Journal of Universal Computer Science, vol. 24, no. 3 (2018), 277-301
submitted: 3/10/17, accepted: 5/3/18, appeared: 28/3/18 © J.UCS

Linking User Online Behavior across Domains with Internet Traffic

Yuanyuan Qiao[1]
(Center for Data Science, Beijing University of Posts and Telecommunications
Beijing, China
yyqiao@bupt.edu.cn)

Yan Wu
(Center for Data Science, Beijing University of Posts and Telecommunications
Beijing, China
yanwu@bupt.edu.cn)

Yaobin He
(Smart City Research Institute of China Electronics Technology Group Corp.
Shenzhen, China
heyaobin@cetccity.com)

Libo Hao
(Center for Data Science, Beijing University of Posts and Telecommunications
Beijing, China
hlb@bupt.edu.cn)

Wenhui Lin
(Technology Research Institute, Aisino Corporation
Beijing, China
linwenhui@aisino.com)

Jie Yang
(Center for Data Science, Beijing University of Posts and Telecommunications
Beijing, China
janeyang@bupt.edu.cn)

Abstract: We are facing an era of Online With Offline (OWO) in the smart city - almost everyone is using various online services to connect friends, watch videos, listen to the music, download resources, and so on. Our online behaviors are separated by different domains, which may cause serious problem in the area of cross-domain recommendation, advertising, and criminal tracking in online and offline world, since it is a very challenging task to link user online behaviors belonging to the same natural person. Existing methods usually tackle user online behavior linkage problem by estimating the profile content similarity between two different online services. However, the profile contents in heterogeneous online services are unreliable or misaligned, and

[1] Corresponding author.

the proposed methods are always limited to several services in a specific domain. In order to link individual's online behavior across domains, in this paper, we propose user Online Behavior Linkage across Domains (OBLD), a novel hybrid model, to link user online behavior across domains with Internet traffic. It derives several significant attributes from users' online behaviors, such as user digital identity, various fingerprints of terminals and browsers, spatio-temporal behavior of users, and leverages a supervised classification method to discover the relationship between users' online behaviors. Also, the proposed model has unsupervised setting for dataset with non or few label data if a certain percentage of user digital identities can be extracted from original dataset. By using real-world network traffic collected from two large provinces in China, we evaluate the OBLD model and the linkage precision achieves 89% and 97.9% for two datasets respectively. Especially, the inputs of OBLD, i.e., network traffic flows, cover all online behavior of users who connect with Internet through monitored networks, which makes it possible to link online behaviors of users in whole online world.

Key Words: Online Behavior Linkage, User Digital Identity, User Identity Linkage, Internet Traffic, Across Domains

Category: L.7.0

1 Introduction

Nowadays, Internet connects human, machines, and objects together. We live in online world (virtual world in Internet) and offline world (real world with physical entities) at the same time, which records our "Digital Footprints" contain online and offline behavior, and current time and location [Irani et al., 2009, Zhang et al., 2015]. Our online digital footprints capture our online behavior and whatever we do on the web becomes part of our Digital IDentity (DID) forever. Existing researches demonstrate that online and offline behaviors of human have strong influence on each other [Hong and Jung, 2018, Dunbar et al., 2015, Hristova et al., 2016, Qiao et al., 2016]. For example, friends in real world will follow each other on online social networks and share opinions, thought, and photos at anytime; those who have similar interests online may happy to arrange a meeting with each other in a coffee shop. It is well known that human mobility behavior is far from random, they tend to follow daily pattern in a limited space due to long duration of travel [Jung, 2017]. However, online behavior is not limited by space nor by distance, it has accelerated the spread of information in an unprecedented manner, and expanded people's activities, which not only facilitate the daily life of the people, but also trigger some adverse issues in the field of public safety, and online recommendation system.

In the field of public safety, according to the 2016 Internet Crime Report by Federal Bureau of Investigation Internet Crime Complaint Center (FBI IC3), IC3 received 298,728 complaints with a monetary loss of 1.33 billion in 2016, including some common cybercrimes such as, Malware/Scarewarem, Virus, and Phishing/Vishing/Smishing/Pharming, and some cyber related crimes as Crimes Against Children, Identity Theft, Harassment/Threats of Violence, Terrorism, and so on. How to discover potential criminal behavior from massive "digital footprints", and then track individual behind these online behaviors, have become an urgent task. In addition, in the field of online recommendation system, for a online service, the online behavior of a specific user can be connected together by account, and then user's online interests and attributes are available by analyzing his/her personal data. However, since different services don't share data, and domains with various services are heterogeneous, in order to fully understand users' interests and provide better recommendations or services, connecting online behavior belonging to unique users across different websites has also become a crucial barrier to accurately estimating behaviors and statistics at the user level. Furthermore, aggregating online behaviors from different domains reveals more information about users and is beneficial for personalization and cross-domain recommendations - particularly for solving cold-start problems where systems suffer from sparse user profiles [Abel et al., 2010]. Also, we can leverage the integrated results to help analyze the patterns of user migration between multiple online services [Shu et al., 2017]. Thus, a fundamental question arises: can we link all online behavior of the same user across domains to get a more comprehensive view about the user?

There are more than 4.6 billion known website pages in Internet currently [Hong and Jung, 2016], it is very challenging (even impossible) to link users' behavior online, since users can access to different domains anonymously, with multiple accounts, or as a visitor, even using several terminals. Although several researches have been carried out in recent 2 years focusing on user identity linkage across online social networks [Shu et al., 2017, Nie et al., 2016, Lee and Jung, 2017, Zhang and Yu, 2016], across domains with location data [Riederer et al., 2016], and visitor stitching on cross-device web logs [Kim et al., 2017], linking user online behavior across domains still remains an open problem, primarily for two reasons: first, online behavior linkage across domains is harder than both classifying [Calabrese et al., 2011], and distinguishing [Onnela et al., 2007] users, and it may have been considered impractical at scale. Second, many existing methods are designed for a specific domain [Shu et al., 2017], or at best domains that are semantically similar [Riederer et al., 2016]. In order to build a more convenient, efficient, safe, and reliable network environment, and obtain an integrated profile for each individual, it is necessary to link his/her online behavior across multiple domains together [Nguyen and Jung, 2017, Ma et al., 2014]. In contrast, our goal is to address the most general case in which data across domains is separately generated and has obvious differences in characteristics.

To address this problem, this paper proposes OBLD, a novel hybrid model with unique digital identity based and probability based correlation methods, to link online behavior of users across various domains. We extend user identity linkage problem to user online behavior linkage problem, which aims to correlate online behaviors of user that belong to the same real person with Internet traffic. First, we extract features, such as user digital identity, online fingerprint, and spatio-temporal behavior of users, to evaluate the similarity and dissimilarity between online behaviors. Then, by using the decision tree, we define the problem of user linkage of different online behavior across domains as a binary classification problem. Finally, we apply the Top1 selection method to optimize the results of decision tree. Based on the massive network data traffic collected by the biggest Internet Service Provider (ISP) that covering northern and southern provinces of China, we validate the effectiveness of our proposed model. In summary, the contributions of our paper are as follows:

(1) We propose OBLD, a novel hybrid model with digital identity based and probability based correlation methods, to formalize our problem as a unified framework. The unique digital identity based correlation method can achieve 100% accuracy in a small scale. The probability based correlation method converts the user online behavior linkage issue to a bipartite problem for all datasets, which is improved by an user online behavior feature based unsupervised function to achieve higher accuracy and coverage [Lee and Jung, 2017].

