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

Huandong Wang, *Student Member, IEEE,* Yong Li, *Senior Member, IEEE,* Yang Chen, *Senior Member, IEEE,* and Depeng Jin, *Member, IEEE*

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 these services through multiple online identifiers (IDs), and interact with each other in both the cyberspace and the physical world. Thanks to the rapid development of wireless and mobile technologies, nowadays these two kinds of interactions are highly relevant with 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 "encountering" events when people come across each other. By analyzing real data collected from a mainstream ISP in China, which contains 32.7 million IDs across the most popular web services, we build a large-scale CLSN, and explore its unique properties from various aspects. The results indicate that the CLSN is quite different from existing online and offline social networks in terms of classic graph metrics. Moreover, we propose a community-based user identification algorithm to find all online IDs belonging to the same physical user. Using some ground-truth data, we demonstrate that our proposed algorithm achieves a high accuracy in user identification. Finally, we perform a user-centric analysis, and we demonstrate the behavioral difference among different types of users.

Index Terms—social networks, graph analysis, cyber-physical systems.

1 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 do 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, not only for IDs on the same online platform, but also for IDs coming from multiple platforms. 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 papers 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, where each ID uniquely corresponds to a user, and is attached to all entries generated by the user in the corresponding service. 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.

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 static 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. Furthermore, we look at the forming of the CLSN from a dynamic perspective.

- *H. Wang, Y. Li and D. Jin are with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (E-mail: liyong07@tsinghua.edu.cn).*
- *Y. Chen is with the School of Computer Science, Fudan University, Shanghai 200433, China.*

Based on the constructed network, we aim to further map the online IDs in the cyberspace to users in the physical world, i.e., detecting physical users from how these cyber ID interact with each other. This is a very challenging problem, as two IDs might come across each other no matter they belong to the same user or not. To achieve an accurate mapping, we propose a community-based user identification algorithm. By checking against the ground-truth data, we validate that our algorithm achieves a high accuracy in finding online IDs belonging to the same user. We also perform a detailed investigation about the impact of key parameters on its performance.

Based on the discovered users, we finally perform a series of user-centric analysis. By running K-means clustering algorithm we divided users into three types, i.e., business-dominated users, entertainment-dominated users, and comprehensive users. Our study shows the difference among these types from different aspects, e.g., temporal behaviors, and ID composition. For example, we quantitatively characterize the difference between entertainment-dominated users and business-dominated users in terms of the frequency of using online social networks, instant messengers, and e-commerce services, during weekdays and weekends.

The rest of the paper is organized as follows. After presenting related works in Section 2, we introduce our data set and formally define CLSN in Section 3. We examine the statistic network structure and dynamic properties of CLSN in Section 4. Then, we investigate the user identification problem in the CLSN in Section 5 and further investigate the links between IDs of physical world and cyberspace in Section 6. Finally, we draw our conclusion and discuss the potential future works in Section 8.

2 Related Work

The recent growth and popularity of online social networks (OSNs) such as Facebook, Twitter, and Linkedin has lead to a surge in measurement and analysis of OSNs [4], [6]–[9]. Existing work includes the analysis of static network structures [4], [10], dynamic network growth patterns [11], [12], link strength modeling [8], and social link prediction [13]. Moreover, even before the emergence of OSNs, offline social networks in the physical world has been an important research topic for tens of years [14]. Existing work include studying the origin of social relationships [15], and detecting social interactions [16]. However, none of these literatures has considered the online and offline social networks as an integrated whole, while they are highly related to each other. On the other hand, some online systems have recently started to integrate users' offline activities. There are two existing types, i.e., location-based social networks (LBSNs) [5], [17], [18], and event-based social networks (EBSNs) [19]. In LBSNs such as Foursquare and Gowalla, an individual user can share his latest location to the platform by conducting a "check-in". However, LBSNs do not record users' interactions in the physical world. In EBSNs, users can organize offline events through online platforms such as Meetup (www.meetup.com) and Plancast (www.plancast.com). As a result, the platform can track users' co-attendance of a certain event. Still, other types of offline interactions organized offline, for example, events not registered in the platform, cannot be tracked at all. Moreover, for both LBSNs and EBSNs, users' online information and interactions are restricted to a single site, while in practice, people might communicate with each other via multiple online platforms. In our study, we aim to link the physical world and cyberspace using massive data collected from a mainstream ISP. Our work differs from an earlier work published by Cranshaw et al. [20] significantly. First, we involve multiple types of online IDs to accurately capture users' online activities from different aspects, while they only use the user information from Facebook. Second, in terms of scale, our study involves 32.7 million online IDs, while their study covers 389 Facebook users with very infrequent interactions among users. In short, we adopt richer online-offline information to investigate a large group of millions of users, and our study can draw a comprehensive picture of online-offline social interactions.

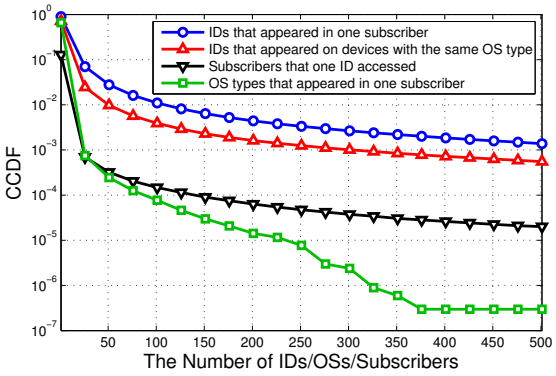
TABLE 1
Dataset summary.

# Records	# Online IDs	# Subscribers	# Services
470 million	32.7 million	3.4 million	4

In terms of user identification, linking accounts of the same user across datasets are recognized as an important open problem studied in diversity contexts [21]–[30]. Most existing solutions rely on either utilizing different portions of the same dataset [21]–[23] or observing the same behavior across thematically similar domains [24], [28]. Specifically, Korula et al. [21] focused on linking user IDs based on the friendship graph. Similarly, Kazemi et al. [22] presented another graph matching algorithm that relies on smaller seeds than other approaches. Goga et al. [23] used user profile attributes such as user name, profile photos. Narayanan et al. [24] linked users of Netflix and IMDB based on the similarity of their movies ratings. Zhang et al. [25] focused on linking users based on their social graph. Naini et al. [26] focused on linked users by matching their statistics. Mu et al. [27] described a multi-platform user ID linkage (MUIL) problem, and used "latent user space" for linking user profiles. Another work from Goga et al. described a set of similarity features for ID linking such as timestamp of posts and writing styles [28]. Gao et al. [29] proposed an unsupervised method to link users based on their attributes and social features. Riederer et al. [30] proposed a generic and self-tunable algorithm of cross-domain user linking which utilizes the temporal-spatial behaviors of humans to build a new type of maximum weight matching. However, most of existing approaches can only match user identifiers of two domains, while there are multiple types of online IDs involved in our analysis to accurately capture users' online activities from different aspects. What's more, we focus on linking online IDs of the same user by utilizing their offline social interactions, which differs from existing approaches significantly.

3 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,



(a) The relations between online IDs, OSs and subscribers

Items	Mean	Standard Deviation
IDs in one subscriber	12.70	122.28
IDs on devices with the same OS type	6.06	69.59
Subscribers accessed by one online ID	1.31	41.68
OS types appeared in one subscriber	2.45	2.57

(b) Related statistics

Fig. 1. Basic properties about the collected data set.

we introduce the concept of co-location social networks (CLSN) and give its formal definition.

3.1 Data Collection and Processing

The data set we used is collected by a mainstream Internet service provider (ISP) in China. It records users' accessing activities via broadband subscribers, which are associated to a physical locations, e.g., a WiFi access point or a broadband interface. 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 ones 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.

Table 1 presents a summary of the dataset. We can observe that 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 in our data set. Let us look at a sample entry: <Weibo, 123456, 789, iOS, 2015112113>. This entry 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. Over 3.4 million subscribers and 32.7 million online IDs for different types of services are involved in our dataset. In addition, to preserve user privacy, the online ID and subscriber identity are anonymized. Overall, we denote L as the set of all locations (broadband subscriber) and denote T as the set of time bins, of which the size is 1 hour in our dataset. In addition, we define V as the set of all online IDs, and we let S represent the set of types of online IDs. For each ID $v \in V$, we denote $s(v)$ as its type (service). Further, $\forall$ location $l \in L$, we use binary variable $X_{vl}^t \in \{0, 1\}$ to represent whether user v appeared in location l on the time bin t.

TABLE 2
Services involved in our study.

