IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 20, 2018, accepted January 8, 2019, date of publication January 23, 2019, date of current version February 20, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2894267

Coupled Tensor Decomposition for User Clustering in Mobile Internet Traffic Interaction Pattern

KE YU[1], LIFANG HE[2], (Member, IEEE), PHILIP S. YU[3], (Fellow, IEEE), WENKAI ZHANG[1], AND YUE LIU[1]

[1]School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA 19104, USA
[3]Department of Computer Science, University of Illinois at Chicago, Chicago, IL 60607, USA

Corresponding author: Ke Yu (yuke@bupt.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61601046, Grant 61503253, and Grant 61171098, in part by the 111 Project of China under Grant B08004, and in part by the EU FP7 IRSES Mobile Cloud Project under Grant 612212.

ABSTRACT The rapid developments in mobile Internet are reshaping our lives and activities. Understanding the user behaviors and dynamics in such a large-scale network is essential for better system design, service provisioning, and network management. In this paper, we focus on the interaction pattern between mobile users and servers based on the traffic flow data. Real traffic flow data is collected from the public network of ISPs by high-performance network traffic monitors. Traffic flow-based heterogeneous information network (TF-HIN) is introduced to represent the traffic interaction pattern, and node correlation characteristics are mined from TF-HIN. Based on the empirical analysis of traffic interaction pattern, we propose the coupled flow tensor to represent the relations among the user, server and time, by incorporating correlations of user and server as auxiliary information. Two iterative algorithms, i.e., FTD and FTD-NFS, are proposed for coupled flow tensor decomposition and the latent factors are used for user clustering. We evaluate the proposed user clustering algorithms by using benchmark datasets and also analyze the user clustering results from real traffic flow dataset. The numerical experiments show that the use of coupled flow tensor with auxiliary information provides a novel and scalable user clustering method and improves the clustering accuracy.

INDEX TERMS Internet traffic, tensor decomposition, heterogeneous information network, user behavior.

I. INTRODUCTION

Nowadays mobile Internet with smart devices and innovative applications is reshaping our worlds and lives. We search information on Google, watch user-created video on YouTube, and share everything with friends on Facebook. Mobile Internet provides a brand-new medium for more convenient access to information and for more instant exchange of information. However, the increasing availability and diverse communication patterns lead to more complicated behaviour of users on Internet, which may incur potential risks of network management and network security. Therefore, it is important for Internet Service Providers (ISPs) to study the user behaviors on mobile Internet, for the purpose of network control and service provisioning.

Internet traffic flow data is the basic data which can be used to investigate the user behaviors on Internet. The traffic flow data is collected from network routers or traffic monitoring systems. A single traffic flow is defined as one or more packets sent from a particular source host and port, to a particular destination host and port, using a specific protocol, over some time interval [1]. Different from the server/client/proxy log mined on single web site or application system, which is normally used to analyze the detailed user behaviors of a specific type of application, the traffic flow data is de facto big data and provides a global view of user behaviors across different applications. Essentially the traffic flow is generated by user interaction and does not exist independently. The correlations among traffic flows reflect the relations among various network entities, such as user, server, application protocol, physical link, and so on. Mining traffic flow correlations is critical for understanding network operations and user behaviors, which helps promote

The associate editor coordinating the review of this manuscript and approving it for publication was Yi-Zhe Song.

network planning, protocol design, and differentiated service provisioning.

In this study, we focus on interaction pattern between mobile users and servers based on the traffic flow data. Real traffic flow data is collected from the public network of ISPs by high-performance network traffic monitors. The first challenge is the representation and characterization of traffic interaction pattern. We introduce heterogeneous information network [2] to represent the traffic interaction pattern, which is called Traffic Flow-based Heterogeneous Information Network (TF-HIN). Two types of nodes in TF-HIN are user node and server node respectivey, with differrent features. The traffic interaction between user node and server node forms the edge of TF-HIN. TF-HIN retains all information about user, server, traffic and time from real traffic flow data, and we further analyze the statistical and dynamic characteristics of TF-HIN, as well as node correlations.

Based on TF-HIN, user clustering in traffic interaction pattern is one of the basic data mining problems, which is useful for traffic and behavior prediction, application recommendation and targeting advertisement. Considering user clusters are not only related to their preferred application types, but also depend on the time when they use the specific application. The second challenge is how to exploit user similarity on application usage and time to improve the clustering accuracy. In addition to introducing a three-order tensor to represent the relations among user, server, and time, we propose to incorporate correlations of user and server as auxiliary information. Actually, the auxiliary matrices incorporate both the application categories of servers, and the correlations between user and server mined from TF-HIN. We design two iterative algorithms for coupled flow tensor decomposition. FTD is a basic algorithm, while FTD-NFS is an extended algorithm by considering sparsity and non-negative constraints. Experiments prove that both FTD-based and FTD-NFS-based user clustering algorithm can significantly improve the clustering accuracy. To our knowledge we are the first to propose coupled tensor decomposition for user clustering in traffic interaction pattern, and this modeling and analysis framework can be easily extended to other data mining problems on traffic flow data. The main contributions of this paper are as follows:

1) Based on real traffic flow data, Traffic Flow-based Heterogeneous Information Network (TF-HIN) is proposed to represent the interaction pattern between mobile users and servers. The statistical and dynamic characteristics of TF-HIN are provided.
2) A coupled flow tensor is proposed, by combining the relations of user/server/time and user/server correlations. Two decomposition algorithms of coupled flow tensor (FTD and FTD-NFS) are proposed, and the latent factors are used for user clustering.
3) Experiments based on benchmark datasets and real sampled dataset are investigated. The performance of the proposed coupled flow tensor method is validated, and the user clutering results are discussed.

The rest of the paper is organized as follows. Section II introduces the related work. In Section III, the collection of traffic flow data and the construction of TF-HIN are described, and the statistical characteristics of TF-HIN are presented. In Section IV, the coupled flow tensor is introduced, and the decomposition method and iterative algorithms are proposed. Section V provides the experiments and results to prove the performance of the coupled tensor decomposition method. Section VI concludes the paper.

II. RELATED WORK

The traffic flow data is widely used to investigate the characteristics of network layer and application layer in the literature. The former focuses on the packet-level statistical characteristics, such as packet size and packet arriving interval. The latter focuses on the stream-level characteristics of a specific application, such as request and response length, message size, and connection interval. Most recent researches make efforts on the latter. In [3], based on five-year real Web traffic of over 70,000 daily users from 187 different countries, major changes in Web traffic characteristics and structure of Web pages were presented. In [4], flow-level data was used to deduce the YouTube data center locations and load balancing strategies, and the effect on traffic dynamics across YouTube and the tier-1 ISP was analyzed. Reference [5] investigated the impact of the network access technologies and the application scenarios on the traffic of online gaming, i.e. World of Warcraft (WoW). Reference [6] investigated the traffic on a popular low-latency anonymous communication system Tor, and proposed TorWard system to discover malicious traffic over Tor. Reference [7] proposed an automatical mobile App identification system FLOWR by continually learning the Apps distinguishing features via HTTP traffic analysis.

Other than considering individual traffic flow as the above mentioned researches, there are a few researches paying attention to the correlations among traffic flows and user interaction behaviors. Reference [8] constructed a Call graph from the Call Detail Records of a mobile operator, and determined the evolution and the structural properties of these graphs to understand the social behavior of mobile phone customers. Reference [9] proposed a novel nonparametric approach for traffic classification, by incorporating correlated flow information into the classification process. In [10], for Traffic Activity Graphs (TAGs) in which nodes represent hosts and edges are observed flows among the hosts, a new sampling-based low-rank approximation method was proposed for investigate the topological properties of the TAGs. In [11], one-mode projections and bipartite graphs were used to model host communications, and clustering algorithm on the similarity matrices was applied to discover clusters of end-hosts in the same network prefixes. Reference [12] studied the interactions of traffic flows from a complex network perspective, and distinct correlative behaviors of six types of applications were discussed.

