User Behavior Fingerprinting with Multi-Item-Sets and Its Application in IPTV Viewer IdentificationMulti-Item-Set Fingerprinting of User Behaviors

Article *in* IEEE Transactions on Information Forensics and Security · January 2021

DOI: 10.1109/TIFS.2021.3055638

CITATIONS
8

READS
253

5 authors, including:

Can Yang
South China University of Technology
36 PUBLICATIONS **595** CITATIONS

SEE PROFILE

User Behavior Fingerprinting With Multi-Item-Sets and Its Application in IPTV Viewer Identification

Can Yang, *Member, IEEE*, Lan Wang, Houwei Cao, *Member, IEEE*, Qihu Yuan, and Yong Liu, *Fellow, IEEE*

***Abstract*— User activities in cyberspace leave unique traces for user identification (UI). Individual users can be identified by their frequent activity items through statistical feature matching. However, such approaches face the data sparsity problem. In this paper, we propose to address this problem by multi-item-set fingerprinting that identifies users not only based on their frequent individual activity items, but also their frequent consecutive item sequences with different lengths. We also propose a new similarity metric between fingerprint vectors that combines the advantages of Jaccard distance and relative entropy distance. Furthermore, we develop a fusion decision scheme by consolidating matching candidates generated by different similarity metrics. It improves the precision at the price of extra rejection. Our proposed approaches can be used in both one-by-one matching and bipartite graph group matching. Through extensive experiments on three real user datasets, in particular a large-scale Internet Protocol Television (IPTV) viewer dataset, we demonstrate that the proposed approaches outperform the state-of-the-art methods. The average matching precision reaches 93.8% for a dataset of 1,000 users and 100% for a dataset of 100 users. This work is of significance for information forensics and raises a new challenge for human privacy protection in cyberspace.**

***Index Terms*— User identification, deanonymization, statistical feature matching, frequent item set, IPTV, pattern recognition, user behaviors, user identification.**

I. Introduction

HUMAN behaviors in cyberspace over time inevitably leave digital traces, which can be analyzed for user identification (UI). There are wide uses of UI in many fields, including information forensics, target tracking, personalized recommendations, and personalized advertisements, etc. While user accounts can be used to identify users, the actual user of an account may not be its real owner when the account is stolen, lent, and anonymized intentionally or unintentionally. Recently, Gerhart and Koohikamali reported that the majority of signed accounts in social networks are anonymous [1]. While accounts and certificates can be illegally obtained or forged, the statistical characteristics of a user's digital behaviors in cyberspace would often reveal her true identity. Rather than the traditional user account-based UI, we define the UI problem in this paper as the process of identifying users by the statistical features of their behaviors. The main motivation for the UI research lies in information forensics and personalized recommendation. We are interested in finding out the real identity of the user who leaves the behavioral traces, instead of the user account used to access the online service. Therefore, this paper focuses on UI based on the statistical information of the actual user behaviors. We propose new methods to improve UI accuracy.

One way to characterize a user's behaviors in cyberspace is to identify her *pattern items* which she consumes, generates, or interacts with in various online activities, such as websites visited, web pages browsed, goods browsed, television channels watched, places visited (e.g., supermarkets, restaurants, car parks, etc.), and words and phrases published in articles, etc. We define a pattern item with a frequency higher than a given threshold as a frequent item, and call the set of the frequent items as the frequent item set (FIS). In this work, we find that the FISs of users are relatively stable over a long time period, and can be used for reliable user identification. Specifically, we can calculate the similarity between an unknown user and a user with known identity by taking the intersection of their FISs and comparing the frequency distributions of their common frequent items. Then we can use the identity of the most similar known user to identify the unknown user. In such a process, there are three critical steps: *1) establishing the frequent item sets; 2) calculating the pairwise similarity between users; 3) making the final one-by-one or group matching decision.* We propose new methods for all the three steps for accurate UI.

One common challenge for statistical feature matching based on UI is data sparsity, where the available data are not rich enough to extract reliable features. To address this problem, prior work has focused on finding multi-source digital features that are relevant to the identity of the behavioral subject, for example, the cross-social network information of user-generated content [2], and fusions of surveillance video and wireless signaling information [3], [4], However, in practical scenarios, it is expensive and error-prone to collect and correlate data from multiple sources. Low quality data

Manuscript received June 29, 2020; revised November 30, 2020 and January 16, 2021; accepted January 19, 2021. Date of publication January 29, 2021; date of current version March 12, 2021. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant U1611461 and Grant 61876065, in part by the Guangzhou Science and Technology Program Key Projects, Guangdong, China, under Grant 201704030124, and in part by the Science and Technology Plan Project of Guangdong, China, under Grant 2014B010115002. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Pedro Comesana. *(Corresponding author: Can Yang.)*

Can Yang and Lan Wang are with the College of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China (e-mail: cscyang@scut.edu.cn).

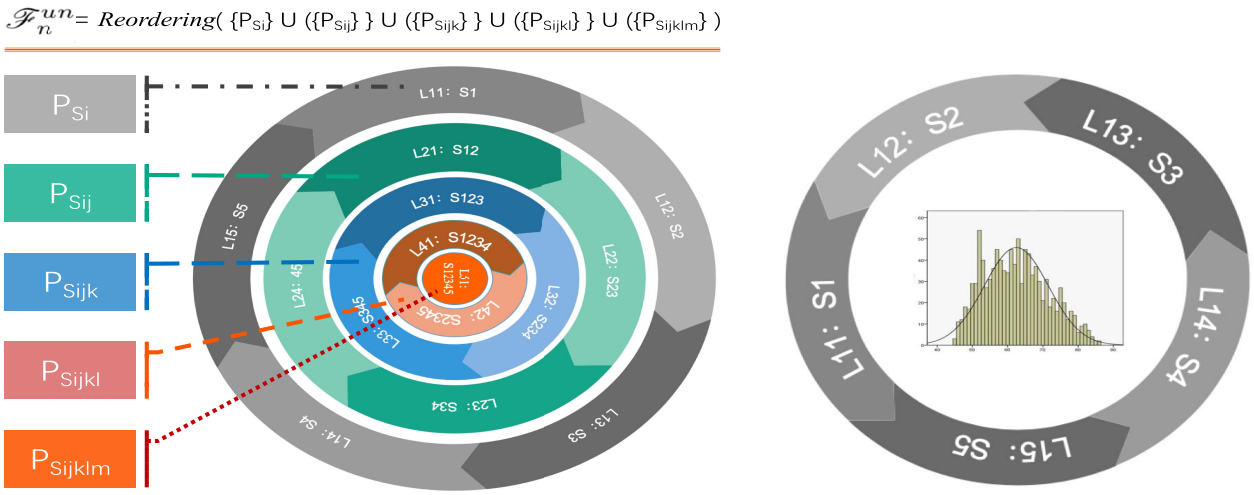
Howwei Cao is with the Department of Computer Science, New York Institute of Technology (NYIT), Old Westbury, NY 10023 USA (e-mail: hcao02@nyit.edu).

Qihu Yuan is with NetEase Computer System Company, Ltd., Guangzhou 510665, China.

Yong Liu is with the Department of Electrical and Computer Engineering, New York University, Brooklyn, NY 11201 USA (e-mail: yongliu@nyu.edu).

The dataset is available online: https://dx.doi.org/10.21227/hp53-wz08

Digital Object Identifier 10.1109/TIFS.2021.3055638



(a) Multi-item-sets Sequential Fingerprint; $\mathcal{F}_n^{un}$(n=5)={{P_{Si}}∪{P_{Sij}}∪{P_{Sijk}}∪{P_{Sijkl}}∪{P_{Sijklm}}}, the symbol of '$P_{S\cdots}$' means the probability of the subsequence of 'S$\cdots$', here m=l+1, l=k+1, k=j+1, j=i+1. $\mathcal{F}_n^{un}$ will be handled by ranking and then tailoring in practice.

(b) Single-item-sets Sequential Fingerprint; $\mathcal{F}_n^{un}$ (n=1) = {Pi}, i=1,2,$\cdots$,M; $\mathcal{F}_n^{un}$ can be represented with the histogram of the original item space.

Fig. 1. Schematic Diagram of a Multi-item-sets Sequential Fingerprint in the case of a 5-level fingerprint.

from one source often introduce noise to multi-source features, leading to low UI accuracy. In this paper, we focus on FIS features from a single information source, and use *frequent item sequences* to address data sparsity. *The key observation is that human activities in cyberspace naturally form temporal sequences, and the transitions between consecutive activities can also be used as fingerprints for user identification.* In essence, we consider not only the frequent individual items, but also frequent consecutive item sequences in user identification. Towards this, we first propose a model on Multi-Item-Set Fingerprinting of User Behaviors, called MISFUB, which aggregates the single-item-sets and extendable multiple-item-sets in sequence, as shown in Figure 1 and described in Section III. We further propose an algorithm called **SURE** (i.e., **S**can, **U**nion, **R**ank, and **E**xtract; see Section III-A for details) to generate the feature vector for the extended fingerprints. Next, to work with multi-item-set fingerprints, we propose a new similarity distance by combining the advantages of Jaccard distance and relative entropy distance. Here, we use the similarity distance for the quantitative representation of user correlation between known set and unknown set. Moreover, we propose a fusion decision scheme that consolidates decisions of multiple matching algorithms for high-precision user matching.

We carry out a systematic experimental investigation and verification of the proposed fingerprint construction, similarity measure, and fusion scheme using three different datasets of real users. In particular, using a large-scale watched-channel dataset of Internet Protocol Television (IPTV) [5], [6], we conduct an application case study of the proposed model and approaches. Extensive experiments show that the proposed approaches significantly outperform the state-of-the-art methods, and the average matching precision reached 93.8% for a dataset of $1,000$ users and 100% for a dataset of 100 users. Moreover, the precision of the fusion of multiple approaches can reach 100% if some decision-making opportunities are abandoned. Our study shows some important characteristics of IPTV viewer identification using watched-channel logs. Given the widespread adoption of IPTV, this case provides a new method of information forensics and subject tracking using basic logs. Most importantly, our study only uses the one-dimensional information source of channels watched by subscribers (namely, the time sequence of watched channel IDs), and does not involve other sensitive attributes or private user information. This makes it more practical and easier to implement than other approaches using multi-source information in research and engineering practice.

This paper proposed the computing framework of MISFUB in methodology for user identification; our main contributions can be summarized as follows:

- We construct n-level multi-item-set fingerprints of user digital behaviors from one sequence of items, i.e., the SURE algorithm.
- We present a new similarity distance combining Jaccard and **a variant** of the Kullback-Leibler divergence function (KL), i.e., JKL, which outperforms other similarity metrics in a large number of experiments using IPTV datasets.
- We propose a novel fusion decision scheme (i.e., FI) to improve the performance of the proposed approaches, which uses the intersection of candidates produced by individual matching approaches and achieve a precision of 100% in many experiments on the IPTV dataset.
- We demonstrate the effectiveness of the proposed n-level fingerprints and similarity distance, not only for one-by-one matching but also for bipartite graph group matching.
- Based on the watched-channel logs of IPTV subscribers, we demonstrate the advantages of the proposed methods, and study characteristics of IPTV UI for information forensics. Additional evaluations on a web-browsing dataset and an online shopping dataset confirm the conclusions drawn on the IPTV dataset.

The rest of the paper is organized as follows: Section II introduces the prior related work. In Section III, we first present the framework of UI and the feature selection approach, and then propose a new similarity distance, illustrate the fusion matching decision scheme. Section IV shows and analyzes a series of experiments in the case of IPTV. Additional evaluation results on web browsing and online shopping datasets are reported in Section V. The paper is concluded in Section VI.

II. Related Work

This work is mainly related to the approaches of UI, similarity distance measures, and fusion decision for user matching. Focusing on the mainstream online technique of UI, Han *et al.* [7] made an overview and divided the prior work into four classes based on the information source, such as individual profiles [8], social relationships [9], authorship [10], and user behavior [11].

In prior work on UI, researchers have usually proposed specific identification solutions based on the attributes and data characteristics of the problem in different scenarios. For example, the author of [12] found a mapping relation of user names between different platforms by analyzing the chain of user social relations to recognize users across social media. In the problem of author identification [10], it was believed that an author's identity could be identified by using the text style. Researchers could perform user recognition by extracting the characteristics of global positioning system (GPS) tracks [13].