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

Characteristics. We now provide an informative overview of the data set. We are interested in the following four metrics, i.e., the number of online IDs that appear in one subscriber, the number of subscribers accessed by one online ID, the number of online IDs that appear on devices with the same OS type, and the number of different OSs that appear in one subscriber, of which the complementary cumulative distribution function (CCDF), the mean and standard deviation are shown in Fig. 1(a) and (b), respectively. We can observe that there are in average 12.70 online IDs that appear in a single subscriber, and a single online ID accesses 1.31 subscribers in average. In addition, there are 6.06 online IDs that appear in devices with the same OS type for one subscriber, and 2.45 different kinds of OSs that appear in one subscriber on average. Though the dimension of subscribers is more coarse-grained compared with the dimension of devices, it is unique and can be corresponding to true locations in physical world. On the other hand, since we can only know the types of OS in the dimension of devices, we still cannot distinguish online IDs of different users. In addition, the sharing of devices is pretty common in physical world, and it may introduce extra noise by using this information. Thus, we ignore the information of devices in the following analysis.

Strengths and Limitations. Our data set includes 3.4 million subscribers and 32.7 million online IDs for different types of services, from which we can observe users roam around in the physical world while accessing online services in cyberspace space. This large-scale data set guarantees the credibility of our analysis of user behaviors in the physical world and cyberspace. On the other, there are also some limitations of data set. First, the timestamps are only accurate to hours in our data set. Then, online IDs and subscriber identities in our dataset are anonymized. Thus, we cannot utilize users' profile information as gender, age in investigating their behavior. Third, though the latitude and longitude of broadband subscribers are also known in our dataset, which arrive accurately decimally hind 4, there is strong hybridity of urban functions in China, e.g., residential areas are very close to business areas. Thus, we cannot obtain the profile information of subscribers (locations) directly from the latitude and longitude. Instead, we use the statistics of user behavior in these subscribers to

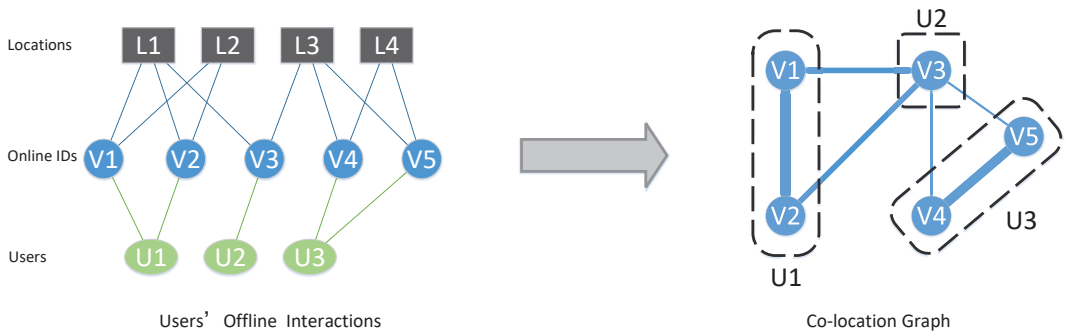


Fig. 2. Illustration of definition and construction of co-location social network.

characterize their attributes.

3.2 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 data 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 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 both the traditional online social networks, which characterizes 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 the 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 bin, we would conclude that they are "co-located", and an edge will be created between them.

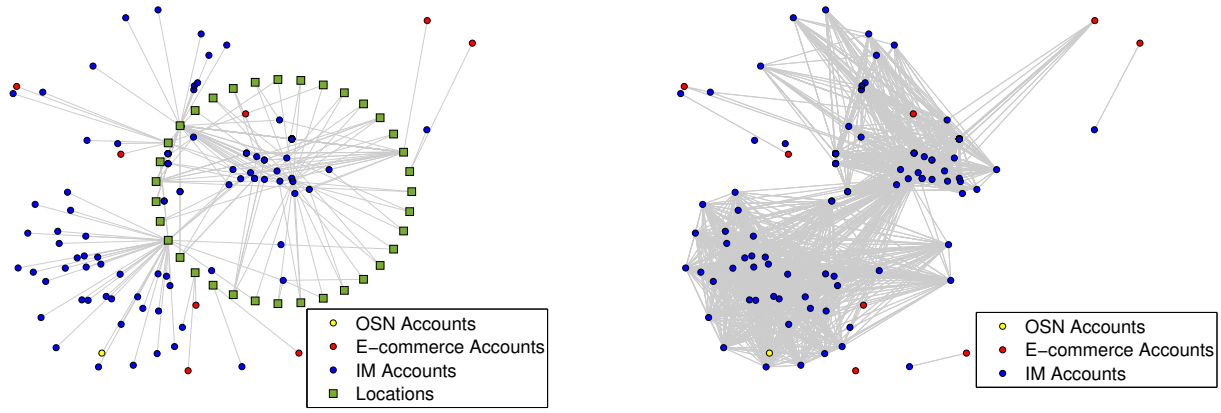
Definition 1 (CLSN) Co-location social network (CLSN) is a weighted undirected graph $G = (V, E)$, in which each node $v \in V$ represents an online ID. For two IDs $v_1 \in V$ and $v_2 \in V$, if they have ever accessed the Internet from the same location in the same time bin, there will be an edge $e = (v_1, v_2) \in E$. The link weight $w(e)$ of edge e between v_1 and v_2 characterizes the frequency of co-location behaviors of v_1 and v_2. If they appear in the same location very often, a large weight will be assigned to the corresponding edge.

Fig. 2 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. 2, we have three users $U1$, $U2$, and $U3$. $U1$ owns the IDs $V1$ and $V2$, $U2$ owns the ID $V3$, and $U3$ owns the IDs $V4$ and $V5$. 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. 2, we can see the edges $(V1, V2)$ and $(V4, V5)$ are very thick. On the other hand, online IDs belonging to the other 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. 2, we can see the edges $(V1, V3)$ and $(V2, V3)$ are much thicker than the edges $(V3, V4)$ and $(V3, V5)$.

To show the constructed CLSN in a visualized example, we sample 50 online IDs with their corresponding locations where they appear in our data set, and plot the graph of the offline interactions between these IDs and locations in Fig. 3(a) as well as the corresponding CLSN graph in Fig. 3(b). This visualized CLSN, which is a part of the whole network, shows a unique structure and properties like the existing of a giant component with small islands. These will be thoroughly analyzed in next section.

4 CLSN: Structure and Dynamics

In this section, we analyze the constructed CLSN from two different angles. On one hand, we look at the aggregation view, i.e., the CLSN constructed by the entire data set, to study its graph structure by referring to several classic graph metrics. On the other hand, we examine the CLSN from a



(a) Online IDs' offline interactions

(b) Co-location graph

Fig. 3. Examples for the constructed CLSN with 50 nodes.

dynamic view by analyzing a series of daily snapshots and its evolution.

4.1 Constructed CLSN: A Static Perspective

In this section, we analyze the CLSN by evaluating several classic graph metrics, including the component size, path length, node degree, and clustering coefficient. A number of works have studied these metrics in different social networks, e.g., Renren [2], Facebook [1], Twitter [31], and MSN [4]. In our work, we aim to reveal the main similarities and differences between the CLSN and existing social networks in terms of these metrics.