In this paper, we also consider the correlations among traffic flows, which can be represented by a dynamic directed and weighted graph. At the same time, we consider the inherent features of users and servers. As a result, we introduce heterogeneous information network [2] to represent the traffic interaction pattern, i.e. Traffic Flow-based Heterogeneous Information Network (TF-HIN). For the user clustering purpose, we propose coupled flow tensor to represent the relations among user, server, and time, by incorporating user and server similarity as auxiliary matrices. The coupled flow tensor decomposition considers both sparsity and non-negative constraints.

The heterogeneous information network (HIN) is an abstraction of the real world network, in which nodes represent objects of different entity types and links repesent different relationships between objects [2]. Many interconnected large-scale datasets, ranging from scientific, engineering, social to business applications can be used to construct heterogeneous information networks. For example, in the bibliographic information network derived from DBLP, papers are linked together via authors, venues and terms; in the Twitter information network, tweets are linked together via users, hashtags and terms. Currently mining in heterogeneous information networks is a promising research frontier in data mining research [13]. Reference [14] proposed a clustering method called RankClus in HIN, by integrating with ranking information of objects. In [15], a ranking-based iterative classification method called RankClass in HIN was proposed. In [16], the relationship prediction problem in HIN was solved by using a meta path-based approach. In [17], a probabilistic model was proposed to link the Web text named entities with a heterogeneous information network, consisting the entity object model and the entity popularity model.

In data science, real data was collected and analyzed, and features are extracted for learning model [18]–[21]. The classical nonnegative matrix factorization (NMF) was proposed for reducing the dimensionality of the data, and has been widely used in many applications [22]. In [23], a method based on Gaussian process priors was presented for nonnegative matrix factorization. Reference [24] proposed a Bayesian matrix factorization method, the beta-gamma nonnegative matrix factorization (BG-NMF) model for the continuous data with bounded support. A tensor is a multidimensional array [25]. Higher-order tensor ($order \geq 3$) is a suitable representation for multi-object relationships in real world, and is applied in many scientific fields such as chemometrics, psychometrics, computer vision, signal processing and data mining. There are two popular low rank decomposition methods for higher-order tensor. One is the CANDECOMP/PARAFAC (CP) decomposition, which decomposes a tensor into a sum of component rank-one tensors [26]. The other is Tucker decomposition, which decomposes a tensor into a core tensor multiplied by a matrix along each mode [27]. Recently new tensor decomposition methods are developed by incorporating important constraints or auxiliary information of data. In [28], a new constrained tensor decomposition method with non-negativity as a baseline constraint is proposed, based on Alternating Direction Method of Multipliers (ADMoM). In [29], two regularization approaches were proposed to improve the tensor decomposition quality, by using auxiliary information induced from the relationships of data.

III. TRAFFIC FLOW-BASED HETEROGENEOUS INFORMATION NETWORKS

A. INTERNET TRAFFIC FLOW DATA

In our study, the Internet traffic flow data was collected by proprietary high-performance traffic monitors in an ISPs' operational mobile network, which serves millions of users in a southern province of China. As shown in Fig.1, an ISP's mobile network can be divided into the access network and the core network. A mobile device communicates with a Base Transceiver Station (BTS) in the access network, which forwards its data traffic to the core network, under the control of Base Station Controller (BSC). In the core network which supports General Packet Radio Service (GPRS), a Serving GPRS Support Node (SGSN) establishes a tunnel on the Gn interface with a Gateway GPRS Support Node (GGSN), which provides connectivity to the Internet via the Gi interface. Through this path, the requesting messages of a mobile device enter the Internet and reach the serving server. Messages responding from the server to the mobile device traverse in the reversed path.

The traffic monitors are deployed between SGSN and GGSN, to capture packets at various network interfaces, such as Gb, Gn, and Gi. In addition to supporting line-speed packet parsing and accurate flow composition, the traffic monitors provide real-time traffic classification by combining port-based application deduction, deep packet inspection (DPI), and deep flow inspection (DFI) technologies.

The traffic flow data used in this paper was collected within a 24-hour period on December, 2013. There were more than 1.5 billion traffic flows recorded in the 24-hour period (0:00 am-11:59 pm). The detailed entries in each flow record include occurring time, flow duration, source and destination IP addresses or mobile phone number, total number of packets and bytes in the flow, application type, and mobile phone brand and model, etc. For application type, there are more than 20 categories, such as Web, Search, Advertising, E-commerce, Music, Video, and more than 100 sub-categories identified, including: Wechat, BaiduSearch, Headline, Weibo, QQMusic, Tencent Video and so on. For mobile phone, there are more than 20 brands and more than 100 models identified, such as Apple, Samsung, and Huawei, which are popular in Chinese market.

B. HETEROGENEOUS INFORMATION NETWORK CONSTRUCTION

Based on the traffic flow data collected, the heterogeneous information network [2] can be used to represent the traffic interaction pattern, which is named Traffic Flow-based Het-

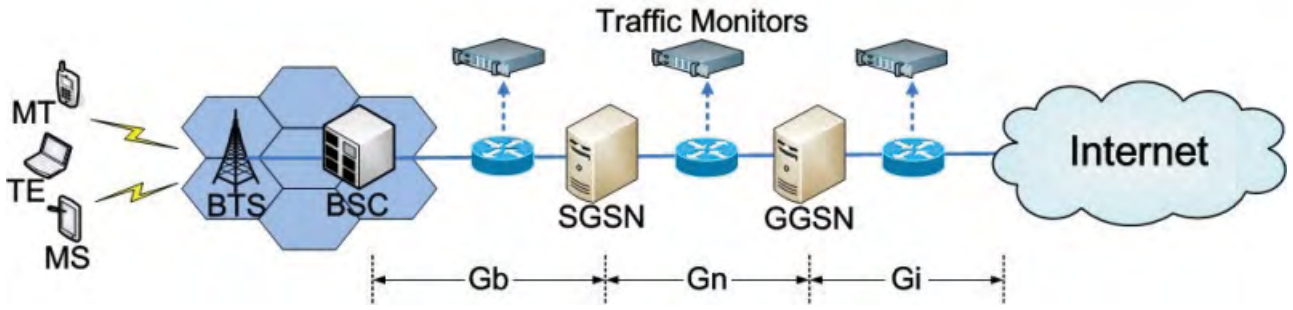


FIGURE 1. Mobile Internet Traffic Collection.

erogeneous Information Network (TF-HIN). In the TF-HIN, there are two types of nodes, i.e. the user node (denoted by mobile phone number) and the server node (denoted by specific IP address). The user node has the features such as mobile phone brand and model, Operating System, and Browser. The server node has the features such as application category and sub-category, implemented protocol, and location. The edge between the user node and the server node is formed by the actual data transfer, which is revealed by each traffic flow record. The features of an edge include start time, end time, packets, bytes, direction (from user to server is upstream, from server to user is downstream), application category label (determined by the server node's category at the time).

As a result, the TF-HIN is illustrated by Fig. 2, which essentially is a directed bipartite graph, denoted by $G = (V, E)$, $|V| = N$, $|E| = M$. The most notable characteristic of the TF-HIN distinguished from other heterogeneous information network is its dynamics, that means the edge is added and deleted frequently, due to the dynamics of Internet traffic. Moreover, the edges in the TF-HIN are identified by different colors, which mean different application category labels. Since a mobile user can use more than one application simultaneously, one user node may link more than one colored edge at the same time. The server node may link more than one colored edge at the same time, but this situation is not too much, because more than 80 percent servers support the only one service according to our actual statistical analysis.

