and standard deviation of the number of user u's check-ins on Foursquare per day. We use $\Delta c^{(m)}_{FS}(u)$ to measure u's activity level on

Foursquare, and $\Delta c_{FS}^{(s)}(u)$ to measure the fluctuation in u's activities on Foursquare. Similarly, we denote $\Delta p_{FB}^{(m)}(u)$ and $\Delta p_{FB}^{(s)}(u)$ as the average and standard deviation of the number of photos u uploads to Facebook per day, and denote $\Delta t_{TW}^{(m)}(u)$ and $\Delta t_{TW}^{(s)}(u)$ as the average and standard deviation of the number of tweets u posts on Twitter per day. On average, a user checks-in to 0.8 venues on Foursquare, uploads 0.2 photos to Facebook, and posts 2.0 tweets on Twitter every day.

Again, we conducted a statistical analysis based on highly active and moderately active user groups. Users are highly active if their $\Delta c_{FS}^{(m)}$, $\Delta p_{FB}^{(m)}$, and $\Delta t_{TW}^{(m)}$ are larger than those average values for all sampled users. Otherwise, they belong to the moderately active group. We observe that 16 percent of sampled users belong to the highly active group. Table 3 shows the SRCCs for user activity dynamics on Facebook, Twitter, and Foursquare. We can see that $\Delta c_{FS}^{(m)}$ is more correlated with $\Delta t_{TW}^{(m)}$ than with $\Delta p_{FB}^{(m)}$, and $\Delta c_{FS}^{(s)}$ is more correlated with $\Delta t_{TW}^{(s)}$ than with $\Delta p_{FB}^{(s)}$. This indicates that users' check-in activities on Foursquare correlate more strongly with their tweeting activities on Twitter than with their photo uploading activities on Facebook.

Table 2. Spearman's rank correlation coefficients (SRCCs) for user content attributes.

SRCC	c_{FS}	p_{FS}	l_{FS}	l_{FB}	p_{FB}
p_{FS}	0.58				
l_{FS}	0.35	0.30			
l_{FB}	0.20	0.16	0.36		
p_{FB}	0.28	0.25	0.19	0.51	
t_{TW}	0.40	0.21	0.18	0.26	0.26

Table 3. Spearman's rank correlation coefficients (SRCCs) for user activity dynamics.

	Highly active users		Moderately active users	
SRCC	$\Delta c_{FS}^{(m)}$	$\Delta p_{FB}^{(m)}$	$\Delta c_{FS}^{(m)}$	$\Delta p_{FB}^{(m)}$
$\Delta p_{FB}^{(m)}$	0.26		0.22	
$\Delta t_{TW}^{(m)}$	0.46	0.24	0.43	0.21
	$\Delta c_{FS}^{(s)}$	$\Delta p_{FB}^{(s)}$	$\Delta c_{FS}^{(s)}$	$\Delta p_{FB}^{(s)}$
$\Delta p_{FB}^{(s)}$	0.22		0.22	
$\Delta t_{TW}^{(s)}$	0.43	0.23	0.41	0.22

Privacy Leakage across OSNs

Our findings confirm that users' activities across Facebook, Twitter, and Foursquare are greatly correlated. We next show that these correlations cause personal information leakage.

User Privacy Settings

Facebook and Twitter both let users manage their privacy settings to protect their personal information. For example, Facebook users can make their friend lists and photos available only to their friends. Twitter gives users a simple option to protect their tweet profiles by hiding lists of their tweets, followers, and who they follow from the public. We observe that 51 percent of sampled users hide their friend lists from the public, 19 percent conceal their gender, 49 percent conceal their hometown, 51 percent conceal their work and education, 91 percent conceal their birthdays, 51 percent conceal living places, and 17 percent protect their Twitter profiles. This indicates that people are more likely to protect their information and posts on Facebook than on Twitter. Figure 2 shows that a large number of users hide their attributes on Facebook, but don't protect their Twitter profiles. Thus, an adversary could obtain their hidden Facebook attributes from their Twitter profiles. For example, among sampled users who don't reveal their place of residence on Facebook, 80 percent don't protect their Twitter profiles; 57 percent of these users explicitly list their place of residence on Twitter.