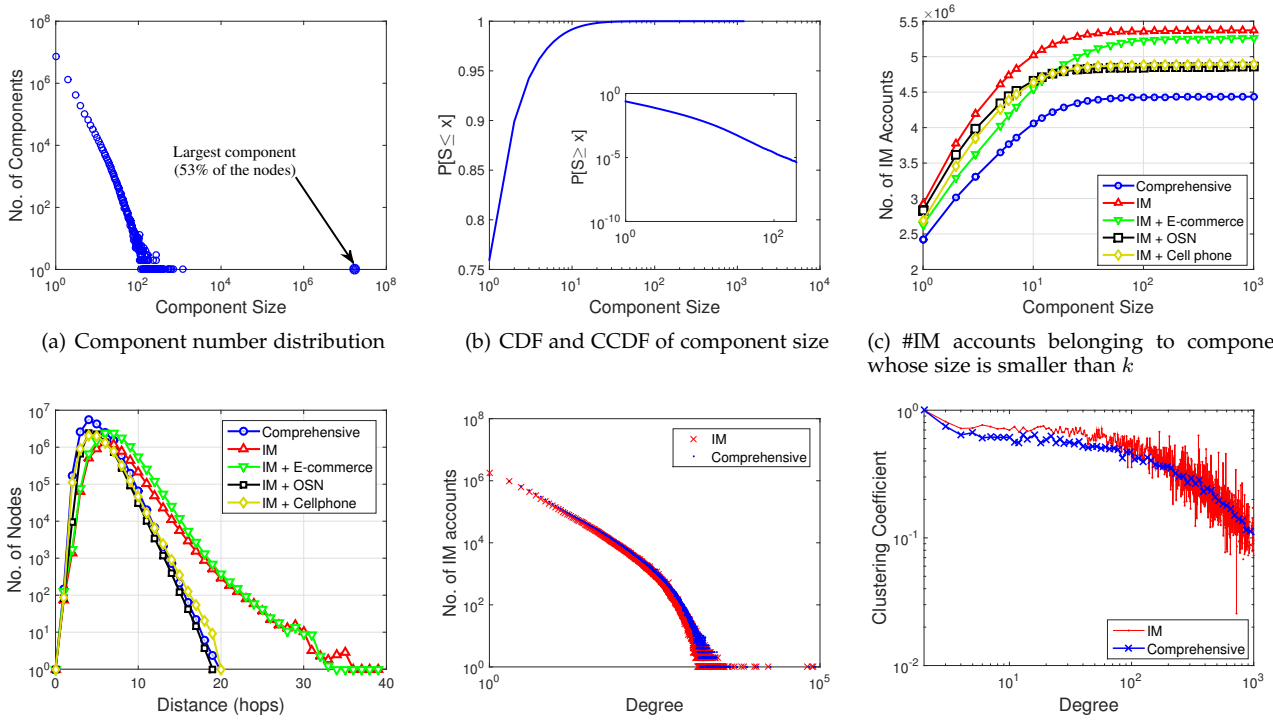
Using the number of component distribution in Fig. 4(a), a complementary cumulative distribution function (CCDF) and a cumulative distribution function (CDF) in Fig. 4(b), we show the connectivity of CLSN. As shown in Fig. 4(a), in this network, about 17.3 million nodes (53% of nodes in the network) 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. 4(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. Although there is a huge connected component covering more than 50% of users, there are still a large number of small-scale connected components in the CLSN, indicating the relatively large number of users tend to use online services in private places rather than public places.

As a unique feature of CLSNs, there are multiple types of online IDs. 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.

By studying our data set, we find that 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(c). Specifically, for the IM-only network, there 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 getting the network connected. The main reason for that phenomenon might be that people are cautious in using their E-commerce accounts, and they would prefer to use them in private places such as home. Differently, they are more willing to use their OSN and cell phone accounts in public places.

In Fig. 4 (d), (e) and (f), and Table 3, we present three key static properties of the network, that is, the diameter of the network, and the distribution of the node degree and the clustering coefficient. In these figures and tables, we compare between the IM-only network and the comprehensive network.

The distribution of the path length of the largest connected component is shown in Fig. 4(d). Due to the giant size of this component, we do not calculate path length of every node pair. Instead, we randomly select 10 nodes, and calculate the path lengths from all the other nodes to them, and obtain its distribution. From the result, we can observe that by adding other types of online IDs, the average path length can be reduced. The impact of different types of online IDs is similar to that shown in Fig. 4(c). That is, the influence of the OSN accounts and cell phones are quite similar to each other, and much larger than the E-commerce accounts, also indicating that people tend to use the E-commerce accounts more in private places compared with other kind of online services. As for the comprehensive



(a) Component number distribution (b) CDF and CCDF of component size (c) #IM accounts belonging to component whose size is smaller than k

(d) Path length of different kinds of online IDs (e) Node degree (f) Clustering coefficient

Fig. 4. Static properties of the constructed co-location social network.

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 larger 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 physical world, which reflects the small world phenomenon in the CLSN.

The distribution of the node degree is shown in Fig. 4(e). For the comprehensive network, we can find that there are about 42.00 edges per nodes on average, which means that each ID appeared at the same places in 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 account 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 except for the existing edges between IM accounts, there are also edges between IM accounts and other kind of online IDs in the comprehensive network.

As for the distribution of clustering coefficient, which is shown in Fig. 4(f), 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 [31]. 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 networks. 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.

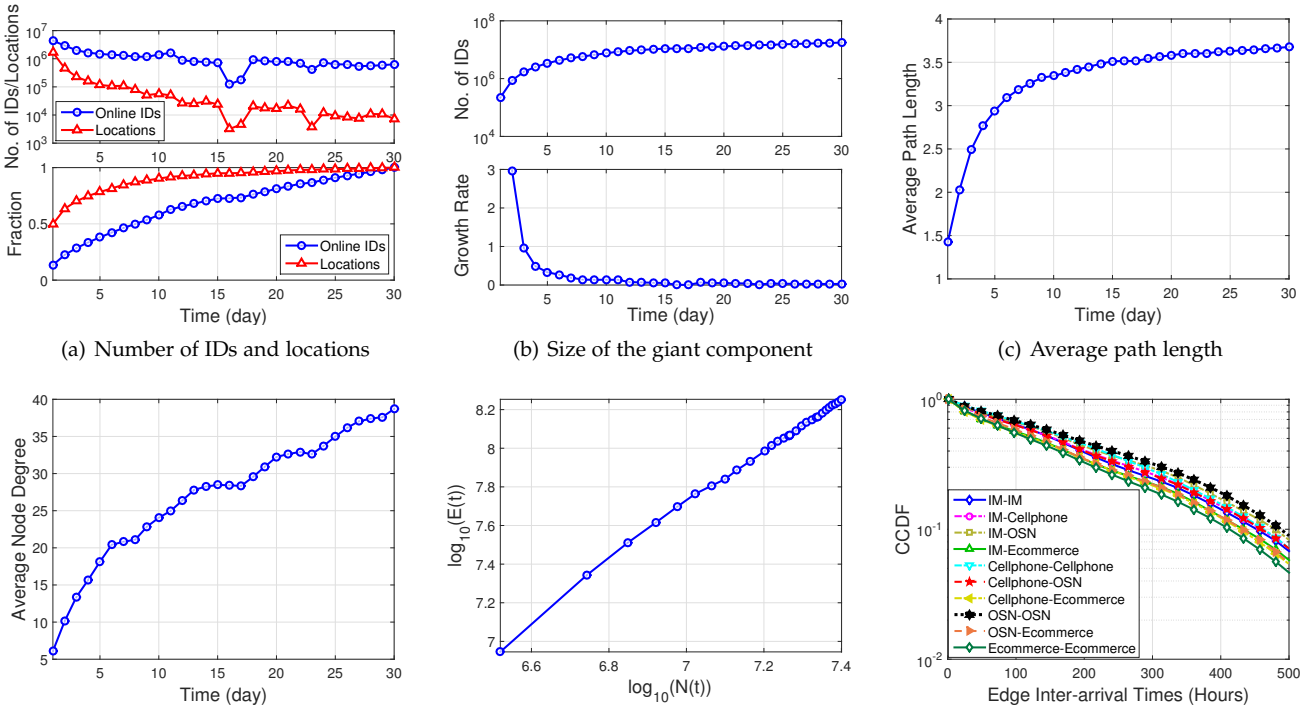
4.2 CLSN Evolution: A Dynamic Perspective

The CLSN analyzed in the last section is constructed by aggregating users' data for one month. To study the growth and evolution of this network, we look at the daily snapshots and evolution of the CLSN, and examine how this network is formed in a gradual way. In our study, we are interested in the following key metrics, i.e., the number of nodes, the size of the largest connected component, average path length, and average node degree. In addition, we also investigate the relationship between the number of nodes and the number of edges from a dynamic perspective, and the distribution of the time gap between two successive co-location events of the same node pair.

Fig. 5(a) shows the fraction of IDs and locations that have appeared as well as the number of newly added IDs and locations each day. We observe that about 58% IDs and 90%

TABLE 3
Statistics of Static Structure of Co-location Social network.

Parameter	IM-only network	Comprehensive network
Average diameter	6.71	4.85
Average node degree	35.19	42.00
Average clustering coefficient	0.4137	0.3824



(a) Number of IDs and locations (b) Size of the giant component (c) Average path length

(d) Average node degree (e) E(t) $vs.$ N(t) (f) Time gaps between online IDs

Fig. 5. Dynamic properties of the constructed co-location social network.

locations have appeared in the first 10 days. Afterwards, the order of the magnitude of the number of newly added IDs is decreased from 10^7 to 10^6 , and the number of the newly added locations also decreases from 10^6 to 10^5 . After 20 days, the number of newly added IDs and locations exhibits a very small change, indicating the network is tending to be stable.