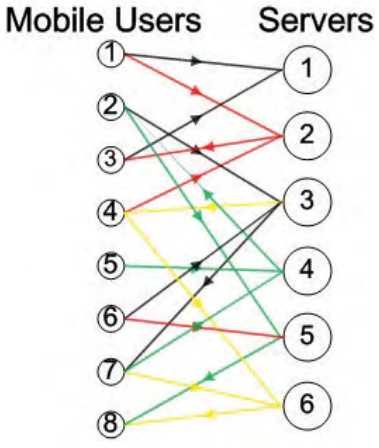


FIGURE 2. Traffic Flow-based Heterogeneous Information Network (TF-HIN).

1) Basic Analysis: Considering the large-scale of the traffic flow data, we investigate the statistical characteristics of the traffic on the basis of per monitor and per day. We choose one typical dataset, and construct the corresponding TF-HIN. We use minute as the time scale, and there are totally 1440 timeslots during one day. Table 1 presents the basic properties of the real dataset.

TABLE 1. Basic parameters of dataset.

	Real Dataset
Time	12/28/2013
Monitor	Monitor 1
Num of Flow Records	174,505,576
Num of User Node	1,051,224
Num of Server Node	61,355

Firstly, we investigate the active user node, active server node and active edge in the TF-HIN. In the Fig. 3, the x-axis is the timeslot, and the y-axis is the number of active user node, active server node, and active edge respectively. We can see that the three curves have the similar trends during a 24-hour duration, i.e. the active nodes and edges approach the lowest point at 2-5 am while remaining at the peak point from 10 am to 10 pm. This is reasonable and it confirms that Internet traffic fluctuates in a similar way as people's daily activity patterns. Another observation is that even if there are more than 1 million users during a day, the active user number at most reaches 8 percent of the total users, about 80,000 users, at every timeslot in the daytime. While in the daytime, the active server number reaches more than 40 percent of the total servers. This implies the diversity of user intersts as well as the universality of distributed services on mobile Internet.

Secondly, we investigate the degree and strength of the user and server nodes. A node's degree in a graph means the number of connections it associates with. In a directed graph, a node's in-degree means the number of incoming edges, while out-degree means the number of outgoing edges. In the TFHIN, the edge also has weight or strength, which

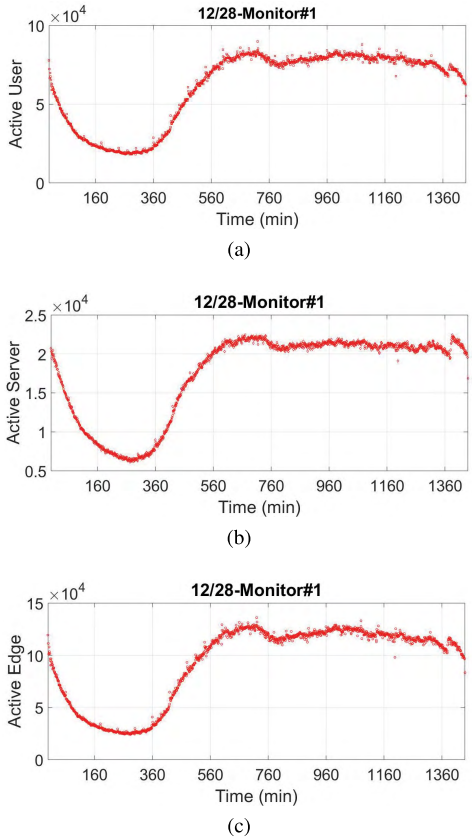


FIGURE 3. Active Node and Edge in TF-HIN. (a) Active User Node. (b) Active Server Node. (c) Active Edge.

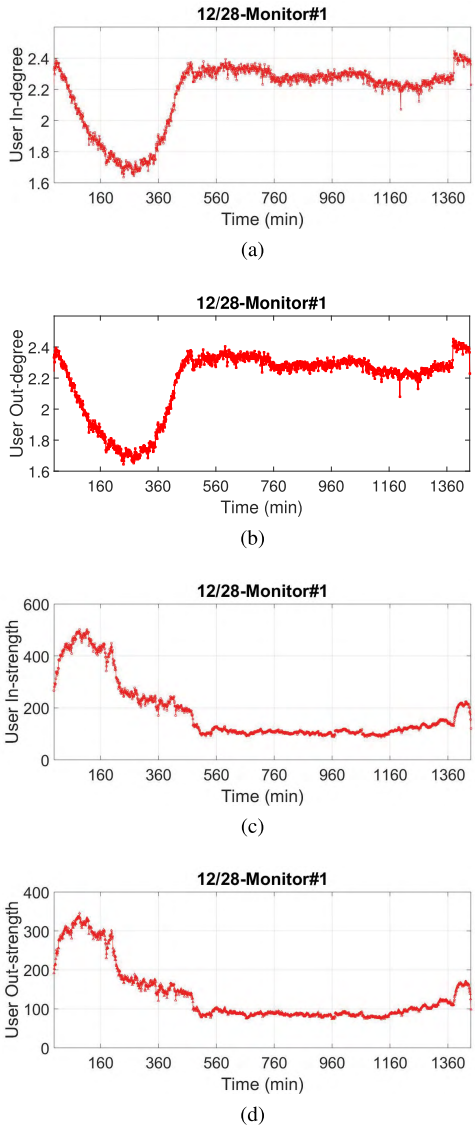


FIGURE 4. User Degree and Strength Dynamics in TF-HIN. (a) User Average In-degree. (b) User Average Out-degree. (c) User Average In-strength. (d) User Average Out-strengthh.

is defined as the number of packets transferred through the edge. A node's in-strength means the total incoming packets from its neighboring nodes, and out-strength means the total outgoing packets to its neighboring nodes.

Fig. 4 shows the degree and strength characteristics of user node in TF-HIN. Fig. 5 shows the characteristics of server node in TF-HIN. In these figures, x-axis is the timeslot, and y-axis is the in-degree, out-degree, in-strength and outstrength respectively. From Fig. 4a, it can be seen that the in-degree of user approaches the lowest point at 2 to 5 am, i.e. 1.7, while remains at the peak point from 10 am to 10 pm, i.e. 2.4. The out-degree of user node in Fig. 4b has the similar trends. However, for the in-strength and out-strength of user in Fig. 4c and 4d, there are different trends. At 1 to 3 am, the in-strength and out-strength reach the peak point; from 10 am to 10 pm, they remain at lower point. The possible reason is that the number of applications activated by a mobile user remains constant from day to night, but the traffic and type of applications change. So we further investigate the difference of connection and traffic of each application category in the day and night.

We choose the most popular seven application categories, including Web, Search, Advertising, E-commerce, Music, Video, Game, and analyze the connections and traffic volumes at 1 to 2 am and at 9 to 10 am. By comparing Fig. 6a and 6b, we find that though each application category involves more connections in the day than those at night, some application categories transfer more traffic volumes at night, such as category 3, category 4, and category 7. For example, category 4 is Web application, which is the dominant application in mobile Internet, and accounts for more than 70 percent of the total traffic. We can see that Web traffic is much larger at night. This can be explained by Web user behaviors. In the day, users mostly browse Web pages due to busy work or study; but at night, users tend to watch online video or download pictures. This is the reason why the average in/out strength of user and server node at night are larger than those in the day, as in Fig. 4 and 5.

