An Empirical Study of the Usage of the Swarm App's Cross-Site Sharing Feature

Shihan Lin[1], **Rong Xie**[1], **Yang Chen**[1], **Yu Xiao**[2], **Pan Hui**[3,4]
[1]School of Computer Science, Fudan University, China
[2]Department of Communications and Networking, Aalto University, Finland
[3]Department of Computer Science, University of Helsinki, Finland
[4]CSE Department, Hong Kong University of Science and Technology, Hong Kong

{shlin15,xieronglucy,chenyang}@fudan.edu.cn,
yu.xiao@aalto.fi, panhui@cse.ust.hk

UbiComp/ISWC'18 Adjunct, October 8–12, 2018, Singapore, Singapore

ACM ISBN 978-1-4503-5966-5/18/10...$15.00
https://doi.org/10.1145/3267305.3274170

Abstract
With the rapid development of online social networks (OSNs), many people have linked their accounts of multiple OSN sites and share contents across them. In this work, we conduct an empirical study of the usage of the Swarm app's cross-site sharing feature, i.e., the feature that enables Swarm users to share their check-ins to Twitter, and reveal factors that impact Swarm users' sharing behavior. We classify factors into two groups, i.e., check-in-related factors and profile-related factors, and dedicate to figure out their individual and combined influence on Swarm users' sharing behavior. Our work can provide a reference for researchers who collect Swarm check-ins from Twitter to study the characteristics of Swarm check-ins, assisting them to identify that whether their Twitter-collected check-ins are representative of the randomly selected check-ins collected directly from Swarm. The OSN sites can also improve their design of the sharing feature through the findings of this work.

Author Keywords
Online social network; cross-site linking; sharing feature; user behavior; Swarm App.

ACM Classification Keywords
[Human-centered computing]: Empirical studies in collaborative and social computing; [Social and professional topics]: User characteristics.

Introduction

As online social networks (OSNs) have become increasingly popular, many people have accounts on multiple OSNs [16]. To assist users to manage their different accounts, a couple of major OSNs, such as Swarm, Pinterest and Quora, have introduced the function of "cross-site linking" [8]. With this function, users are able to share their generated contents from one OSN site to its linked sites. For instance, if a user is going to post a check-in on Swarm and she has linked her Swarm account to Twitter, she can share this check-in to Twitter by using Swarm's "sharing" feature to show the check-in to her Twitter followers. Meanwhile, since an increasing number of users own multiple OSN accounts, many researchers have analyzed several OSNs together from different perspectives. However, most of these works focus on inferring knowledge about users from several OSNs [6], exploring the consistency of user behavior among different OSNs [16], and discovering the privacy security problem caused by the information aggregation of multiple OSNs [11]. The research about the factors that affect users' sharing behavior among multiple OSNs is still lacking.

In this paper, we seek to reveal the factors that affect users' cross-site sharing behavior. Namely, how users use an OSN's "sharing" feature. For example, we wonder whether users prefer to share their generated contents to other linked OSNs during a certain time period of a day? Whether the contents created at a certain category of places are more likely to be shared? We choose two popular OSNs, Swarm and Twitter, as examples to study.

This problem is important due to the following reasons: First, by encouraging users to share contents to established OSNs, newly built OSNs are able to enlarge the visibility. The critical step for successful encouragements is to know more about users' sharing behavior. Second, it is a common practice for researchers to collect Foursquare/Swarm check-ins from Twitter due to the difficulty of collecting the check-ins directly from Foursquare/Swarm [10, 4, 5, 18]. However, they can only get check-ins that are shared to Twitter instead of the randomly sampled check-in dataset gathered directly from Foursquare/Swarm. Thus, this trick raises an interesting question: Can the check-ins crawled from Twitter represent the randomly selected check-ins collected directly from Foursquare/Swarm? If we know the factors that affect users' sharing behavior, we are able to know the common features of the check-ins that are shared to Twitter. By going a step further, the representativeness of these shared-to-Twitter check-ins is clear. For instance, if a check-in's venue category is a critical factor affecting users' sharing behavior, and check-ins with the venue category of "Food" have a much larger fraction in shared-to-Twitter check-ins than in the complete set of Swarm check-ins, the shared-to-Twitter check-ins are not able to represent Swarm check-ins from the venue category perspective.