We also analyze the growth patterns of the largest connected component of the network, and plot its size as well as the growth rate as the function of time in Fig. 5(b). We can observe that the size of the largest connected component also increases quickly in the first several days, while it becomes stable about after the 10th day.

As shown in Fig. 5(c), the average path length exhibits a similar trend as the number of IDs, locations and the size of the largest connected component. At the beginning, most of the paths have only two hops, and the average path length keeps increasing during the first 10 days. Afterwards, it changes very little. Differently, according to Fig. 5(d), the average node degree shows a different trend, as it keeps growing steadily during the whole month.

In Fig. 5(e), we plot the logarithm of $E(t)$ as a functions of the logarithm of $N(t)$, where $E(t)$ and $N(t)$ are the number of edges and nodes on the t_{th} day. From the results, we can observe they are linearly dependent, i.e., $E(t) \propto N(t)^{\alpha}$, with the slope α of 1.41. To evaluate how well the data fits this model, we use coefficient of determination, i.e., R^2 value. The obtained value of the R^2 metric is 0.9959, which is relatively high and confirms the accuracy of the model.

We show the distribution of the time gap between two successive co-location events by the same node pair in Fig. 5(f). The time gap between online IDs with different types follows an exponential distribution. If we assume that the login behavior of IDs follows Poisson distribution and is uncorrelated with the place and login behavior of other accounts, the time gap follows an exponential distribution. Meanwhile, the parameters for the exponential distribution for online IDs of different types are different. The parameter for edges between E-commerce accounts is the largest, indicating the highest frequency for them to be co-located, while the parameter for edges between OSN accounts is smallest. However, the parameter for edges between Cell phone accounts is higher than that between OSN accounts, indicating people tend to use cell phone more in public places compared with OSN accounts. Overall, the edge inter-arrival times is quite large, e.g., inter-arrival time of over 50% online ID pairs of all types is larger than 100 hours, indicating temporal granularity of one hour is enough to capture the "co-location" of IDs.

Overall, the CLSN is rapidly evolved, and it is formed almost completely within about 10 days, which guarantees the reliability of study based on the data set of one month. In addition, online IDs with different types exhibit different patterns of activity in terms of time gap of their co-location events.

5 USER IDENTIFICATION

In the last section, we have analyzed the structure and evolution of CLSN, and gained understandings from different aspects. The CLSN is a social network constructed based on virtual online IDs and physical co-location interactions. Thus, the basic and fundamental problem here is how to map the cyberspace IDs to physical world users, i.e., finding all online IDs of one user by using the co-location behaviors of online IDs. Thus, in this section, we first elaborately

define the link weight in CLSN to characterize the co-location behaviors of online IDs comprehensively. Then, we develop a community-based algorithm to detect online IDs belonging to the same user based on their topological structure.

5.1 Measure Link Weight

We first exploit the expression of link weight to characterize the co-location behaviors of online IDs. For each pair of online IDs v_1 and v_2, a direct solution is to use the number of times that they are "co-located" as their link weight as in approaches [32], [33], which are formally expressed as following:

$$w_0(v_1, v_2) = \sum_{t\in T}\sum_{l\in L} X^t_{v_1 l} \cdot X^t_{v_2 l}.$$

However, there are two important problems that have not be addressed in this definition of link weight:

- *Sparsity of co-location*: As we can observe from Fig. 5(f), inter-arrival time of over 50% online ID pairs of all types is larger than 100 hours. Thus, the link weight characterized by the number of sparse co-location, i.e., w_0, is very likely to be biased.
- *Diversity of locations*: Co-locations at different places bring different information to us. For example, co-locations at private places such as home bring more information to us than public place such as Starbucks or KFC. However, different locations are equally dealt with, which is unreasonable.

In order to solve the first problem, we further consider the appearing of nodes in the same location at different time bins by modelling the mobility of users, which will be introduced in detail in Section 5.1.1. Then, in order to solve the second problem, we further weight co-locations at different places by considering the location context in terms of entropy, which will be introduced in detail in Section 5.1.2.

5.1.1 Modelling Human Mobility

The "co-location" of IDs might be sparse. For example, as we can observe from Fig. 5, inter-arrival time of over 50% online ID pairs of all types is larger than 100 hours. Thus, the link weight characterized by the number of sparse co-location is very likely to be biased. To deal with this problem, we consider the appearing of nodes in the same location at different time periods. Different with "co-location", i.e., appearing at the same location in the same time period, we assign a weight to the number of appearing in the same location at different time periods, which is decreasing with the time difference and in proportion to the probability they are co-located in that place. Specifically, there have been a number of works studying the mobility of humans. In [34], there is a theorem describing the probability distribution of the time duration that human stay in one place, which is expressed in detail as follow:

Theorem 1 (Waiting time of humans) [34] Let $\triangle t$ denote waiting time, that is the time a user spent at one location. Then $P(\triangle t)$ follows $P(\triangle t) \sim |\triangle t|^{-1-\tau}$ with $\tau = 0.8 \pm 0.1$ and a cutoff of $\triangle t = 17h$.

Combining this result, we define the weight to the appearing in the same location with time difference $\triangle t$ to be $|\triangle t|^{-1-\tau}$ with a cutoff of $\triangle t = 17h$. Then, the link weight of edges in CLSN is modified as follow:

$$w_T(v_1, v_2) = \sum_{l\in L}\sum_{t_1\in T} X^{t_1}_{v_1 l} \cdot \max_{t_2\in T} X^{t_2}_{v_2 l} \cdot (|t_2 - t_1| + 1)^{-1-\tau}.$$

By this way, the appearing of nodes in the same location at different time periods is involved in the weight, which helps to reduce the impact of sparsity of IDs' co-location.

5.1.2 Considering Location Context

On the other hand, we treat different locations equally in the link weight defined in w_T. However, in fact, co-locations at different places bring different information to us. For example, if the location is some public place with many physical users such as Starbucks or KFC, two online IDs which are co-located at this location have a large probability of belonging to different users. We say these co-location events are less important. However, if the location is a private place with few physical users such as home, the two IDs are very likely to belong to the same user. We conclude that these co-location events are more important. Thus, to improve the performance of the algorithm, we further modify the link weight by taking the difference of locations in consideration. Specifically, we use an important concept in information theory, the entropy, to measure the importance of a co-location at one place. Specifically, $\forall$ location $l \in P$, the entropy of l, $H(l)$ can be calculated as:

$$H(l) = -\sum_{v\in N(l)} P_l(v)\log P_l(v),$$

in which $N(l) = \{v|X^t_{vl} > 0, t \in T\}$ is the set of online IDs appeared at location l. $P_l(v)$ is the probability of any online ID appeared in l is v, which can be calculated as:

$$P_l(v) = \frac{\sum_{t\in T} X^t_{vl}}{\sum_{b\in N(l)}\sum_{t\in T} X^t_{bl}}.$$

Based on the concept of entropy, we further modify link weight as follows:

$$w_{ST}(v_1, v_2) = \sum_{l\in L}\sum_{t_1\in T} \frac{1}{H(l)} L^{t_1}_{v_1 l} \cdot \max_{t_2\in T} L^{t_2}_{v_2 p} \cdot (|t_2 - t_1| + 1)^{\tau}.$$

For locations of public places, the appeared online IDs each day are unstable. Though there are many online IDs appearing each day, online IDs appearing at two different days have little in common between each other, leading to a larger entropy. Thus, multiplied by $1/H(p)$, the co-locations in public places have less contribution to the link weight. While for locations of private places, the number of IDs appearing at each day is small, and the online IDs appearing at different days are almost the same, i.e., they are very stable, leading to a smaller entropy. Thus, the co-locations in private places have larger contribution to the link weight.

5.2 Community-based Algorithm

As discussed in Fig. 2, online IDs belonging to the same user will have a high chance to access the Internet from the same locations. From the CLSN's point of view, these IDs will have more edges and larger weight among each other, which can be further demonstrated by the large clustering coefficient of the CLSN. Therefore, we model the user identification as an existing community detection problem. As defined in [35], a community in a weighted network is defined as a cluster of nodes which have more and stronger connections with nodes in the same cluster, and comparatively fewer and weaker connections with nodes belonging to others. Now, we adopt a community detection algorithm for user identification.

Using community detection algorithms to detect users in CLSNs is not trivial, since most of the existing community detection algorithms are designed for binary networks, which cannot be extended to weighted networks. Moreover, many of the existing algorithms have high computational complexity, and cannot be used in our large-scale network. Considering these two issues, we design the user identification algorithm based on the community detection algorithm proposed by Lu et al. in [36], which is designed for weighted networks with fast computation.