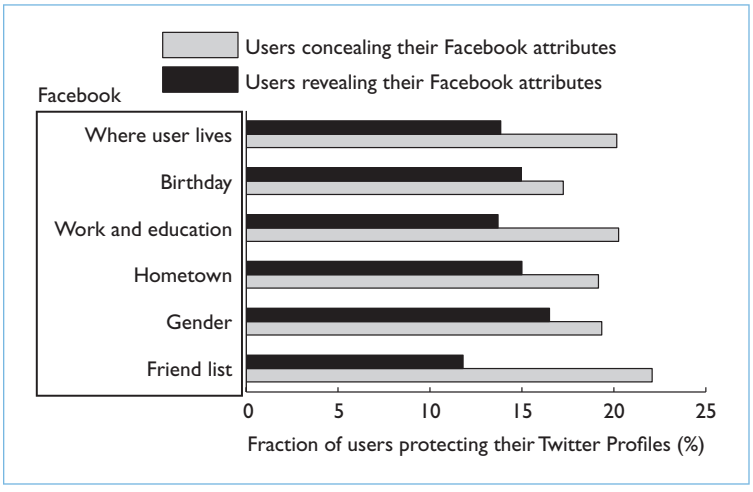

Figure 2. Correlations of user privacy settings across Facebook and Twitter. A large fraction of users protect their information and posts on Facebook but not on Twitter.

Correlations of Registered Names

We found that 99 percent of users' first names and the first letters of their last names are revealed on Foursquare profiles. Among these users, 81.0 percent and 64.4 percent have the same first names as on their Facebook and Twitter profiles, respectively. Among users revealing both their Facebook and Twitter accounts on their Foursquare profiles, 42.7 percent provide the exact same full names on their Facebook and Twitter profiles. These findings indicate that people tend to use the same name on different OSNs.

Account Association Leakage

As we've seen, people usually use the same or similar names on both Facebook and Foursquare, and friends on Foursquare are also friends on Facebook with a high probability. Using these two observations, we have a good chance of inferring the Facebook accounts of Foursquare users who otherwise are hiding these accounts. We randomly sampled 300 Foursquare users who didn't reveal their Facebook accounts. For each sampled user u, we used the following steps to infer his or her Facebook account:

1. Retrieve C, the set of u's friends on Foursquare.
2. Obtain C^*, the set of users in C who reveal their Facebook accounts on Foursquare.
3. For each user in C^* that makes his or her Facebook friend list publicly available, search the user's Facebook friend list to determine whether a friend v exists whose first name and last name's first letter are the same as that of u on Foursquare.
4. For each Facebook user v detected in step 3, manually verify whether u and v are registered by the same person based on their profile pictures and photos on Facebook and Foursquare.

Using this method, we successfully identified 54 percent of sampled users' Facebook accounts. We didn't get any results for 27 percent of the sampled users, possibly because they have no Facebook accounts, or they register different names on Facebook and Foursquare. For the other sampled users, we obtained their suspected Facebook accounts, but couldn't determine whether our results were correct for various reasons — for instance, the sampled users hadn't uploaded any profile pictures or photos to Foursquare. Similarly, we randomly sampled 300 Foursquare users who didn't reveal their Twitter accounts, and then inferred those accounts. We obtained 31 percent and 17 percent of sampled users' verified and suspected Twitter accounts, respectively. We thus inferred a large fraction of Foursquare users' Facebook and Twitter accounts.

Network Connection Leakage

Some Facebook users conceal attributes such as age and gender, information that's important to third parties such as insurance and advertising companies. However, these private attributes can be inferred from users' neighbors' publicly available information.[10] One study demonstrates that the properties of eight Facebook friends are usually enough to reveal a user's personal information.[13] This motivates a significant fraction of Facebook users to make their friend lists unavailable to the public. For example, we find that 51 percent of our sampled Facebook users hide their friend lists. To infer friends of a user u hiding his or her list, Ratan Dey and his colleagues propose a reverse lookup method that looks at all other Facebook users who disclose their friend lists and indicate they are friends with u.[14] However, their method is impractical because Facebook's size makes it prohibitive to query all users' friend lists. To address this problem, we propose a more efficient method for identifying u's Facebook friends, when u is also a Foursquare user. Friends on Foursquare are usually friends on Facebook, so we limit the reverse lookup operation to Facebook users who are friends of u on Foursquare. We randomly sampled 300 users who reveal their Facebook accounts on Foursquare but protect their Facebook friend lists. For each sampled user u with a Facebook account, we used the following steps to infer Facebook friends:

1. Retrieve C, the set of u's friends on Foursquare.
2. Obtain C^*, the set of Facebook accounts of users in C. Let $C_1 \in C$ be the set of users in C who explicitly reveal their Facebook accounts on Foursquare. Then, we can directly obtain the Facebook accounts of users in C_1. For other users in C, we detect their Facebook accounts using our previously described method.
3. For each user in C^* that makes his or her Facebook friend list publicly available, we search this list to determine whether u is a friend.