In this work, we collected the complete set of check-ins from more than 6,000 Swarm users who have linked their Swarm accounts to Twitter, and identified their shared-to-Twitter check-ins. Based on their more than 10 million check-ins, we investigate the factors that impact Swarm users' sharing behavior from both the check-in perspective and the user's profile perspective. Besides the individual factors, we also measure the impact of different combinations of these factors by machine learning techniques. Specifically speaking, we use the selected factors and the machine learning model to predict whether a check-in will be shared or not. The high prediction performance proves that those combinations of factors do affect users' sharing decisions. The F1-score and Area Under Curve (AUC) can reach 0.90 and 0.95, respectively.

Overall, our contributions are summarized as below.

- To the best of our knowledge, this is the first work that analyzes the factors impacting users' sharing behavior. We revealed that the following four factors affect the usage of Swarm's sharing feature: length of text attached to a check-in, whether the last check-in was shared, the user's gender, and number of the user's Twitter followers. With these influential factors, we are able to predict whether a check-in will be shared to Twitter or not with an F1-score of 0.90. Our findings can facilitate OSNs designing their sharing feature better. For instance, they are able to predict users' sharing behavior more accurately and provide targeted users with sharing shortcuts.
- Our findings can be referred by researchers who use shared-to-Twitter check-ins instead of check-ins directly collected from Foursquare/Swarm. They are able to determine the representativeness of their shared-to-Twitter check-ins. For example, it is reasonable for researchers to utilize Twitter-collected check-ins to analyze user behavior related to time, venue categories, and the number of Swarm friends, since these factors hardly affect users' sharing decisions. Their research results will be approximate to those obtained from check-ins that are directly collected from Foursquare/Swarm.

In the following parts of this paper, we first introduce related work and the dataset used in this work. Then we discuss the individual effect of factors, and investigate the effect of their combinations before we make a conclusion.

Related Work

A variety of literatures have examined multiple OSNs together and drawn conclusions from perspectives of cross-site linking [8, 20], information aggregation [6, 2], behavior and profile consistency [15, 19], and user privacy security [11]. Gong et al. [8] investigated the proportion of the adoption of different linking options in different Foursquare user groups and studied the relationship between filling optional fields of profile and enabling cross-site linking function. Farseev et al. [6] revealed that multiple OSNs' data mutually complement each other and if we fused their data appropriately, we were able to predict other attributes of a user more accurately. Pasquale et al. [15] aimed to answer the question that whether users behave similarly across OSNs or they behave in terms of the characteristics of an OSN. As for the user privacy, Irani et al. [11] showed that by aggregating a user's information from her different OSN accounts, there existed an unintended personal-information leakage problem. However, to the best of our knowledge, the factors that impact users' sharing behavior among multiple OSNs have not been examined yet.

The other line of literatures aimed to study the Foursquare users' behavior pattern through their check-ins. Since a Foursquare user's check-ins are visible to her friends only, they are unavailable for crawling. A popular strategy to circumvent this problem is to collect the check-ins from Twitter that are shared from Foursquare [10, 4, 5]. Brent and Monica [10] collected the Foursquare check-ins from Twitter, together with the geo-tagged posts of Twitter and Flickr, they examined the urban biases in these data which they called "volunteered geographic information". Cheng et al. [4] gathered Foursquare check-ins from the public Twitter feed and analyzed their spatial, temporal and social characteristics. Cranshaw et al. [5] utilized the Foursquare check-ins collected from Twitter to study the urban dynamics. However, there is no literature showing that the check-ins collected from Twitter are representative of the randomly selected check-in set on Foursquare/Swarm. In this work, by inves-

tigating how users use the sharing feature of the Swarm app, an OSN split from Foursquare in 2014, we are able to know whether those check-ins shared to Twitter have common characteristics and the representativeness could be examined.