Thirdly, we investigate the flow duration, which is egde lifetime in TF-HIN. We find that the flow duration distribution $p(d)$ follows a power-law distribution as $p(d) \sim d^{-\gamma}$,

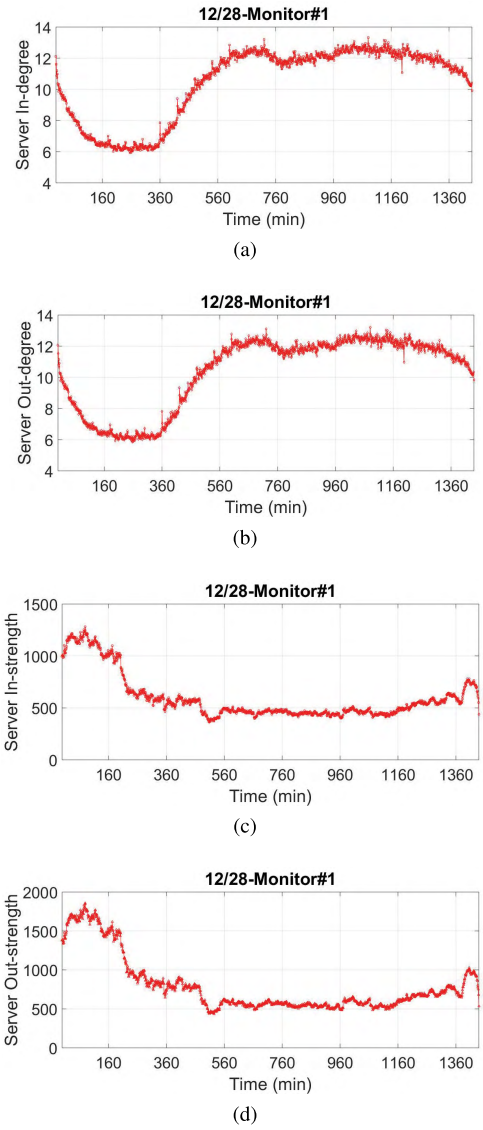


FIGURE 5. Server Degree and Strength Dynamics in TF-HIN. (a) Server Average In-degree (b) Server Average Out-degree. (c) Server Average In-strength. (d) Server Average Out-strength.

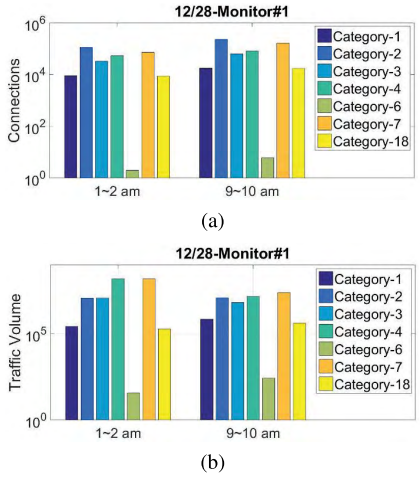


FIGURE 6. Connection and Traffic Comparison. (a) Connections per App Category (b) Traffic Volume per App Category.

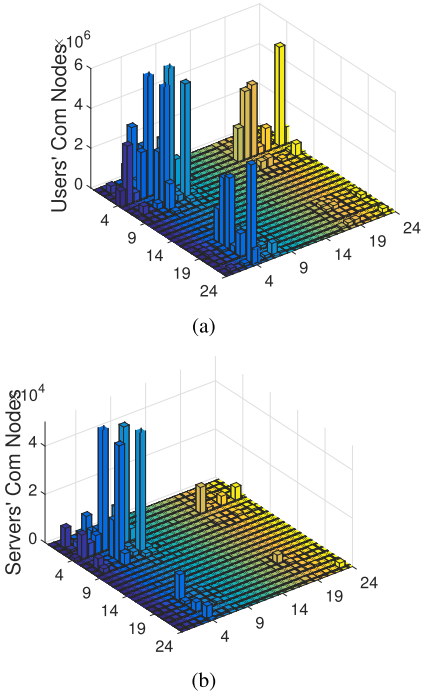


FIGURE 7. Common Neighboring Node. (a) User's Common Neighboring Node. (b) Server's Common Neighboring Node.

with the scaling exponents γ. This implies that most flows continue a relatively short time, which is shorter than 1 second; while there are several flows lasting very long time, which is longer than 1000 seconds. In order to reduce the memory and calculations, we compromise to use minute as time-scale in TF-HIN and the following flow tensor.

2) Node Correlation Analysis: As the above mentioned, TF-HIN represents the actual traffic interaction pattern since it captures the information of mobile user, server and traffic. In addition to basic statistical analysis of TF-HIN, it is important to investigate the node correlation of TF-HIN, which is helpful for more accurate node clustering.

Considering users using the same application category tend to connect the same servers, and conversely servers with the same application category tend to have similar user interest group, we investigate the common neighboring nodes of users and servers in TF-HIN.

In Fig. 7a, considering TF-HIN with one-minute in the day (i.e. 9:00 am to 9:01 am, 25,653 user nodes and 4,977 server nodes), x-axis and y-axis is the application category (i.e. Category 0 to Category 24), and z-axis is the number of common neighboring nodes of users who are using the same application category. We can see that for several application categories, such as category 4, category 7, and category 1, users in the same category have much more common neighboring nodes, which implies strong correlation among users

with the same application category. These users may have similar behaviors and form a cluster.

Meanwhile, there are some users with different application categories having more common neighboring nodes, such as users of categories 4 and 7, categories 1 and 4, categories 18 and 4. This implies users of these categories are more correlated, which may form one larger cluster.

Fig. 7b shows the common neighboring nodes of servers. It can also be seen that servers with the same application category have more common nodes. But for different application categories, it seems that server correlation is less than user correlation.

We also investigate the common neighboring node of TF-HIN at night and find the similar trend as Fig. 7a and 7b. So for a 24-hour period, user correlation between category i and category j can be represented by the ratio of average common neighboring nodes of user belong to category i and j over the total common neighboring nodes. Fig. 8 shows the user correlation during 24-hour period. We observe that users of category 4 have the strongest correlation, and users of category 7, as well as categories 4 and 7, categories 4 and 18, categories 4 and 1, are more correlated.

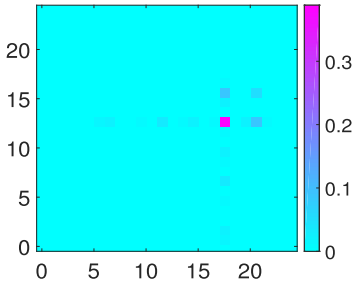


FIGURE 8. User Correlation based on Common Neighboring Node.

IV. COUPLED TENSOR DECOMPOSITION FOR USER CLUSTERING

As introduced in Section I, user node clustering in TF-HIN implies user interest group in traffic interaction pattern, which is useful for traffic and behavior prediction, application recommendation and targeting advertisement. In this Section, we use three-order tensor to model TF-HIN, which represents the relations among user, server, and time. In order to improve the clustering accuracy, we propose to incorporate correlations of user and server as auxiliary information,which can be obtained from TF-HIN. Then two iterative algorithms for coupled flow tensor decomposition are proposed, by considering sparsity and non-negative constraints of the coupled flow tensor.

A. COUPLED MATRIX AND TENSOR

A three-order tensor $\chi \in R^{I\times J\times K}$ is defined to represent the TF-HIN, where I, J, K are the number of user nodes, server nodes, timeslots in a day respectively. The element $x_{i,j,k}$ is defined as follows:

$$x_{i,j,k} = \begin{cases} b_{ij} & if\ u_i\ visits\ s_j\ at\ t_k \\ 0 & otherwise \end{cases} \tag{1}$$

$$b_{ij} = \sum_f \frac{p_{ij,f} + p_{ji,f}}{d_y}, \quad for\ each\ edge\ (i,j)\ at\ t_k \tag{2}$$

where $p_{ij,f}$ and $p_{ji,f}$ are number of packets from user i to server j (uplink) and from server j to user i (downlink) respectively, in the flow f, and d_y is the duration of flow f (second).