In the process of user recognition, the first goal is to represent user behavior patterns. Thus, we need to design a "fingerprint" to represent a user behavior pattern [14], and heterogeneous "fingerprints" provide different validation for heterogeneous scenarios of UI. As stated in [11], a histogram of user features can be represented by a feature vector and then used for various methods of pattern recognition, which can be used to match "fingerprints". For instance, neural networks can be used for UI from keystrokes [15]. In the case of a communication research community, Soltani *et al.* [16] investigated network flow fingerprinting and used a queuing model to conduct an in-depth analysis of its limits in multiple presented cases.

According to the above classification, our work is quite close to user-behavior-based methods, especially [11], and the main difference between our work and the prior literature is that we use only a one-dimensional original item sequence, and we expand it to a multi-item-set sequence to use a limited information source effectively. In this paper, we first present how to construct a multilevel extendable item set fingerprint, which has not been clearly reported in previous work on user behavior matching. In practice, a one-dimensional item accessed by a user is easier to collect and involves fewer personal privacy concerns [17] in data investigation than some approaches to UI that need multiple information sources [8]. Moreover, a significant difference from the majority of prior work is that we use a relatively large-scale IPTV dataset to test our approaches in multiple dimensions, such as the number of users, the frequency of accessing the item set, the similarity distance, and a fusion of multiple approaches. To the best of our knowledge, this is the first work to investigate user matching in the case of IPTV scenarios; it is different from the studies on IPTV channel recommendation [5], [6] and video recommendation [18], but strongly related to their research background.

Next, we propose a similarity metric, called JKL, combining the KL distance proposed in [11], [19] and the Jaccard distance. The work of [20], [21] and [22] analyzed the advantages and limitations of Jaccard similarity in recommendation systems. For a study of UI, the authors of [23] presented a similarity distance based on Jaccard to identify the same person across two similar social networks. But, the proposed JKL in this paper differs from the previous ones.

In this work, we find that frequent item sets are superior to infrequent ones for the presented fingerprint pattern matching. It is well known that frequent item sets are widely used for associated rule mining; for example, Hu *et al.* [24] used item sets for the detection of frequent alarm patterns in industrial alarm floods. Chee *et al.* [25] compared different algorithms for frequent pattern mining in terms of computing consumption. We therefore introduce this concept into user matching using a feature vector comprising frequent item sets.

Recently, Li, Yongjun *et al.* [26] studied the UI of user-generated content (UGC) by supervised machine learning using three kinds of similarity—of spatial, temporal and content dimensions—on two UGC accounts, and they used a fusion classifier with a supervised learning mode combining the time, location, and word information of a UGC account. Similarly, we provide a fusion decision scheme based on intersection of the results of different similarity distances. Unlike the supervised approaches in [26] and [15], all of our approaches in this paper are unsupervised.

For matching decision methods, the authors of [13] deanonymize users one-by-one based on a mobility model called the mobility Markov chain, whereas those of [11] deanonymize all the users. In this work, we first use one-by-one matching to evaluate the present approach and algorithm for deanonymization and then perform a comparison between bipartite graph matching and one-by-one matching, similar to [11]. The shared code on the maximum weight bipartite matching (bipartite_matching) in [27] uses a state-of-the-art algorithm and reduces the time complexity to $O(N^3)$ to solve this problem, and we find that the innovation is also valid for graph matching. Moreover, we discuss the performance improvement and time cost of bipartite graph matching.

In addition, we recently focused on studies related to IPTV scenarios, such as IPTV channel recommendation [5], [6], [28] and fast channel switching [29]. If we can know the identifier of a real user whenever he/she uses a nickname or fake account, this is also of significance in channel recommendation and other online promotion of digital businesses [30]. A similar work [31] investigated how to recommend YouTube videos to Facebook users and designed a top-k video recommendation method using the social interaction information between users for niche videos. In particular, in a peer-to-peer application [32], [33], the identifier of each user is usually anonymous, and it is difficult to identify the same user by accessed content. However, this work can make it possible and can provide information forensics in peer-to-peer scenarios.

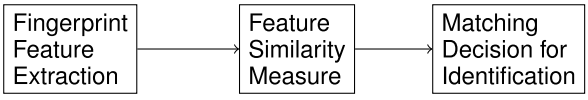
TABLE I
SYMBOLS AND NOTATION

Sym.	Notation	Description
UI	User Identification	IPTV Viewer Identification
C	Item	TV channel
m	Index of an item	Index of a channel
N	Total number of users	Total number of viewers
M	Total number of items	Total number of channels
r	Index of a record	Index of a watched channel
V_x	Observed/test user set	User ID in V_x is unknown
u_{x_i}	user in V_x	u_{x_i} is a member of V_x
V_y	Reference user set	User ID in V_y is known
u_{y_j}	user in V_y	u_{y_j} is a member of V_y
$\|\cdot\|$	Modulus of an item set	The counting function
S_{V_x}	the size of unknown user set	the size of V_x
S_{V_y}	the size of known set	the size of V_y
n	Number of levels or rings	Extendable channel types
N_r	Number of records	Viewed-channel records
d	Index of period of time	Date of the IPTV dataset
$\mathscr{F}_n^{un}$	feature vector of union	Union of all the item sets
$\mathscr{F}_n$	Used feature vector	A part cut out of $\mathscr{F}_n^{un}$
r_{top}	Ratio of the top item sets	$r_{top} = \|\mathscr{F}_n\| / \|\mathscr{F}_n^{un}\|$
K	Number of approaches	for fusion decision
k	Index of approach	for fusion decision
KL	A Variant of Kullback-Leibler	The similarity metric [11] [19]
JKL	A new similarity metric	Proposed in this paper
TP	True positives	Number of Successful matches
FP	False positives	Number of Incorrect matches
Pre	Precision Rate	Precision = TP/(TP+FP)
RJ	Rejection number	Number of rejections
RR	Rejection Rate	RR= Rj/N
NM	one-by-one matching	Matching users one by one
GM	Bipartite Graph Matching	Matching for all the users

In summary, the key characteristics of our work for UI are one input sequence of single items, n-level extendable multi-item-sets fingerprint derived from user behaviors, JKL similarity distance, fusion decisions, and unsupervised solutions.

III. METHODOLOGY AND PROPOSED APPROACHES

For the convenience of description, Table I presents the nomenclature. The general workflow of UI is shown as follows:



In this section, we first introduce the basic methods of matching decisions for UI and then propose our work in three parts. For simplicity, similar to the work in [11], we adopt a complete bipartite graph to illustrate the user matching problem. Let V_x be the set of users to be identified, and V_y be the set of known users. The user matching problem can be modeled by a bipartite graph $G =< V, E >$ in Figure 2, where $V = V_x \cup V_y$, $V_x \cap V_y = \varnothing$, and an edge connects the vertices between V_x and V_y, there is no edge between two vertices in the same partition. If each vertex in V_x is connected to a vertex in V_y, G is a complete bipartite graph. The weight of the edge, denoted as w_{ij}, represents the possibility that u_{x_i} matches u_{y_j}. Based on a certain similarity measure, the weights are computed as $w_{ij} = similarity(u_{x_i}, u_{y_j})$, where $j = 1, 2, \cdots, N_y$, N_y denotes the number of members in V_y. There are two matching methods to de-anonymize a group of users. One conducts one-by-one matching with each

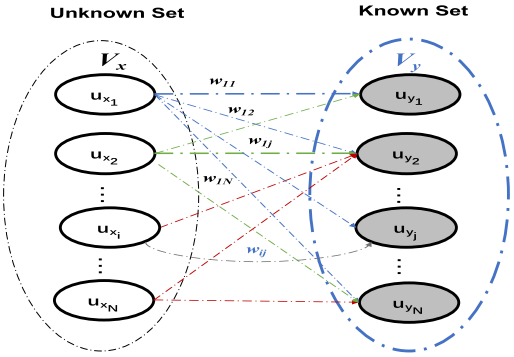


Fig. 2. Schematic Diagram of Fingerprint Feature Matching: The problem of matching fingerprints across two sets can be visualized as a matching problem on a weighted complete bipartite graph. $\hat{u}_{x_i} = \mathscr{D}(w_{ij}) = \arg\max_{j \in \{V_y\}}(w_{ij})$.

user, where the matching decision is based on the maximal similarity, i.e.,

$$\hat{u}_{x_i} = \arg\max_{j \in \{V_y\}} \quad w_{ij}. \tag{1}$$

If $\hat{u}_{x_i} = u_{x_i}$, the matching is true; otherwise, it is false. The other method uses complete bipartite graph matching in batching mode for all the users. The matching, noted as $\mathcal{M}$, is a one-one mapping between V_x and V_y, $\mathcal{M} = \{< i, f(i) >, \forall i \in V_x\}$, where $f(i) \in V_y$ is the matched item for the original item i, here $f(i) \neq f(j)$, if $i \neq j$. The decision function of GM is:

$$\hat{U}_{V_x, V_y} = \arg\max_{\mathcal{M}_k \in \mathbf{M}} \sum_{<i,j> \in \mathcal{M}_k} w_{ij}, \tag{2}$$

where $\mathbf{M}$ is the set of possible one-one mappings. The algorithms to solve maximum weight bipartite matching problem was comprehensively surveyed in [11]. Our proposed fingerprint features and similarity metric can be used for both methods to improve their matching performance. Graph matching usually outperforms one-by-one matching in terms of accuracy (see Section IV-G and Figure 15), whereas it has much higher computation overhead. We will discuss the computational complexity in detail in Section IV-G.2.

To improve the performance of user matching, our work focuses on three tiers: constructing an extensible multi-item-set fingerprint feature vector, exploring a new similarity metric, and developing a fusion decision approach using multiple similarity measures.

A. The Extensible Multi-item-set Feature Fingerprint

The primary cause of the failure of user matching is data sparsity, where too few features are available to recognize the real user identity. To acquire enough user behavioral features from the limited data sources, we present the MISFUB, a Multi-Item-Set Feature fingerprint of User Behaviors, which uses the single item sets in the original data to generate multi-level sequence feature item sets, as illustrated in Figure 1a. In comparison, Figure 1b illustrates the conventional features of user matching using the original items.

Given the original sequence of items $S_1, \cdots, S_{N_r}$, we can study the sequential patterns with different lengths, which consist of consecutive multiple items. We define the number of items in the multi-item-set as the modulus. In Figure 1, each ring represents the behavior trajectory of a progressively increasing modulus of the item set. As shown in Figure 1a, from outside to inside, the first ring is the set of individual items, i.e. `level-1`, $C_1 = \{S_i \mid i = 1, 2, \cdots, N_r\}$, the second ring consists of all the `level-2` consecutive subsequences, i.e., $C_2 = \{\langle S_i, S_{i+1}\rangle, \mid i = 1, 2, \cdots, N_r - 1\}$, $\cdots$, and the k-th ring is the set of `length-k` consecutive subsequences, $C_k = \{\langle S_i, \cdots, S_{i+k-1}\rangle \mid i = 1, 2, \cdots, N_r - k + 1\}$. The symbol of 'Lij' in Figure 1 represents the label of an item set, here 'i' means the level of the ring, and 'j' denotes the index of an item set in the current ring. The modulus of a subsequence item set in the k-level ring is k. The perimeter of the k-level ring, called PC_k, equals to $N_r - k + 1$; i.e., the ring at level k has PC_k subsequences. Therefore, the total number of subsequences among the n rings is:

$$OP_n = \sum_{i=1}^{n}(N_r - k + 1) = n \cdot N_r + \frac{n - n^2}{2}.$$

The above illustration in Figure 1a is the preliminary pattern of the presented fingerprints. If this pattern is used directly, the ratio of the enlarged n-level sample space to the original space is $\eta_o = OP_n/OP_1 = n(1 + (1 - n)/(2 \cdot N_r))$, here, $OP_1 = N_r$. In essence, what we are truly interested in is the unique item sets and their frequencies. Therefore, we propose a feature vector construction algorithm to generate user fingerprints, called **SURE**.

1) Scan: First, we scan the one-dimensional input sequence of the original items to construct n levels of rings of the fingerprint graph, i.e., $C_r, r = 1, 2, \cdots,$ n, according to the approach mentioned above.