Dataset Description

Foursquare and Swarm

Foursquare has been a representative OSN since 2009. Its users leave tips about places they have visited and make check-ins to show their current locations. In 2014, Foursquare separated the check-in function to form a new application called Swarm [3] and they share the same user account system. A check-in records the time, location and the venue category of the place that the creator visits. Users can also attach text with a check-in. If a Swarm user shares a check-in to his Twitter account's timeline, the user's Twitter followers know that this tweet is shared from the Swarm app. Since Twitter is a more popular OSN than Swarm, the sharing feature helps improve the visibility of Swarm.

On Foursquare/Swarm, venue categories are organized as a hierarchic category tree. A higher level category is divided into several lower level categories. We aggregate all the lower level categories into their corresponding top level categories. There are 10 top level venue categories in all, i.e., Arts, Events, Travel, Shops, Food, Outdoors, Nightlife, Residence, Colleges, and Professional.

Data collection

Since users' check-ins are only visible to their friends on Swarm, we use the same method as [14] to gather check-ins with users' consent. We registered Swarm accounts with the following description in profiles: "We are from Fudan University, China. We add friends for human behavior modeling. We respect your privacy and your data will only be used for our research." From February to April 2018, we randomly enumerated Swarm users' IDs and checked their Foursquare profiles [1] first to see whether they have linked their accounts to Twitter and what their Twitter IDs are if applicable. Then we sent Swarm friend requests to those who have linked their Swarm accounts to Twitter. Users were able to see our description in the profile when they received the friend request. For those who accepted our friend request, we crawled their tweets according to their Twitter IDs through Twitter API [2]. Finally, filtering out users that have not made any check-in in recent three months, we have the permits from 6,050 Swarm users and collected their complete set of check-ins.

As in [10], we call those tweets on Twitter that are actually check-ins shared from Foursquare/Swarm as *"check-in tweets"*. Since only the most recent 3,000 tweets are allowed to be accessed from the Twitter API, we synchronized the check-in dataset and tweet dataset by timestamp. Specifically speaking, for each user, we removed the check-ins generated before the creation time of the user's oldest tweet we gathered. Check-in tweets can be detected with the "source" attribute of tweets. If the "source" field shows "Foursquare" [3], then we know that this is a check-in tweet. Finally, 10,703,371 check-ins were left for our following research and 1,053,602 of them were shared to Twitter.

In this paper, we use *"whole group"* to represent the complete check-in set that were directly collected from Swarm (i.e., 10,703,371 check-ins). In the "whole group", we call the check-ins that were shared to Twitter (i.e., 1,053,602

[1] The cross-site linking information is only accessible through users' Foursquare profiles instead of their Swarm profiles.

[2] https://developer.twitter.com/content/developer-twitter/en.html

[3] Each check-in shared from Swarm is tagged with "from Foursquare" on Twitter.

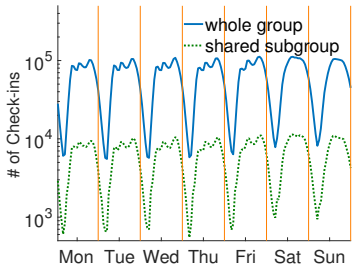

Figure 1: Number of check-ins per hour during a week

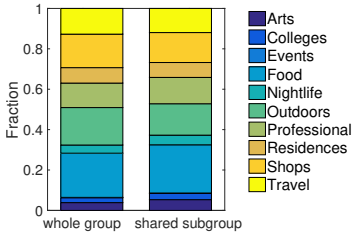

Figure 2: Fractions of check-in with different venue categories

check-ins) as *"shared subgroup"* and the rest of them as *"unshared subgroup"*.