At the same time, we introduce two auxiliary matrices, i.e. user correlation matrix and server correlation matrix to represent the user and server similarity. Let $p_{i_1 i_2}$ be the probability of having common server nodes between user i_1 and user i_2 the user correlation matrix is $A \in R^{I\times I}$ with element $pr_{i_1 i_2}$. The server correlation matrix is $B \in R^{J\times J}$, which includes two types of information, i.e. the application category label and the common user nodes. Let server label matrix as $Z \in R^{J\times L}$, where L is the number of application category labels. $z_{j,l} = 1$ if s_j belongs to application category l, otherwise $z_{j,l} = 0$. Then element in B is defined as follows:

$$b_{j_1,j_2} = \alpha z^T_{j_1:} z_{j_2:} + pr_{j_1,j_2} \tag{3}$$

where $pr_{i_1 i_2}$ is the probability of having common user nodes between server j_1 and server j_2. The coupled flow tensor is shown as Fig. 9, which represents the actual traffic interaction as the TF-HIN.

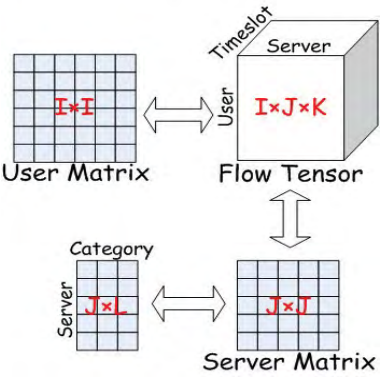


FIGURE 9. Flow tensor.

B. PARAFAC-BASED DECOMPOSITION METHOD

As illustrated in [25], the PARAFAC decomposition of a tensor χ is:

$$\chi \approx \sum_{r=1}^{R} u_r \circ v_r \circ w_r. \tag{4}$$

where $U = [u_1, u_2, \ldots u_R] \in R^{I\times R}$, $V = [v_1, v_2, \ldots v_R] \in R^{J\times R}$ and $W = [w_1, w_2, \ldots w_R] \in R^{K\times R}$ are the factor matrices and can be thought of as the principal components in each mode.

Considering the coupled flow tensor, the objective function is defined as follows:

$$f(U,V,W) = \frac{1}{2} \parallel \chi - \sum_{r=1}^{R} u_r \circ v_r \circ w_r \parallel_F^2 + \frac{\lambda_1}{2} Tr(U^T L_A U + V^T L_B V). \quad (5)$$

where $\parallel \chi - \sum_{r=1}^{R} u_r \circ v_r \circ w_r \parallel_F^2$ is to control the error of decomposition of χ, Tr$(\cdot)$ denotes the matrix traces, λ_1 is the parameter controlling the contribution of each part during the coupled decomposition. $L_A = D_A - A$ is the Laplacian matrix of user correlation matrix A, D_A is a diagonal matrix with element $d_{i_1 i_2} = \sum_{i_1} a_{i_1 i_2}$. Similarly, $L_B = D_B - B$ is the Laplacian matrix of user correlation matrix B, D_B is a diagonal matrix with element $d_{j_1 j_2} = \sum_{j_1} a_{j_1 j_2}$. $Tr(U^T L_A U)$ is obtained by considering two users i_1 and i_2 with higher similarity (i.e. $a_{i_1 i_2}$ is bigger) should have a closer distance between the vector u_{i_1} and u_{i_2} in the matrix U:

$$\begin{aligned} &\frac{1}{2} \sum_{i_1,i_2} \parallel u_{i_1} - u_{i_2} \parallel_2^2 a_{i_1} a_{i_2} \\ &= \sum_{i_1,i_2} u_{i_1} a_{i_1 i_2} u_{i_1}^T - \sum_{i_1,i_2} u_{i_1} a_{i_1 i_2} u_{i_2}^T \\ &= \sum_{i_1} u_{i_1} a_{i_1 i_1} u_{i_1}^T - \sum_{i_1,i_2} u_{i_1} a_{i_1 i_2} u_{i_2}^T \\ &= Tr(U^T (D_A - A) U) = Tr(U^T L_A U). \end{aligned} \quad (6)$$

$Tr(V^T L_B V)$ is obtained in the similar way.

The objective function is not jointly convex to all the variables U, V, W. So it is very hard to get closed-form solutions to minimize the objective functions. We propose an iterative method, which optimize one of U,V,W with fixing the others to the current values, and alternately update them by changing the factor matrix. This method is named Flow Tensor Decomppsition (FTD). More specifically, to optimize U, the objctive function as Equation (5) is transformed to the following equation:

$$U = arg \min_U f(U,V,W). \quad (7)$$

$$\begin{aligned} &f(U,V,W) \\ &= \frac{1}{2} \parallel X_{(1)} - U(W \odot V)^T \parallel_F^2 + \frac{\lambda_1}{2} Tr(U^T L_A U) \\ &= \frac{1}{2} Tr((X_{(1)} - U(W \odot V)^T)^T (X_{(1)} - U(W \odot V^T))) \\ &\quad + \frac{\lambda_1}{2} Tr(U^T L_A U). \end{aligned} \quad (8)$$

where $X_{(1)}$ denotes the mode-1 matricization of χ, and $\odot$ denotes the Khatri-Rao product.

We can obtain an exact solution to U by differentiating Equation (8) with respect to U and setting it to be zero. Actually Equation (9) is Sylvester Equation, and its solution is U:

$$\begin{aligned} &U(W \odot V)^T (W \odot V) + \lambda_1 L_A U \\ &\quad = U(V^T V * W^T W) + \lambda_1 L_A U = X_{(1)}(W \odot V). \end{aligned} \quad (9)$$

where $*$ denotes the Hadamard product. Similarly, we can obtain V and W. Since the coupled flow tensor has the properties of sparsity and non-negative, the coupled flow tensor decomposition method FTD can be extended to FTD-NFS by considering sparsity and non-negative constraints. In FTD-NFS method, the objective function is as follows:

$$\begin{aligned} f(U,V,W) = &\frac{1}{2} \parallel \chi - \sum_{r=1}^{R} u_r \circ v_r \circ w_r \parallel_F^2 \\ &+ \frac{\lambda_1}{2} Tr(U^T L_A U + V^T L_B V) \\ &+ \frac{\lambda_2}{2} (\parallel U \parallel_{2,1} + \parallel V \parallel_{2,1} + \parallel W \parallel_{2,1}) \\ &subject\ to\ U \geq 0, V \geq 0, W \geq 0. \end{aligned} \quad (10)$$

where 0 is the zero matrix of appropriate dimensions, and the inequalities are element-wise. As in [28], the NTF problem as Equation (10) can be transfered to Alternating Direction Method of Multipliers (ADMM) form. By introducing auxiliary variables $\overline{U} \in R^{I \times R}$, $\overline{V} \in R^{J \times R}$ and $\overline{W} \in R^{K \times R}$, we consider the equivalent optimization problem:

$$\begin{aligned} &\min_{U,V,W,\overline{U},\overline{V},\overline{W}} f(U,V,W) + g(\overline{U}) + g(\overline{V}) + g(\overline{W}) \\ &subject\ to\ U - \overline{U} = 0, V - \overline{V} = 0, W - \overline{W} = 0. \end{aligned} \quad (11)$$

where, for any matrix argument M, if $M \geq 0$, $g(M) = 0$, otherwise $g(M) = \infty$.