Before describing the algorithm, we first provide the following two definitions as supplement knowledge for the user identification.

Definition 3 (Belonging Degree [36]) In a weighted network $G = (V, E)$, the belonging degree of a node u to a cluster C, denoted as $B(u, C)$, is defined as the ratio of the sum of weights of edges between u and C to the sum of weights of all edges connected to u, which can be formulated as:

$$B(u, C) = \frac{\sum_{v \in C} w(u, v)}{K_u}, \tag{1}$$

in which $K_u = \sum_{v \in G} w(u, v)$. $B(u, C)$ can represent the probability of the node u to be included in cluster C. For a node u, if all its neighbors are included in cluster C, we have $B(u, C) = 1$.

Definition 4 (Conductance [36]) In a weighted network $G = (V, E)$, the conductance of a cluster C, i.e., $\Phi(C)$, is defined as the ratio of the sum of weights of all cut edges of C to the sum weights of all edges connected to nodes in C, which can be represented as:

$$\Phi(C) = \frac{\mathrm{cut}(\mathrm{C}, \mathrm{G} \backslash \mathrm{C})}{w(C)}, \tag{2}$$

where $\mathrm{cut}(C, G \backslash C)$ denotes the weights of the cut edges of C, which can be calculated as $\sum_{v \in G \backslash C} \sum_{u \in C} w(u, v)$. $w(C)$ denotes the sum of weights of all edges in cluster C, including the cut edges, which can be calculated as $\mathrm{cut}(C, G \backslash C) + \sum_{u,v \in C} w(u, v)$. Conductance is a natural and widely-adopted notion of community goodness and is also known as the normalized cut metric [37]. With a lower conductance, there are more and stronger edges between nodes of the community, and thus the community is suitable.

The community-based algorithm is shown in Algorithm 1, in which C is the target cluster, i.e., the set of online IDs of one user. N_C is the set of the neighbors of C, and T_C is the set of types of online IDs involved in C. The algorithm starts from one or two initial nodes, and works in an iterative way to discover prospective nodes belonging to the cluster. In each iteration round, among all nodes adjacent to C, it picks the node with the highest belonging degree. This node will be added to C and a new cluster C' will be formed. After that, it compares $\Phi(C)$ and $\Phi(C')$. If $\Phi(C') < \Phi(C)$, it will continue to select another node. Otherwise, C is designated as a detected cluster. In addition, if there is no neighbor of C or all types of online IDs are all involved in C, it also stops iterating.

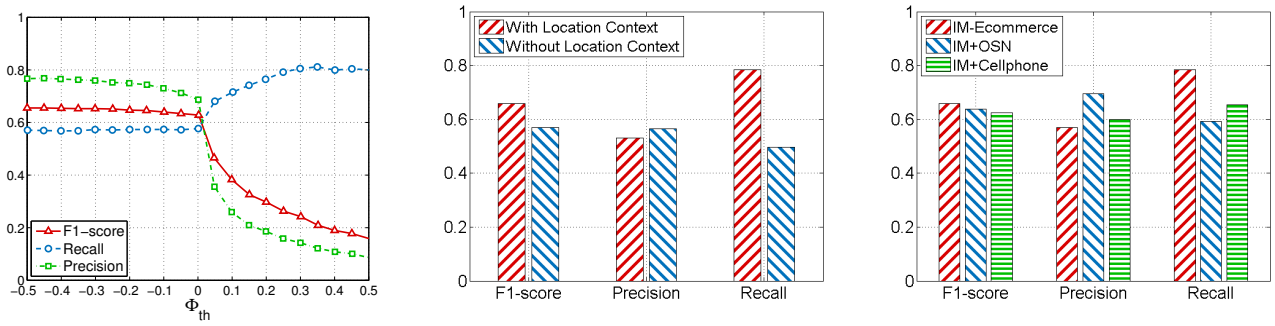
We have made the following three adjustments to tailor this algorithm to fit our problem better. First, we change the method of finding the initial nodes. In [36], the initial nodes are the pair of nodes (u, v) with the biggest edge weight w_{uv}. However, in a CLSN, such two nodes can still belong to two different users, if they have frequent offline interactions. Therefore, in our solution, we just select one initial node with the biggest K_u, as this node must be the most active online ID of some user. Second, we assume that C can only involve one online ID of each type at most. In reality, one physical user may have multiple IDs of the same type. However, they tend to mainly use one ID, which contributes to their behavior most. In addition, considering owning multiple IDs of one type for one user also introduces much complexity and noise to our problem. Thus, in our algorithm, we assume the target cluster can only involve one ID of each type at most. Third, we modify the condition of convergence of $\Phi(C') < \Phi(C)$ to be $\Phi(C') < \Phi(C) + \Phi_{th}$, where Φ_{th} is an adjustable parameter. A larger Φ_{th} can efficiently reduce the false positive results, while a smaller Φ_{th} can reduce the false negative results. Thus, by selecting a suitable value of Φ_{th}, we can trade off between the false positive rate and false negative rate.

5.3 Compared Algorithm

We compare our algorithm with four state-of-the-art algorithms as follows:

Algorithm 1: Community-based Algorithm

Input: Network $G = (V, E)$, the set of types of online IDs S, the type $s(v) \in S$ for all online ID $v \in V$, and an initial online ID u_0.
Output: C, the cluster of online IDs belonging to the owner of u_0.
Initialize:
  $C \leftarrow \{u_0\}$;
  $S_C \leftarrow S$;
  $N_C \leftarrow \{v | (u_0, v) \in E\}$;
while $N_C \neq \emptyset$ *and* $S_C \neq \emptyset$ **do**
  $u_{\max} = \mathrm{argmax}_{u \in N_C} B(u, C)$;
  $C' = C \cup \{u_{\max}\}$;
  if $\Phi(C') < \Phi(C) + \Phi_{th}$ **then**
    $C = C'$;
  else
    break;
  $S_C = S_C \backslash \{s(u_c)\}$;
  $N_C = \{v | (u, v) \in E, u \in C, v \notin C\} \backslash \{v | s(v) \notin S\}$.



(a) Threshold of conductance (b) Location context (c) Types of online ID

Fig. 6. Performance vs. different parameters for the community-based algorithm.

5.3.1 NE Algorithm

Cecaj et al. [32] link IDs belonging to the same user based on the number of "encountering" events of these IDs, which is equal to w_0 introduced in Section 5.1 as follow:

$$w_0(v_1, v_2) = \sum_{t\in T}\sum_{l\in L} X^t_{v_1 l} \cdot X^t_{v_2 l}.$$

Specifically, by assuming a Dirichlet prior, they find that the probability of ID pairs (v_1, v_2) belonging to the same person is in proportion to the number of their "encountering" events, i.e., $w_0(v_1, v_2)$.

5.3.2 POIS Algorithm

Riederer *et al.* [30] also consider using the "encountering" events to match the same users. They assume the number of visits of each user to a location during a time period follows Poisson distribution, and an action (*e.g.* login) on each service occurs independently with Bernoulli distribution. Based on this mobility model, the algorithm computes a weight for every candidate pair of IDs, which can be calculated as follows,

$$w_{\mathrm{POIS}}(v_1, v_2) = \sum_{t\in T}\sum_{l\in L} \phi_{l,t}(X^t_{v_1 l}, X^t_{v_2 l}),$$

where ϕ measures the importance of an "encountering" event in location l at time slot t. By defining $Y(v_1, v_2)$ as the binary variable representing whether online IDs v_1 and v_2 belong to the same physical user, ϕ can be given as follows,

$$\phi_{l,t}(X^t_{v_1 l}, X^t_{v_2 l}) = \frac{P(X^t_{v_1 l}, X^t_{v_2 l} | Y(v_1, v_2) = 1)}{P(X^t_{v_1 l})P(X^t_{v_2 l})}.$$

It can be calculated based on their mobility model with the assumptions of Poisson visits and Bernoulli actions. In addition, this algorithm filters out IDs by the "eccentricity" factor ϵ, which is defined as the threshold for the weight gap between the best and second-best IDs.

5.3.3 CS-based Algorithm

Another solution to link IDs is based on their topological structure in social network. Since online IDs belonging to the same physical user tend to appear in the same places, they are usually co-located with same online IDs and have similar distribution of weight between these IDs. On the other hand, cosine similarity (CS) is a widely adopted metric to measure the similarity of nodes in social network [28], [38]–[40]. It is formally defined as follow:

$$w_{\mathrm{CS}}(v_i, v_j) = \frac{\sum_{v_k\in V} w(v_i, v_k)\cdot w(v_j, v_k)}{\sqrt{\sum_{v_k\in V} w^2(v_i, v_k)}\cdot\sqrt{\sum_{v_k\in V} w^2(v_j, v_k)}}.$$

A larger cosine similarity indicates the two nodes have similar distribution of co-locations between their social neighbors. Thus, they tend to belong to the same physical user.