2) Union: Then, we construct the union set $\mathscr{F}_n^{un}$ of all the item sets of n rings in the fingerprint. Thus, the enlarged ratio of the matching space becomes $\eta_o^{union} = \|\mathscr{F}_n^{un}\|/\|\mathscr{F}_1^o\|$, which contributes many more opportunities than a single-item-set feature. On the other hand, too many feature elements might lead to computing overhead and noise that interfere with matching. Although minimal feature selection is an NP-complete problem [34], we can still extract valid feature item sets in $\mathscr{F}_n^{un}$ by the following steps.

3) Rank: We rank subsequences in $\mathscr{F}_n^{un}$ by their frequencies, i.e., $\mathscr{F}_n^{rk} = sort(\mathscr{F}_n^{un} \mid_{frequency})$, in descending order.

4) Extract: We extract the most frequent subsequences in $\mathscr{F}_n^{rk}$ to form the final version of the fingerprint feature vector, i.e., $\mathscr{F}_n$. In the experimental investigation of this paper, we use a normalized parameter r_{top} to represent the ratio of $Length(\mathscr{F}_n)$ to $Length(\mathscr{F}_n^{un})$, instead of selection using a minimum cutoff threshold $support\_min$, to gain insight into the effect of the frequency on the matching performance.

Each element in the constructed feature vector $\mathscr{F}_n$, is a tuple of a subsequence of items and its frequency. The SURE algorithm therefore yields the statistical histogram of the extended multiple item subsequences for computing the similarity distances, which differs from that of the original single item set. The creation of the feature vector $\mathscr{F}_n$ is the primary foundation of user matching for the fingerprint presented in this work and is shown in detail as Algorithm 1. The effect of r_{top} on the matching performance is investigated in Section IV-D.

Algorithm 1: SURE: Building the Feature Vector of the **MISFUB** by Item Set Popularity

Input: n, r_{top}, C_1
Output: $\mathscr{F}_n$
Function $\mathscr{F}_n = \textbf{SURE}(n, r_{top}, C_1)$
$\mathscr{F}_n^{un} = \varnothing$
For r = 1:n
  $S_r = \textbf{Scan}(C_1, r)$
  $\mathscr{F}_n^{un} = \textbf{Union}(\mathscr{F}_n^{un}, S_r)$
End For
$fr_n^{un} = count(unique(\mathscr{F}_n^{un}))$
$\mathscr{F}_n^{rk} = \textbf{Rank}(\mathscr{F}_n^{un}) = \text{sort}(\mathscr{F}_n^{un}, \text{ by } fr_n^{un}, \text{'descend'})$
Lo = $length(\mathscr{F}_n^{rk})$;
$\mathscr{F}_n = \textbf{Extract}(\mathscr{F}_n^{rk}) = \mathscr{F}_n^{rk}(index \le r_{top} \cdot Lo)$
Function $S_r = \textbf{Scan}(S, r)$
LS = length(S)
If $r == 1$ $S_r = S$
Else $S_r = \varnothing$
  For k = 1:LS − r + 1
    Forl = 1: r
      $S_r = [S_r, S\{k + l - 1\}]$
    End For
    $S_r = S\{k + 1\}$
  End For
End If

B. The Proposed Similarity metric

Although the input features are vital for user matching, similarity metrics also have important impact on the matching performance. The similarity distance is a quantitative metric of how similar the behaviors of an unknown user and known users are. The previous studies have used several similarity distance functions, such as Jaccard, Pearson, Cosine, L1-norm, L2-norm, and KL [11]. Patrick Thiran *et al.* [11] conducted an in-depth investigation on the matching performance of KL, L1-norm, and Cosine for UI in scenarios similar to our work, and found that their proposed KL distance outperforms the other similarity distances. However, we find that the matching performance of Jaccard can also outperform the majority of the existing similarity distances for our IPTV dataset. Moreover, we propose a new similarity distance inspired by Jaccard and achieve good user matching performance. The design of similarity distance is a sophisticated problem, and a good design should be relatively stable across distinct scenarios. For the derivation of our proposal, we first adopt the standard Jaccard distance as the weight of the bipartite graph model in Figure 2, using the feature vector proposed

in Section III-A, as follows:

$$w_{ij}^{J} = Jaccard(u_{x_i}, u_{y_j}) = \frac{\|\mathscr{F}_n^{u_{x_i}} \cap \mathscr{F}_n^{u_{y_j}}\|}{\|\mathscr{F}_n^{u_{x_i}} \cup \mathscr{F}_n^{u_{y_j}}\|} = \frac{L_{ij}^{in}}{L_{ij}^{un}}, \quad (3)$$

where $\mathscr{F}_n^{u_{x_i}}$ and $\mathscr{F}_n^{u_{y_j}}$ are the feature vectors for user u_{x_i} and u_{y_j} respectively, and L_{ij}^{in} denotes the number of common subsequences shared by the two users, i.e., $L_{ij}^{in} = \|\mathscr{F}_n^{u_{x_i}} \cap \mathscr{F}_n^{u_{y_j}}\|$, and L_{ij}^{un} denotes the union of the subsequences between the two users, i.e., $L_{ij}^{un} = \|\mathscr{F}_n^{u_{x_i}} \cup \mathscr{F}_n^{u_{y_j}}\|$. The Jaccard distance can be used directly to find matching according to (1) or (2). Similar to prior works [11], [19], and [35], we further consider a similarity metric using KL divergence as follows:

$$w_{ij}^{KL} = \mathcal{D}\left(h_{u_{x_i}}^{in} \| \frac{h_{u_{x_i}}^{in} + h_{u_{y_j}}^{in}}{2}\right) + \mathcal{D}\left(h_{u_{y_j}}^{in} \| \frac{h_{u_{x_i}}^{in} + h_{u_{y_j}}^{in}}{2}\right), \quad (4)$$

where $h_{u_{x_i}}^{in}$ and $h_{u_{y_j}}^{in}$ denote the frequency histograms of user u_{x_i} and u_{y_j} over the set of their common subsequences $\mathscr{F}_n^{u_{x_i}} \cap \mathscr{F}_n^{u_{y_j}}$, respectively, and $\mathcal{D}(\cdot\|\cdot)$ is the KL divergence function [36], defined as

$$\mathcal{D}(\cdot\|\cdot) = \mathcal{D}_{KL}(\alpha||\beta) = \sum_{l=1}^{L^{in}} \alpha(l) log(\frac{\alpha(l)}{\beta(l)}),$$

where L^{in} is the number of common subsequences of u_{x_i} and u_{y_j}.

However, Jaccard does not consider the frequency value of items, while KL only cares about the frequencies of the common items in the matching targets and does not take the global item information into account, in this paper. To make full use of the advantages of Jaccard and KL and avoid their limitations, we propose a new similarity distance, called **JKL**, which combines equation 4 and equation 3, and formulates as follows:

$$w_{ij}^{JKL} = w_{ij}^{J} / w_{ij}^{KL} = \frac{L_{ij}^{in}/L_{ij}^{un}}{\sum_{l=1}^{L^{in}} h_{u_{x_i}}^{in}(l) log(\frac{h_{u_{x_i}}^{in}(l)}{\beta(l)}) + \sum_{l=1}^{L^{in}} h_{u_{y_j}}^{in}(l) log(\frac{h_{u_{y_j}}^{in}(l)}{\beta(l)})}, \quad (5)$$

where $\beta(l) = \frac{1}{2}(h_{u_{x_i}}^{in}(l) + h_{u_{y_j}}^{in}(l))$.

The proposed JKL similarity is simply the product of the Jaccard similarity and the inverse of KL divergence. While more complicated combinations of the two similarity metrics are possible, we didn't find significant performance improvement in maximal matching based UI. Using the presented fingerprint feature vector $\mathscr{F}_n$, all the similarity metrics used in this work show an improved matching performance in our experiments, and the proposed JKL outperforms the other similarity metrics on the IPTV, web browsing and online shopping datasets. In essence, the similarity distance is a kind of quantitative representation of correlation, and the presented extendable level item-set could be helpful to enhance the degree of the correlation. In the experiments, we mainly compare JKL, KL and Jaccard in Section IV, and we consider several other traditional similarity metrics in Section IV-G.3.

C. The Fusion Matching Decision Scheme

To further improve the precision of fingerprinting, we propose a fusion decision approach to enhance the confidence of matching at the price of a certain number of rejections. Most importantly, the proposed approach has no need for additional outside information sources. The fusion approach with rejection can abandon decisions when the decision-making condition is not satisfied, which can improve the matching precision or confidence. Improving the precision of user matching is helpful in information forensics to reduce the error rate of user matching, although the rejection rate is a price of approaches with rejection.

Given a set of K matching approaches, we provide a fusion matching approach based on the intersection of the candidate sets generated by each approach, called **FI**, and we formulate the fusion decision function as follows:

$$\hat{U}_{x_i}(\{\hat{u}_{x_i}^k\}_{k=1}^K) = \bigcap_{k=1}^{K} \hat{u}_{x_i}^k \quad (6)$$

where k denotes the index of the matching approach used, $\hat{u}_{x_i}^k$ is the matching result returned by approach k. There is at most one element in the intersection because each matching approach produces one matching candidate. In other words, if all K matching approaches return the same result, the common matching result will be used as the final fusion matching; otherwise, no fusion matching will be returned. To evaluate this method, we define the precision as follows:

$$Precision = \frac{\sum_{i=1}^{|V_x|} \mathbb{1}(u_{x_i} \in \hat{U}_{x_i})}{\sum_{i=1}^{|V_x|} \mathbb{1}(\hat{U}_{x_i} \neq \varnothing)}, \quad (7)$$

where $\mathbb{1}(\cdot)$ is the indicator function. The rejection rate can be defined as follows:

$$RejectRate = \frac{1}{|V_x|} \sum_{i=1}^{|V_x|} \mathbb{1}(\hat{U}_{x_i} = \varnothing). \quad (8)$$

We provide a metric called the net true rate (NTR) to evaluate the performance of the fusion approaches.

$$NTR = \frac{1}{|V_x|}(TP \cdot \mathscr{W}_{TP} + FP \cdot \mathscr{W}_{FP} + RJ \cdot \mathscr{W}_{RJ}), \quad (9)$$

where RJ denotes the number of rejections, namely how many times the intersection set is empty, here, TP $= \sum_{i=1}^{|V_x|} \mathbb{1}(u_{x_i} \in \hat{U}_{x_i})$, FP $= \sum_{i=1}^{|V_x|} \mathbb{1}(u_{x_i} \notin \hat{U}_{x_i})$, and $\mathscr{W}_{**}$ respectively denotes the weight of TP, FP and RJ; in general, we set $\mathscr{W}_{TP} = 1$, $\mathscr{W}_{FP} = -1$, and $\mathscr{W}_{RJ} = 0$, that is,

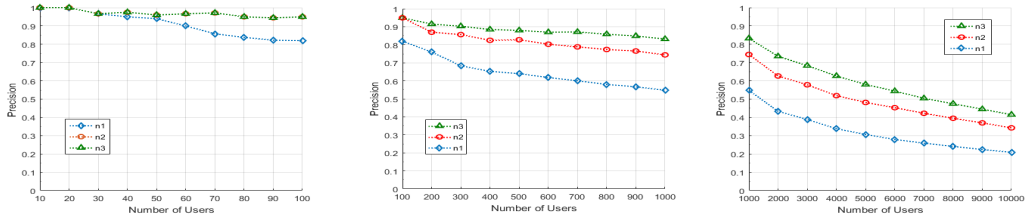
$$NTR = \frac{TP - FP}{TP + FP + RJ} = \frac{TP - FP}{|V_x|}. \quad (10)$$

Based on the MISFUB, we design two rounds of fusion with rejection and gain a higher precision than that of a single approach in the case of IPTV experiments. The first round of fusion is to fuse matching results using Jaccard and KL similarity metrics respectively:

$$\hat{U}_{x_i}^{J\&KL} = \arg\max_{j \in \{V_y\}}(w_{ij}^{J}) \cap \arg\max_{j \in \{V_y\}}(w_{ij}^{KL}).$$

The second round of fusion is applied to matching results obtained from three different feature vector rings,

$$\hat{U}_{x_i}^{J\&KL}(\mathscr{F}_1, \mathscr{F}_2, \mathscr{F}_3) = \bigcap_{n=1}^{3} \hat{U}_{x_i}^{J\&KL}(\mathscr{F}_n),$$



(a) Jaccard (10, 100) (b) Jaccard (100, 1000) (c) Jaccard (1000, 10000)

Fig. 3. Precision of the **MISFUB** for n1, n2 and n3 with Jaccard, with a top ratio of 50% ($r_{top} = 0.5$) during testing days from 17 to 31, where the number of users is in the range [10, 100], [100, 1000], and [1000, 10000].

where $\hat{U}_{x_i}^{J\&KL}(\mathscr{F}_n)$ is $J\&KL$ fusion matching result obtained from frequent subsequences up to length n, i.e., ring n as illustrated in Figure 1a. Importantly, we find that all the precision values in the experiments are better than that of any single similarity distance, and the precision after the second round fusion can reach 100% in many cases. Our experiments show that the use of FI could further improve the precision of user identification, which is of great significance in specific application scenarios, such as enhancing the validation of information forensics.