Then, by introducing the dual variables $Y_U \in R^{I \times R}$,$Y_V \in R^{J \times R}$ and $Y_W \in R^{K \times R}$, and the vector of penalty terms $\rho = [\rho_U \quad \rho_V \quad \rho_W]^T$, the augmented Lagrangian is given as follows:

$$\begin{aligned} &L_\rho(U,V,W,\overline{U},\overline{V},\overline{W},Y_U,Y_V,Y_W) \\ &= f(U,V,W) + g(\overline{U}) + g(\overline{V}) + g(\overline{W}) \\ &\quad + Y_U * (U - \overline{U}) + \frac{\rho_U}{2} \parallel U - \overline{U} \parallel_F^2 \\ &\quad + Y_V * (V - \overline{V}) + \frac{\rho_V}{2} \parallel V - \overline{V} \parallel_F^2 \\ &\quad + Y_W * (W - \overline{W}) + \frac{\rho_W}{2} \parallel W - \overline{W} \parallel_F^2 . \end{aligned} \quad (12)$$

By substituting $f(U,V,W)$ in Equation (10), the iteration solution for variable U in Equation (12) is as follows:

$$\begin{aligned} U^{k+1} = &arg \min_U (\frac{1}{2} \parallel X_{(1)} - U(W^k \odot V^k)^T \parallel_F^2 \\ &+ \frac{\lambda_1}{2} Tr(U^T L_A U) + \frac{\lambda_2}{2} \parallel U \parallel_{2,1} \\ &+ \frac{\rho_U}{2} \parallel U - \overline{U}^k + \frac{Y_U^k}{\rho_U} \parallel_F^2) \end{aligned} \quad (13)$$

$$\overline{U}^{k+1} = U^{k+1} + \frac{1}{\rho_U} Y_U^k \quad (14)$$

$$Y_U^{k+1} = Y_U^k + \rho_U (U^{k+1} - \overline{U}^{k+1}) \quad (15)$$

In order to solve U in Equation (13), we can differentiate it with respect to U and setting it to be zero, then solve the

Sylvester Equations:

$$U(W \odot V)^T(W \odot V) + (\lambda_1 L_A + \lambda_2 Q_U + \rho_U I_U)U = X_{(1)}(W \odot V) + \rho_U \overline{U} - Y_U \quad (16)$$

where I_U is Identity Matrix. At the same time, we let $\| P \|_{2,1} = Tr(P^T QP)$, $Q = Diag(q)$, and q is the auxiliary vector of the $l_{1,2}$ norm. The elements of q are computed as follows:

$$q_i = \frac{1}{2\sqrt{\| p_i \|_2^2 + \varepsilon_1}}$$

Here ε_1 is a smoothing term that avoids division by zero.We can obtain V and the W in the similar way.

C. USER CLUSTERING ALGORITHM

Based on the proposed FTD and FTD-NFS method, we design two user clustering algorithm, as the following Algorithm1 and Algorithm 2.

Algorithm 1 FTD-Based User Clustering Algorithm

Input:
flow tensor χ;
user correlation matrix A;
server correlation matrix B;
error threshold ε; max iteration times $IterMax$;
Output:
low rank matrices U,V,W;
initializing U,V,W with small random values;
calculating Laplacian matrices L_A and L_B;
while $loss^{t+1} - loss^t > \varepsilon$ and $t < IterMax$ **do**
  calculating U^{t+1};
  calculating V^{t+1};
  calculating W^{t+1};
  calculating $loss^{t+1}$;
end while
running K-means by input U;
return U,V,W;

Algorithm 2 FTD-NFS-Based User Clustering Algorithm

Input:
flow tensor χ;
user correlation matrix A;
server correlation matrix B;
error threshold ε; max iteration times $IterMax$;
Output:
low rank matrices U,V,W;
initializing $U,V,W,\overline{U},\overline{V},\overline{W},Y_U,Y_V,Y_W$ with small random values;
calculating Laplacian matrices L_A and L_B;
calculating Q_U,Q_V,Q_W;
while $loss^{t+1} - loss^t > \varepsilon$ and $t < IterMax$ **do**
  calculating U^{t+1},V^{t+1},W^{t+1};
  calculating $\overline{U}^{t+1},\overline{V}^{t+1},\overline{W}^{t+1}$;
  calculating $Y_U^{t+1},Y_V^{t+1},Y_W^{t+1}$;
  calculating $loss^{t+1}$;
end while
running K-means by input U;
return U,V,W;

V. EXPERIMENTS AND RESULTS

A. EXPERIMENT DATASET

To evaluate the performance of the proposed user clustering algorithms, we use three experimental benchmark datasets from the real traffic flow data as in Section III. We choose the most popular five types of applications (i.e. Web, Search, Advertising, E-commerce and Video), randomly extract a part of user nodes and server nodes according to the percentage of traffic volume of each application, and select the corresponding flow records from the real datasets to construct TF-HIN as well as the coupled flow tensor. In order for easy evaluation, each user node and server node belong to only one application category, so actually there are five non-overlapped clusters in TF-HIN. The three benchmark datasets are shown in Table 2.

To further analyze the proposed clutering method on the real flow data, we combine DPI data and App Crawler data, and constrcut the Network Footprint Data (NFP). NFP data records the information when a user visits an App, and the detailed data fields include User id, App id, App category, Time (hour) and PV (the number of times which users use the App at this hour). We choose a randomly sampled dataset from NFP data on a week of 2017, as shown in Table 3. Top 10 Apps visited by users are wechat, BaiduSearch, TencentVideos, QQ, Headline, DIAN PING, Alipay, weibo, Gaode Map, Ctrip Travel.

TABLE 2. Benchmark datasets.

	Users	Servers	Timeslots	Nonzeros
Benchmark Dataset 1	568	355	791	11562
Benchmark Dataset 2	1328	717	925	27397
Benchmark Dataset 3	1470	855	1265	24765

TABLE 3. Sampled dataset.

	Users	APPs	Timeslots	Nonzeros
Sampled Dataset	1000	55	96	35715

B. ALGORITHM VALIDATION

We focus on three clustering algorithms. The baseline is basic K-means algorithm. We use an application vector for each user to record its visit. The application category information is also included in the application vector. The other two algorithms are our proposed FTD-based and FTD-NFS-based user clustering algorithm. We investigate the clustering algo rithm performance based on three typical clustering accuracy metrics, i.e., Normalized Mutual Information (NMI), Purity Accuracy (ACC) and Rand Index. It should be noted that we do not use CP decomposition of flow tensor (without auxiliary matrices) as baseline, because we did experiments

on it with different parameters and found NMI was less than 0.1. This proves indirectly the necessity of incorporating auxiliary information to solve the user clustering problem.

Firstly, we analyze the accuracy performance of different user clustering algorithms. Fig. 10 shows the performance comparison based on three experimental datasets. Both FTD and FTD-NFS algorithm use the typical parameter setting, i.e., $R = 5$ and $R = 8$, while iteration times are 18-20. It can be seen that overall FTD-NFS performs slightly better than FTD, and both of them surpass the basic K-means.

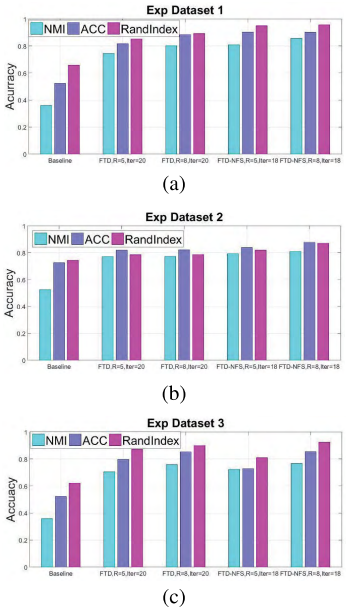


FIGURE 10. Comparison of Different Clustering Algorithms. (a) Experiment Dataset 1. (b) Experiment Dataset 2. (c) Experiment Dataset 3.

Furthurmore, we investigate the effect of iteration times on the performance of proposed FTD and FTD-NFS algorithms. For FTD algorithm, the InterMax is set to 20, 30, 50, 100 and 200. From Fig. 11a, we can see that for $R = 3$, there is little change for the performance of FTD even under different iteration settings. While for $R = 5$ and $R = 8$, with a large number of iterations, there is performance degradation. This is partially because the overfitting during tensor decomposition.

