Co-Location Social Networks: Linking the Physical World and Cyberspace

Huandong Wang*,Yong Li*, Yang Chen†, Yue Wang*, Jian Yuan*, Depeng Jin*
* Tsinghua National Laboratory for Information Science and Technology
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
† School of Computer Science, Fudan University, Shanghai 200433, China

***Abstract*—Various dedicated web services in the cyberspace, e.g., social networks, e-commerce, and instant communications, play a significant role in people's daily-life. Billions of people around the world access them through multiple online identifiers (IDs), and interact with each other in both the cyberspace and the physical world. These two kinds of interactions are highly relevant to each other. In order to link between the cyberspace and the physical world, we propose a new type of social network, i.e., co-location social network (CLSN). A CLSN contains online IDs describing people's online presence and offline interactions when people come across each other. By analyzing real data collected from a mainstream ISP in China, which contains 32.7 million IDs across most popular web services, we build a large-scale CLSN, and evaluate its unique properties. The results verify that the CLSN is quite different from existing online and offline social networks in terms of different classic graph metrics. This paper is the first research to study CLSN at scale and paves the way for future studies of this new type of social network.**

I. Introduction

Thanks to the rapid development of the Internet and web technologies, various dedicated web services in the cyberspace, i.e., social networking services (SNS), e-commerce, and instant communication, are growing quickly. They have attracted billions of users around the world, and have become an important part of people's daily life. It is quite normal for an individual user to have multiple identifiers (IDs) in the cyberspace, such as SNS accounts, email addresses, and instant messenger accounts. By referring to these accounts, people interact with each other online, for example, sending messages to each other. Moreover, more and more offline social events, ranging from informal get-togethers (e.g., movie night and dining out) to professional activities (e.g., technical conferences and business meetings), use online platforms to perform the organizing. Therefore, online interactions will trigger people's movement and interactions in the physical world. Meanwhile, people's offline interactions will further boost the interactions in the cyberspace. Given these facts, to understand Internet users' behavior, it is important to make observations from both online and offline perspectives, and consider them as an integrated whole.

On one hand, there are numerous works studying online social networks by investigating data from SNS platforms [1], [2], e-mail networks [3], and instant messenger networks [4]. However, all these work do not take the users' offline interactions into account. On the other hand, researchers have also investigated a lot on users' social interactions in the physical world, for example, human mobility patterns [5]. However, little has been done to systematically explore the links between the online and offline social networks. In this paper, we propose a new type of social networks, known as *co-location social networks (CLSN)*. In a CLSN, we put the users' online IDs and their offline social interactions together. As a result, this new network can capture the face-to-face social interactions in participating events in the offline physical world when the online IDs are appearing in the same locations. In this paper, we conduct a data-driven investigation for CLSNs. It is based on real data collected from a mainstream Internet service provider in China with 32.7 million online IDs across different popular online services in one month. The constructed network can not only demonstrate the presence of online IDs, but also get insights from their mobility and co-existence in the physical world.

To the best of our knowledge, this is the first work to create and investigate a large-scale co-location social network with millions of nodes, and link between online IDs in the cyberspace and the offline interactions among users in the physical world. Our study reveals many unique aspects of the constructed CLSN, which are different from conventional pure online or offline social networks. In terms of the network structure, our findings include the existence of a giant connected component, a high average degree, and a strong locality of social interactions. In addition, we found that adding more types of online IDs to a CLSN can significantly increase the network connectivity, which implies a synergistic relationship among these different types of IDs.

The rest of the work is organized as follows. In Section II, we introduce our data set and formally define CLSN. Then, we examine the network structure of CLSN in Section III. Finally, we draw our conclusion and discuss the potential future works in Section IV.

II. Co-location Social Networks: Data Collection and Construction

In this section, we first introduce the collection of data that records massive users' online and offline behaviors. Then, we introduce the concept of the co-location social network (CLSN) and give its formal definition.