Analysis of Individual Factors

We classify factors into two classes, i.e., check-in-related factors and profile-related factors.

For a check-in-related factor f, let V_f be the set of its different values. In the "whole group" and "shared subgroup" respectively, for each factor f and its v in V_f, let p_v be the fraction of check-ins with factor value v in all check-ins of the group. We define "*the check-in fraction distribution of a factor f*" as the set $\{p_v|v \in V_f\}$. If "the check-in fraction distribution of a factor" in these two groups show great difference, we consider that this factor has a great impact on the usage of the sharing feature.

Besides, we divide Swarm users into different groups according to the profile-related factors. We define a user's "*sharing fraction*" as the fraction of shared-to-Twitter check-ins in all check-ins created by this user. Then, we explore the Cumulative Distribution Function (CDF) of "sharing fraction" of those user groups. If the CDF curves of different user groups vary much, we consider that the factor, by which these user groups are divided, can greatly affect users' sharing behavior.

Check-in-related factors

For each check-in, check-in-related factors include its creation time, the venue category of the place it belongs to, the length of its attached text, and whether its' last check-in was shared. Figure 1-4 show the analysis results of the check-in-related factors.

Figure 1 shows the number of check-ins per hour during a week in the whole group and shared subgroup, respectively. Figure 2 demonstrates the check-in fractions of different venue categories in those two groups. From these two figures we can know that the check-in fraction distributions over both creation time and venue category in the whole group and shared subgroup are similar. In other words, check-in's creation time and its venue category can hardly affect users' sharing behavior.

Besides, to quantify the difference of check-in fraction distributions between two groups, we regard the fraction distribution in each group as a probability distribution, and calculate these two distributions' *Kullback-Leibler divergence* (KL divergence) [13]. For two probability distributions P and Q, the KL divergence $D_{KL}(P||Q)$, also called *relative entropy*, is a measure of the information loss when Q is used to approximate P. Therefore, it can be used to measure the difference between P and Q, and it ranges in $[0, +\infty)$. if its value is approximate to 0, it means that two distributions are very similar. On the contrary, the larger the KL divergence is, the more different manner two distributions behave. Although $D_{KL}(P||Q) \neq D_{KL}(Q||P)$, we only consider $D_{KL}(whole\ group\ ||\ shared\ subgroup)$ in this work since we care about to what extent the shared subgroup can be used to approximate the whole group.

The KL divergence values calculated from the two distributions over time in Figure 1 and the two distributions over venue category in Figure 2 are both less than 0.02, presenting that no salient distinction exists between these two groups with regard to those two factors.

We define the length of text as the number of bytes in the text. The reason why we use a byte instead of a word as the unit of the text length is that plenty of text in our dataset is not English. Considerable text is in the language whose sentences cannot be easily split into words, such as Chinese and Japanese. Therefore, with this definition, we are able to calculate the length of text in any language.