5.3.4 KLD-based Algorithm

Kullback-Leibler divergence (KLD) is another important metric widely used in ID linking [26], [41], which can be formally defined as follow:

$$w_{\mathrm{KL}}(v_1, v_2) = -D_{\mathrm{KL}}(\boldsymbol{p}_{v_1}|\frac{\boldsymbol{p}_{v_1}+\boldsymbol{p}_{v_2}}{2}) - D_{\mathrm{KL}}(\boldsymbol{p}_{v_2}|\frac{\boldsymbol{p}_{v_1}+\boldsymbol{p}_{v_2}}{2}),$$

where $\boldsymbol{p}_{v_1}$ is a $|V|$-sized vector with each element $p_{v_1}(v_x) = w(v_1, v_x)/\Sigma_{v\in V} w(v_1, v)$. It represents the distribution of co-locations between node u and other nodes in the social network. In addition, D_{KL} is the Kullback-Leibler divergence function [41], which can be calculated as follow:

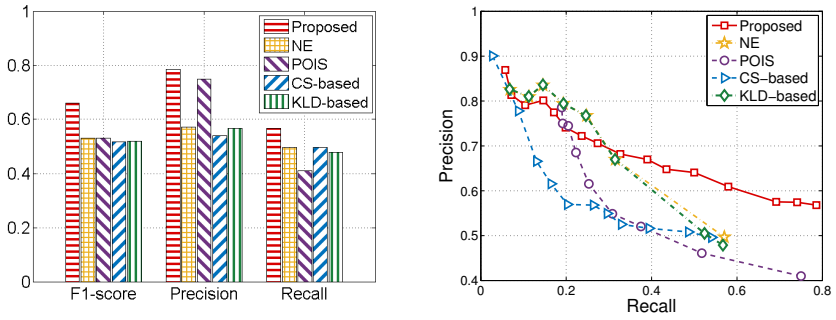
$$D_{\mathrm{KL}}(\boldsymbol{p}|\boldsymbol{q}) = -\sum_{v\in V} p(v)\log\frac{q(v)}{p(v)}.$$

Similarly with cosine similarity, a larger w_{KL} also indicates that the two nodes have larger probability to belong to the same physical user.

All these algorithms are designed for pair-wise ID matching between two services. However, there are IDs of multiple services involved in our problem. Thus, we apply them by adding the matched ID one by one.

5.4 Performance Evaluation

To evaluate the accuracy of our proposed user identification algorithm, we need some ground-truth data for the validation. We use the information of devices to obtain the ground-truth. Specifically, compared with PC, cell phones are more private devices. Therefore, we extract the list of online IDs on each cell phone device. To ensure they are not mixed with the IDs on other devices with the same type of OS in the same subscriber, we further remove devices with more than 1 online IDs of the same type, and use them as the ground truth to evaluate our user identification algorithm. For each



(a) Best performance (b) Precision-recall plot

Fig. 7. Performance of different algorithms.

user, we pick up one online ID for a selected service type as the initial node, and use different algorithms to detect all online IDs belonging to this user. After obtaining a set of online IDs, we compare it with the ground-truth data by looking at another selected type of IDs to obtain the accuracy.

We use three key metrics in binary classification to quantify the detection accuracy, i.e., precision, recall, and F1-score [42]. For each user, the precision is defined as the fraction of online IDs selected by our algorithm that are included in the ground-truth data, and the recall is defined as the fraction of online IDs in the ground-truth data that are successfully retrieved. The F1-score is the harmonic mean of precision and recall, and it can be calculated as $F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$.

We first select the IM accounts and E-commerce accounts as an example to show the performance of our proposed community-based algorithm and the influence of different parameters. Specifically, we tune the parameter Φ_{th} and obtain the value of F1-score, Recall and Precision, which are shown in Fig. 6(a). We can observe that with the increasing of Φ_{th}, precision shows an upward trend while recall shows a downward trend. However, the decreasing of recall is faster than the increasing of precision, leading to the decreasing of F1-score.

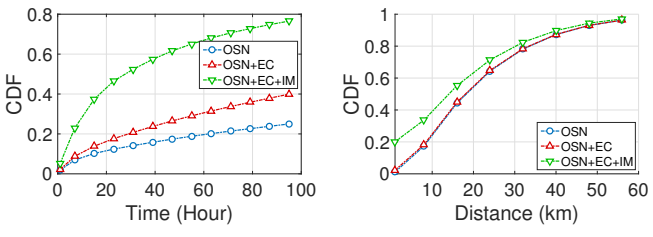
Then, we study the impact of modified link weight. Specifically, we show the best performance of the community-based algorithm with modified link weight and without modified link weight in Fig. 6(b). As we can observe, by using the modified link weight, the F1-score is increased by about 0.13, and the precision is increased by over 0.22, indicating the effectiveness of modified link weight.

In Fig. 6(c), the best performance of different pairs of online IDs is compared. As we can observe, the user identification based on IM accounts to E-commerce accounts shows the best F1-score of 0.66 and the best recall of 0.79 and the worst precision of 0.58. The user identification based on IM accounts shows the best precision of 0.70 and the worst recall of 0.60, while the user identification based on cell phone numbers shows the worst F1-score of 0.57. Overall, the performance of our proposed algorithms of different pairs of online IDs do not show much difference.

Then, we compare the performance of the community-based algorithm against the four state-of-the-art algorithms. Specifically, we show their best performance in Fig. 7(a). As we can observe, our proposed community-based algorithm outperforms other algorithms. The performance gaps between our proposed algorithm and baselines are over 0.12 in terms of F1-score. Then, we further show their precision-recall plot in Fig. 7(b). Similarly, the performance of our proposed community-based algorithm outperforms other algorithms, since it lies in the top right corner, meaning it has larger precision when recall is equal and larger recall when precision is equal. Take the CS-based algorithm as an example. Their performance gap is small when recall or precision is close to 1, while it is large when recall and precision are balanced. For example, the gap of precision between two algorithms is only about 0.04 when recall is 0.1, while it is 0.15 when recall is 0.4, indicating the effectiveness of our proposed algorithm.

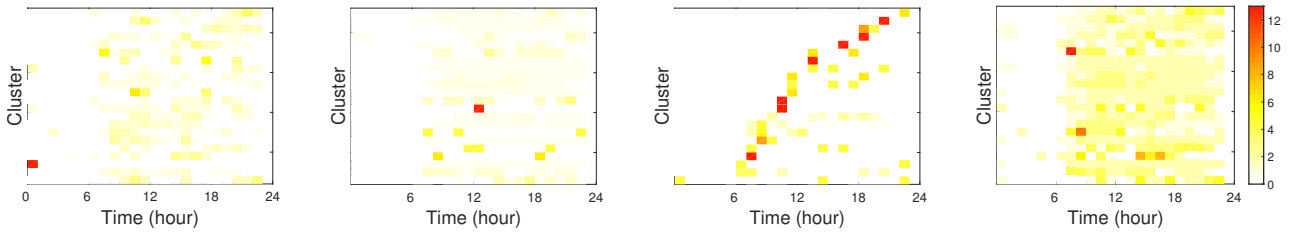
6 User Perspective Analysis

By detecting the online IDs of one user, we link the cyberspace and physical world, which enables characterizing the user's behavior from these two aspects. Specifically, as a single user might have multiple online IDs, by clustering them together, we can study their behaviors in terms of different usage of online IDs. For example, e-mail accounts are mainly used for business and professional scenarios, and they are more active in weekdays and in office buildings. Also, online shopping accounts are more active during after-work hours and in apartments. Thus, by detecting all online IDs of a user and combining them together, a user's cyberspace activities can be characterized in a comprehensive way. In this section, we focus on analyzing the user's properties based on the results of user identification by investigating the relationship of the behaviors between



(a) (b)

Fig. 8. Statistics of traces in terms of spatial and temporal dimension.



(a) IM accounts (b) OSN accounts (c) EC accounts (d) Overall

Fig. 9. Daily behavior of users for using different kinds of online IDs, where the same row represents behavior of same cluster of users and deep color indicates a higher distribution density.

these two kinds of IDs of the cyberspace and physical world. Specifically, we focus on a subset of our dataset, which consists of 2,425 users correctly identified by the proposed user identification algorithm. We first present basic statistics of users in Section 6.1. Then, we analyze the behavior of users in terms of their daily patterns and weekly patterns in Section 6.2 and Section 6.3, respectively.