In summary, to enhance the performance of UI, we propose an extendable multi-item-set fingerprint of user behaviors (**MISFUB**), a similarity metric (**JKL**), and fusion decision-making schemes (**FI**). We will evaluate the performance of the proposed methods using real user datasets in the next two sections.

IV. Experimental Evaluation on IPTV Dataset

To investigate the feasibility and performance of our proposals, we conducted a series of experiments in the case of an IPTV subscriber behavior dataset. In IPTV scenarios, an item refers to a live TV channel watched by a viewer, and an item set represents a sequence of channels watched by of a viewer. We designed several groups of experiments to investigate the performance impacts of different factors, involving the levels of the feature fingerprint, the similarity metric, the size of the training set, the number of users, and the fusion scheme of decision-making. We first use one-by-one matching to evaluate the performance of multi-item-set fingerprint, the feature selection in $\mathscr{F}_n$, and the impact of the size of the known dataset, respectively. Then, we study the intersection-based fusion approaches. Furthermore, we conduct graph matching and compare their performance and computation overhead with one-by-one matching.

A. The IPTV Dataset

In the original version of the IPTV dataset,[1] there are more than 4 million records generated by more than 10,000 subscribers to watch 157 TV channels during one month. For convenience and simplicity in research, we split the records of each user in the dataset into two parts according to time: the data before the splitting day comprise the training/known set, and the rest comprise the test/unknown set. We reorder them in descending order of the frequency of user access; that is, a smaller channel index represents that the user had more frequent access. The user identifiers are visible in the training set but invisible in the test set for evaluation.

In the data preparation for experiments, the splitting operation guarantees the formation of a perfect bipartite graph of the user space for a known set and test set. Note that the splitting operation makes the two sets bijective, and the decision-making function in equation 1 leads to each false positive being associated with one false negative. Each member of the unknown set has a corresponding matching target in the known set. The objective of our work is to re-identify the bijection. In this matching mode, the three metrics of $Precision$, $Recall$, and $F1-score$ are equal. $Precision$ is therefore used for the main evaluation metric in the following experiments.

B. Effect of the Level of the Presented Fingerprint

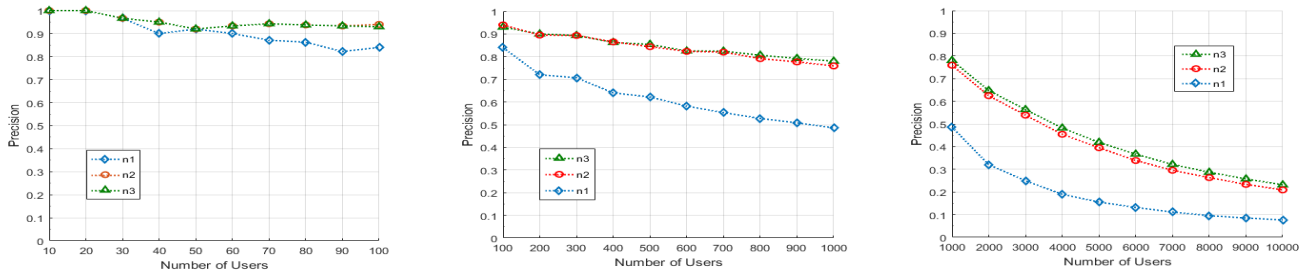
The number of rings in the MISFUB represents the extension level in the presented fingerprint; e.g., n2 denotes the inclusion of 2 rings, n3 denotes 3 rings, and so on. First, we use Jaccard and KL (proposed in [11]) similarity to investigate the effectiveness of the n-level extended fingerprint compared with the original, i.e., n1. To the best of our knowledge, the state-of-the-art work related to user matching by statistics is based on n1. We divide the number of users into three scales, i.e., a small scale [10 100], a medium scale [100 1000], and a large scale [1000 10000]. The experimental results in the case of the IPTV dataset are shown in Figure 3 and Figure 4, from which we can determine the following characteristics:

1. Using the presented extendable fingerprint, e.g., n2 and n3, can significantly improve the accuracy of matching based on Jaccard or KL over the original version (n1).

2. The overall trend of precision in user matching decreases with the number of users. That is, more users can lead to a lower precision, which is consistent with the findings in [11].

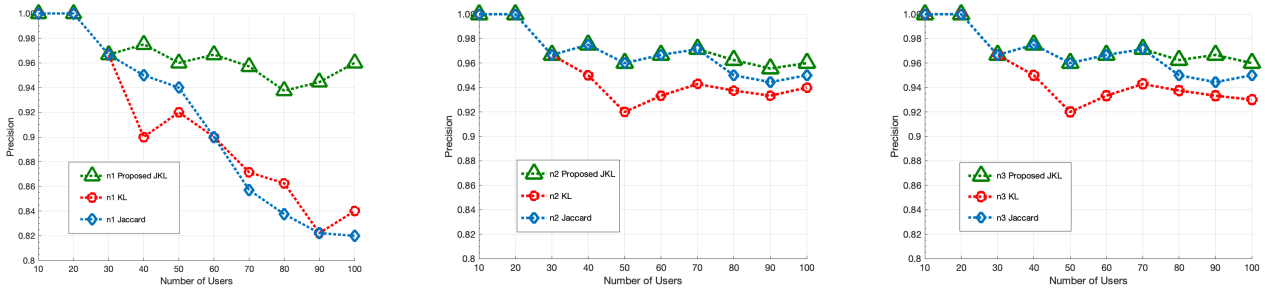
3. In the vast majority of cases, the greater the number of rings in the presented fingerprint, i.e., the larger n is, the higher the precision will be (see Figure 3c and Figure 4c). However, this is not absolute; particularly when the number of users is

[1]The original data come from a metropolitan TV station in southern China (the sensitive information has been deleted), the subset of which is available at [37], https://dx.doi.org/10.21227/H2396N; more detailed data titled "UCND: IPTV records of 10K users" are available on IEEE Dataport, https://dx.doi.org/10.21227/hp53-wz08.



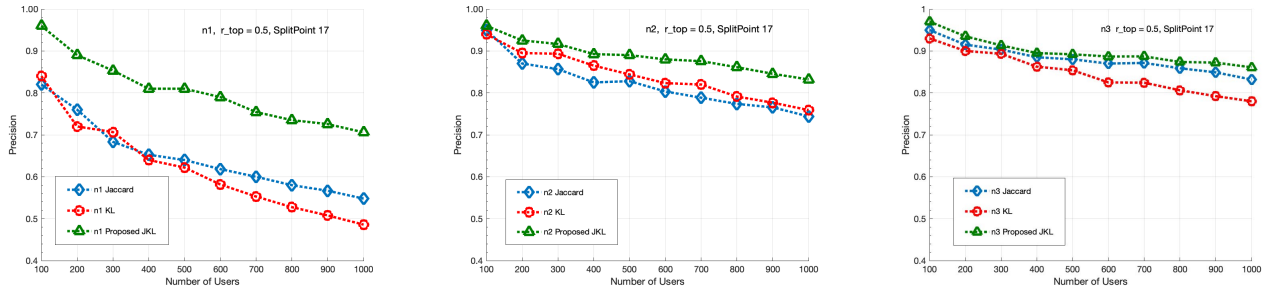
(a) KL (10, 100) (b) KL (100,1000) (c) KL (1000, 10000)

Fig. 4. Precision of combining KL (proposed in [11]) with our proposed **MISFUB** for n2 and n3, respectively, with a top ratio of 50% ($r_{top} = 0.5$) during testing days from 17 to 31, where the number of users is in the range [10, 100], [100, 1000], and [1000, 10000], respectively. Note that the highlighted blue curve n1 shows the KL approach for the IPTV dataset, and the red curve n2 and green curve n3 are based on the union of the KL approach proposed in [11] and our presented fingerprints.



(a) n1 (10, 100) (b) n2 (10, 100) (c) n3 (10, 100)

Fig. 5. Precision of n1, n2 and n3, the three curves in each subfigure represent Jaccard, KL and the proposed JKL respectively, with a top ratio of 50% during testing days from 17 to 31 and a number of users from 10 to 100.



(a) n1 (100, 1000) (b) n2 (100, 1000) (c) n3 (100, 1000)

Fig. 6. Precision of n1, n2 and n3, the three curves in each subfigure represent Jaccard, KL and the proposed JKL respectively, with a top ratio of 50% during testing days from 17 to 31 and a number of users from 100 to 1000.

small, the precision of n3 may be equal to that of n2 or even less than that of n2 (see Figure 3a and Figure 4a).

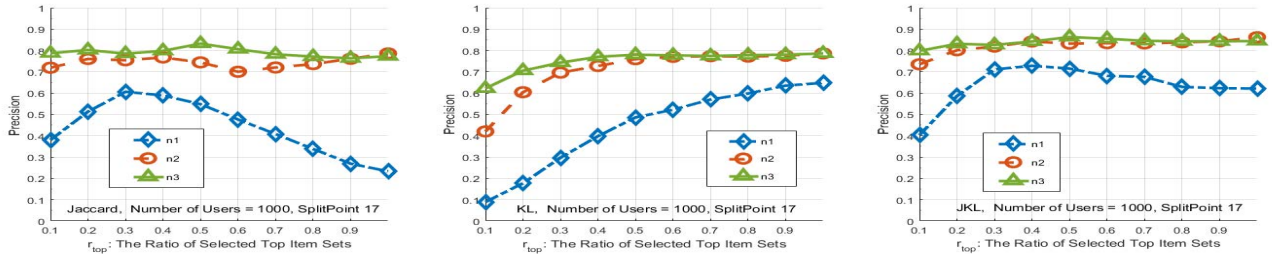
C. Proposed Similarity Distance Vs Jaccard and KL

The above experiment verified that the presented fingerprint via n-level, $n > 1$, extensions of frequent item sets can outperform the original version (n1) using existing similarity metrics. To demonstrate the performance of our proposed similarity distance, shown in equation (5), we compare the performance of our similarity distance with KL and Jaccard in Figure 5 and Figure 6. We yield the following findings:

1. Using the proposed similarity distance and the presented frequent item set fingerprint, the performance of UI can be improved over that obtained by independently using the KL and Jaccard distances, except in the case of a small number of users. In addition, using the proposed similarity distance, n3 and n2 outperform n1.

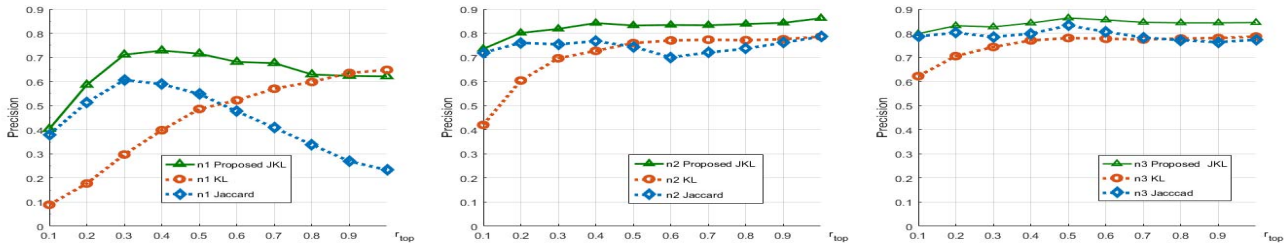
2. As the number of users increases, the performance of a higher n-level fingerprint is superior to that of a lower n-level fingerprint in most cases.