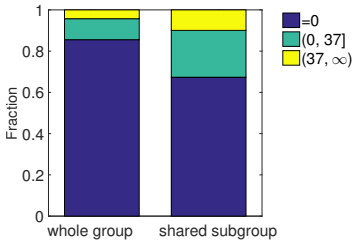

Figure 3: Fractions of check-in with different text length

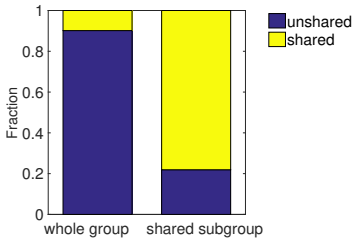

Figure 4: Fractions of check-in with different last check-in's sharing status

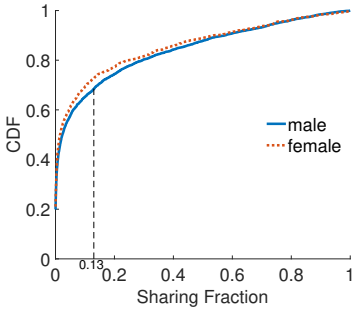

Figure 5: CDF of the sharing fraction according to users' gender

Figure 3 shows the check-in fractions of different text length intervals, i.e., no bytes, no more than 37 bytes, and more than 37 bytes. The threshold 37 is the average of non-zero text length in the whole group. Apparently, although the majority of check-ins are made without text in both of the whole group and shared subgroup, check-ins with text are more likely to be shared to Twitter, and the longer the text is, the more likely it is shared. It makes sense since it is more laborious to create a longer text, and users naturally prefer sharing it to other platforms so that she can attract attentions from more people. We also apply the KL divergence to the text length, and the value is 0.13, which is larger than values of temporal and venue category factors presented before (<0.02). Therefore, we conclude that longer text encourages a user to share check-ins.

Figure 4 presents the fractions of check-ins having different last check-in's sharing status. In the legend of Figure 4, "shared" represents check-ins whose previous check-in was shared to Twitter and "unshared" represents those whose previous one was not. It can be clearly observed that check-ins with their last check-in shared to Twitter possess a much larger fraction in the shared subgroup than in the whole group, and the KL divergence can reach 1.55 . It means that if a user shares a check-in to Twitter then it is highly probable that she shares her next check-in to Twitter as well.

One probable reason is the fact that Swarm by default shares a user's current check-in to Twitter if she has used the sharing feature in her last check-in. Although users can change this setting when they post check-ins, this setting encourages users to experience its sharing feature and helps Swarm attract more users from other OSNs. The other probable reason is that users' sharing behavior may have the characteristics of time locality. Namely, during a period of time, they enjoy the Swarm's sharing feature, so they share nearly each check-in to Twitter. However, once they feel bored of it, they seldom use it.

To sum up, among the check-in-related factors, whether the last check-in was shared shows great power on influencing users' sharing behavior, while the text length has influence to some extent. The factors of check-in's creation time and venue category hardly have influence. Besides, we can draw a conclusion that shared-to-Twitter check-in dataset is an alternative to Swarm check-ins if it is used for the analysis related to check-ins' creation time and venue category. However, researchers are not encouraged to adopt the attributes of text length and last check-in's sharing status when they are extracted from the shared-to-Twitter check-ins.

Profile-related factors

Profile-related factors include the check-in creator's gender, his number of friends on Swarm, and his number of followers on Twitter [4]. Figure 5-7 illustrate the relationship between profile-related factors and users' sharing behavior.

Figure 5 presents the CDF of male's and female's sharing fraction. Recall that a user's "sharing fraction" is defined as the fraction of shared-to-Twitter check-ins in all his check-ins. We can observe that the CDF curves is slightly different between each other, i.e., there are fewer males than females with their sharing fraction smaller than 13%. Thus, males mildly tend to share check-ins to Twitter in comparison with females.

For the factors of the number of Swarm friends and Twitter followers, we divide users into three groups according to each factor. We divide the number of Swarm friends into

[4] On Twitter, you follow someone means that you can see her tweets. If user A follows user B, then user A is called a follower of user B.

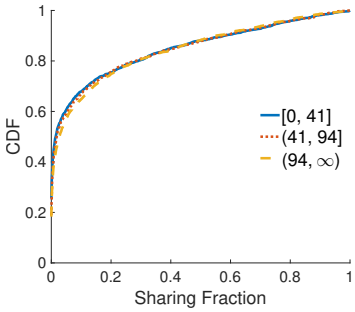

Figure 6: CDF of the sharing fraction according to users' number of Swarm friends