For FTD-NFS algorithm, the InterMax is from 10 to 18. From Fig. 11b, we can see that as iteration times increase, FTD-NFS performs better and better, and when InterMax approaches 18, FTD-NFS under different parameter settings ($R = 3, 5, 8$) achieve similar good performance. Although each iteration for FTD-NFS has more complexity than FTD, but the total complexity of FTD-NFS is equivalent to that of FTD, since for the same accuracy requirement, fewer iterations are needed by FTD-NFS.

C. CLUSTERING RESULTS AND DISCUSSIONS

We use FTD-NFS on the sampled dataset to investigate the user clustering results. For the parameter setting of FTD-NFS,

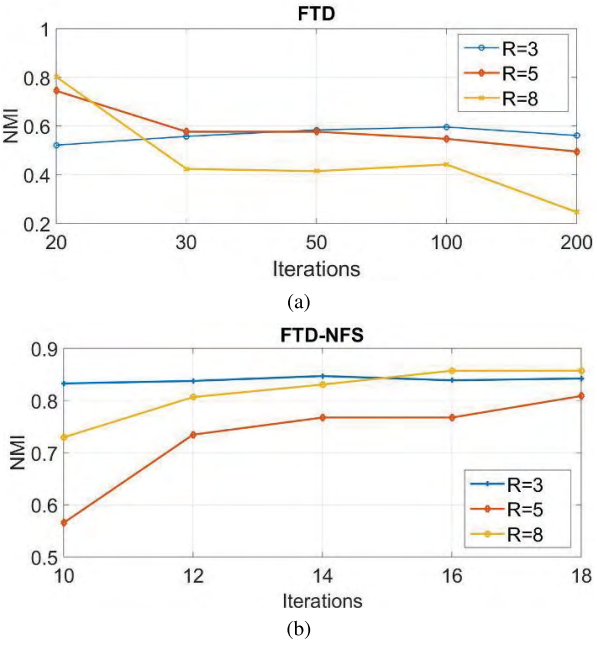


FIGURE 11. Algorithm Iteration. (a) FTD Iteration. (b) FTD-NFS Iteration.

let $R = 8$ and iterations be 20. Based on the users' feature vectors, We cluster users into 5 groups by using K-means.

Fig. 12 shows the statistics information of users in five clusters. We split one day into four time slots: 1:00-6:00, 7:00-12:00, 13:00-18:00 and 19:00-24:00, and the App visiting information during each time slot was shown in a histogram. In the histogram, x-axis represents five user clusters, y-axis is the number of visits, and one color represents one App. We select the most frequently visited Apps for each cluster to display.

By analyzing the user behavior of each cluster, we can infer user group's interests. Note that each user belongs to only one cluster. From the figure, a very obvious feature can be seen that wechat is widely used in each user group, which is very suitable for the users of APP in China. Users in Cluster 0 tend to use Search App at work shift, use entertainment App at rest time and also use news App. Because of the obvious division of working hours and entertainment hours, and the group often watches news, they might be office workers. Users in Cluster 1 use Search App at all time slot and order takeaway food in the early morning. They have longer working hours than office workers and might be overtime workers. Users in Cluster 2 and Cluster 3 look like students. But users in Cluster 2 use more Search App (might for scientific research), while users in Cluster 3 use more kinds of applications. So users in Cluster 2 are more likely to be graduate students, and users in Cluster 3 are more likely to be college students. Users in Cluster 4 are loyal users of Wechat and fans of video and news. This behavior is more like an older person's behavior, so they might be retirees.

From this experiment, We can find that the users' clusters are formed based on several factors (i.e. App category, time,

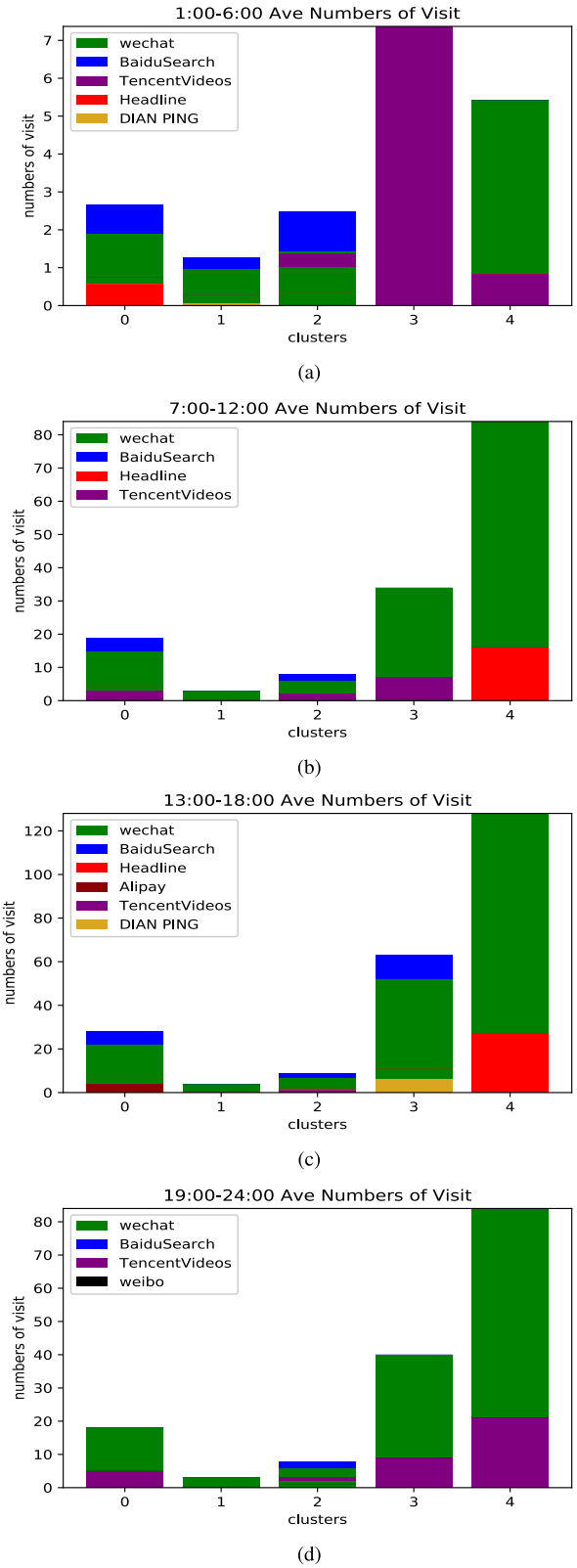


FIGURE 12. **User Clsutering based on App Visit. (a) App Visit 1:00-6:00. (b) App Visit 7:00-12:00. (c) App Visit 13:00-18:00. (d) App Visit 19:00-24:00.**

traffic, etc) simultaneously. Our proposed clustering method is scalable to more than one factor and can find new user groups in the mobile Internet.

VI. CONCLUSION

In this paper, we investigate user clustering in traffic interaction pattern based on real traffic flow data. We introduce Traffic Flow-based Heterogeneous Information Network (TF-HIN) to represent the traffic interaction pattern between users and servers. After analyzing the characteristics of TF-HIN, we use the coupled flow tensor by incorporating user and server correlation as auxiliary information, and propose two iterative decomposition motheds, i.e. FTD and FTD-NFS, for user clustering. Experiments validate the performance of the proposed FTD-based and FTD-NFS-based user clustering algorithm. It is proved that the coupled flow tensor decomposition method improves the accuracy of user clustering.

For the future work, we will design parrellel and distributed coupled flow tensor decomposition methods and investigate user clustering on actual large-scale traffic flow data.

REFERENCES

[1] T. Zseby, J. Quittek, B. Claise, and S. Zander, *Requirements for IP Flow Information Export (IPFIX)*, document IETF RFC 3917, 2004.

[2] Y. Sun, J. Han, X. Yan, and P. S. Yu, "Mining knowledge from interconnected data: A heterogeneous information network analysis approach," *Proc. VLDB Endowment*, vol. 5, no. 12, pp. 2022–2023, 2012.