(2) This model has fully modeled the similarity and dissimilarity between users' online behaviors from several aspects from Internet traffic point of view. Online fingerprints and offline spatio-temporal patterns are extracted as features to train the model [Tan et al., 2014]. The attributes of user's online behavior used here are not dependent on a specific domain, which can be applied to data sets from different domains and different network environments. What's more, the proposed method can be easily scaled to adapt massive data traffic in networks [Bui and Jung, 2018].

(3) We validate the effectiveness of our proposed model with two real big data sets, which are extracted from real network data traffic of Internet collected from typical provinces in China covering millions of people over a month. The real network traffic data is generated by factual users while connecting with Internet, and various users' online behaviors of services across domains can be found and extracted from the data to evaluate the mode.

The rest of the paper is organized as follows. In Section 2, related works in the field of user online behavior linkage are introduced. Section 3 provides the problem definition for user online behavior linkage. In Section 4, we introduce the data set used in our experiment. Feature extraction from network data traffic is illustrated in Section 5. Section 6 introduces the proposed hybrid model to link user online behavior across domains. Experimental results and analyses based on real data traffic are given in Section 7. Conclusions are drawn in Section 8.

2 Related Work

The increasing popularity of users accessing to multiple domains with many accounts, or anonymous makes user online behavior linkage problem of critical importance to business intelligence by gaining from user's online behavior a deeper understanding and more accurate profiling of users. A general framework for user online behavior linkage is usually composed of two major phases: (1) Feature extraction and (2) Model construction [Shu et al., 2017]. In the feature extraction phase, features that could distinguish users' online behaviors are extracted from content [Zafarani and Liu, 2013, Kong et al., 2013],users' profile [Lee and Jung, 2017, Zhang et al., 2015, Perito et al., 2011], network structures [Korula and Lattanzi, 2014, Zafarani et al., 2015, Bartunov et al., 2012, Man et al., 2016], trajectories [Riederer et al., 2016, Cho et al., 2011], web logs [Kim et al., 2017], and online interest [Liu et al., 2014, Nie et al., 2016]. Then, extracted features [Man et al., 2016, Zafarani and Liu, 2013] are then used as inputs for training supervised [Nie et al., 2016, Perito et al., 2011], semi-supervised [Zhang et al., 2015] or unsupervised [Riederer et al., 2016] model. As last, the proposed model [Liu et al., 2014] estimates the correlation of users' online behaviors to solve the user online behavior linkage problem.

In 2016, a review surveys advancements in user identity linkage across online social networks, and introduces a unified framework for the user identity linkage problem [Shu et al., 2017]. It suggests that, in the future, methods should be adjusted and applied in cross network scenarios [Liu et al., 2016], and more practical problem settings can be further explored. In the latest papers, some researchers tend to link user digital identity by mapping heterogeneous networks to a homogeneous space, they apply network embedding techniques [Man et al., 2016, Liu et al., 2016], propose latent user space model [Lee and Jung, 2017], or solve network alignment problem [Zhang and Yu, 2016], to learn the follower-ship / followee-ship of each user [Liu et al., 2016], to obtain social network structures with low dimension space [Man et al., 2016], to learn a projection function [Lee and Jung, 2017], to explore multiple user and location anchor link prediction [Zhang and Yu, 2016]. Some researchers try to understand the content of users' behavior in different domains and link users' behavior by comparing the mined contents [Nie et al., 2016, Riederer et al., 2016]. In paper [Nie et al., 2016], core interests of users are extracted by topic modeling, to connect user in different social networks. Aligning [Riederer et al., 2016] and spectral co-clustering [Han et al., 2017] algorithms are applied to users' trajectories, and then finds the most likely matching user behaviors by utilize the maximum weighted matching scheme. In *WWW* 2017, Sungchul Kim and et. al proposed probabilistic soft logic learning based framework to solve visitor stitching problem on noisy and incomplete cross-device web logs [Kim et al., 2017]. Internet Protocol (IP) addresses, geo-graphic coordinates, and user-agent information are considered as features to train the model [Tan et al., 2014]. The studies also pointed out that, in the future, proposed methods should adapt to exponentially large number of data source [Lee and Jung, 2017, Kim et al., 2017], unsupervised learning framework [Lee and Jung, 2017, Shu et al., 2017], and fully mine contents in the dataset [Shu et al., 2017, Riederer et al., 2016, Han et al., 2017].

In this paper, in order to extract the features with high distinguishability from Internet traffic, we select features with universal existence, uniqueness, and reliability property, to reduce the data skews and noise caused by diversity and heterogeneity of users' online behavior in network environment [Ma et al., 2018]. We also try to understand the contents by considering online fingerprint and offline spatio-temporal patterns as features. The proposed hybrid model, OBLD, links users' online behaviors with unique digital identity and probability based correlations, which has a unsupervised setting and can be easily scaled to adopt real network environment with massive imbalanced data traffic. Particularly, the input of our model is traffic flows that aggregated by data packets collected from core networks of ISP, which may cover millions of domains and population [Liu et al., 2013].

Nowadays, smart devices, carried by users as sensors wherever they go, bring us ubiquitous mobile access to the Internet. In this digital age, our paper trails and digital trails coexist, an important component of our footprints are our online digital footprints [Irani et al., 2009]. Every time we connect with Internet, data is "delivered" in packets, i.e., our device and the web servers we access exchange tens of thousands of data packets through Internet, which contain online digital footprints with rich information, such as current time, current location, visited web services, Uniform Resource Identifier (URI), device type, data traffic size, and so on. With the explosion in data traffic amount [Dunbar et al., 2015], increasing number of studies emerge in the area of data traffic analysis [Naboulsi et al., 2015, Blondel et al., 2015, Calabrese et al., 2015], including social, mobility, and network analysis. In recent years, ISP has opened more and more Internet traffic to public, to encourage the researchers discover the potential of data traffic on solving city traffic problem, reducing energy consumption, improving human health in the city, and understanding poverty [Nguyen and Jung, 2018, Bello-Orgaz et al., 2016, Hong et al., 2017]. As a result, data traffic of Internet has become a unique data source to study population, society, and development on a scale not yet seen. It presents both opportunities and challenges for linking online behavior of users. On the one hand, focusing on Internet traffic makes it possible to connect all online behavior of users together in all domains, and provides a comprehensive view of users' online interests. On the other hand, how to link user's behavior based on PB or EB level big data traffic with four Vs characteristics [Chen et al., 2014], is an un-avoidable problem, which can be handled by us at technical level [Qiao et al., 2018] and method level proposed in this paper.

While obtaining a full picture of users' online behavior has many applications in industry and research, it also raises legitimate and serious concerns about the privacy of users online. Along with the encryption technology development, we may not able to collect users' behavior online, however, spatio-temporal trajectories [Han et al., 2017] and fingerprints extracted from data traffic [Rahmati et al., 2011] are enough to track and identify users. Even though, many researchers believe that formulating users' digital footprints and linking multiple digital identities can help to find out security and privacy breaches with dire consequences [Gundecha et al., 2014], detect and protect users from various privacy and security threats arising due to vast amount of publicly available user information [Riederer et al., 2016], keep the user informed about such threats and suggest her preventive measures [Malhotra et al., 2012], and verify ages online to protect children [Zafarani and Liu, 2013]. In general, leverage new methods of user behavior linkage for a better tomorrow is important and necessary.