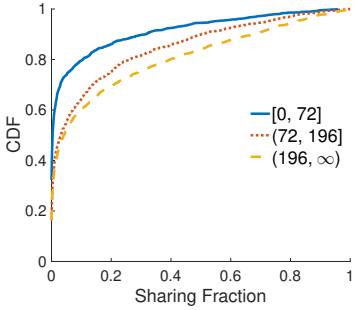

Figure 7: CDF of the sharing fraction according to users' number of Twitter followers

three intervals, i.e., no more than 41, between 41 and 94, and more than 94. The thresholds 41 and 94 are the quartile and median of users' friend count, respectively. We perform the same operation on the number of Twitter followers, with quartile of 72 and median of 196. Figure 6 and Figure 7 show the CDF of sharing fractions in different user groups divided by the above two factors, respectively. We can see that the CDF curves in Figure 7 vary much among one another while CDF curves in Figure 6 almost coincide. It means that users with more Twitter followers are more likely to share their Swarm check-ins to Twitter. However, the amount of their Swarm friends does not affect their sharing decision. The different results of these two factors demonstrate that users care more about connections with Twitter friends when they are using Swarm's sharing feature.

To sum up, males share more check-ins to Twitter, and users who have more Twitter followers are more likely to share check-ins to Twitter. Besides, since the check-in creators' number of Swarm friends do not affect users' sharing behavior, check-ins shared to Twitter are representative of the check-ins randomly collected from Foursquare/ Swarm from this perspective.

Analysis of the Combinations of Factors

We have clarified the effect of individual factors on users' sharing behavior, but it remains unclear that how the combinations of these factors will affect users' sharing behavior. Will those factors that individually contribute little to users' sharing behavior become influential when they are combined together? Can we construct a model to predict users' sharing actions accurately with the help of those influential factors? This model can be used by the Swarm app to improve user experience on its sharing feature, such as providing sharing shortcuts for those who is predicted to share a check-in.

In this section, we utilize machine learning algorithms to investigate the impact of factors' combinations. According to the analysis results in the previous section, we call the factors of "check-in's creation time", "venue category", and "the number of Swarm friends" as "*weak factors*"; we call the factors of "the last check-in's sharing status", "text length","gender", and "the number of Twitter followers" as "*strong factors*". A combination of factors is considered to be influential if a machine learning model, which uses these factors as features for input, can accurately predict whether a check-in will be shared to Twitter. All individual factors mentioned before are adopted as a check-in's features, and we set the profile-related features of a check-in as its creator's profile-related factors. Therefore, each check-in has both check-in-related features and profile-related features. We conduct the experiments with three combinations as shown below:

Case 1: weak factors only
Case 2: strong factors only
Case 3: weak factors and strong factors

In Case 1, we combine the weak factors to see whether their combination has a different performance from when they are used separately. Meanwhile, to further validate the factors that we reveal to be influential in the previous section, in Case 2, we combine them to see whether their combination performs the same. Finally, in Case 3, we aggregate weak and strong features to find out how accurately we can predict users' sharing behavior.

Obviously, it is an imbalanced learning task with 1,053,602 shared check-ins and 9,649,769 unshared check-ins. Therefore, we perform a simple randomly undersampling [9] on the major class: we randomly select 1,053,602 check-ins

from the unshared subgroup, and merge them to the shared check-ins. Finally, a new dataset with balanced classes is used for learning. We employ Decision Tree, Random Forest [17], LightGBM [12], and XGBoost [1] to make the classification, and five-fold cross-validation is used for evaluation. The results are shown in Table 1.

In Table 1, three metrics are used for evaluation: accuracy, F1-score, and AUC (Area Under Curve). We regard shared check-ins as positive instances and unshared ones as negative instances. Accuracy is the fraction of correctly predicted instances. F1-score is the harmonic mean of precision and recall. Precision is the fraction of correctly predicted positive instances in all predicted positive instances, and recall is the fraction of correctly predicted positive instances in all really positive instances. According to [7], AUC means the probability that a classifier will rank a randomly selected positive instance higher than a randomly selected negative one.