3. The approach using the KL distance can possibly underperform the Jaccard distance when it is combined with the presented n2 or n3 fingerprint, and it can outperform Jaccard, e.g., when it is combined with n1. However, we did not find that either of them outperforms the similarity distance proposed in our experiments. Although the performance obtained by using KL directly can slightly underperform that obtained by using Jaccard in the given IPTV dataset, the computing idea of KL and Jaccard inspired a new combination in our proposed similarity distance, which is significant from the viewpoint of knowledge fusion.



(a) Jaccard, 1000 users (b) KL, 1000 users (c) Proposed JKL, 1000 users

Fig. 7. The precision using Jaccard, KL and JKL, when r_{top} is varied from 0.1 to 1 during testing days from 17 to 31, where the number of users is 1000.



(a) n1, 1000 users (b) n2, 1000 users (c) n3, 1000 users

Fig. 8. The precision of n1, n2 and n3, when r_{top} is varied from 0.1 to 1 during testing days from 17 to 31, where the number of users is 1000.

D. Effect of the Top Ratio of Selecting Item Sets in $\mathscr{F}_n$

An open problem is which top ratio of the selected item sets versus all the item sets in the proposed fingerprint (r_{top}) is best. A similar problem to this was proven to be an NP-complete problem in [34]. In the experiments, unless stated otherwise, we let the metric of r_{top} be 0.5, which represents the first half of the feature item sets used for user matching. In this subsection, we gain insight into the effect of r_{top} via the experiment below for the range of $r_{top} \in [0.1, 1]$; we also show the test results of the 1000-user dataset in Figure 7 and Figure 8, where the popularity of each item set in the user fingerprint feature vector is ranked from hot to cold, and the legend of Figure 7 is the n-level fingerprint while that of Figure 8 is the similarity distance. From the experiment, we can determine how the evolution of the precision changes with r_{top} for each abovementioned approach.

1. The different values of r_{top} have different effects on the matching performance using a given similarity distance, and we find that $r_{top} = 0.5$ is a relatively fair and reasonable selection for the evaluation of system performance.

2. A too low value of r_{top} can lead to worse user matching performance, but a too great value may not be better; the r_{top} that is too small can lead to a lack of information fed into the feature matching process, while a too great r_{top} can bring interference due to information overload.

3. The experiments show that the proposed similarity distance (**JKL**) outperforms Jaccard and KL when using multi-item-set fingerprints, e.g., n3 and n2, and it is somewhat better than them in most cases even when using the original item version (i.e., n1).

E. Effect of the Size of Known Set and Unknown Set

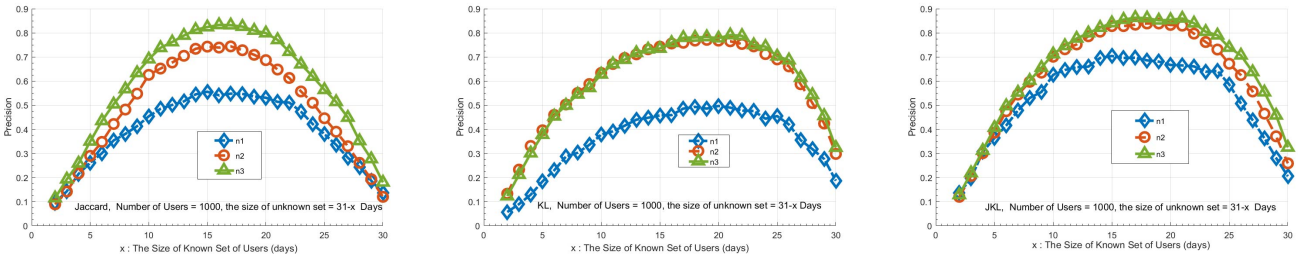
To simplify the description, we let S_{V_x} denote the size of unknown/testing set while S_{V_y} denote the size of known/training set. To maximally use our available data, we first partition our dataset so that $S_{V_x} + S_{V_y}$ equals to the total data size. Figure 9 shows how the precision varies when the splitting point of S_{V_x} and S_{V_y} slides. The x-axis denotes is the known set size, and (31-x) is then the unknown set size. We next conduct experiments by fixing the size of the known or the unknown dataset and varying the other one. More specifically, in Figure 10 (a), (b) and (c), S_{V_x} is fixed at five days while S_{V_y} increases from 1 to 26 days. In Figure 10 (d), (e) and (f), S_{V_y} is fixed at five days, and S_{V_x} increases from 1 to 26 days. Experiments show the following characteristics of IPTV user matching:

1. In the vast majority of cases, the performance is relatively optimal when S_{V_x} and S_{V_y} are similar; n3 and n2 respectively outperforms n1, and JKL outperforms Jaccard and KL.

2. When the size of the unknown set S_{V_x} is fixed, if the known set size S_{V_y} is less than or slightly larger than S_{V_x}, the matching performance increases with S_{V_y}. But if S_{V_y} is much larger than the fixed S_{V_x}, the matching performance saturates, and even starts to decline as S_{V_y} increases further, as shown in figure 10 (a) and (c). The decline trend might be due to that user behavior patterns shift over time and old user data might actually introduce noise to user matching.

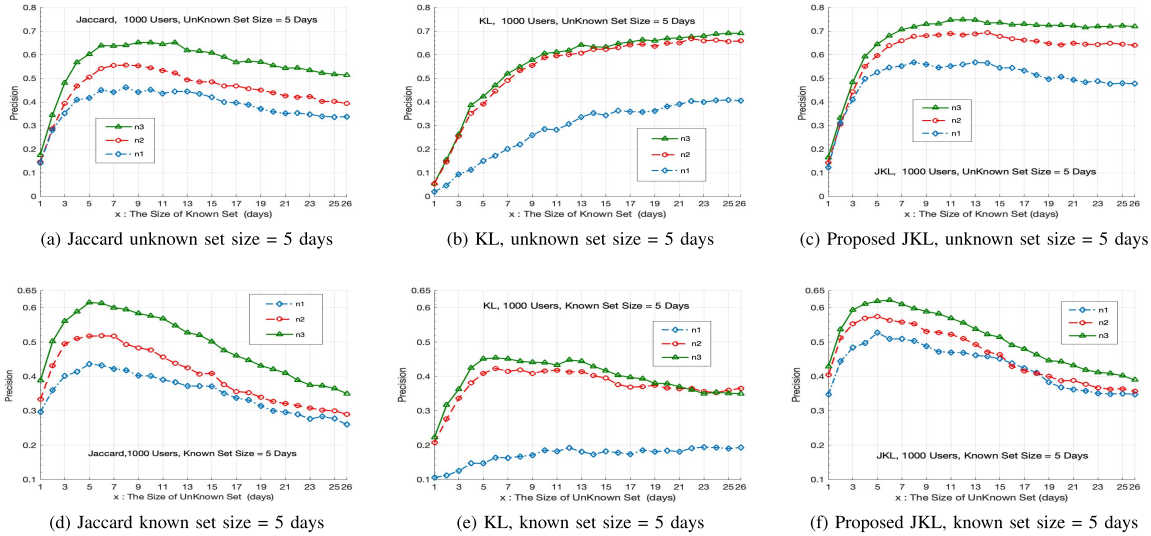
3. With a fixed known set size S_{V_y}, the matching performance increases with unknown set size S_{V_x} when S_{V_x} is less than or slightly larger than S_{V_y}. But as S_{V_x} increases further and become much larger than S_{V_y}, the performance will not improve and even decline when S_{V_x} is too large, as shown in figure 10 (d), (e), (f).

Our UI approach is based on matching statistical features of known and unknown users. The variances of those statistical features are determined by the sizes of the datasets. It is ideal that both the known and unknown datasets are large enough so that the obtained features have low variances for good



(a) Jaccard, 1000 users (b) KL, 1000 users (c) Proposed JKL, 1000 users

Fig. 9. The precision of Jaccard, KL, and JKL for a known set (x) from day 2 to day 30 and an unknown set (31-x) from day 29 to 1, where the number of users is 1000 (one curve for each level of n1, n2 and n3 respectively.).



(a) Jaccard unknown set size = 5 days (b) KL, unknown set size = 5 days (c) Proposed JKL, unknown set size = 5 days

(d) Jaccard known set size = 5 days (e) KL, known set size = 5 days (f) Proposed JKL, known set size = 5 days

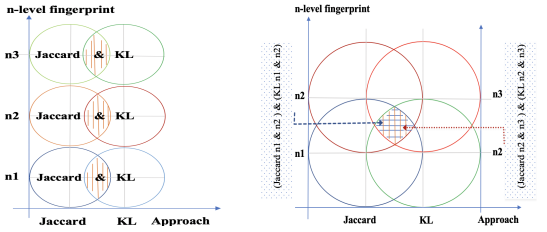
Fig. 10. The precision of n-level fingerprinting, respectively using Jaccard, KL and JKL, x means varying known set and fixed unknown set in (a), (b) and (c), while varying unknown set and fixed known set in (d), (e) and (f).

matching. From the experiments on IPTV datasets, too large difference between known set and unknown set in size usually degrade the performance. In practice, the timeliness of user matching is also critical. Our results and analysis above can be used to guide the decision on how long one should wait to collect behavior data of unknown users before our matching algorithms can be for high precision user matching. We will explore this timing issue in our future study.

F. Intersection-Based Approach With Rejection

The above approaches are based on one given similarity in performing user matching. Each matching will produce one candidate for a possible user target. In this subsection, we investigate two so-called intersection-based approaches to using joint voting with rejection. In this subsection, we use "A & B" to represent the intersection of A and B, e.g., "KL & Jaccard" means the intersect of the result of KL distance (Equation 4) and that of Jaccard (Equation 3).

1) Intersection of KL and Jaccard: As shown in Figure 11a, the idea of this approach is to recommend a candidate in the intersection of the candidates produced by each metric. If the intersection is empty, we do not provide a candidate; i.e., we reject matching the user. We show the precision and rejection rate of the fusion approach using the intersection of KL and Jaccard in Figure 12. For comparison to the approaches without rejection, we also illustrate the results using the similarity



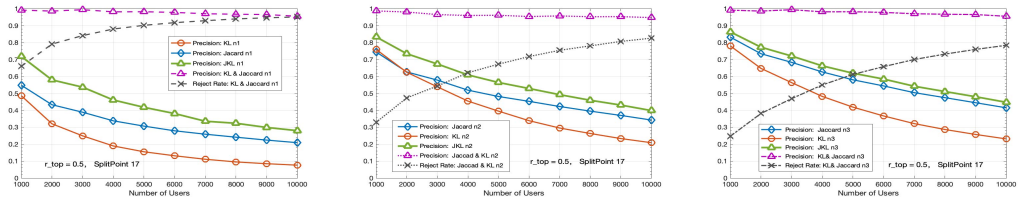
(a) Fusion of approaches at the same level of fingerprint to enhance the matching performance, where $k_1 = Jaccard{+}n*, k_2 = KL{+}n*$.

(b) Fusion of approaches at different n-levels of fingerprint to enhance the matching performance, where $k_1 = Jaccard + n1, k_2 = KL + n1, k_3 = Jaccard + n2, k_4 = KL + n2$.

Fig. 11. Fusion of Jaccard and KL approaches and n-level fingerprints to enhance the matching performance; $\hat{u}_{x_i} = \mathscr{D}_k(Approach(k)) = \mathscr{D}_k(\arg\max_j(w_{ij}^k))$.

distances of Jaccard, KL and JKL in Figure 12. We find good precision, an inevitable rejection rate and some fundamental characteristics of this intersection approach, as follows:

1. The intersection approach (**FI**) of KL and Jaccard can achieve much higher precision than any one independent approach with rejection, which is shown in Figure 12. The mean precision of n1, n2, and n3 using the **FI** of KL and Jaccard reaches 97.17%, where the number of users ranges over [1000, 10000]. From the experiment, the precision of **FI** is greatly superior to any individual approach, such as JKL, KL or Jaccard.