3 Problem Definition

Let P denotes the set of all online services in real life. For an online service S, C_s denotes the set of all digital identities in online service S, and each one belongs to a distinct user. ϕ_s is a method mapping each digital identity in service S to a natural person.

Our user **O**nline **B**ehavior **L**inkage (OBL) problem is defined as follows:

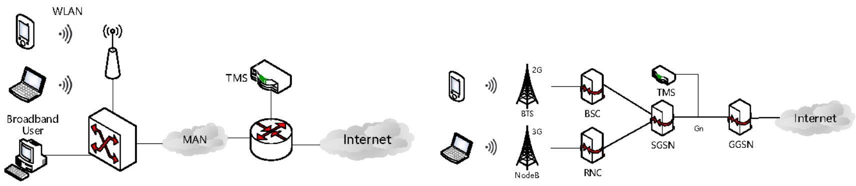
Given two online service platforms S and $S^{'}$, the OBL is to design a linkage method f to decide whether the two digital identities in S and $S^{'}$ respectively correspond to the same natural person, i.e., $f\colon C_s \times C_{s'} \mapsto \{0,1\}$ such that for any pair of digital identities $(u_i, u_{i'}) \in C_s \times C_{s'}$, we have

$$f(\langle u_i, S\rangle, \langle u_{i'}, S'\rangle) = \begin{cases} 1 & \textit{if } (\phi_S(u_i) = \phi_{S'}(u_{i'})), \\ 0 & \textit{otherwise} . \end{cases}$$

For simplicity, this work assumes that one user has at most one digital identity in an online service.

4 Data Description

In this paper, we use two data sets which are collected from different network environments: fixed network and mobile (wireless) network. The two data sets are collected by the high-speed Traffic Monitoring System (TMS) [Qiao et al., 2018] developed by our group.



(a) *Fixed network* (b) *Mobile network*

Figure 1: The network architectures and deployment of TMS.

For the fixed network, as shown in [Fig.1(a)], User Equipment (UE) connect with Internet through Local Area Networks (LAN) or Wireless Local Area Networks (WLAN). TMS collects the HyperText Transfer Protocol (HTTP) packets between Metropolitan Area Network (MAN) and Internet, and then aggregate them into HTTP records. The collected dataset comprises a sequence of time-stamped records, which contain IP address, Uniform Resource Identifier (URI), cookie, user-agent, etc.

For the mobile network, as shown in [Fig.1(b)], by deploying TMS at the core network edge connecting to the 2G/3G network interfaces, traffic data generated by UE is collected. A UE communicates with a base transceiver station (BTS) or Node B, which transmits its network traffic to a base station controller (BSC) or radio network controller (RNC). The controllers (BSC/RNC) then deliver the network traffic to a serving GPRS support node (SGSN) that establishes a tunnel on Gn interface (interface between the GGSN and the SGSN) with a gateway GPRS support node (GGSN) through which the data enters the Internet (GPRS represents "General Packet Radio Service"). We collect mobile Internet traffic from Gn interface and store the data traffic as flow records, which contain base station ID, user's anonymized ID, IP address, URI, user-agent, etc.

The online services used by users are distinguished by keywords in URI, such as "*y.qq.com*", "*jd.com*", and "*taobao.com*", etc. The digital identities of users can be found in URI or Cookie. Cookie is a small piece of data sent from a website and stored on the user's computer by the user's web browser while the user is browsing. It usually contains user's digital identity. User-agent contains user's online fingerprints, e.g., mobile phone operation system (mobile-os) and mobile phone brand for mobile phone users, as well as Personal Computer operation system (PC-os) and browser version (BV) for PC users.

5 Feature Extraction

Every time users connect with Internet, much information, such as current time, IP, URI, User-agent, and so on, is transmitted between user's device and servers with data traffic. In order to identify who generated which data traffic, and then trace all the online behavior of users, features that widely available in Internet traffic and highly discriminative between different users should be extracted.

We can extract user's online behavior from each flow record and form a vector $\overline{UBlog}$:
$\overline{UBlog} = (UTC, IP, service, \overline{DID}, os, BV, brand)$.

If several digital identities have been extracted from one record, it will generate a digital identity relationship vector as follows:
$\overline{DID} = (DIDtype_1, DID_1, DIDtype_2, DID_2, ...)$.

The elements in the vector $\overline{UBlog}$ respectively stands for the UTC time when the record is generated, the IP used by the user when the record is generated, the online service's name, vector for digital identity, terminal's operation system as well as its version, the type and version of browser and mobile phone brands. If the online fingerprint $finger \in \{os, brand, browser\}$ can not be extracted, the value is assigned to *Null*. In $\overline{DID}$, *DIDtype* refers to the type of *DID* (like cell phone number, email account, user name, and etc). Usually, if *DIDs* in $\overline{DID}$ of two $\overline{UBlog}$ belong to the same actual person, the possibility that these two $\overline{UBlog}$ belong to the same users is very high.

In this part, we illustrate three kinds of features used to train our model, i.e., user digital identity, online fingerprint, and spatio-temporal behavior of user.

5.1 User Digital Identity

Digital IDentity (DID) is the representation of a human identity that is adopted or claimed in cyberspace to interact with machines or people [Sorrentino, 2009]. Users may also project more than one digital identity through multiple communities. In Internet traffic, many digital identities can be extracted, for example, IP address, identifier in Cookie, and online service account of user. However, above digital identity is not universal unique, maybe unavailable (except for IP), and usually changes with time or domain. Therefore, in order to link user's online behavior across domain, specific rules for user digital identity based feature extraction are required.

5.1.1 IP based features

When users get access to online services, each network and terminal will be assigned with an IP address. Firstly, we give some definitions:

Definition 1: LAN users. It represents the users in a local area network.

Definition 2: IP sharing. It represents the situation that multiple terminals share the same IP over a LAN.

Definition 3: public IP. Due to the setting of LAN, the same IP used by multiple users is called *public IP*.

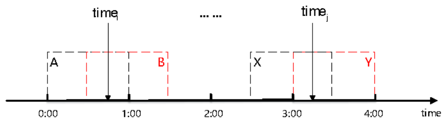


Figure 2: Time discretization.

IP relevance can be very effective in characterizing the relationship between different online behaviors. In order to quantify the relevance of IP usages between online behaviors, we discretize the start time UTC of the flow record. By setting the time window with the *width* and the moving *step*, the start time UTC can be discretized and mapped into the discrete time *window*. As shown in [Fig.2], the *width* selected in the experiment is 1 hour, and the *step* between two time windows is 0.5 hour. Each $\overline{UBlog}$ records the start time and IP of a digital identity when user browsing a specific service. The IP usage of a digital identity can be represented as a 2-tuple $(ip, window)$. During a long time, the replacement path of a user's IP can be traced by the HTTP records and then be abstracted as a sequence *IPSeq* as: $IPSeq = \langle IP_1, IP_2, ..., IP_n \rangle$, where $IP_k = (ip_k, window_k)\,, k \in (1, n)$.

Heuristic knowledge 1: If there are two digital identities for $service_1$ and $service_2$ respectively, i.e., $DID_1 = (idtype_1, id_1)$, $DID_2 = (idtype_2, id_2)$, and the intersection of their *IPSeq* is: $IPintersec = IPSeq_1 \bigcap IPSeq_2$.

