This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2020.3021987, IEEE Transactions on Mobile Computing

To What Extent We Repeat Ourselves? Discovering Daily Activity Patterns Across Mobile App Usage

Tong Li, *Student Member, IEEE,* Yong Li, *Senior Member, IEEE,*
Mohammad A. Hoque, Tong Xia, Sasu Tarkoma, and Pan Hui, *Fellow, IEEE*

Abstract—With the prevalence of smartphones, people have left abundant behavior records in cyberspace. Discovering and understanding individuals' cyber activities can provide useful implications for policymakers, service providers, and app developers. In this paper, we propose a framework to discover daily cyber activity patterns across people's mobile app usage. We first segment app usage traces into small time windows and then design a probabilistic topic model to infer users' cyber activities of each window. By exploring the coherence of users' activity sequences, the daily patterns of individuals are identified. Next, we recognize the common patterns across diverse groups of individuals using a hierarchical clustering algorithm. We then apply our framework on a large-scale and real-world dataset, consisting of 653,092 users with 971,818,946 usage records of 2,000 popular mobile apps. Our analysis shows that people usually obey yesterday's activity patterns, but the patterns tend to deviate as the time-lapse increases. We also discover five common daily cyber activity patterns, including afternoon reading, nightly entertainment, pervasive socializing, commuting, and nightly socializing. Our findings have profound implications on identifying the demographics of users and their lifestyles, habits, service requirements, and further detecting other disrupting trends such as working overtime and addiction to the game and social media.

1 Introduction

Individuals exhibit diverse activities both in physical and cyberspace. Discovering and understanding patterns of such activities are fundamental to promote and support healthier lifestyles for individuals and further improve people's well-being. Although prior research has made significant progress towards understanding activity patterns, they only focus on small groups or specific activities. With limited data and controlled studies, it is only possible to characterize the patterns of an individual or a few individuals. In most cases, the investigated groups are either narrowly focused or so diverse that it is difficult to find their common activity patterns. The common patterns among large-scale individuals may not only hint at their demographics and lifestyles but also expose disrupting trends in our society. However, this requires exploring the activity patterns of a large-scale population, e.g., millions of people.

The difficulties in accessing data from many individuals are one of the main obstacles in large-scale pattern studies. Traditional data collection methods, like surveys and questionnaires, need a lot of human effort, which is usually inefficient and delayed. In recent years, the multifaceted usage of smartphones in daily lives has established them as a necessity, which records various users' cyber activities. In addition to traditional uses, e.g., communication and web browsing, people use smartphones in more complex activities such as ordering food, shopping online, and managing health [1]. A great number of research studies in recent years have applied app usage data to investigate user behavior [2]–[5]. These studies include app energy drain [6], app signatures [3], how app usage varies with different kinds of users [4], and how unusual events disrupt the app engagement [5].

In this work, we focus on discovering daily activity patterns in cyberspace from large-scale app usage data. More specifically, we study the following research problems.

- What activities can be discovered from app usage data?
- What common patterns we share with others in our daily cyber activities?

There are three challenges to address these two problems. **1)** The complexity of mobile app usage behavior. App usage behavior varies across app categories, usage time, and users. Due to the unbalanced usage of different categories, some useful and critical app usage, like ordering food, may be overwhelmed by popular categories and cannot be detected. Besides, the same app usage may imply different activities for different users in different contexts. Hence, how to extract user activity features from such complicated data is fundamental but difficult. **2)** The lack of ground truth labels of user cyber activities. All large-scale app usage datasets lack meaningful labels to map user activities to app usage records. Finding the appropriate activity labels

- *T. Li and P. Hui are with the System and Media Laboratory (SyMLab), Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong. They are also with the Department of Computer Science, University of Helsinki, Helsinki, Finland. E-mail: t.li@connect.ust.hk, panhui@cse.ust.hk*
- *Y. Li and T. Xia are with Beijing National Research Center for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing, China. E-mail: liyong07@tsinghua.edu.cn, xia-t17@mails.tsinghua.edu.cn*
- *M. A. Hoque and S. Tarkoma are with the Department of Computer Science, University of Helsinki, Helsinki, Finland. E-mail: mohammad.a.hoque@helsinki.fi, sasu.tarkoma@helsinki.fi*

Manuscript received XX xx, 2019; revised xx xx, 2019.