(a) n1 (b) n2 (c) n3

Fig. 12. Precision of n1, n2 and n3 for KL, Jaccard, and KL&Jaccard (computed by equation 7) and the rejection rate of KL&Jaccard (computed by equation 8), with a top ratio of 50% during testing days from 17 to 31.

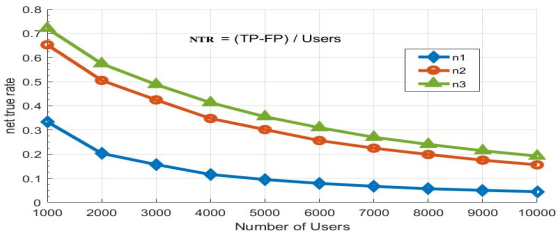


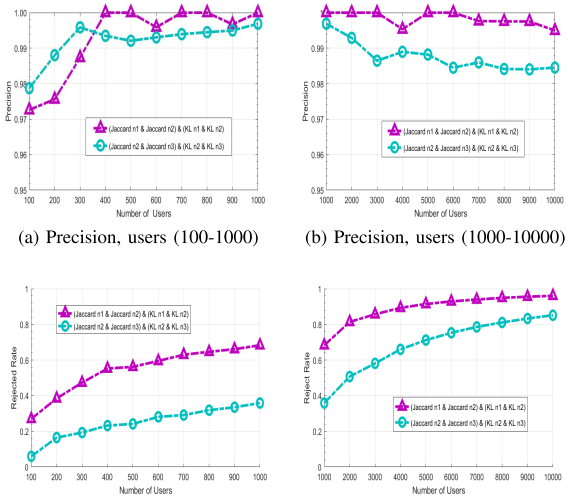
Fig. 13. The net true rate of n1, n2 and n3 for the intersection of Jaccard and KL, with a top ratio of 50% during testing days from 17 to 31, where the number of users is from 1000 to 10000, here, net true rate (NTR) is calculated from equation (10), where $|V_x|$ represents the number of users.

2. The precision of n1, n2, and n3 has an overall descending trend with an increasing user scale. However, their rejection rates also increase, as shown in Figure 12.

3. To comprehensively evaluate the overall performance by considering true positives, false positives, and rejections, we show the net true rates (NTRs) of the methods in Figure 13, where the NTR is computed by equation 10. As shown in Figure 13, the overall performance increases with the number of extended levels in the fingerprint. The overall performance can still be ranked as $n3 > n2 > n1$.

*2) Multi-Intersection of the Similarity Distance and Number of Rings in the **MISFUB**:* To further improve precision, as shown in Figure 11, we present the multi-intersection approach, which is to recommend a candidate from the intersection of approaches using distinct similarity distances and multiple levels of fingerprints. In short, the multi-intersection approach is an iteration of the approach proposed in Section IV-F.1; we will also reject performing matching if the multi-intersection is empty. We show the precisions and rejection rates of the two proposed multi-intersection approaches based on the intersection of similarity distances and n-level rings. One is based on the intersection of $(Jaccard_{n1}$ & $Jaccard_{n2})$ & $(KL_{n1}$ & $KL_{n2})$, which is called **FI12**, and the other is based on $(Jaccard_{n2}$ & $Jaccard_{n3})$ & $(KL_{n2}$ & $KL_{n3})$, which is called **FI23**. From Figure 14, we find that the mean precision of the two approaches is more than 98%, and the first multi-intersection approach, of $(Jaccard_{n1}$ & $Jaccard_{n2})$ & $(KL_{n1}$ & $KL_{n2})$, can reach 100% in many subsets of user datasets. Based on Figure 14, the detailed precision and reject rate are shown in Table II.

In most cases, we find that the proposed fusion approaches based on intersections have a higher precision and lower rejection rate for a small number of users than for large ones. In addition, the reject rate of the intersection between n2 and n3 is much lower than that of the intersection between n1 and n2. This is interesting in user matching research, which may make deterministic identification possible at a certain rejection rate. Of course, higher precision could lead to higher rejection rate in the fusion methods, which could provide a way to control the precision and rejection by the demand in actual application scenarios. How to make the tradeoff between



(a) Precision, users (100-1000) (b) Precision, users (1000-10000)

(c) Rejection rate, users (100-1000) (d) Rejection rate, users (1000-10000)

Fig. 14. The precision and rejection of n1&n2 and n1&n2&3 with the intersection of Jaccard and KL, with a top ratio of 50% during testing days from 17 to 31 and a number of users from 1000 to 10000.

TABLE II
THE RESULT OF MULTI-INTERSECTION APPROACHES

Approach	User Data	M.R.R.	Precision		
			Mean	Median	Mode[2]
FI12	DS1000	0.5458	0.9928	0.9984	**1**
FI12	DS10000	0.8894	0.9983	0.9988	**1**
FI23	DS1000	0.2479	0.9922	0.9937	0.9787
FI23	DS10000	0.6854	0.9877	0.9862	0.984

DS1000: the number of users is in the range of [100:100:1000]
DS10000: the number of users is in the range of [1000:1000:1000]
M.R.R. : mean of reject rate

[2]The number that appears most frequently in a set of data.

TABLE III

COMPUTATIONAL COMPLEXITY

Symbol	Meaning	1 user	N users
t_f	to generate the histogram of the fingerprint	$O(N)$	$O(2N)$
t_w	to compute an edge weight matrix	$O(N)$	$O(N^2)$
T_{om}	to identify a user using one-by-one matching	$O(N)$	$O(N^2)$
T_{gm}	to identify a user using the graph matching	$O(N^3)$	$O(N^3)$

the precision and rejection rate can be an interesting topic for future work.

G. Comparison of One-by-One and Graph Matching

The matching decision in all the above experiments is based on one-by-one matching according to Equation (1). It is possible that two users are matched to the same user. At the same time, we perform graph matching according to Equation (2) by using the maximum-weight bipartite matching (bipartite_matching), referring to [27], and we find that our proposed ideas still can improve graph matching accuracy.

1) Matching Precision: Similar to previous work [11], we use bipartite graph matching to verify the validation of our innovation in n-level fingerprinting and of the JKL similarity metric. As shown in Figure 15, graph matching achieves a positive gain over one-by-one matching in terms of precision. From the experimental results, we find that the gain of using the baseline fingerprint (n1) is the most significant in the majority of cases, which is mainly because n1 has the lowest precision when using one-by-one matching. In other words, the denominator of the gain of n1 is relatively small compared to those of n2 and n3. Similarly, the experiment shows that the average gain of using JKL is smaller than that of using KL and Jaccard, mainly because JKL outperforms Jaccard and KL in most cases of one-by-one matching. One-by-one matching can save more computing time than graph matching. Figure 15 shows the performance improvement of using graph and one-by-one matching compared with the approach proposed in [11].

2) Computing Time Complexity: While graph matching outperforms one-by-one matching, it incurs higher computation time than one-by-one matching. The time-complexity of naively implemented graph matching is $O(N!)$, and can become $O(N^3)$ when we use Hungarian algorithm discussed in [11]. Low computation time is a critical requirement in the majority of application scenarios, such as suspect tracing and real-time recommendation systems. For detailed analysis, we define the symbols and give the computational complexity as follows,

We therefore have $T^1_{om} = O(N)$, $T^N_{om} = O(N^2)$, and $T^1_{gm} = T^N_{gm} = O(N^3)$. Based on the above analysis, we know that the time consumption of graph matching (T^N_{gm}) is more than that of one-by-one matching (T^N_{om}) and increases with N. For the individual matching requirement, the time cost of graph matching for 1 target user (T^1_{gm}) is much higher than that of one-by-one matching (T^1_{om}). In the case of the IPTV dataset for 100 users to 1000 users, we first visualize the main components of time consumption in Figure 16a and then show the time cost increment of graph matching compared with one-by-one matching in Figure 16b and finally illustrate the ratio of the precision gain to the time cost increment in Figure 16c. The detailed description is shown in the caption of the corresponding figure. In short, the performance price ratio of graph matching to one-by-one matching decays with the number of users (N).

TABLE IV

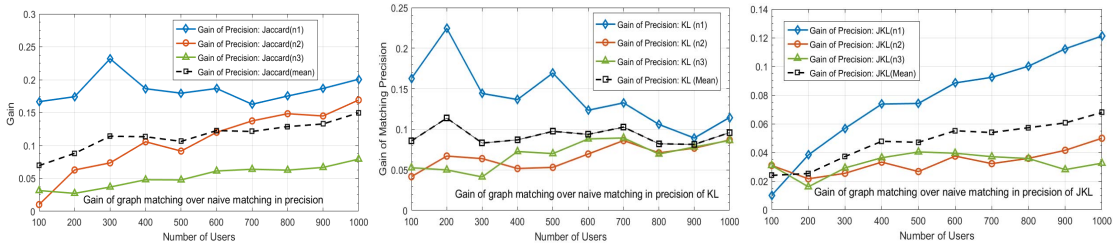
AVERAGE PRECISION OF OUR METHOD (PRO_*) AND THE STATE-OF-THE-ART APPROACH, WHERE THE NUMBER OF USERS (NOTED AS n) RANGES OVER [100 1000]

Approach No.		Method	C. C.	Sim. Dist.	n-level	Avrg. Precision	
						$r_{top}1$	$r_{top}0.5$
Ours	Pro_1	Graph	$O(N^3)$	JKL*	n2^	0.938	0.925
	Pro_2	Graph	$O(N^3)$	KL	n2^	0.920	0.907
	Pro_3	Naive	$O(N^2)$	JKL*	n2^	0.907	0.896
	Pro_4	Graph	$O(N^3)$	JKL*	n1	0.868	0.892
	Pro_5	Naive	$O(N^2)$	KL	n2^	0.856	0.853
	Pro_6	Naive	$O(N^2)$	JKL*	n1	0.771	0.830
State-of-the-art		Graph	$O(N^3)$	KL	n1	0.828	0.702
Baseline		Naive	$O(N^2)$	KL	n1	0.739	0.618

Here, C. C. denotes "Computing Complexity", * denotes our similarity distance, ^ denotes our fingerprint generated by SURE in **MISFUB**.

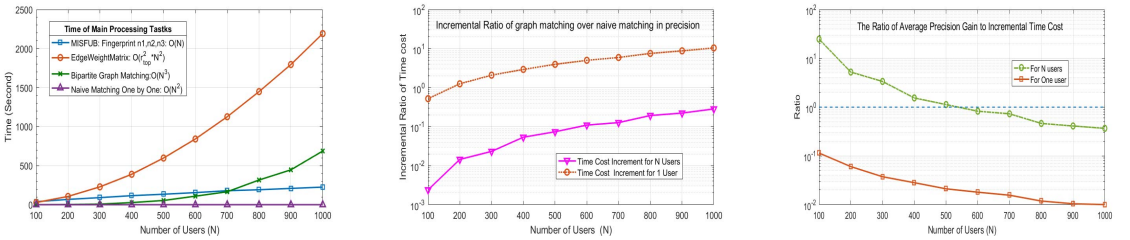
3) Comparison With the State-of-the-Art Approach: The innovation of the **MISFUB** is to extend the feature from the original version of a single item (i.e., n1) to a multi-item-set (e.g., n2, n3). Another contribution is the proposed similarity distance, i.e., JKL $= \frac{Jaccard}{KL}$ (see Equation 5). Using the proposed approach, we can achieve a precision of 100% in a small dataset, where the number of users increases from 10 to 100 at step-size of 10. At the same time, the proposed methods outperform the state-of-the-art methods in a relatively large dataset, where the number of users increases from 100 to 1000 at step-size of 100. For a more thorough comparison, in addition to the KL and Jaccard distances, we also compare the proposed **JKL** with 3 traditional similarity metrics—$Cosine$, $1/L1_{norm}$, and $1/Euclidean$—and another derived distance, $\frac{Jaccard}{Euclidean}$, similar to **JKL** and the idea of [21]. Here, let **LF** be $|V_{x_i} \cap V_{y_j}|$; we have Cosine $= 1 - \frac{V_{x_i} \cdot V_{y_j}}{\sqrt{|V_{x_i}|^2 |V_{y_j}|^2}}$, $L1_{norm} = \sum_{l=1}^{\mathbf{LF}} |V^l_{x_i} - V^l_{y_j}|$, and $Euclidean^2 = \sum_{l=1}^{\mathbf{LF}} (V^l_{x_i} - V^l_{y_j})^2$, and $Jaccard$ refers to Equation 3. The result using bipartite graph matching is shown in Figure 17. As shown in Figure 17, the gain of the matching performance over that of the state-of-the-art approach increases with the number of users. The comparison between the state-of-the-art approach and ours in terms of average precision is shown in Table IV.