In Case 1, none of these algorithms can predict sharing behavior accurately, and F1-scores and AUCs are around 0.60 and 0.65, respectively. However, all algorithms in Case 2 achieve quite good results, with all scores around 0.90. The results are almost the same to the results in Case 3 but with fewer features. The results in Case 1 indicate that even the combination of those weak factors can hardly impact users' sharing behavior. Besides, the algorithms perform well with features in Case 2, confirming that the factors of text length, whether the last check-in is shared, the user's gender, and the number of her Twitter followers have great influence on a user's sharing willingness.

To sum up, the combination of weak factors has little impact on users' sharing behavior and the combination of strong factors performs much better. For researchers, these results justify the method of collecting check-ins from Twitter,

Algorithms \ Metrics		Accuracy	F1	AUC
Case 1	Decision Tree	0.63	0.62	0.69
	Random Forest	0.61	0.61	0.67
	LightGBM	0.61	0.59	0.65
	XGBoost	0.66	0.65	0.73
Case 2	Decision Tree	0.90	0.89	0.95
	Random Forest	0.90	0.89	0.95
	LightGBM	0.89	0.89	0.94
	XGBoost	0.90	0.89	0.95
Case 3	Decision Tree	0.91	0.90	0.95
	Random Forest	0.90	0.89	0.94
	LightGBM	0.90	0.89	0.95
	XGBoost	0.91	0.90	0.96

Table 1: Machine learning results

not only when they use the weak factors separately, but also when the combinations of those factors are considered. Furthermore, with only the strong features, we are able to accurately predict whether a Swarm check-in will be shared to Twitter or not.

Conclusion and Future Work

In this work, we investigate OSN users' sharing behavior among their linked accounts. By taking Swarm and Twitter as examples, we find out the influential factors that affect the usage of the Swarm app's cross-site sharing feature. Specifically, check-ins with longer text slightly tend to be shared, and males share more check-ins to Twitter to a small extent. Besides, users' sharing behavior is closely related to whether the last check-in is shared, and users who have more Twitter followers are much more likely to share check-ins to Twitter. By using machine learning algorithms, the above factors' combination is able to predict whether a

check-in will be shared to Twitter. The prediction's F1-score and AUC can reach 0.90 and 0.95, respectively.

For OSNs who have cross-site sharing feature, our findings can help them understand how this feature is used by users and design the feature in a better way. For example, OSNs are able to improve user experience on the sharing feature by predicting their sharing willingness more accurately. For researchers who collect the check-ins from Twitter instead of directly from Foursquare/Swarm, they are able to know the representativeness of their data from different perspectives. For instance, it is valid to use Twitter-collected check-ins in the analysis related to time, venue categories, the number of Swarm friends, or their combinations.

This work is a first step into the investigation of users' sharing behavior across multiple OSNs, providing references for the future work. Since limited factors and OSNs are discussed in this paper, future work can focus on exploring the relationship between other factors and users' sharing behavior, such as the geographic location that check-ins are created at or the semantic meaning of check-in text. Researchers can also work on case studies to better justify the conclusions.

Acknowledgement

This work is sponsored by National Natural Science Foundation of China (No. 61602122, No. 71731004), Natural Science Foundation of Shanghai (No. 16ZR1402200), Shanghai Pujiang Program (No. 16PJ1400700), Projects 26211515 and 16214817 from the Research Grants Council of Hong Kong. Yang Chen is the corresponding author.