If DID_1 and DID_2 belong to a same actual user, the number of elements in set *IPintersec* is at least 1. In view of this, we customize some indicators from multiple dimensions: *IPSeq similarity*, *Uniqueness of shared IP*, *Correlation of IPSeq combines with the number of LAN users*, and *IP dissimilarity*.

1) *IPSeq Similarity*

The indicator is designed to assess the IP usage similarity between two digital identities in different services in any time window. We define the *IPSeq* similarity ρ as follows:

$$\begin{aligned} \rho &= Jaccard(IPSeq_1, IPSeq_2) \\ &= \frac{|IPSeq_1 \bigcap IPSeq_2|}{|IPSeq_1 \bigcup IPSeq_2|}. \end{aligned} \tag{1}$$

The large value of ρ indicates that the two digital identities almost have the same IP in any time window, so they highly likely belong to a same user. If $\rho = 0$, the two digital identities are considered to belong to different users. However, the indicator ignores the case of *IP sharing*, which brings error to the results. In this case, two digital identities refer to different users may also have a high value of ρ.

2) *Uniqueness of shared IP*

Considering the case of *IP sharing*, we define *uniq* to measure the uniqueness of *IP sharing*:

$$\begin{aligned} uniq &= uniq\,(IPSeq_1 \cap IPSeq_2) \\ &= \frac{\sum_{i=1}^{M=|IPintersec|} \frac{1}{n_{IPintersec(i)}}}{|IPintersec|}. \end{aligned} \tag{2}$$

where $IPintersec\,(i)$ represents the i^{th} element in *IPintersec*, and $n_{IPintersec(i)}$ represents the amount of users who use ip_i in the time window $window_i$. The large value of *uniq* indicates that the *public IP* used by the two digital identities is shared by very few digital identities at the same time. In other words, these two digital identities appear within a small LAN.

3) *Correlation of IPSeq combines with the number of LAN users*

Taking the number of digital identities within the LAN into consideration, we define the indicator *corr* to describe the IP similarity between two digital identities in different services in any time window:

$$\begin{aligned} corr &= corr\left(n_{IPSeq_1}, n_{IPSeq_2}\right) \\ &= \frac{n_{IPSeq_1} \cdot n_{IPSeq_2}}{\left|n_{IPSeq_1}\right| \cdot \left|n_{IPSeq_2}\right|}. \end{aligned} \tag{3}$$

where n_{IPSeq_1} represents a vector formed by the reciprocal of the amount of users using ip_i in the time window $window_i$. The large value of *corr* indicates that the digital identity pair always has the same IP in any time window, and the degree of *IP sharing* is low. So this pair is more likely to refer to a same user.

4) *IP dissimilarity*

In order to fully discover the relationship between digital identities, this part defines a feature from a new perspective of dissimilarity.

Heuristic knowledge 2: There are two digital identities DID_1 and DID_2 which refer to a same user in the real world, and they are from online $service_1$ and $service_2$ respectively. The two digital identities may not appear with different IP at the same time.

In the view of this, we define *dissim* to measure the dissimilarity of IP usage between digital identities in different online services:

$$\begin{aligned} dissim &= dissim\,(IPSeq_1, IPSeq_2) \\ &= \frac{|T_{diff}\,(IPSeq_1, IPSeq_2)|}{|T_{all-windows}\,(IPSeq_1, IPSeq_2)|}. \end{aligned} \tag{4}$$

where $T_{all-windows}$ represents a sequence of time windows when both DID_1 and DID_2 appear in Internet traffic. T_{diff} satisfies $T_{diff} \subseteq T_{all-windows}$ and represents the set of time windows in which DID_1 and DID_2 use different IP. The value of *dissim* ranges from 0 to 1. The large value of *dissim* indicates a weak correlation between identities, and they are less likely to belong to the same user. For instance, the value larger than 0.5 indicates the two digital identities appear with different IP for more than half of the time windows.

5.1.2 Unique digital identity based features

When a user registers for an account on a service, digital identity is created. Usually, user may input a username that is universal unique in current service as an account. Also, phone number (PN) or email (*Email*) account is bound with the account, which can be seen as a unique identifier across all domains. Here, we define *UniDID*=$\{$*PN, Email*$\}$, which refers to digital identity that can be used to identify the identity of people uniquely. In our experiments, phone number and email account have been map to a hash number in order to protect users' privacy information. As a result, *UniDID* are unique digital identity based features in our experiments, which only cover a part proportion of traffic data since some users may not visit service with account, or account information of some services is encrypted.

5.2 Online Fingerprint

Online fingerprints are important clues to link different online behaviors by distinguishing digital identity that belongs to different users. This part introduces several features extracted from online fingerprints.

Heuristic knowledge 3: There are two digital identities DID_1 and DID_2 which refer to a same user in the real world, and they are in online $service_1$ and $service_2$ respectively. When DID_1 and DID_2 access services through terminals, the online fingerprints extracted from their flow records are always identical or related. If the online fingerprints left by the two digital identities are completely different, they could not belong to a same user.

Each $\overline{UBlog}$ records the online fingerprint while using a specific service. Within the observation time, when a *DID* appears, the online fingerprint can be represented by a sequence as:

$Finger = \{finger_1, finger_2, ..., finger_n\}$.

We define an indicator σ to measure the similarity of online fingerprints between *DID*s for different online services, which is applied as below.

1) Firstly, we measure the importance of each online fingerprint $finger_i$ in sequence *Finger*. Since some types of fingerprints appear more frequently in general, we first calculate the Term Frequency-Inverse Document Frequency (TF-IDF) of its online fingerprint information for each *DID*.

We define $f = (v_1, v_2, ..., v_n)$, where v_i represents the value of TF-IDF of the i^{th} online fingerprint, n indicates the distinct number of online fingerprints. The definition of v_i is as follows:

$$v_i = \frac{n_i}{N_i} \times log \frac{count}{\|\{Finger|finger_i \in Finger\}\|}, \tag{5}$$

where n_i indicates the number of occurrences time of $finger_i$ in *Finger*, N_i indicates the number of fingerprints in *Finger*, *count* indicates the distinct number of users, $\|\{Finger|finger_i \in Finger\}\|$ indicates the number of users who have the online fingerprint of $finger_i$.

2) Based on vector *f*, we calculate the online fingerprint correlation between different digital identities. The indicator σ is defined as follows:

$$\sigma = \sigma(f, f') = \frac{f \cdot f'}{|f| \cdot |f'|}. \tag{6}$$

If $\sigma = 0$, the two digital identities are considered to belong to different users. Otherwise, we consider they are likely to belong to the same user.

5.3 Spatio-temporal Behavior of User

Users participate in different activities at different times of the day, and tend to follow similar patterns in different days. In the evening, users are generally active at "home", where they generate amounts of Internet traffic. In contrast, in the day of weekday, users usually stay in the "workplace". In order to distinguish different times of a day, we define the time period of "work hour" and "home hour" as follows:

1) "work hour": 8:00 to 19:00 on weekday.

2) "home hour": 20:00 to 7:00 of the next day on weekday and the whole day of weekend.

Based on "work hour" and "home hour", we divide the set of user online behavior vectors into two complementary sets. If the set of vectors $\overline{UBlog}$ in which UTC is in the time period of "work hour" or "home hour", then we define it as $\overline{UBlog}_{work}$ or $\overline{UBlog}_{home}$ respectively. The whole set of $\overline{UBlog}$ is represented as $\overline{UBlog}_{all}$. These three logs record user's behaviors in different time periods.