Using this method, on average, we detected 26 Facebook friends for a sampled user, with a standard deviation of 7.5. Our detected Facebook friends don't include any public accounts, such as celebrities. Similarly, we randomly sampled 300 users who reveal their Twitter accounts on Foursquare but protect their Twitter profiles, and then inferred their Twitter friends (following or followers). On average, we detected 15 Twitter friends for a sampled user, with a standard deviation of 6.2.

Clearly, user behaviors are highly correlated among different OSNs, and users' interactions across OSNs result in personal information leakage. We strongly recommend addressing these problems in the following ways. First, OSNs should let users remove their profiles from public search engines. Second, OSNs should disable the reverse lookup function — that is, they must remove users who conceal their friend lists from the publicly available friend lists of other users. Finally, OSNs such as Foursquare should let users make their attributes, such as friend lists, unavailable to the public.

References

1. P.J. Mucha et al., "Community Structure in Time-Dependent, Multiscale, and Multiplex Networks," *Science*, vol. 328, no. 5980, 2010, pp. 876–878.
2. J. Gómez-Gardeñes et al., "Evolution of Cooperation in Multiplex Networks," *Scientific Reports*, vol. 2, 2012, article no. 620.
3. S. Gómez et al., "Diffusion Dynamics on Multiplex Networks," *Physical Rev. Letters*, vol. 110, no. 2, 2013; doi: 10.1103/PhysRevLett.110.028701.
4. E. Zhong et al., "Comsoc: Adaptive Transfer of User Behaviors over Composite Social Network," *Proc. 18th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining*, 2012, pp. 696–704.
5. E. Zhong et al., "Modeling the Dynamics of Composite Social Networks," *Proc. 19th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining*, 2013, pp. 937–945.
6. W. He, X. Liu, and M. Ren, "Location Cheating: A Security Challenge to Location-Based Social Network Services," *Proc. 31st Int'l Conf. Distributed Computing Systems* (ICDCS 11), 2011, pp. 740–749.
7. A. Nazir, S. Raza, and C.-N. Chuah, "Unveiling Facebook: A Measurement Study of Social Network-Based Applications," *Proc. 8th ACM SIGCOMM Conf. Internet Measurement*, 2008, pp. 43–56.
8. H. Kwak et al., "What Is Twitter, a Social Network or a News Media?" *Proc. World Wide Web Conf.*, 2010, pp. 591–600.
9. M.A. Vasconcelos et al., "Tips, Dones, and ToDos: Uncovering User Profiles in Foursquare," *Proc. 5th ACM Int'l Conf. Web Search and Data Mining*, 2012, pp. 653–662.
10. A. Mislove et al., "You Are Who You Know: Inferring User Profiles in Online Social Networks," *Proc. 3rd ACM Int'l Conf. Web Search and Data Mining*, 2010, pp. 251–260.
11. N. Sharma et al., "Inferring Who Is Who in the Twitter Social Network," *Proc. 2012 ACM Workshop Online Social Networks*, 2012, pp. 55–60.
12. T. Pontes et al., "We Know Where You Live: Privacy Characterization of Foursquare Behavior," *Proc. 4th Int'l Workshop Location-Based Social Networks*, 2012, pp. 1–8.
13. J. Bonneau et al., "Eight Friends Are Enough: Social Graph Approximation via Public Listings," *Proc. 2nd ACM EuroSys Workshop on Social Network Systems*, 2009, pp. 13–18.
14. R. Dey et al., "Estimating Age Privacy Leakage in Online Social Networks," *Proc. IEEE Int'l Conf. Computer Communications* (INFOCOM 12), 2012, pp. 2836–2840.

Pinghui Wang is a postdoctoral researcher in the School of Computer Science at McGill University, Canada. His research interests include Internet traffic measurement and modeling, traffic classification, abnormal detection, and online social network measurement. Wang has a PhD in automatic control from Xi'an Jiaotong University, China. Contact him at wangpinghui369@gmail.com.

Wenbo He is an assistant professor in the School of Computer Science at McGill University, Canada. Her research interests include social network security, wireless networks, and software-defined networks. She has a PhD in computer science from the Multimedia Operating System and Networking Group at the University of Illinois at Urbana-Champaign. Contact her at wenbohe@cs.mcgill.ca.

Junzhou Zhao is a PhD candidate in the Systems Engineering Institute and Ministry of Education Key Lab for Intelligent Networks and Network Security at Xi'an Jiaotong University, China. His research interests include Internet traffic measurement and modeling, traffic classification, abnormal detection, and online social network measurement. Zhao has an MS in information engineering from Xi'an Jiaotong University. Contact him at jzzhao@sei.xjtu.edu.cn.