[3] S. Ihm and V. S. Pai, "Towards understanding modern Web traffic," in *Proc. ACM IMC*, 2011, pp. 295–312.

[4] V. K. Adhikari, S. Jain, and Z.-L. Zhang, "YouTube traffic dynamics and its interplay with a tier-1 ISP: An ISP perspective," in *Proc. ACM IMC*, 2010, pp. 431–443.

[5] X. Wang, T. Kwon, Y. Choi, M. Chen, and Y. Zhang, "Characterizing the gaming traffic of world of warcraft: From game scenarios to network access technologies," *IEEE Netw.*, vol. 26, no. 1, pp. 27–34, Jan./Feb. 2012.

[6] Z. Ling, J. Luo, K. Wu, W. Yu, and X. Fu, "Torward: Discovery of malicious traffic over Tor," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Apr./May 2014, pp. 1402–1410.

[7] Q. Xu *et al.*, "Automatic generation of mobile app signatures from traffic observations," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Apr./May 2015, pp. 1481–1489.

[8] A. A. Nanavati *et al.*, "Analyzing the structure and evolution of massive telecom graphs," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 5, pp. 703–718, May 2008.

[9] J. Zhang, Y. Xiang, Y. Wang, W. Zhou, Y. Xiang, and Y. Guan, "Network traffic classification using correlation information," *IEEE Trans. Parallel Distrib. Syst.*, vol. 24, no. 1, pp. 104–117, Jan. 2013.

[10] Y. Liu, W. Chen, and Y. Guan, "Monitoring traffic activity graphs with low-rank matrix approximation," in *Proc. IEEE LCN*, Oct. 2012, pp. 59–67.

[11] K. Xu, F. Wang, and L. Gu, "Behavior analysis of Internet traffic via bipartite graphs and one-mode projections," *IEEE/ACM Trans. Netw.*, vol. 22, no. 3, pp. 931–942, Jun. 2014.

[12] X. Wu, X. Wang, K. Yu, and F. Y. Li, "A measurement-based study on the correlations of inter-domain internet application flows," *Comput. Netw.*, vol. 58, pp. 127–140, Jan. 2014.

[13] J. Han, "Mining heterogeneous information networks: The next frontier," in *Proc. ACM KDD*, 2012, pp. 2–3.

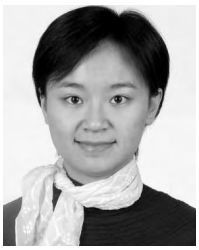
[14] Y. Sun, J. Han, P. Zhao, Z. Yin, H. Cheng, and T. Wu, "Rankclus: Integrating clustering with ranking for heterogeneous information network analysis," in *Proc. EDBT*, 2009, pp. 565–576.

[15] M. Ji, J. Han, and M. Danilevsky, "Ranking-based classification of heterogeneous information networks," in *Proc. ACM KDD*, 2011, pp. 1298–1306.

[16] Y. Sun, J. Han, C. C. Aggarwal, and N. V. Chawla, "When will it happen?: Relationship prediction in heterogeneous information networks," in *Proc. ACM WSDM*, 2012, pp. 663–672.

[17] W. Shen, J. Han, and J. Wang, "A probabilistic model for linking named entities in web text with heterogeneous information networks," in *Proc. ACM SIGMOD*, 2014, pp. 1199–1210.

[18] A. Ramisa, F. Yan, F. Moreno-Noguer, and K. Mikolajczyk, "BreakingNews: Article annotation by image and text processing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1072–1085, May 2018.
[19] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. ICASSP*, 2018, pp. 1–5.
[20] Z. Ma, H. Yu, W. Chen, and J. Guo, "Short utterance based speech language identification in intelligent vehicles with time-scale modifications and deep bottleneck features," *IEEE Trans. Veh. Technol.*, vol. 68, no. 1, pp. 121–128, Jan. 2019.
[21] Z. Ma, Y. Lai, W. B. Kleijn, Y.-Z. Song, L. Wang, and J. Guo, "Variational Bayesian learning for Dirichlet process mixture of inverted Dirichlet distributions in non-Gaussian image feature modeling," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 2, pp. 449–463, Feb. 2019.
[22] D. S. Seung and D. D. Lee, "Algorithms for non-negative matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, pp. 556–562.
[23] M. N. Schmidt and H. Laurberg, "Nonnegative matrix factorization with Gaussian process priors," *Comput. Intell. Neurosci.*, vol. 2008, no. 3, Jan. 2008.
[24] Z. Ma, A. E. Teschendorff, A. Leijon, Y. Qiao, H. Zhang, and J. Guo, "Variational Bayesian matrix factorization for bounded support data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 4, pp. 876–889, Apr. 2015.
[25] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Rev.*, vol. 51, no. 3, pp. 455–500, 2009.
[26] A. H. Kiers, "Towards a standardized notation and terminology in multiway analysis," *J. Chemometrics*, vol. 14, no. 3, pp. 105–122, 2000.
[27] L. R. Tucker, "Some mathematical notes on three-mode factor analysis," *Psychometrika*, vol. 31, no. 3, pp. 279–311, 1966.
[28] A. P. Liavas and N. D. Sidiropoulos, "Parallel algorithms for constrained tensor factorization via alternating direction method of multipliers," *IEEE Trans. Signal Process.*, vol. 63, no. 20, pp. 5450–5463, Oct. 2015.
[29] A. Narita, K. Hayashi, R. Tomioka, and H. Kashima, "Tensor factorization using auxiliary information," *Data Mining Knowl. Discovery*, vol. 25, no. 2, pp. 298–324, 2012.



KE YU received the B.S. degree in computer science and the Ph.D. degree in signal and information processing from the Beijing University of Posts and Telecommunications (BUPT), China, in 2000 and 2005, respectively. In 2011, she held visiting position at the University of Agder, Norway. From 2015 to 2016, she was a Visiting Scholar with the University of Illinois at Chicago, USA. She is currently an Associate Professor with the School of Information and Communication Engineering, BUPT. Her current research interests include communication network theory, network data mining, mobile Internet application, machine learning, and human–machine intelligence.

LIFANG HE (GS'12–M'14) received the B.S. degree in information and computer science from Northwest Normal University, in 2009, and the Ph.D. degree from the School of Computer Science and Engineering, South China University of Technology, in 2014. She was a Postdoctoral Researcher in computer science with the University of Illinois at Chicago, and also with the Department of Healthcare Policy and Research, Weill Cornell Medical College, Cornell University. She is currently a Postdoctoral Associate with the Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania. Her research interests include tensor mining, machine learning, with major applications in biomedical data and neuroscience. She is a member of the IEEE.

PHILIP S. YU received the B.S. degree in electrical engineering from National Taiwan University, the M.S. and Ph.D. degrees in electrical engineering from Stanford University, and the M.B.A. degree from New York University. He was with IBM, where he was a Manager of the Software Tools and Techniques Department, Watson Research Center. He is currently a Distinguished Professor in computer science with the University of Illinois at Chicago and also holds the Wexler Chair in information technology. He has published more than 1,000 papers in refereed journals and conferences. He holds or has applied for more than 300 U.S. patents. His research interests include big data, including data mining, data stream, database, and privacy. He is a fellow of the ACM and the IEEE.

WENKAI ZHANG received the B.E. degree from Shandong University, Jinan, in 2017. She is currently pursuing the master's degree with the Beijing University of Posts and Telecommunications, China. Her research interests include data mining, machine learning, and recommendation in heterogeneous information networks.

YUE LIU received the B.E. degree from the Harbin Institute of Technology, Weihai, in 2017. She is currently pursuing the master's degree with the Beijing University of Posts and Telecommunications, China. Her research interests include data mining, machine learning, and the Internet service provisioning.