6.1 Basic Statistics Analysis

In order to measure the benefit obtained from merging different online IDs, we compare the merged traces and unmerged traces in terms of spatial and temporal dimension. Specifically, in the temporal dimension, we investigate the distribution of time gap between each time bin and its most recent record. The results are shown in Fig. 8(a). As we can observe, by merging login records of multiple online IDs, the time gap is obviously reduced. In terms of the spatial dimension, we show the distribution of spatial distance between adjacent records in Fig. 8(b). We can observe that the spatial distance is also reduced. For example, the average spatial distance is reduced by about 19.8%. These results demonstrate that by combining different types of online IDs together, the quality of the obtained traces is significantly improved, which benefits the analysis of user behavior.

6.2 Daily Behavior Analysis of Users

We first analyze the daily behavior of online IDs by clustering. Using access frequency across one day of different kinds of online IDs as the feature and applying hierarchical clustering algorithm, we successfully divide users into clusters with the same preference of online service usage. That is, users who prefer to use the same kind of online IDs in the same time period are divided into one cluster. The results are shown in Fig. 9. Specifically, Fig. 9(a) shows the usage of IM accounts, Fig. 9(b) shows the usage of EC accounts, Fig. 9(c) shows the usage of OSN accounts, and Fig. 9(d) shows the overall usage of different services. The x-axis shows the time in hours and y-axis shows the obtained clusters of online IDs, i.e., each row in the figure represents the temporal distribution of one cluster. In addition, rows of identical height in different figures correspond to the same cluster of users. The color presents the distribution density of login records, where deep color indicates a higher distribution density and light color stands for a lower distribution density.

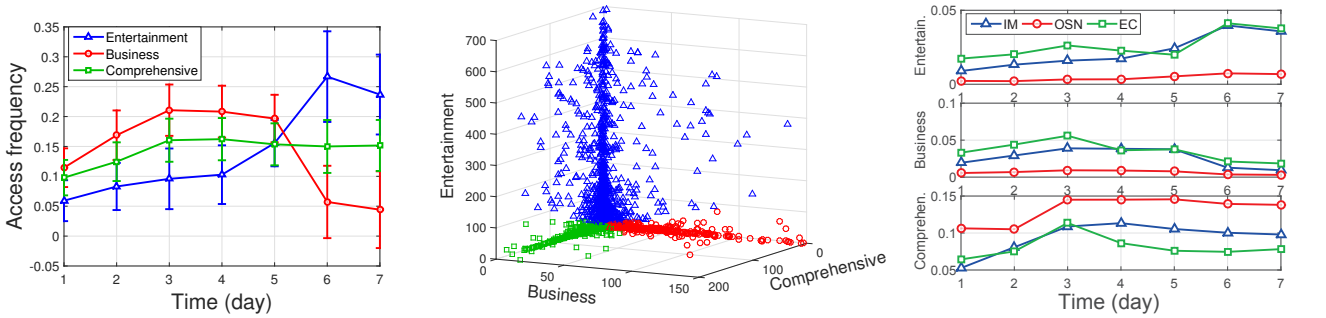
From the result, we can observe that most people tend to use their EC accounts in only one hour throughout the day, while they tend to use their IM accounts in all hours over the day, reflecting the immediacy of the IM service. In addition, the temporal distribution of login records of both IM and EC accounts has an interrupt between 12PM and 6AM, reflecting the daily sleep schedule of humans. As for the OSN accounts, the temporal distribution of their login records mainly concentrates on daily rest time, such as 7AM (wake-up time), 1PM (lunch break) and 5PM (quitting time). In addition, from Fig. 9(d), we can observe the overall usage of users is most similar with the usage of the IM service. Although login records of different kinds of online IDs show different temporal patterns, the peak of their usage frequency is very close for the same clusters of users, reflecting the consistency of the users they belong to.

6.3 Weekly Behavior Analysis of Users

We next analyze the weekly behavior of online IDs by clustering. Using the number of login records per day across one week as the feature and applying K-means clustering algorithm, we find the online IDs can be divided to three types, which we refer as business, entertainment and comprehensive accounts, respectively. Their frequency as a function of time along a week is shown in Fig. 10(a). The business accounts tend to be more active on weekdays, while the entertainment accounts tend to be more active in weekends. In addition, the peak of access frequency of entertainment accounts is usually in Saturday, and there has been an obvious increasement since Friday. On the other side, the access frequency of the comprehensive accounts are almost keeping constant throughout the week.

Then, we use the utilization frequency of online IDs of entertainment, business and comprehensive accounts to analyze user behavior. By clustering, we also divide users to three types, of which the result is shown in Fig. 10(b). They use the online IDs of entertainment, business and comprehensive most frequently, respectively. We refer these users as entertainment-dominated, business-dominated and comprehensive users.

These two dimensions of dividing users and online IDs are closely relevant. In Fig. 10(c), we plot the patterns of access frequency of different types of IDs for different kinds of users. From the results, we can observe the entertainment-dominated users tend to use e-commerce services most frequently in weekends. The access frequency of IM services in weekends is also increased compared with that on weekdays, which indicates that they are using the e-commerce and IM services in their entertainment. As for



(a) Three clusters of online IDs in terms of time patterns for their access frequency

(b) Distribution for the users in terms of aggregate access frequency of IDs with different time patterns

(c) Normalized access frequency of IDs with different types for different kinds of users

Fig. 10. Weekly behavior of users for using different kinds of online IDs.

the business-dominated users, access frequency of IM and e-commerce services is obviously reduced in weekends compared with that on weekdays, while that of their OSN services is almost not changed, indicating that they are using the IM and e-commerce services in their business. As for the comprehensive users, it is not surprising that access frequency of all their online IDs is little changed throughout the week. However, different with business-dominated and entertainment-dominated users, their access frequency of OSN accounts is higher than IM and EC accounts, These results further confirm the close relevance of the two dimensions of dividing users and online IDs.

7 Implication and Application

In this work, we have explored a new type of social network, i.e., co-location social network (CLSN), which shows a significantly different perspective of users' interaction than LBSN [5], [17], [18] and EBSN [19]. For example, Cho et al. [5] mainly focused on the relationship between users' online friendship network and offline mobility, while offline interaction between users was ignored. Liu et al. [19] focused on the difference of users' online and offline interactions restricted to a single site, while users' cross-site interactions was ignored. Different with them, CLSNs capture users' offline interaction cross multiple services. Specifically, through static and dynamic perspective analysis, we find that CLSNs exhibit stronger locality and faster evolution than existing social networks [1], [2], [31]. Further by implementing a community detection technique, we extract the community structure formed by cross-site online IDs belonging to the same user, and we then analyze the users' behavior in terms of different services in detail. Our work mainly reveals the relationship of users' online behavior cross multiple sites and offline interactions.

On the other hand, we can further adopt the CLSNs to address a very broad of important problems. We list five possible applications as follows.

Analyzing online-offline social network interactions: Previously, most of the existing literatures treat the online social networks and offline social networks differently, and existing interaction models make distinctions between these two types of networks. However, the cyberspace and the physical world are linked, and they are influenced by each other. We can dig deeper to further analyze the relationship between online activities and offline interactions.

Dynamic population sensing: Knowing where people are and how they are distributed is important for a number of critical social problems, such as public health, natural disasters, and traffic management. Our work also introduces a possibility of obtaining the up-to-date population distribution in an accurate way. Instead of following traditional practices such as population census, our sensing can be done in real-time. We are able to estimate temporal variations of the population density in emergency and data-scarce situations.

Tracking of human mobility and social structure: By using our CLSNs, we are able to perform a large-scale human mobility tracking. We can observe physical interactions among real users from a dynamic perspective. More importantly, such tracking can be done by utilizing existing network infrastructure, instead of deploying a large number of tracking devices, e.g., cameras.

Link prediction: Link prediction is another important problem in social network, which has been widely utilized in many application areas, such as recommendation [43], security [44], etc. Specifically in CLSNs, link prediction is to predict whether and how many times two users encounter each other in a certain time period, which further helps us reveal the hidden relationship between users.

Community structure analysis: Through community detection techniques, we have found the cluster of online IDs belonging to the same user. Further community analysis can dig deeper to reveal the relationship of user, e.g., a family or a company. Compared with existing social networks, more attribute information of communities can be obtained in CLSNs, based on the link weight distribution of users' IDs of different services.