1536-1233 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: CAMBRIDGE UNIV. Downloaded on November 16,2020 at 21:46:11 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2020.3021987, IEEE Transactions on Mobile Computing

is challenging, and annotating the extracted cyber activities with semantic terms is also tricky. **3)** The complexity of user activities in cyberspace. Activity patterns can be diverse in both individual and group granularity. For individuals, her/his cyber activity patterns may change over days. How to identify a typical daily cyber activity pattern for each individual is complicated. Also, whether there are any common patterns across individuals and discovering these common patterns is another challenge.

To overcome the above challenges, we propose a framework to discover daily activity patterns in cyberspace by leveraging the following three key designs. **First**, we devise a probabilistic topic model based method to identify the activity features of app usage windows obtained using a time-based segmentation approach. The proposed method regards a window as a document, a user as an author, an activity as a topic, and app usage logs related to the window as words. In this way, the probabilistic topic model can characterize the relationships between users, app usage, and cyber activities in a cohesive manner. **Second**, to annotate extracted cyber activities with semantic terms, we explore the semantic information of app categories and prior knowledge of activity temporal patterns. More specifically, we compute an average app category feature vector for each extracted activity. In terms of the density in the calculated vector, we rank app categories in an activity termed as internal ranking. Besides, we also rank all activities for each app category termed as external ranking. By jointly considering internal ranking, external ranking, and temporal distribution of activities, we determine each extracted cyber activity with a meaningful label. **Third**, we look into the regularity of individuals' cyber activity patterns by measuring their average intra-distance of patterns across different days. We then identify a daily cyber activity pattern for every individual. By applying the agglomerative hierarchical clustering with edit distance, we recognize the common patterns across diverse groups, which address the third challenge.

The contributions of our work can be summarized as follows.

- We investigate the problem of discovering daily activity patterns in cyberspace of a large-scale population by using their mobile app usage data. We propose a novel activity discovery framework based on a probabilistic topic model, which can characterize the relationships between users, apps, and activities in a cohesive manner.
- We apply our framework on a large-scale and real-world app usage dataset. We discover people prefer to imitate yesterday's activity patterns and provide evidence that the population follows five common patterns in their daily cyber activities, including afternoon reading (8.50%), nightly entertainment (18.45%), pervasive socializing (7.56%), commuting (29.29%), and nightly socializing (36.20%).
- We verify our findings via a small-scale dataset with users' occupation information. We find that different common patterns correspond to different demographic groups. More specifically, freelancers have a nightly socializing pattern. White-collar workers have a commuting pattern. Advertisers and socializers have a pattern of pervasive socializing. Also, we infer that senior citizens are mostly involved in afternoon reading, and the younger generation is primarily involved in nightly entertainment.
- Based on the discovered cyber activity patterns, we detect several social issues, e.g., around 35% of workers always work overtime, and nearly 42% of the younger generation is addicted to gaming applications. We further explore the implications of our framework and findings for policymakers and government, researchers, service providers, and app developers.

The rest of this paper is organized as follows. In section 2, we present an overview of our dataset. In section 3, we present how to discover activities from app usage traces. In section 4, we elaborate on the schemes of identifying the common daily patterns. We then discuss the implications and limitations of our work in section 5. Related works are presented in section 6. Finally, we briefly conclude the paper in section 7.

2 Dataset Overview

To understand the daily activity patterns of people in a large city, we explore a city-scale app usage dataset provided by a primary Internet Service Provider (ISP) in China. The dataset was collected during one week in April 2016, covering the whole metropolitan area of Shanghai, one of the world's largest cities. The dataset includes over 2 million users and their app usage records during the data collection period. The app usage dataset is characterized by the ISP with an anonymized user ID, timestamp, and app ID.