ASONAM 2016, August 18-21, 2016, San Francisco, CA, USA
978-1-5090-2846-7/16/$31.00

A. Data Collection and Processing

The data set we used is collected by a mainstream Internet service provider (ISP) in China. To observe the online behavior of users, our study focuses on a series of representative online services in China, i.e., QQ (online instant messenger), Weibo (online social network), Taobao (online shopping site), and cell phone, which are summarized in Table 1 with their website URLs and total number involved in our data set. All of them are the leading and most popular web services among the corresponding categories in China. By sniffing millions of broadband subscribers in Shanghai city, the ISP performs deep packet inspection (DPI) to capture users' login actions to aforementioned online services from each subscriber. As soon as a user accesses one of these services, the login action will be recorded. The data collection was from Nov. 1 to Nov. 30, 2015, and the collected data trace is as large as 50 GB.

There are 470 million entries in our data set. Each entry contains following fields: name of the online service, online ID, identity of the broadband subscriber, Operating System (OS) the user used, and login time, which is accurate to hours. Let us look at a sample entry: <Weibo, 123456, 789, iOS, 2015112113>. It represents a user launched an iOS-based Weibo application at 13PM Nov. 21, 2015 with ID 123456, and the identity of the corresponding subscriber is 789.

To preserve user privacy, the online IDs and subscriber identities are anonymized. Our data set includes 3.4 million subscribers and 32.7 million online IDs for different types of service. This large-scale data set guarantees the credibility of our analysis of user behaviors in the physical world and cyberspace.

B. Co-Location Social Network: Definition and Construction

We aim to build a network to link users' online and offline activities. Thanks to our massive data set, we are able to accurately determine whether two online IDs are co-located, i.e., getting connectivity from the same subscriber in a certain time period, of which the duration is one hour in our dat set. A subscriber can be either the broadband interface of an apartment, which includes both the wired and WiFi traffic of a family, or the broadband interface of a company, which includes traffic from several subnetworks. Thus, different subscribers are corresponding to different locations. If two online IDs get connectivity from the same subscriber in the same time period, they are at the same location in that time period. Then, we can look into the frequency of their co-location behaviors. For online IDs belonging to the same person, or a small group of people who are living or working together, i.e., family members or close friends, they would have a high probability

TABLE I: The services and types involved in our study.

Services	Types	Website	Number
QQ	Instant messengers (IM)	qq.com	11M
Taobao	E-commerce	taobao.com	15M
Weibo	Online social networks (OSN)	weibo.com	2M
Cell phone	-	-	4M

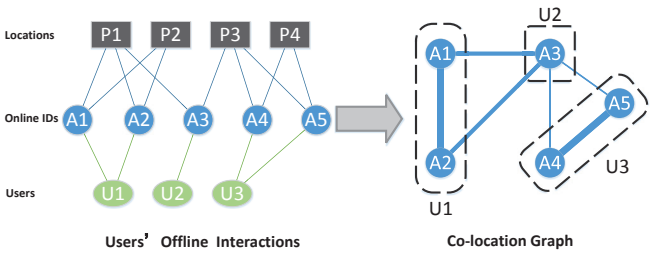


Fig. 1: Illustration of definition and construction of CLSN.

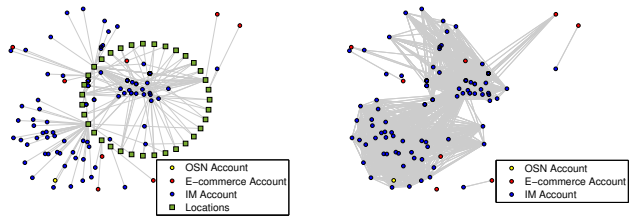
to "meet", i.e., accessing the Internet from the same subscriber at the same location. Based on this intuition, we build a new social network to characterize the relationship between online IDs by referring to their offline co-location activities.