6 OBLD: The Proposed Hybrid Model

Our goal of linking online behavior across domains is to establish a model which can determine whether a pair of digital identities refers to the same user [Ma and Leijon, 2011]. This section gives a hybrid model to address this problem. The structure of our model is illustrated in [Fig.3], and the two main parts of the model are unique digital identity based correlation and probability based correlation. For the former, we define a precision linkage rule to precisely link

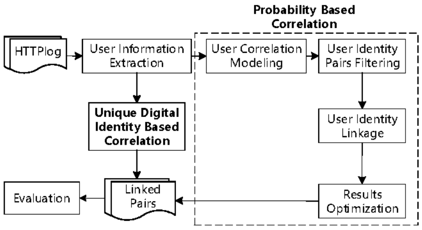


Figure 3: The architecture of OBLD model.

some unique digital identities. However, this method can only apply to a small number of digital identities, so the probability based correlation, a more general method, is proposed. We design several conditions to reduce the amount of input data by filtering some digital identity pairs first. Then, combining the supervised classification with Top1 selection, the correct-linked pairs are identified.

6.1 Unique Digital Identity Based Correlation

We have defined unique digital identity based feature *UniDID*={*PN, Email*}, i.e., phone number and email account, as unique identifier of users across all domains. In our proposed hybrid model, firstly, we use *UniDID* to link user's online behavior with high precision that covers very limited proportion of data traffic. Furthermore, this step labels some source data as "linked pairs", which is the input of probability based correlation module to train and test the model. In this way, the proposed hybrid model is trained in an unsupervised manner. We link user's online behavior with digital identity in the following way.

1) For user online behavior vector $\overline{UBlog_i}$ extracted from data traffic flow record i, get digital identity vector, i.e.,

$$\overline{DID_i} = (DIDtype_1, DID_{1i}, DIDtype_2, DID_{2i}, ...).$$

Here, $DIDtype_1 \in (DID - UniDID)$ and $DIDtype_2 \in UniDID$.

2) For two data traffic flow records i and i', we have

$$\overline{DIDR_i} = (DIDtype_1, DID_{1i}, DIDtype_2, DID_{2i}, ...),$$
$$\overline{DIDR_{i'}} = (DIDtype_1, DID_{1i'}, DIDtype_2, DID_{2i'}, ...).$$

From Section 3 we know that
$\phi_{service}(DIDtype_1, DID_{1i}) = \phi_{service}(DIDtype_2, DID_{2i})$,
$\phi_{service'}(DIDtype_{1'}, DID_{1i'}) = \phi_{service'}(DIDtype_{2'}, DID_{2i'})$.

3) If $(DIDtype_2 == DIDtype_{2'})$ && $(DID_{2i} == DID_{2i'})$, we have $\phi_{service}(DIDtype_1, DID_{1i}) = \phi_{service'}(DIDtype_{1'}, DID_{1i'})$. In other words, *UniDID* is a bridge to link different *DIDs* of the same actual person. If same

UniDID is found in different flow records, different *DIDs* of the same person can be identified, then user's online behavior carried out by these *DIDs* are linked together.

However, in our experiments, only less than 2.8% flow records have *UniDID* that can be extracted. Therefore, other linkage method that can be applied to all flow records is required.

6.2 Probability Based Correlation

In each flow record, IP based features, online fingerprint, and spatio-temporal behavior of user can be extracted. These universal available features give us the opportunity to link all user online behavior. However, in real network environment, users usually connect with Internet by dynamic or public IP, using equipment with same brand, which makes it difficult to distinguish online behaviors between different people. In order to handle huge amount of data, and provide as high as possible linkage precision to link user online behavior, in our model, we apply filtering and correlation methods for probability based correlation.

6.2.1 User Linkage Pairs Filtering

If we extract all features from each flow record for all input data, then examine the correlation between them, for two service domains S and S' with N_1 and N_2 digital identities respectively, we will face a dataset with T times of matching,

$$\overline{T} = \sum_{n=1}^{min(N_1,N_2)} \frac{N_1!N_2!}{n!(N_1-n)!n!(N_2-n)!}$$

which grows exponentially as the size of input dataset grows. What's more, the vast majority of wrong-linked pairs in the results of the IP and fingerprints correlation may lead to a serious data skew. Hence, we define some rules to filter some pairs out:

1) Remove the $\overline{UBlog}$ contains wrong-formatted *DID*.
2) Remove the $\overline{UBlog}$ contains inactive *DID*. If a *DID* generated Internet traffic in less than 12 time windows during a month, we call it an *inactive user*.
3) Ignore the $IP_k = (ip_k, window_k)$ whose ip_i is used by more than 1000 *LAN users* in the $window_i$.
4) Remove the pairs whose ρ=0 in the $\overline{UBlog}_{work}$ or $\overline{UBlog}_{home}$.
5) Remove the pairs whose *dissim*>0.5 in the $\overline{UBlog}_{work}$, $\overline{UBlog}_{home}$, or $\overline{UBlog}_{all}$.
6) Remove the pairs whose σ=0 in the $\overline{UBlog}_{all}$.

Features	Definition
ρ	*IPSeq* Similarity, distinguished by $\overline{UBlog}_{work}$, $\overline{UBlog}_{home}$, and $\overline{UBlog}_{all}$.
uniq	uniqueness of shared IP address, distinguished by $\overline{UBlog}_{work}$, $\overline{UBlog}_{home}$, and $\overline{UBlog}_{all}$.
corr	the correlation measure of *IPSeq* combines with the number of LAN users, distinguished by $\overline{UBlog}_{work}$, $\overline{UBlog}_{home}$, and $\overline{UBlog}_{all}$.
dissim	IP dissimilarity, distinguished by $\overline{UBlog}_{work}$, $\overline{UBlog}_{home}$, and $\overline{UBlog}_{all}$.
σ_{os}	The similarity of terminal's operation systems in $\overline{UBlog}_{all}$.
$\sigma_{browser}$	The similarity of browsers in $\overline{UBlog}_{all}$.
σ_{brand}	The similarity of brands in $\overline{UBlog}_{all}$.

Table 1: Features calculation

6.2.2 User Linkage Pairs Correlation

In order to distinguish user's different behaviors at different times of the day, based on the three logs defined before, i.e., $\overline{UBlog}_{work}$, $\overline{UBlog}_{home}$, and $\overline{UBlog}_{all}$, we calculate some features, which represent the user's online behaviors in different times. All features calculated are summarized in [Tab.1].

To achieve a better linkage performance, we first apply Chi-Squared Statistic to select some significant features. Then, we use the decision tree classification to obtain pairs that are likely to be a same user. Finally, a self-defined Top1 selection method is used to optimize the results of decision tree. Details are listed as below.

Scoring the selected features. We apply Chi-Squared Statistic evaluation to give score value to each feature. In our case, features with high score value are selected to quantify the relationship between two different *DID*s.

Decision tree classification. Selected features are evaluated by adopting the decision tree to decide whether the *DID* pair is linked correctly.