In detail, the ISP identified app usage traces based on users' network access records collected from gateways, generated when users issue network connection requests. In terms of the data format, each network access record contains a user ID, timestamp, and the connection's metadata. To determine the corresponding app ID for each network access record, the ISP inspected the HTTP head and used the destination domain and user-agent as the app identifier. By adopting a systematic tool, SAMPLE [7], the ISP constructed conjunctive rules to match specific apps. SAMPLE applies a supervised learning algorithm over a small set of labeled data streams to automatically generate the conjunctive rules, which can identify over 90% of apps with an average accuracy of 99% [7]. In practice, the ISP built the conjunctive rules by manually operating a small set of apps to generate data streams. They then crawled the 2,000 most popular apps across app stores and matched network traffic records to these apps. Also, the ISP manually verified the correctness of the matched apps. In terms of the statistics from the ISP, more than 95% of network traffic used HTTP at the time of data collection, and they could map up to 90% of the network traffic to specific apps. During data collection, although some apps used HTTPS for critical functions, e.g., log-in, most parts of their traffic still used HTTP. Also, we notice most apps use HTTPS in recent years. Some existing studies [8], [9] have demonstrated that app usage traces can also be identified from encrypted data traffic. Overall, the app usage dataset provided by the ISP,

1536-1233 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: CAMBRIDGE UNIV. Downloaded on November 16,2020 at 21:46:11 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2020.3021987, IEEE Transactions on Mobile Computing

TABLE 1
The numuber of apps and usage records for each app category.

No.	Category	# of apps	# of usage records	No.	Category	# of apps	# of usage records
1	Game	342	44,553,941	13	Photography	26	2,082,726
2	Finance	27	22,421,634	14	Lifestyle	79	55,949,873
3	Stock	106	24,724,632	15	Health & fitness	73	4,616,153
4	Shopping	152	48,973,459	16	Sports	33	7,385,860
5	Parent & child	34	631,729	17	News	86	79,210,799
6	Education	67	7,135,838	18	Reading	98	17,182,331
7	Weather	34	4,181,546	19	Media & video	105	45,994,839
8	Travel	66	5,353,283	20	Music & audio	86	70,127,620
9	Navigation	68	89,816,498	21	Business	59	5,116,745
10	Transportation	78	39,915,957	22	House & home	26	1,412,869
11	SON & IM	185	827,784,029	23	Car	39	4,156,542
12	Food & drink	53	69,783,067	24	Tools & others	78	14,337,945

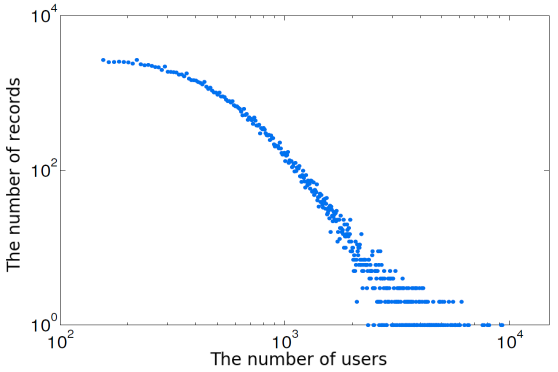
although not covering all traffic, is sufficient for our analysis of mainstreams of usage behavior modeling.

Next, we classify the 2,000 most popular apps into different categories. An app category has an inherent semantic meaning, and apps in the same category are more likely to be involved in a similar cyber activity [5]. As our data is collected from network operators, the dataset includes both Android and iOS users. Due to the difference in app categorization systems of Apple Store (iOS apps of 25 categories) and Google Play (Android apps of 30 categories), we cannot directly adopt the category information of app stores. For some apps, they may belong to different categories in different app stores. For example, Firefox belongs to utilities in Apple Store while it is in the communication category in Google Play. Therefore, we re-categorize apps. We first crawl app descriptions from app stores and generate an app-description matrix using Jieba [10]. With the app-description matrix, we apply the Latent Dirichlet Allocation [11] model to cluster apps and extract their category information. To determine the optimal categories, we gradually vary it from 20 to 30 and manually check the coherence of the outputs. We find that 24 categories are optimal and with the minimum perplexity. We then count the apps and usage records for each app category and show them in Table 1. It is worth noting that, in practice, some apps may fall into multiple categories. For example, the specific category of YouTube may vary depending on the content viewed by users. It belongs to the news category when users watch news videos while belonging to the education category when users watch education videos. However, due to the privacy restrictions, we cannot obtain the information of the content accessed by users. Therefore, in this work, we only consider a hard categorization scheme, where one app solely belongs to one major category. For example, YouTube belongs to the category of media and video in terms of its app description.