Our newly introduced network can be represented by a network $G = (V, E)$. As we use "co-location" information to construct the network, we denote our newly introduced network as *co-location social network (CLSN)*. Similarly to existing social networks, CLSNs capture social interactions among users. In addition, different from traditional online social networks, which characterize online interactions between virtual IDs in the cyberspace (e.g., exchanging messages, sharing photos), and offline social networks, representing offline interactions between human beings in the physical world, CLSNs incorporate two important social elements as follows.

- Online virtual IDs as nodes: In network G, the node set V are online virtual IDs. As one physical user might have several online IDs on one or multiple online services, one user might own more than one node in the network.
- Offline social interactions as edges: In network G, the edge set E includes observed offline social interactions among the nodes. If two online IDs appear at the same location in the same time period, we would infer that they are "co-located", and an edge will be created between them.

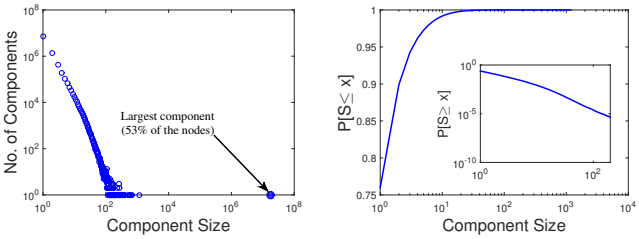
Definition 1 (CLSN) Co-location social network (CLSN) is a weighted undirected graph $G = (V, E)$, in which each node $a \in V$ represents an online ID. For two IDs $a_1 \in V$ and $a_2 \in V$, if they have ever accessed the Internet from the same location in the same time period, there is an edge $e = (a_1, a_2) \in E$. The link weight $w(e)$ of edge e between a_1 and a_2 characterizes the frequency of co-location behaviors of a_1 and a_2. If they appear in the same location very often, a large weight is assigned to the corresponding edge.

Fig. 1 demonstrates how we extract the online IDs and the corresponding users' offline interactions to build the graph G. As shown in the left part of Fig. 1, we have three users $U1$, $U2$, and $U3$. $U1$ owns the IDs $A1$ and $A2$, $U2$ owns the ID $A3$, and $U3$ owns the IDs $A4$ and $A5$. On one hand, online IDs belonging to the same user always have a high probability to login from the same place. According to the right part of Fig. 1, we can observe the edges $(A1, A2)$ and $(A4, A5)$ are very thick. On the other hand, online IDs belonging to different



(a) Online IDs' offline interactions (b) Co-location graph

Fig. 2: An example of the constructed CLSN with 50 nodes.



(a) Component number distribution (b) CDF and CCDF

Fig. 3: Connectivity and component size of the CLSN.

users might still login from the same place. In our example, we assume that $U1$ and $U2$ are very close friends, and they have a high chance to meet from time to time. Meanwhile, we assume that $U2$ and $U3$ are ordinary friends, and they meet occasionally. According to the right part of Fig. 1, we can observe the edges $(A1, A3)$ and $(A2, A3)$ are much thicker than the edges $(A3, A4)$ and $(A3, A5)$.

To show the constructed CLSN in a visualized example, we sample 50 online IDs with their corresponding locations from our data set, and plot the graph of the offline interactions between these IDs and locations in Fig. 2(a) as well as the corresponding CLSN in Fig. 2(b). This visualized CLSN, which is a part of the whole network, shows a unique structure with properties like the existing of a giant component with small islands. These will be thoroughly analyzed in next section.

III. Network Structure of CLSN

In this section, we study the graph structure of the CLSN constructed by the entire data set by referring to several classic graph metrics, including component size, path length, node degree, and clustering coefficient.

Using the number of component distribution in Fig. 3(a), a complementary cumulative distribution function (CCDF) and a cumulative distribution function (CDF) in Fig. 3(b), we show the connectivity of CLSN. As shown in Fig. 3(a), in this network, about 17.3 million nodes (53% of nodes in the network) are connected with each other, forming the largest connected component. The other 15.4 million nodes form 2.31 million small connected components and 7.33 million isolated nodes. Fig. 3(b) displays the distribution of the connected components. From the results, we can observe that the size of small connected components follows a power-law distribution. Compared with other social networks, for example, the instant-messaging network in [4], whose giant component covers about 99.9% of the nodes, the giant component of CLSN is relatively smaller. Moreover, there are much more small-scale connected components in the CLSN, indicating relatively large number of users tend to use online service in private places rather than public places.