In summary, not only various new applications (e.g. population sensing) can be achieved through CLSNs, but many existing applications (e.g. link prediction, community analysis) are strengthened in CLSNs as well, indicating the usefulness and effectiveness of the CLSN.

8 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. Our study covers both the static network structure and dynamic evolution properties of the constructed network. Our analysis shows that the CLSN has small diameter, large average node degree and clustering coefficient, indicating the existence of significant small world phenomenon and tight connection. In addition, it is rapidly evolved and has different patterns of activity in terms of different services. Further, we demonstrate that using the constructed network, we are able to judge which online IDs belong to the same person, i.e., mapping the cyberspace IDs back to the physical world users with high accuracy. Last but not least, we perform a user-centric analysis, where a significant different access time and service behaviors are revealed among different types of users.

ACKNOWLEDGMENT

This work is supported by National Basic Research Program of China (973 Program) (No. 2013CB329105), National Natural Science Foundation of China (No. 61301080, No. 61171065, No. 61273214, No. 61602122, and No. 71731004), 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 *Proceedings of the 4th ACM European conference on Computer systems (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 *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement (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 *Proceedings of the 17th international conference on World Wide Web*, 2008, pp. 915–924.

[5] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: user movement in location-based social networks," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, 2011, pp. 1082–1090.

[6] C. M. Cheung, P.-Y. Chiu, and M. K. Lee, "Online social networks: why do students use facebook?" *Computers in Human Behavior*, vol. 27, no. 4, pp. 1337–1343, 2011.

[7] N. B. Ellison, C. Steinfield, and C. Lampe, "The benefits of facebook friends: social capital and college students use of online social network sites," *Journal of Computer-Mediated Communication*, vol. 12, no. 4, pp. 1143–1168, 2007.

[8] R. Xiang, J. Neville, and M. Rogati, "Modeling relationship strength in online social networks," in *Proceedings of the 19th international conference on World Wide Web (WWW)*, 2010, pp. 981–990.

[9] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida, "Characterizing user behavior in online social networks," in *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference (IMC)*, 2009, pp. 49–62.

[10] N. B. Ellison *et al.*, "Social network sites: Definition, history, and scholarship," *Journal of Computer-Mediated Communication*, vol. 13, no. 1, pp. 210–230, 2007.

[11] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins, "Microscopic evolution of social networks," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, 2008, pp. 462–470.

[12] X. Zhao, A. Sala, C. Wilson, X. Wang, S. Gaito, H. Zheng, and B. Y. Zhao, "Multi-scale dynamics in a massive online social network," in *Proceedings of the 2012 ACM conference on Internet measurement conference (IMC)*, 2012, pp. 171–184.

[13] H. Kashima and N. Abe, "A parameterized probabilistic model of network evolution for supervised link prediction," in *Proceedings of the Sixth IEEE International Conference on Data Mining (ICDM)*, 2006, pp. 340–349.

[14] I. de Sola Pool and M. Kochen, "Contacts and influence," *Social networks*, vol. 1, no. 1, pp. 5–51, 1979.

[15] M. E. Newman, D. J. Watts, and S. H. Strogatz, "Random graph models of social networks," *Proceedings of the National Academy of Sciences (PNAS)*, vol. 99, no. suppl 1, pp. 2566–2572, 2002.

[16] K. Faust, "Centrality in affiliation networks," *Social networks*, vol. 19, no. 2, pp. 157–191, 1997.

[17] D. J. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg, "Inferring social ties from geographic coincidences," *Proceedings of the National Academy of Sciences (PNAS)*, vol. 107, no. 52, pp. 22 436–22 441, 2010.

[18] S. Scellato, A. Noulas, and C. Mascolo, "Exploiting place features in link prediction on location-based social networks," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, 2011, pp. 1046–1054.

[19] X. Liu, Q. He, Y. Tian, W.-C. Lee, J. McPherson, and J. Han, "Event-based social networks: linking the online and offline social worlds," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, 2012, pp. 1032–1040.

[20] J. Cranshaw, E. Toch, J. Hong, A. Kittur, and N. Sadeh, "Bridging the gap between physical location and online social networks," in *Proceedings of the 12th ACM international conference on Ubiquitous computing (UbiComp)*, 2010, pp. 119–128.

[21] N. Korula and S. Lattanzi, "An efficient reconciliation algorithm for social networks," *Proceedings of the VLDB Endowment (PVLDB)*, vol. 7, no. 5, pp. 377–388, 2014.

[22] E. Kazemi, S. H. Hassani, and M. Grossglauser, "Growing a graph matching from a handful of seeds," *Proceedings of the VLDB Endowment (PVLDB)*, vol. 8, no. 10, pp. 1010–1021, 2015.

[23] O. Goga, P. Loiseau, R. Sommer, R. Teixeira, and K. P. Gummadi, "On the reliability of profile matching across large online social networks," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2015, pp. 1799–1808.

[24] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Proceedings of the IEEE Symposium on Security and Privacy (SP)*, 2008, pp. 111–125.

[25] J. Zhang and S. Y. Philip, "Multiple anonymized social networks alignment," in *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, 2015.

[26] F. M. Naini, J. Unnikrishnan, P. Thiran, and M. Vetterli, "Where you are is who you are: User identification by matching statistics," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 2, pp. 358–372, 2016.

[27] X. Mu, F. Zhu, E. P. Lim, J. Xiao, J. Wang, and Z. H. Zhou, "User identity linkage by latent user space modelling," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016.

[28] O. Goga, H. Lei, S. H. K. Parthasarathi, G. Friedland, R. Sommer, and R. Teixeira, "Exploiting innocuous activity for correlating users across sites," in *Proceedings of the 22nd international conference on World Wide Web (WWW)*, 2013, pp. 447–458.

[29] M. Gao, E.-P. Lim, D. Lo, F. Zhu, P. K. Prasetyo, and A. Zhou, "Cnl: collective network linkage across heterogeneous social platforms," in *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, 2015.

[30] C. Riederer, Y. Kim, A. Chaintreau, N. Korula, and S. Lattanzi, "Linking users across domains with location data: Theory and validation," in *Proceedings of the 25th International Conference on World Wide Web (WWW)*, 2016, pp. 707–719.

[31] A. Java, X. Song, T. Finin, and B. Tseng, "Why we twitter: understanding microblogging usage and communities," in *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis (WebKDD/SNAKDD)*, 2007, pp. 56–65.

[32] A. Cecaj, M. Mamei, and F. Zambonelli, "Re-identification and information fusion between anonymized cdr and social network data," *Journal of Ambient Intelligence and Humanized Computing*, vol. 7, no. 1, pp. 83–96, 2016.

[33] M. Srivatsa and M. Hicks, "Deanonymizing mobility traces: Using social network as a side-channel," in *Proceedings of the ACM conference on Computer and communications security (CCS)*, 2012, pp. 628–637.
[34] C. Song, T. Koren, P. Wang, and A.-L. Barabási, "Modelling the scaling properties of human mobility," *Nature Physics*, vol. 6, no. 10, pp. 818–823, 2010.
[35] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3, pp. 75–174, 2010.
[36] Z. Lu, Y. Wen, and G. Cao, "Community detection in weighted networks: Algorithms and applications," in *Proceedings of the IEEE International Conference on Pervasive Computing and Communications (PerCom)*, 2013, pp. 179–184.
[37] J. Leskovec, K. J. Lang, and M. Mahoney, "Empirical comparison of algorithms for network community detection," in *Proceedings of the 19th international conference on World wide web (WWW)*, 2010, pp. 631–640.
[38] M. Newman, *Networks: an introduction*. Oxford university press, 2010.
[39] L. Rossi and M. Musolesi, "It's the way you check-in: identifying users in location-based social networks," in *Proceedings of the second ACM conference on Online social networks (COSN)*, 2014, pp. 215–226.
[40] A. Narayanan and V. Shmatikov, "De-anonymizing social networks," in *Proceedings of the IEEE Symposium on Security and Privacy (SP)*, 2009, pp. 173–187.
[41] J. A. Thomas and T. M. Cover, *Elements of information theory*. John Wiley & Sons, 2006.
[42] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," 2011.
[43] H. Chen, X. Li, and Z. Huang, "Link prediction approach to collaborative filtering," in *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries (JCDL)*.
[44] M. Al Hasan, V. Chaoji, S. Salem, and M. Zaki, "Link prediction using supervised learning," in *SDM06: workshop on link analysis, counter-terrorism and security*, 2006.