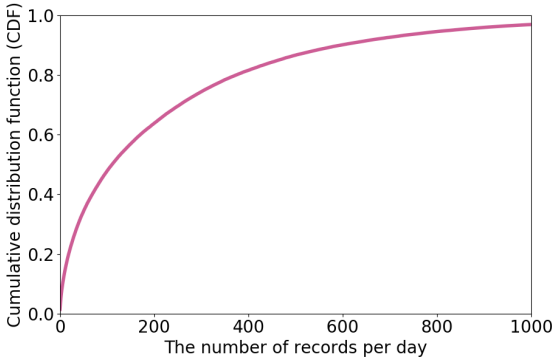
Since we focus our analysis on the collected 2,000 most popular apps, we first filter out the outlier users who do not have any matched app usage. After the filtering, the remaining dataset contains 1,699,386 users and 1,492,849,915 usage records. Fig. 1 shows the statistics of the dataset without outlier users. From Fig. 1(a), we find the records for each unique user still follows the power-law distribution, which represents that we still keep the characteristics of app usage behavior, e.g., the long tail, when we only consider the 2,000 apps. Further, we solely keep those users who are always active during the data collection period. In terms of the statistics, more than 90% of users have average daily records exceeding five, as shown in Fig. 1(b). Hence, we define the active day for each user is the one with at least five app usage records. After these two-step filtering operations, we finally obtain 971,818,946 valid app usage records and 653,092 unique active users. Table 2 presents a summary of the filtered app usage dataset.

TABLE 2
Dataset summary.

# of Records	# of Users	# of Identified Apps	# of App Categories
971,818,946	653,092	2,000	24



(a)



(b)

Fig. 1. This figure shows (a) the records for each unique user. (b) The cumulative distribution function of the average number of records per day.

1536-1233 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: CAMBRIDGE UNIV. Downloaded on November 16,2020 at 21:46:11 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2020.3021987, IEEE Transactions on Mobile Computing

Meanwhile, we are very aware of the privacy implications of using the dataset for research and our research findings. We and the ISP have taken adequate measures to safeguard the privacy of mobile users according to [12]. **First**, the ISP has the consent to collect mobile data, and stripped all the personally identifiable information from the traces and enforced strict non-disclosure agreements for all researchers. The dataset is stored in a server protected by authentication mechanisms and firewalls in the ISP network. **Second**, the ISP only gave us the anonymized user IDs. We never had access to their actual identifiers. Also, we did not have location information of users. The discovered pattern of an individual alone does not leak the privacy, as it cannot be associated with the user's actual ID. **Finally**, this work has received approval from both the ISP and the authors' local institute.

3 Discovery of User Activities

In this section, we identify users' cyber activities based on their app usage traces logged over a whole week. For each user, we first divide her/his app usage traces into several small time windows to capture short-term activities. By using a probabilistic topic model, we obtain the activity label for each window. We then determine the semantic terms of each cyber activity in terms of internal ranking and external ranking of app categories. Finally, we recognize seven unique activities across large-scale app usage data.

3.1 App Usage Trace Representation

In this paper, we assume that users' app usage records reflect their cyber activities. To capture users' short-term activities, we first divide app usage traces of each day into multiple small time windows. In practice, we use a time-based segmentation mechanism. Considering the diurnal pattern of app usage [13], we make a trade-off between non-uniform time grids and uniform time grids. In our case, a window refers to an app usage block composed of app usage records during a specific time period.

More specifically, we first look into the cumulative distributions of window size, the number of app usage records for different candidate time grids, as shown in Fig. 2. With half an hour and one-hour time grids, 20% of windows have less than 15 records, which are too short of capturing user activities. On the other hand, setting time grids as four and six hours, the windows become too large where multiple activities will be mixed. Hence, in terms of Fig. 2, the time grid of two hours strikes a balance between having enough app usage records within each window and having more stationary windows during the usual active hours of the day, i.e., 5.00 to 21.00.

Moreover, app usage has a typical diurnal pattern, as shown in Fig. 3. The app usage sharply declines after 21.00, and there is very limited app usage during the usual inactivity period, i.e., 0.00 - 5.00. This is expected, as people usually begin to take rest and sleep during these hours. To get a more stable size for windows, similar to [14], we aggregate the windows of low app usage during these two periods. In the end, we divide app usage traces of each day for each user into ten windows. Therefore, in our dataset, each user has 70 windows, i.e., 7 (# of days) $\times$ 10 (# of windows for each day). Also, we can apply the usual notions to some time slots, such as morning rush hour (7.00 to 9.00), noon break (11.00 to 13.00), and evening rush hour (17.00 to 19.00). After segmentation, we get 45,716,440 windows for the 653,092 unique active users. Among all these windows, only 9,196,661 windows are unique.