As a unique feature of CLSNs, there are multiple types of online IDs, as mentioned before. To evaluate the difference among these types of IDs in terms of their impact, we study different subsets of the network by referring to different combinations of ID types. The results are shown in Fig. 4.

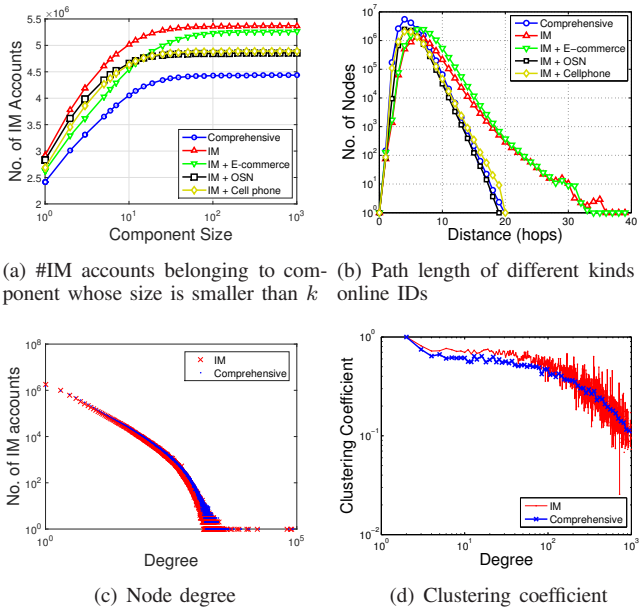
In our data set, 34.14% of all IDs are IM ones, with the most highest frequency of appearing. Therefore, we study CLSN based on these IDs, and show different kinds of combinations in Fig. 4(a). As we can observe, for the IM-only network, these are 5.37 million isolated IM accounts, i.e., they do not belong to the largest connected component. However, when we consider all online IDs, the number of isolated IM accounts will be reduced to 4.44 million. To compare the impact between other types of IDs, we study the "IM+E-commerce", "IM+OSN" and "IM+Cell phone" networks. From the result, we can observe that if we add the OSN accounts or cell phone to the IM-only network, the number of isolated IM accounts will be reduced by 0.52 and 0.48 million, respectively. Differently, if we add the e-commerce accounts to the IM-only network, it will be reduced by 0.11 million only. Therefore, the OSN and cell phone accounts play a more significant role in connecting the network node. The main reason for this phenomenon might be that people tend to use their OSN and cell phone accounts in public places, while they tend to use their e-commerce accounts in private places such as home.

Fig. 4 (b), (c) and (d), and Table II show three other key static properties of the network, i.e., the path length, node degree, and clustering coefficient of the network. In these results, we compare between the IM-only network and the comprehensive network with all kinds of online IDs.

The distribution of the path length of the largest connected component is shown in Fig. 4(b). Due to the giant size of this component, we do not calculate path length of every node pair. Instead, we randomly selected 10 nodes, and calculate the path length from all the other nodes to them to obtain the distribution. From the result, we can observe that by adding other types of online IDs, the average path length

TABLE II: Statistics of Static Structure of Co-location Social network.

Parameter	IM-only network	Comprehensive network
Average path length	6.71	4.85
Average node degree of IM accounts	35.19	42.00
Average Clustering Coefficient	0.4137	0.3824



(a) #IM accounts belonging to component whose size is smaller than k

(b) Path length of different kinds of online IDs

(c) Node degree

(d) Clustering coefficient

Fig. 4: Static properties with different kind of online IDs.

can be reduced. The impact of different types of online IDs is similar to that shown in Fig. 4(a). That is, the influence of the OSN account and cell phone is almost the same, and much larger than the e-commerce account, also indicating that people tend to use the e-commerce account more in private places compared with other kind of online service. As for the comprehensive network, we find that the distribution reaches the peak at 5 hops, and the average path length is 4.85. The co-location of online IDs is corresponding to users' "encounter" in physical world, and the distance between online IDs must be longer than that of their corresponding users in the physical world. It further indicates that, for example, a virus can be infected from an arbitrary user to another by less than 6 times of "encounter" on average in the physical world, which reflects the small world phenomenon in the CLSN.