REFERENCES

1. Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A Scalable Tree Boosting System. In *Proc. SIGKDD*. ACM, 785–794.
2. Terence Chen, Mohamed Ali Kaafar, Arik Friedman, and Roksana Boreli. 2012. Is More Always Merrier?: A Deep Dive into Online Social Footprints. In *Proc. WOSN*. ACM, 67–72.
3. Yang Chen, Jiyao Hu, Hao Zhao, Yu Xiao, and Pan Hui. 2018. Measurement and Analysis of the Swarm Social Network With Tens of Millions of Nodes. *IEEE Access* 6 (2018), 4547–4559.
4. Zhiyuan Cheng, James Caverlee, Kyumin Lee, and Daniel Z Sui. 2011. Exploring Millions of Footprints in Location Sharing Services. In *Proc. ICWSM*. 81–88.
5. Justin Cranshaw, Raz Schwartz, Jason Hong, and Norman Sadeh. 2012. The Livehoods Project: Utilizing Social Media to Understand the Dynamics of A City. In *Proc. ICWSM*. 58–65.
6. Aleksandr Farseev, Liqiang Nie, Mohammad Akbari, and Tat-Seng Chua. 2015. Harvesting Multiple Sources for User Profile Learning: A Big Data Study. In *Proc. ICMR*. ACM, 235–242.
7. Tom Fawcett. 2006. An Introduction to ROC Analysis. *Pattern Recognition Letters* 27, 8 (2006), 861–874.
8. Qingyuan Gong, Yang Chen, Jiyao Hu, Qiang Cao, Pan Hui, and Xin Wang. 2018. Understanding Cross-site Linking in Online Social Networks. *ACM Transactions on the Web* (2018).
9. Haibo He and Edwardo A Garcia. 2008. Learning from Imbalanced Data. *IEEE Transactions on Knowledge & Data Engineering* 21, 9 (2008), 1263–1284.
10. Brent J Hecht and Monica Stephens. 2014. A Tale of Cities: Urban Biases in Volunteered Geographic Information.. In *Proc. ICWSM*. 197–205.

11. Danesh Irani, Steve Webb, Kang Li, and Calton Pu. 2011. Modeling Unintended Personal-information Leakage from Multiple Online Social Networks. *IEEE Internet Computing* 15, 3 (2011), 13–19.

12. Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Proc. NIPS*. 3146–3154.

13. Solomon Kullback. 1997. *Information Theory and Statistics*. Courier Corporation.

14. Shihan Lin, Rong Xie, Qinge Xie, Hao Zhao, and Yang Chen. 2017. Understanding User Activity Patterns of the Swarm App: A Data-driven Study. In *Proc. UbiComp*. ACM, 125–128.

15. Pasquale de Meo, Emilio Ferrara, Fabian Abel, Lora Aroyo, and Geert-Jan Houben. 2013. Analyzing User Behavior Across Social Sharing Environments. *ACM Transactions on Intelligent Systems and Technology* 5, 1 (2013), 14:1–14:31.

16. Raphael Ottoni, Diego B Las Casas, Joao Paulo Pesce, Wagner Meira Jr, Christo Wilson, Alan Mislove, and Virgílio AF Almeida. 2014. Of Pins and Tweets: Investigating How Users Behave Across Image-and Text-Based Social Networks. In *Proc. ICWSM*. 386–395.

17. Fabian Pedregosa, Gaël Varoquaux, and et al. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12, Oct (2011), 2825–2830.

18. Daniel Preoţiuc-Pietro and Trevor Cohn. 2013. Mining User Behaviours: A Study of Check-in Patterns in Location Based Social Networks. In *Proc. WebSci*. ACM, 306–315.

19. Pinghui Wang, Wenbo He, and Junzhou Zhao. 2014. A Tale of Three Social Networks: User Activity Comparisons Across Facebook, Twitter, and Foursquare. *IEEE Internet Computing* 18, 2 (2014), 10–15.

20. Changtao Zhong, Mostafa Salehi, Sunil Shah, Marius Cobzarenco, Nishanth Sastry, and Meeyoung Cha. 2014. Social Bootstrapping: How Pinterest and Last. fm Social Communities Benefit by Borrowing Links from Facebook. In *Proc. WWW*. ACM, 305–314.