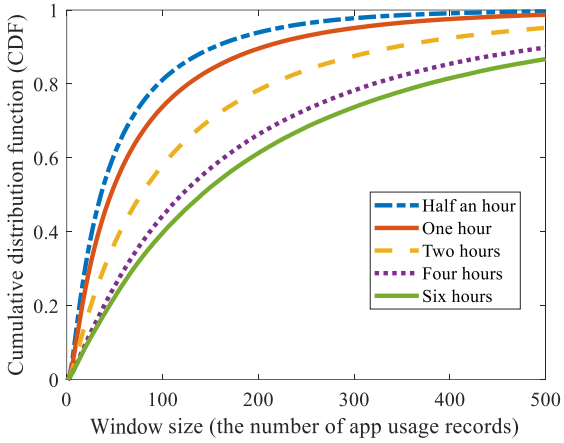


Fig. 2. Quantity of windows with respect to the number of app usage records.

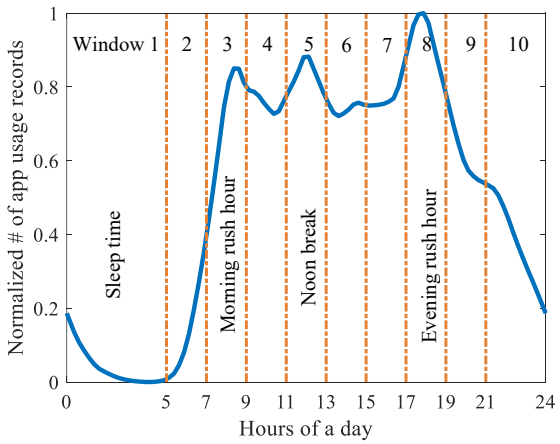


Fig. 3. Min-max normalized number of app usage records during one day.

3.2 Activity Discovery

To characterize the activities of windows, we explore the power of the author-topic model [15]. As a probabilistic topic model, the author-topic model has been successfully used for discovering the hidden topic structure in large documents [15]–[17]. In this model, each document exhibits multiple topic features. Each word of a document supports topics in probability, and the authors of the document determine the mixture weights for different topics as well. Given all words and authors of each document as observations, the author-topic model is trained to infer the hidden topic of each document.

In the author-topic model, the generative model for documents is depicted in Fig. 4. There is one latent variable z for topics. Assuming that there are D documents in the corpus, document d is a sequence of N_d words and has a group of authors $\boldsymbol{a_d}$. x indicates the author responsible for

1536-1233 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: CAMBRIDGE UNIV. Downloaded on November 16,2020 at 21:46:11 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2020.3021987, IEEE Transactions on Mobile Computing

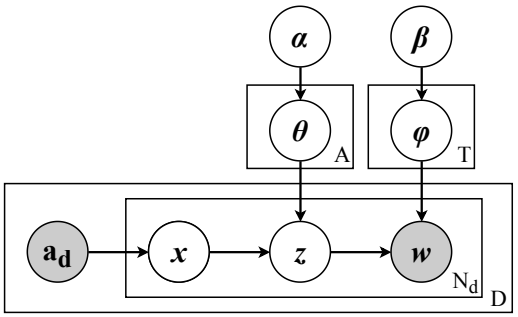


Fig. 4. Generative model for documents. Observations are represented by shadowed nodes.

a given word w_d^i and chosen from $\boldsymbol{a_d}$. Let α and β be the prior parameters for the Dirichlet author-topic distribution and topic-word distribution. φ denotes a $T \times V$ matrix of topic distributions, with a multinomial distribution over V vocabulary items, i.e., all words in D documents for each of T topics. Each φ_t is drawn independently from the symmetric Dirichlet(β) prior. θ_a stands for the topic proportion for author a. z denotes the topic responsible for generating the word w. By using the above notations, we illustrated the generative process as follows.

1) For each author in the A authors, choose the vector of topic proportions $\theta_a \sim$ Dirichlet(α).
2) For each topic in the T topics, choose the vector of word proportions $\varphi_t \sim$ Dirichlet(β).
3) For each of the word w_d^i in document d,
   - choose an author $x \sim$ Uniform($\boldsymbol{a_d}$);
   - choose a topic $z \sim$ Multinomial(θ_a);
   - choose a word $w_d^i \sim$ Multinomial(φ_z).