Top1 selection. In the result set of decision tree classification, a *DID* of the target online services may be linked to more than one identities of another online service, which means there are some wrong-linked pairs. To address the problem, we define an indicator *score* to find pairs that are most likely to be correct-linked based on IP usage, which is the most easily accessible information in Internet traffic.

$$score = 0.5 \times (\rho + \sigma)\,. \tag{7}$$

Pair with highest *score* value is considered as a correct linkage.

7 Experimental Results

In this section, the effectiveness of the OBLD model is evaluated based on the ground truth.

7.1 Data Collections

In order to validate the universality of the proposed model, we evaluate it on data sets of fixed network and mobile network respectively. DataSet_1 is collected from fixed network in a large province of southern China, while DataSet_2 is collected from mobile network in a large province of northern China. The duration of both datasets last for one month, and DataSet_2 contains each user's anonymized ID. We select four popular online services based on three conditions:

1) These services are in different domains,

2) Users need to register an account before using services,

3) Services have a high user coverage.

In view of these matters, the four online services we selected are as follows:

1) QQ: the most commonly used online social service platform in China, which provides instant messaging service for users.

2) SinaWeibo: a popular social service platform in China, like Twitter.

3) JD: a famous online e-commerce service platform, like amazon.

4) Taobao: the biggest e-commerce service platform in China. Similar to JD, users can purchase multiple and diverse goods on the platform.

7.2 The Analysis of User Behavior Linkage Characteristics

Before we testing the model with real massive data traffic, in this section, we examine the characteristic of our experimental datasets to answer the following questions: How many digital identities can be extract from different service domains? What percentage of online behaviors are linked by applying unique digital identity based correlation? The answer of above questions may help us predict the amount of data to be processed and how the hybrid model work in each step.

service	*DIDtype*	*DID* count	user count	proportion
Taobao	PN	339	3381978	0.01%
JD	email	4007	417433	0.96%
JD	PN	11145	417433	2.67%
Weibo	email	125307	737543	16.99%

Table 2: The proportion of UniDID *for selected services in* $DataSet_1$

Firstly, we examine the distribution of *DID* in two dataset. In our hybrid model, the unique digital identity based correlation method links user's online behavior with unique digital identity, i.e., phone number and email account, refers to *UniDID*=$\{$*PN, Email*$\}$. [Tab.2] presents how many unique digital identities can be extracted from selected services. Only 2.8% digital identities are unique digital identities.

Above unique identifiers can help us link some user behaviors with nearly 100% accuracy. However, in order to link user's online behavior as many as possible, based on all the extracted digital identities, we apply the probability based correlation method in the next step. In [Fig.4,] we examine how many digital identities a user has in $\mathrm{DataSet}_2$, which has labeled unique identifier for each user. Most of users own less than 10 digital identities. At the same time, 0.01% of users have more than 100 digital identities, which implies that abnormal users and data exception exist in our dataset.

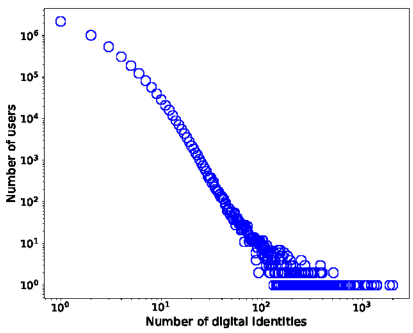


Figure 4: The distribution of the number of user's digital identities.

Note that, in our model, we assume that data noise is very common in data traffic of Internet. As a result, we don't deal with data exception, data noise, or data skew, instead, we try to find "linked pairs" with highest probability based on digital identity, online fingerprint, and spatio-temporal behavior of user.

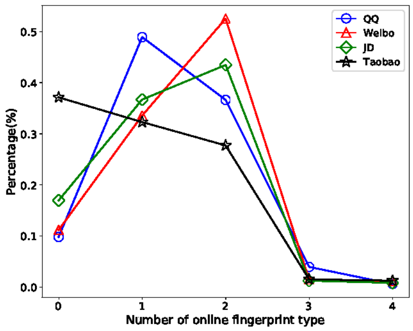


Figure 5: The distribution of online fingerprint.

What's more, we extract the online fingerprints of users from four services (QQ, Weibo, Taobao and JD) and compare the coverage of fingerprints extracted from different services to see whether data misalignment exists. We define the coverage of fingerprints as the number of identities with online fingerprints divided by the total number of *DID*s extracted in a certain service. The coverage of online fingerprints in different online services is shown in [Fig.5]. For the four services in different services, almost 70% of *DID*s have at least one online fingerprint. The coverage of online fingerprints in different services is similar. This fact indicates that the online fingerprints do not depend on a particular service. Data misalignment caused by the differences between services doesn't exist in our dataset.

7.3 Evaluation Metrics

We consider the problem of online behavior linkage as a binary classification task that, given two *DID*s, estimate whether they are matching or not:
1) True Positive (TP): These predicted matched *DID*s (f=1) actually belong to same natural person;
2) True Negative (TN): These predicted unmatched *DID*s (f=0) actually belong to different natural persons;
3) False Negative (FN): These predicted unmatched *DID*s (f=0) actually belong to same natural person;
4) False Positive (FP): These predicted matched *DID*s (f=1) actually belong to different natural persons.

Based on aforementioned possible classification results, we can define following metrics, $Precision = \frac{|TP|}{|TP|+|FP|}$, $Recall = \frac{|TP|}{|TP|+|FN|}$, $F1\ Score = 2 \cdot \frac{Precision \cdot Recall}{Precision+Recall}$, and $Accuracy = \frac{|TP|+|TN|}{|TP|+|TN|+|FP|+|FN|}$.

UniDID	Services	Linked pairs	Precision
Email	JD, Weibo, QQ	29417	100%
Phone number	JD, Taobao	118	100%

Table 3: Results of unique digital identity based correlation

7.4 Performance Analysis

We apply Chi-Squared Statistic evaluation to give score value to each feature. As is shown in [Tab.1], there are 12 features of IP usage and 4 features of fingerprints. [Tab.4] shows us the Chi-Squared values of features of two data sets. The larger Chi-Squared value indicates the more relevant the feature is to the data. So we select the top 9 features for DataSet_1 and the top 4 features for DataSet_2 as the most significant features.

Feature	Score (Dataset_1)	Score (Dataset_2)
$\rho\,(work)$	**4338.71**	3.96
$\rho\,(home)$	**3640.88**	**41.11**
$\rho\,(all)$	**3197.07**	7.82
$uniq\,(work)$	**11177.7**	**181.27**
$uniq\,(home)$	**13202.51**	**106.41**
$uniq\,(all)$	**10493.27**	**58.28**
$corr\,(work)$	**6775.46**	20.37
$corr\,(home)$	**7292.65**	30.17
$corr\,(all)$	**4317.35**	24.69
$dissim\,(work)$	797.77	9.02
$dissim\,(home)$	738.9	0.82
$dissim\,(all)$	156.26	3.64
$\sigma\,(os)$	326.78	16.45
$\sigma\,(browser)$	562.61	
$\sigma\,(brand)$	9	0

Table 4: Chi-Squared value of the selected features

DID pairs in each test set are divided into ten folds, we randomly pick seven folds as a training set and the left three folds as a testing set. We leverage decision tree to identify the relationships between the two *DID*s. The results are listed in [Tab.5]. If only part of the significant features are used to identify the correct-linked pairs, results are not good enough. The combination of all selected features achieves 92% accuracy, 82% precision, 77% recall and 79% F1 score for DataSet_1.