In summary, a large number of user matching experiments using a real IPTV viewer behavior dataset show the advantages of n-level extendable fingerprinting, the JKL similarity distance, and the fusion decision of multi-approach joint voting. Experiments show that the proposed fingerprint and similarity distance can outperform the state-of-the-art approach (KL+n1). Moreover, the proposed multi-approach fusion decision approach can achieve very high, even up to 100%, precision at the price of a certain rejection rate.



(a) Gain of Jaccard (b) Gain of KL (c) Gain of JKL

Fig. 15. The gain in precision using GM compared with NM; here, Gain = (Precision(GM)-Precision(NM)/Precision(NM)).

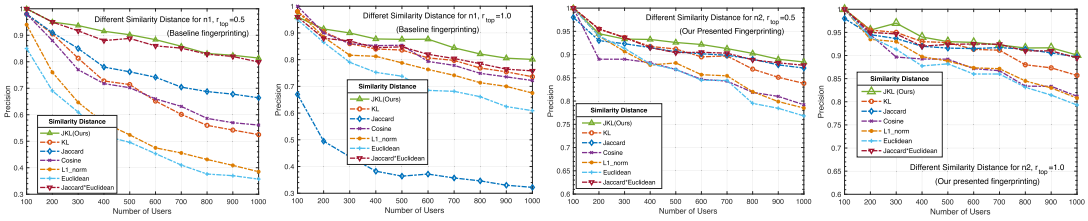


(a) Computation Time of Graph Matching vs. One-by-one Matching

(b) Computation Time Ratio between Graph Matching and One-by-one Matching, where y is shown in logarithmic form; $TCI^1 = (T^1_{gm} - T^1_{om})/T^1_{om}$, and $TCI^N = (T^N_{gm} - T^N_{om})/T^N_{om}$

(c) Ratio of the precision gain to computation time increase, where the y coordinate is shown in logarithmic form; $Ratio^1 = PreGain/TCI^1$, and $Ratio^N = PreGain/TCI^N$

Fig. 16. Computation Time Comparison of Graph and One-by-one Matching.



(a) n1 ($r_{top} = 0.5$) (b) n1 ($r_{top} = 1.0$) (c) Presented n2 ($r_{top} = 0.5$) (d) Presented n2 ($r_{top} = 1.0$)

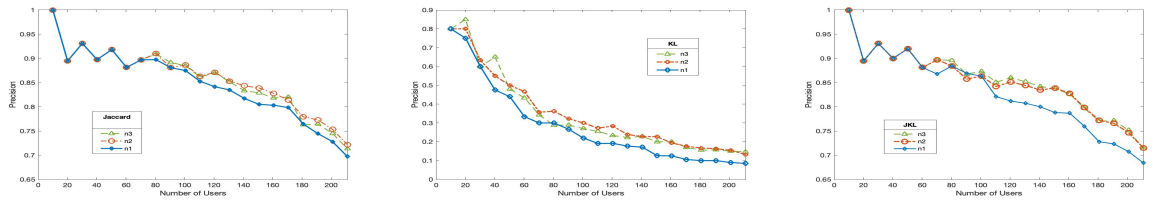
Fig. 17. Comparison of bipartite graph matching with our proposed similarity distance of JKL and the other distances.

Additionally, we verified that the proposed fingerprint and similarity distance can be used for bipartite graph matching in the same way. We further compared the performance and computational complexity of bipartite graph matching and naive one-by-one matching for UI and found that the former can outperform the latter in matching precision at the price of additional computing consumption. In the end, we recommend using one-by-one matching for time-sensitive applications to identify one or many users from a large-scale user dataset due to the complexity of $O(N^3)$, whereas we can use graph matching for the high precision on a relatively small user scale.

V. Evaluation on Web Browsing and Online Shopping Datasets

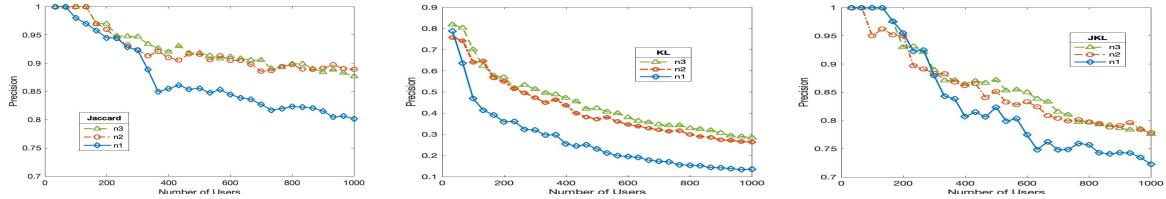
To verify whether the conclusions from the IPTV experiments are generalizable, we further study the effectiveness of the proposed multi-item-set fingerprint and similarity distance on two additional datasets. One is the user web browsing history (WBH) dataset [38], [39] through the Firefox browser, which includes the accessed website records of 211 users during 31 consecutive days from January 2007 to July 2014. There are 32,851 websites in total, ranked by their accessed frequency in descent order. This dataset is similar to that used in [11], but different in respect of the selected user IDs, which are not disclosed in [11]. Here, we can regard a website as an item, similar to a channel in the IPTV datasets. The other is an online shopping record dataset [40] provided by the Application Technology Department of Business Division in Alibaba Group. It contains the user shopping records from November 18, 2014 to December 18, 2014. The dataset consists of records of the goods that each user browses and the goods category. In the experiment, we adopted the top 1,000 users ranked by the number of browsed records. There are 5,937 categories of goods. We treat each category as an item.

In order to verify the impact of sequence length on matching effect, we adopted the graph matching to conduct experiments on the two data sets. The results on the length of itemset are shown in Figure 18 and Figure 19, respectively. The result on similarity distance is shown in Figure 20 and Figure 21, respectively. From the experimental results, the effect of item sequence length on the matching performance is consistent with the conclusion drawn on the IPTV dataset, namely, the extended itemsets (e.g. n2 and n3) outperform the single itemset (n1). Moreover, the impact of similarity distance on the matching performance is almost consistent with the IPTV conclusion, i.e., the JKL similarity metric proposed by us is



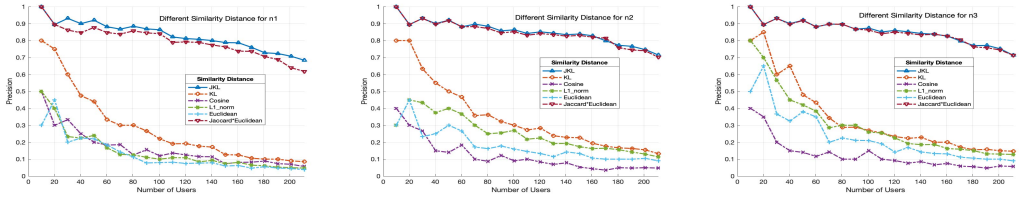
(a) Jaccard (b) KL (c) JKL

Fig. 18. Website Dataset: Precision of n1, n2 and n3, for Jaccard, KL and JKL, $r_{top} = 50\%$, testing days from 17 to 31.



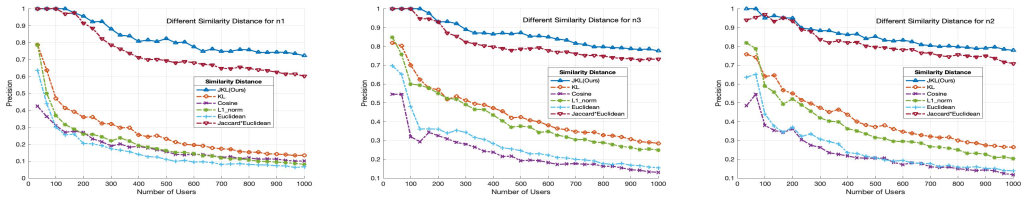
(a) Jaccard (b) KL (c) JKL

Fig. 19. Shopping Dataset: Precision of n1, n2 and n3, for Jaccard, KL and JKL, $r_{top} = 50\%$, testing days from 17 to 31.



(a) n1 (b) n2 (c) n3

Fig. 20. Website Dataset: Precision of n1, n2 and n3 using multiple similarities, $r_{top} = 50\%$, testing days from 17 to 31.



(a) n1 (b) n2 (c) n3

Fig. 21. Shopping Dataset: Precision of n1, n2 and n3 using multiple similarities, $r_{top} = 50\%$, testing days from 17 to 31.

superior to most other similarity metrics. However, unlike the IPTV dataset, JKL does not certainly outperform Jaccard in the shopping dataset all the time, in Figure 19. There are two potential reasons: 1) the catalogs of goods in the shopping dataset are much larger than the TV channels in the IPTV dataset; 2) goods browsing of a user in the shopping dataset is much less than channel switching of a viewer in the IPTV dataset. This might reduce the performance impact of the KL component in JKL metric. Comparing Figure 17, Figure 18 and Figure 19, we can find the precisions for the IPTV and Shopping datasets are better than that of web browsing. This is mostly because the web browsing dataset has less user behavior records than the other two. A large number of user records are needed for accurate user identification.

VI. Conclusion

In this work, we studied how to recognize the real identities of users by analyzing the sequences of their accessed items in cyberspace, and conducted a case study of IPTV user matching. We first presented a user behavior fingerprint matching framework of n-level multi-item-sets for UI. Then, we proposed a new similarity distance by combining two general similarity distances. Finally, we proposed a fusion decision scheme, by which we further improved the precision at the price of extra rejections.

We conducted extensive experiments using a large-scale IPTV dataset. We demonstrated that the presented approach outperforms the state-of-the-art in behavior fingerprints [11] and similarity metrics [19], [21], respectively. In addition, we further verified the main conclusions drawn from the IPTV dataset on other two datasets. In summary, the proposed extendable multi-item-set fingerprint construction algorithm (SURE) outperformed the methods only using single items, but the matching performance was not surely improved when the extension level goes beyond two. The proposed JKL similarity distance is usually superior to the others in the majorities of studied scenarios. The intersection-based joint-voting approach can improve the precision, but increase

the rejection rate. Graph matching outperforms one-by-one matching in terms of precision, but consumes higher computation time.

This multi-item-set behavior fingerprint framework is of significance in information forensics and targeted recommendation for an identified user. It can be used to analyze anonymous and/or camouflaged user behaviors by using sparse information to generate extended features for pattern recognition. Meanwhile, it raises a new challenge for privacy preservation in cyberspace. Based on the proposed framework of extendable multi-item-set fingerprinting, there are open future research problems in targeted recommendation, behavioral pattern recognition, privacy preservation of online users, and so on.

ACKNOWLEDGMENT

The authors would like to thank the editors, the reviewers for their comments and advices for improving this paper, Miss Yingying Zhu for the data processing and testing, the vendor of IPTV for the provision of the experimental dataset, David Gleich for the shared code for bipartite graph matching, and English editing service provided by https://www.aje.com.

REFERENCES

[1] N. Gerhart and M. Koohikamali, "Social network migration and anonymity expectations: What anonymous social network apps offer," *Comput. Hum. Behav.*, vol. 95, pp. 101–113, Jun. 2019.

[2] K. Deng, L. Xing, L. Zheng, H. Wu, P. Xie, and F. Gao, "A user identification algorithm based on user behavior analysis in social networks," *IEEE Access*, vol. 7, pp. 47114–47123, 2019.

[3] J. Lv, W. Chen, Q. Li, and C. Yang, "Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7948–7956.

[4] J. Lv, H. Lin, C. Yang, Z. Yu, Y. Chen, and M. Deng, "Identify and trace criminal suspects in the crowd aided by fast trajectories retrieval," in *Database Systems for Advanced Applications, Part II* (Lecture Notes in Computer Science (LNCS)), vol. 8422, S. S. Bhowmick, C. E. Dyreson, C. S. Jensen, M. L. Lee, A. Muliantara, and B. Thalheim, Eds. Cham, Switzerland: Springer, 2014, pp. 16–30.