The distribution of the node degree is shown in Fig. 4(c). For the comprehensive network, we can find that there are 42.00 edges per nodes on average, which means that each ID appeared at the same places at the same time period with about 42 other IDs on average. If ignoring multiple IDs belonging to same person, we can infer that the number of persons that the owner of the online ID met all over the month is about 42, which is consistent with our priori knowledge. In addition, it is not surprising that the node degree for the IM accounts in the comprehensive network is larger than that in the IM-only network, in which the average node degree is 35.19. It is because that expect for the existing edges between IM accounts, there are also edges between IM accounts and other kind of online IDs in the comprehensive network.

In terms of the distribution of clustering coefficient shown in Fig. 4(d), we can observe the average clustering coefficient is much larger than that in the online social networks, such as 0.063 in Renren [2], 0.164 in Facebook [1], and 0.106 in Twitter [6]. Therefore, online IDs in the CLSN are tightly connected.

Overall, the CLSN is a new kind of social network consisting of a giant connected component and many other small connected components, which is similar to other kind of social network. In addition, it has small diameter, large average node degree and clustering coefficient, indicating the existence of significant small world phenomenon and tight connection in the CLSN.

IV. Conclusion and Discussion

In this work, we propose the idea of co-location social networks (CLSNs). By using a data set covering the login activities of 32.7 million online IDs and 3.4 million locations in one month, we build a large-scale CLSN to link the online IDs and offline co-location interactions. Then, we thoroughly analyze the static structure of the CLSN by referring to its four key properties. We believe we can further adopt the CLSNs to address a very broad of important problems. We list two possible future work as follows.

Dynamic properties analysis: Besides the static network structure our study has covered, dynamic evolution properties of the constructed CLSN is also an important aspect, which can be further investigated by analyzing a series of daily snapshots.

Physical user detection: CLSN is a social network constructed based on virtual online IDs and physical co-location interactions. Thus, another basic and fundamental problem here is how to link the cyberspace IDs to physical world users, i.e., detecting all online IDs of one user by using the co-location behaviors of online IDs.

Acknowledgment

This work is supported by National Basic Research Program of China (973 Program) (No. 2013CB329105), National Nature Science Foundation of China (No. 61301080, No. 61171065 and No. 61273214), Natural Science Foundation of Shanghai (No. 16ZR1402200) and Shanghai Pujiang Program (No. 16PJ1400700).

References

[1] C. Wilson, B. Boe, A. Sala, K. P. Puttaswamy, and B. Y. Zhao, "User interactions in social networks and their implications," in *Proc. 4th EuroSys*, 2009, pp. 205–218.

[2] J. Jiang, C. Wilson, X. Wang, P. Huang, W. Sha, Y. Dai, and B. Y. Zhao, "Understanding latent interactions in online social networks," in *Proc. 10th ACM IMC*, 2010, pp. 369–382.

[3] L. Adamic and E. Adar, "How to search a social network," *Social Networks*, vol. 27, no. 3, pp. 187 – 203, 2005.

[4] J. Leskovec and E. Horvitz, "Planetary-scale views on a large instant-messaging network," in *Proc. 17th ACM WWW*, 2008, pp. 915–924.

[5] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: user movement in location-based social networks," in *Proc. 17th ACM SIGKDD*, 2011, pp. 1082–1090.

[6] A. Java, X. Song, T. Finin, and B. Tseng, "Why we twitter: understanding microblogging usage and communities," in *Proc. 9th WebKDD and 1st SNA-KDD*, 2007, pp. 56–65.