Feature	DataSet1				DataSet2			
	Precision	Recall	Accuracy	F1 Score	Precision	Recall	Accuracy	F1 Score
$\rho\,(work)$	62%	22%	84%	32%				
$\rho\,(home)$	54%	20%	83%	29%	**97.9%**	**100%**	**97.9%**	**98.9%**
$\rho\,(all)$	58%	30%	84%	39%				
$uniq\,(work)$	78%	65%	92%	71%	97.4%	100%	97.4%	98.7%
$uniq\,(home)$	76%	65%	90%	70%	97.5%	100%	97.5%	98.8%
$uniq\,(all)$	78%	67%	91%	73%	97.4%	100%	97.4%	98.7%
$corr\,(work)$	62%	52%	86%	57%				
$corr\,(home)$	53%	37%	83%	44%				
$corr\,(all)$	68.3%	47%	87%	56%				
all features	82%	**77%**	92%	79%	97.7%	99.9%	97.5%	98.74%
all features + Top1 selection	**89%**	73%	**94%**	**80%**	97.7%	99.5%	97.1%	98.55%

Table 5: User online behavior linkage results

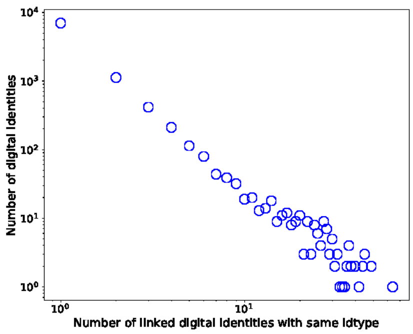


Figure 6: The distribution of linked DID*s with same* DID *type.*

As for DataSet$_2$, feature $\rho_{ipSeq}\,(home)$ alone achieves the highest performance, i.e., 97.9% accuracy, 100% precision, 97.9% recall, and 98.9% F1 score.

Take DataSet$_1$ as an example, in the result set of decision tree, as shown in [Fig.6], more than 1000 *DIDs* are linked to at least two identities of the same service. The result shows that there are still an assignable number of wrong-linked pairs yet. Furthermore, if we combine the decision tree and Top1 selection together, the most likely linked pairs can be selected out. As a result, the precision can achieve 84% at last.

Since all the existing methods focus on user profile or social relationship, and the datasets in different paper is different, we can only compare the precision presented in each paper, as shown in [Tab.6]. For POIS, the performances based on different data sets have obvious differences, so we list all the results in the Table. It is very clear that the proposed hybrid model OBLD outperforms others.

Method	Precision	Recall
CONSET [Zhang et al., 2015]	82.05%	71.10%
MAH [Tan et al., 2014]	85.30%	-
Ulink-CPP [Lee and Jung, 2017]	62%	-
Ulink-APG [Lee and Jung, 2017]	60%	-
POIS [Riederer et al., 2016]	75%/95%/30%	38%/77%/18%
OBLD	89%/97.9%	73%/100%

Table 6: Performance of different methods

8 Conclusions

In this paper, we conducted a systematic and detailed investigation on the problem of linking online behavior of different online services across domains. We precisely defined the problem and proposed a hybrid model OBLD to address it. First, we clarified the practical significance of proposing a model for digital identity linkage across service domains. Second, we described the characteristics of the data set used in the experiment. Next, we introduced the three kinds of features used in our model, i.e., user digital identity, online fingerprint, and spatio-temporal behavior of users, to evaluate the similarity and dissimilarity between online behaviors. Then, a set of candidate digital identity pairs were obtained through the investigation of IP correlation and online fingerprint correlation between pairs. Finally, based on the unique digital identity based correlation and probability based correlation, we combined the decision tree with Top1 selection to identify the correct-linked pairs. Based on factual data, the performance of the proposed model was evaluated and the linkage precision achieved as high as 89%, and 97.9% for two datasets respectively. The result validated the effectiveness of the proposed model. Moreover, the model can be applied to data sets from different domains, different network environments, and different countries. What's more, the algorithm used in our model is simple, and will have a good guidance on the following scientific researches and engineering realization. Considering that user's online behaviors will keep changing or being accumulated as time goes by, in the future a more practical linkage method is expected to be proposed to extract attributes and link online behavior[Irani et al., 2009] dynamically.

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China (61671078, 61701031), Director Funds of Beijing Key Laboratory of Network System Architecture and Convergence (2017BKL-NSAC-ZJ-06), and 111 Project of China (B08004, B17007). This work is conducted on the platform of Center for Data Science of Beijing University of Posts and Telecommunications.

References

[Abel et al., 2010] Abel, F., Henze, N., Herder, E., and Krause, D. "Interweaving public user profiles on the web", In Proceedings of the 18th International Conference User Modeling, Adaptation, and Personalization, pages:16–27.

[Bartunov et al., 2012] Bartunov, S., Korshunov, A., Park, S.-T., Ryu, W., and Lee, H. "Joint link-attribute user identity resolution in online social networks". In Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining, Workshop on Social Network Mining and Analysis. ACM.

[Bello-Orgaz et al., 2016] Bello-Orgaz, G., Jung, J.J., and Camacho, D. "Social big data: Recent achievements and new challenges", Information Fusion, 28:45–59.

[Blondel et al., 2015] Blondel, V. D., Decuyper, A., and Krings, G. "A survey of results on mobile phone datasets analysis", EPJ Data Science, 4(1):10.

[Bui and Jung, 2018] Bui, K.-H. and Jung, J.J. "Internet of agents framework for connected vehicles: A case study on distributed traffic control system". Journal of Parallel and Distributed Computing, 116:89–95

[Calabrese et al., 2015] Calabrese, F., Ferrari, L., and Blondel, V. D. "Urban sensing using mobile phone network data: a survey of research", ACM Computing Surveys, 47(2):25.

[Calabrese et al., 2011] Calabrese, F., Smoreda, Z., Blondel, V. D., and Ratti, C. "Interplay between telecommunications and face-to-face interactions: A study using mobile phone data", PloS one, 6(7):e20814.

[Chen et al., 2014] Chen, M., Mao, S., and Liu, Y. "Big data: A survey", Mobile Networks and Applications, 19(2):171–209.

[Cho et al., 2011] Cho, E., Myers, S. A., and Leskovec, J. "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, pages:1082–1090. ACM.

[Dunbar et al., 2015] Dunbar, R. I., Arnaboldi, V., Conti, M., and Passarella, A. "The structure of online social networks mirrors those in the offline world", Social Networks, 43:39–47.

[Gundecha et al., 2014] Gundecha, P., Barbier, G., Tang, J., and Liu, H. "User vulnerability and its reduction on a social networking site", ACM Transactions on Knowledge Discovery from Data (TKDD), 9(2):12.

[Han et al., 2017] Han, X., Wang, L., Cui, C., Ma, J., and Zhang, S. "Linking multiple online identities in criminal investigations: A spectral co-clustering framework", IEEE Transactions on Information Forensics and Security, 12(9):2242-2255.

[Hong and Jung, 2016] Hong, M. and Jung, J.J. "MyMovieHistory: Social Recommender System by Discovering Social Affinities Among Users", Cybernetics and Systems 47(1-2):88–110.

[Hong and Jung, 2018] Hong, M. and Jung, J.J. "Multi-Sided recommendation based on social tensor factorization". Information Sciences,447:140–156.