[5] C. Yang, S. Ren, Y. Liu, H. Cao, Q. Yuan, and G. Han, "Personalized channel recommendation deep learning from a switch sequence," *IEEE Access*, vol. 6, pp. 50824–50838, 2018.

[6] C. Yu, H. Ding, H. Cao, Y. Liu, and C. Yang, "Follow me: Personalized iptv channel switching guide," in *Proc. 8th ACM Multimedia Syst. Conf.*, Taipei, Taiwan, 2017, pp. 147–157.

[7] X. Han, L. Wang, C. Cui, J. Ma, and S. Zhang, "Linking multiple online identities in criminal investigations: A spectral co-clustering framework," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 9, pp. 2242–2255, Sep. 2017.

[8] Z. Yu, E. Xu, H. Du, B. Guo, and L. Yao, "Inferring user profile attributes from multidimensional mobile phone sensory data," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 5152–5162, Jun. 2019.

[9] W. S. Lin, H. V. Zhao, and K. J. R. Liu, "Behavior forensics with side information for multimedia fingerprinting social networks," *IEEE Trans. Inf. Forensics Security*, vol. 4, no. 4, pp. 911–927, Dec. 2009.

[10] J. Houvardas and E. Stamatatos, "N-gram feature selection for authorship identification," in *Artificial Intelligence: Methodology, Systems, and Applications* (Lecture Notes in Computer Science), vol. 4183, J. Euzenat and J. Domingue, Eds. Berlin, Germany: Springer, 2006, pp. 77–86.

[11] F. M. Naini, J. Unnikrishnan, P. Thiran, and M. Vetterli, "Where you are is who you are: User identification by matching statistics," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 2, pp. 358–372, Feb. 2016, doi: 10.1109/TIFS.2015.2498131.

[12] R. Zafarani, L. Tang, and H. Liu, "User identification across social media," *ACM Trans. Knowl. Discovery Data*, vol. 10, no. 2, pp. 1–30, Oct. 2015.

[13] S. Gambs, M.-O. Killijian, and M. N. del Prado Cortez, "De-anonymization attack on geolocated data," *J. Comput. Syst. Sci.*, vol. 80, no. 8, pp. 1597–1614, Dec. 2014.

[14] A. Brown and M. Abramson, "Twitter fingerprints as active authenticators," in *Proc. IEEE Int. Conf. Data Mining Workshop (ICDMW)*, P. Cui, J. Dy, C. Aggarwal, Z. H. Zhou, A. Tuzhilin, H. Xiong, and X. Wu, Eds. Atlantic City, NJ, USA, 2015, pp. 58–63, doi: 10.1109/ICDMW.2015.223.

[15] S. Mondal and P. Bours, "Combining keystroke and mouse dynamics for continuous user authentication and identification," in *Proc. IEEE Int. Conf. Identity, Secur. Behav. Anal. (ISBA)*, Feb. 2016, pp. 1–8.

[16] R. Soltani, D. Goeckel, D. Towsley, and A. Houmansadr, "Fundamental limits of invisible flow fingerprinting," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 345–360, 2020.

[17] H. Li, G. Ye, X. Liu, F. Zhao, D. Wu, and X. Lin, "URLSight: Profiling mobile users via large-scale Internet metadata analytics," in *Proc. IEEE Trustcom BigDataSE ISPA*, Aug. 2016, pp. 1728–1733.

[18] Y. Zhou, J. Wu, T. H. Chan, S.-W. Ho, D.-M. Chiu, and D. Wu, "Interpreting video recommendation mechanisms by mining view count traces," *IEEE Trans. Multimedia*, vol. 20, no. 8, pp. 2153–2165, Aug. 2018.

[19] J. Unnikrishnan, "Asymptotically optimal matching of multiple sequences to source distributions and training sequences," *IEEE Trans. Inf. Theory*, vol. 61, no. 1, pp. 452–468, Jan. 2015.

[20] H. Yan and Y. Tang, "Collaborative filtering based on Gaussian mixture model and improved Jaccard similarity," *IEEE Access*, vol. 7, pp. 118690–118701, 2019.

[21] S. Bag, S. K. Kumar, and M. K. Tiwari, "An efficient recommendation generation using relevant Jaccard similarity," *Inf. Sci.*, vol. 483, pp. 53–64, May 2019.

[22] M. R. Hamedani and S.-W. Kim, "JacSim: An accurate and efficient link-based similarity measure in graphs," *Inf. Sci.*, vol. 414, pp. 203–224, Nov. 2017.

[23] Z. Zhang, Q. Gu, T. Yue, and S. Su, "Identifying the same person across two similar social networks in a unified way: Globally and locally," *Inf. Sci.*, vols. 394–395, pp. 53–67, Jul. 2017.

[24] W. Hu, T. Chen, and S. L. Shah, "Detection of frequent alarm patterns in industrial alarm floods using itemset mining methods," *IEEE Trans. Ind. Electron.*, vol. 65, no. 9, pp. 7290–7300, Sep. 2018.

[25] C.-H. Chee, J. Jaafar, I. A. Aziz, M. H. Hasan, and W. Yeoh, "Algorithms for frequent itemset mining: A literature review," *Artif. Intell. Rev.*, vol. 52, no. 4, pp. 2603–2621, Dec. 2019.

[26] Y. Li, Z. Zhang, Y. Peng, H. Yin, and Q. Xu, "Matching user accounts based on user generated content across social networks," *Future Gener. Comput. Syst.*, vol. 83, pp. 104–115, Jun. 2018.

[27] *Bipartite Matching Code*. Accessed: Mar. 29, 2020. [Online]. Available: http://github.com/dgleich/gaimc/tree/master

[28] G. Li, L. Qiu, C. Yu, H. Cao, Y. Liu, and C. Yang, "IPTV channel zapping recommendation with attention mechanism," *IEEE Trans. Multimedia*, early access, Apr. 2, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9055049/, doi: 10.1109/TMM.2020.2984094.

[29] C. Yang and Y. Liu, "On achieving short channel switching delay and playback lag in IP-based TV systems," *IEEE Trans. Multimedia*, vol. 17, no. 7, pp. 1096–1106, Jul. 2015.

[30] Y. Liu, Y. Liu, Y. Shen, and K. Li, "Recommendation in a changing world: Exploiting temporal dynamics in ratings and reviews," *ACM Trans. Web*, vol. 12, no. 1, pp. 1–20, Feb. 2018.

[31] B. Nie, H. Zhang, and Y. Liu, "Social interaction based video recommendation: Recommending YouTube videos to facebook users," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Apr. 2014, pp. 97–102.

[32] X. Hei, C. Liang, J. Liang, Y. Liu, and K. W. Ross, "A measurement study of a large-scale P2P IPTV system," *IEEE Trans. Multimedia*, vol. 9, no. 8, pp. 1672–1687, Dec. 2007.

[33] X. X. Chen, C. Yang, and Y. W. Chen, "Banacast: A fast adaptive application layer multicast network for P2P live streaming," in *Proc. 1st Int. Conf. Inf. Sci. Eng.*, 2009, pp. 149–153.

[34] S. Davies and S. Russell, "NP-completeness of searches for smallest possible feature sets," *IEEE AAAI Fall Symp. Relevance*, Nov. 1994, pp. 37–39. [Online]. Available: https://www.aaai.org/Papers/Symposia/Fall/1994/FS-94-02/FS94-02-011.pdf

[35] J. Unnikrishnan and F. M. Naini, "De-anonymizing private data by matching statistics," in *Proc. 51st Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Oct. 2013, pp. 1616–1623.

[36] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: Wiley, 2006,

[37] S. Ren and C. Yang, "Channels switch sequences of 300 IPTV viewers in a month," *IEEE Dataport*. Accessed: Mar. 29, 2020. [Online]. Available: https://ieee-dataport.org/documents/channels-switchsequences-300-iptv-viewers-month-0, doi: 10.21227/H2396N.
[38] *Web History Repository*. Accessed: Nov. 11, 2020. [Online]. Available: http://webhistoryrepository.l3s.uni-hannover.de/download.php
[39] R. Kawase, G. Papadakis, E. Herder, and W. Nejdl, "Beyond the usual suspects: Context-aware revisitation support," in *Proc. 22nd ACM Conf. Hypertext Hypermedia (HT)*, P. D. Bra and K. Grønbæk, Eds., Eindhoven, The Netherlands. New York, NY, USA: ACM, Jun. 2011, pp. 27–36, doi: 10.1145/1995966.1995974.
[40] *Tianchi*. Accessed: Nov. 29, 2020. [Online]. Available: https://tianchi.aliyun.com/competition/entrance/231522/information

Can Yang (Member, IEEE) received the master's degree in pattern recognition and intelligent control from the Image Institute of HUST in June 1997, and the Ph.D. and bachelor's degree from the Department of Electronic Science and Technology, Huazhong University of Science and Technology (HUST), Wuhan, China, in 2002 and 1994, respectively. He held a post-doctoral position with the School of Computer Science and Technology, HUST, from August 2002 to October 2004. He was a Visiting Scholar at the Electrical and Computer Engineering Department, New York University, from 2013 to 2014. He is currently with the School of Computer Science and Engineering, South China University of Technology (SCUT), where he became an Associate Professor in February 2005 and has been a Professor in Advance since May 2012. His current research interests include artificial intelligence, big data, and multimedia networking in the cloud. He is a member of the ACM, a Senior Member of the CCF, and a member of the Council of Guangdong Computer Academy (GDCA). He won the Science and Technology Progressive Award of Guangdong Province in 2011 and 2015 and the Sci. and Tech. Innovation Award of SARFT of China in 2012.

Lan Wang received the B.S. degree from the School of Computer Science and Technology, Guangxi University, Nanning, China, in June 2017. She is currently pursuing the master's degree with the School of Computer Science and Engineering, South China University of Technology. Her current interests include multimedia behavior pattern recognition and deep learning applications.

Houwei Cao (Member, IEEE) received the Ph.D. degree in electronic engineering from the Chinese University of Hong Kong, in 2011. She was an Adjunct Professor at the Computer Science and Engineering Department, Tandon School of Engineering of New York University, before joining NYIT. She was a Post-Doctoral Fellow at the University of Pennsylvania from 2011 to 2014, and at Tufts University from 2014 to 2015. She was also an Insight Data Science Fellow in 2015. She is currently an Assistant Professor with the Department of Computer Science, University of New York Institute of Technology (NYIT). Her main areas of research are signal processing, machine learning, data mining and their applications in human-centric data analytics, with an emphasis on developing computational methods for speech recognition, text analytics, affect detection, and healthcare analysis. She won the audio-visual emotion recognition challenge (AVEC) in 2012. She is a member of the International Speech Communication Association (ISCA) and the Association for the Advancement of Affective Computing (AAAC). She has served as a Reviewer for numerous conferences and journals, including Interspeech, ISCSLP, O-COCOSDA, the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, and the *IET Transactions on Computer Vision and Speech Communications*.

Qihu Yuan received the B.S. degree from Lanzhou University, Lanzhou, China, in June 2015, and the master's degree from the College of Computer Science and Engineering, South China University of Technology, Guangzhou, China, in June 2019. He is currently an Algorithm Engineer in NetEase Computer System Company, Ltd. His current interests include multimedia recommendation systems and deep learning applications.

Yong Liu (Fellow, IEEE) received the Ph.D. degree from the Electrical and Computer Engineering Department, University of Massachusetts, Amherst, MA, USA, in May 2002. He is currently a Professor at the Electrical and Computer Engineering Department, Tandon School of Engineering of New York University. His current research directions include next-generation networks and applications, overlay networks, network measurement, online social networks, and recommender systems. He is a member of the ACM. He is the winner of the Best Paper Award of the ACM/USENIX Internet Measurement Conference (IMC) 2012, the National Science Foundation Career Award in 2010, the Best Paper Award of the IEEE Conference on Computer Communications (INFOCOM) in 2009, and the IEEE Communication Society Multimedia Communications Best Paper Award in 2008. He has served as an Associate Editor for the IEEE/ACM TRANSACTIONS ON NETWORKING and *Elsevier Computer Networks Journal*.