[Hong et al., 2017] Hong, M., Jung, J.J., Piccialli, F., and Chianese, A. "Social recommendation service for cultural heritage", Personal and Ubiquitous Computing 21(2):191–201.

[Hristova et al., 2016] Hristova, D., Williams, M. J., Musolesi, M., Panzarasa, P., and Mascolo, C. "Measuring urban social diversity using interconnected geo-social networks". In Proceedings of the 25th International Conference on World Wide Web, pages:21–30. International World Wide Web Conferences.

[Irani et al., 2009] Irani, D., Webb, S., Li, K., and Pu, C. "Large online social footprints–an emerging threat". In Computational Science and Engineering, 2009. CSE'09. International Conference on, volume 3, pages:271–276. IEEE.

[Jung, 2017] Jung, J.E. "Discovering Social Bursts by Using Link Analytics on Large-Scale Social Networks", Mobile Networks & Applications, 22(4):625–633.

[Kim et al., 2017] Kim, S., Kini, N., Pujara, J., Koh, E., and Getoor, L. "Probabilistic visitor stitching on cross-device web logs". In Proceedings of the 26th International Conference on World Wide Web, pages:1581–1589. International World Wide Web Conferences.

[Kong et al., 2013] Kong, X., Zhang, J., and Yu, P. S. "Inferring anchor links across multiple heterogeneous social networks". In Proceedings of the 22nd ACM international conference on Information & Knowledge Management, pages:179–188. ACM.

[Korula and Lattanzi, 2014] Korula, N. and Lattanzi, S. "An efficient reconciliation algorithm for social networks", Proceedings of the VLDB Endowment, 7(5):377–388.

[Lee and Jung, 2017] Lee, O-J. and Jung, J.E. "Sequence Clustering-based Automated Rule Generation for Adaptive Complex Event Processing", Future Generation Computer Sciences, 66:100–109.

[Liu et al., 2013] Liu, J., Zhang, F., Song, X., Song, Y.-I., Lin, C.-Y., and Hon, H.-W. "What's in a name?: an unsupervised approach to link users across communities". In Proceedings of the sixth ACM international conference on Web search and data mining, pages:495–504. ACM.

[Liu et al., 2016] Liu, L., Cheung, W. K., Li, X., and Liao, L. "Aligning users across social networks using network embedding.". In IJCAI, pages:1774–1780.

[Liu et al., 2014] Liu, S., Wang, S., Zhu, F., Zhang, J., and Krishnan, R. "Hydra: Large-scale social identity linkage via heterogeneous behavior modeling". In Proceedings of the 2014 ACM SIGMOD international conference on Management of data, pages:51–62. ACM.

[Ma and Leijon, 2011] Ma, Z. and Leijon, A. "Bayesian estimation of beta mixture models with variational inference", IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(11):2160–2173.

[Ma et al., 2014] Ma, Z., Rana, P. K., Taghia, J., Flierl, M., and Leijon, A. "Bayesian estimation of dirichlet mixture model with variational inference", Pattern Recognition, 47(9):3143–3157.

[Ma et al., 2018] Ma, Z., Xue, J.-H., Leijon, A., Tan, Z.-H., Yang, Z., and Guo, J. "Decorrelation of neutral vector variables: Theory and applications", IEEE transactions on neural networks and learning systems, 29(1):129–143.

[Malhotra et al., 2012] Malhotra, A., Totti, L., Meira Jr, W., Kumaraguru, P., and Almeida, V. "Studying user footprints in different online social networks". In Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on, pages:1065–1070. IEEE.

[Man et al., 2016] Man, T., Shen, H., Liu, S., Jin, X., and Cheng, X. "Predict anchor links across social networks via an embedding approach.". In IJCAI, pages:1823–1829.

[Naboulsi et al., 2015] Naboulsi, D., Fiore, M., Ribot, S., and Stanica, R. "Mobile traffic analysis: a survey", Université de Lyon, Tech. Rep. hal-01132385.

[Nie et al., 2016] Nie, Y., Jia, Y., Li, S., Zhu, X., Li, A., and Zhou, B. "Identifying users across social networks based on dynamic core interests", Neurocomputing, 210:107–115.

[Nguyen and Jung, 2017] Nguyen, D.T. and Jung, J.E. "Real-time event detection for online behavioral analysis of big social data", Future Generation Computer Systems, 66:137–145.

[Nguyen and Jung, 2018] Nguyen, H.L. and Jung, J.E. SocioScope: A framework for understanding Internet of Social Knowledge, Future Generation Computer Systems, 83:358–365.

[Onnela et al., 2007] Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J., and Barabási, A.-L. "Structure and tie strengths in mobile communication networks", Proceedings of the national academy of sciences, 104(18):7332–7336.

[Perito et al., 2011] Perito, D., Castelluccia, C., Kaafar, M. A., and Manils, P. "How unique and traceable are usernames?". In International Symposium on Privacy Enhancing Technologies Symposium, pages:1–17. Springer.

[Qiao et al., 2018] Qiao, Y., Xing, Z., Md. Fadlullah, Z., Yang, J., and Kato, N. "Characterizing flow, application, and user behavior in mobile networks - a framework for mobile big data", IEEE Wireless Communications, 25(1):40–49.

[Qiao et al., 2016] Qiao, Y., Zhao, X., Yang, J., and Liu, J. "Mobile big-data-driven rating framework: measuring the relationship between human mobility and app usage behavior", IEEE Network, 30(3):14–21.

[Rahmati et al., 2011] Rahmati, A., Shepard, C., Tossell, C., Dong, M., Wang, Z., Zhong, L., and Kortum, P. "Tales of 34 iphone users: How they change and why they are different", arXiv preprint arXiv:1106.5100.

[Riederer et al., 2016] Riederer, C., Kim, Y., Chaintreau, A., Korula, N., and Lattanzi, S. "Linking users across domains with location data: Theory and validation". In Proceedings of the 25th International Conference on World Wide Web, pages:707–719. International World Wide Web Conferences Steering Committee.

[Shu et al., 2017] Shu, K., Wang, S., Tang, J., Zafarani, R., and Liu, H. "User identity linkage across online social networks: A review", ACM SIGKDD Explorations Newsletter, 18(2):5–17.

[Sorrentino, 2009] Sorrentino, F. "The virtual identity, digital identity, and virtual residence of the digital citizen". In Encyclopedia of Information Communication Technology, pages:825–832. IGI Global.

[Tan et al., 2014] Tan, S., Guan, Z., Cai, D., Qin, X., Bu, J., and Chen, C. "Mapping users across networks by manifold alignment on hypergraph.". In AAAI, volume 14, pages:159–165.

[Zafarani and Liu, 2013] Zafarani, R. and Liu, H. "Connecting users across social media sites: a behavioral-modeling approach". In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pages:41–49. ACM.

[Zafarani et al., 2015] Zafarani, R., Tang, L., and Liu, H. "User identification across social media", ACM Transactions on Knowledge Discovery from Data (TKDD), 10(2):16.

[Zhang and Yu, 2016] Zhang, J. and Yu, P. S. "Pct: partial co-alignment of social networks". In Proceedings of the 25th International Conference on World Wide Web, pages:749–759. International World Wide Web Conferences.

[Zhang et al., 2015] Zhang, Y., Tang, J., Yang, Z., Pei, J., and Yu, P. S. "Cosnet: Connecting heterogeneous social networks with local and global consistency". In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages:1485–1494. ACM.