Dirichlet($\cdot$) means the Dirichlet distribution. Uniform($\cdot$) represents the uniform distribution, and Multinomial($\cdot$) is the multinomial distribution. The exact inference of the hidden variables is computationally intractable. Therefore, in practice, approximate inference algorithms are commonly used, such as Laplace approximation [18], Variational approximation [19], and Markov chain Monte Carlo (MCMC) [20]. In our work, we use the Variational Bayes inference algorithm [21], a variant of Variational approximation method.

We aim to find the hidden cyber activity structure of app usage windows for the problem of activity discovery. Specifically, each window is a block of app usage traces, represented as a sequence of app IDs. Each window has multiple activity features, and each app usage of a window supports hidden activities in probability. Hence, the relationships among activities, apps, and windows, are highly similar to the relationships among topics, words, and documents. By considering they also have similar objectives, we build an analogy between the activity discovery of windows and the topic discovery of documents. As shown in Fig. 5, a unique window represents a document, the users of a window represent the authors of a document, and an activity represents a topic. A window has multiple activity features, which is just like a document that has various topics. Apps in one window are deemed as words in a document. The vocabulary is the set of all words in documents, while the collected apps are the set of all apps in windows. Thus, the collected apps are regarded as vocabulary. In practice,

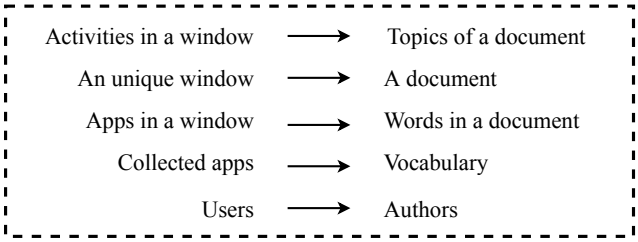


Fig. 5. Analogy between window-activities to document-topics.

the collected apps stand for the app IDs of the 2,000 most popular apps.

Specifically, we applied the author-topic model to extract hidden activity features of app usage windows due to the following three reasons. 1). As shown in Fig. 2, app usage windows vary from 20 records to 500 records. Traditional clustering algorithms, like K-means and hierarchical clustering, have difficulty dealing with inputs of different sizes. Alternatively, the author-topic model has no input size limitation and has proven to perform well for both short texts [22], e.g., twitters, and long texts [16], e.g., articles. 2). As for traditional clustering algorithms, we need to empirically and manually extract features from app usage windows to identify activities. Nevertheless, the author-topic model can characterize the relationships between users, app usage, and cyber activities cohesively and automatically identify activities of app usage windows. 3). The same app usage may imply different activities in different contexts. As a probabilistic topic model, the author-topic model enables app usage to support multiple hidden activities in probability, which solves the semantic ambiguity of app usage.

Like the other topic models [23], [24], the author-topic model requires us to specify the number of topics in advance. Although determining the most appropriate number of topics remains an open issue, we can evaluate the model by measuring how perplexity scores vary with the number of topics. Perplexity is a measurement of how well a probability distribution or probability model predicts samples [25]. Perplexity is a common metric for estimating topic model performance [26]. Mathematically, the perplexity of a set of words, W_d for document d, is defined as follows,

$$\text{Perplexity}(W_d) = \exp\left[-\frac{\ln P(W_d)}{N_d}\right]. \quad (1)$$

Intuitively, perplexity stands for the confusion of the model about its decision. More accurately, perplexity expresses the average number of words that have to be picked to get a correct one, when we randomly choose words from the probability distribution calculated by the author-topic model at each time step [27].

To determine the most appropriate number of topics, we vary it ranging from 2 to 30 and compute perplexity. Fig. 6 shows perplexity relative to the number of topics. The lower value of perplexity implies the better performance of the model. We then find the perplexity score is lowest when the number of topics is 12, while the knee of the curve is where the number of topics is 6. Empirically, the knee of the curve is better. Subsequently, we obtain the cyber activity features of each window by applying the author-topic model and setting the predefined number of topics as 6.

1536-1233 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: CAMBRIDGE UNIV. Downloaded on November 16,2020 at 21:46:11 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2020.3021987, IEEE Transactions on Mobile Computing

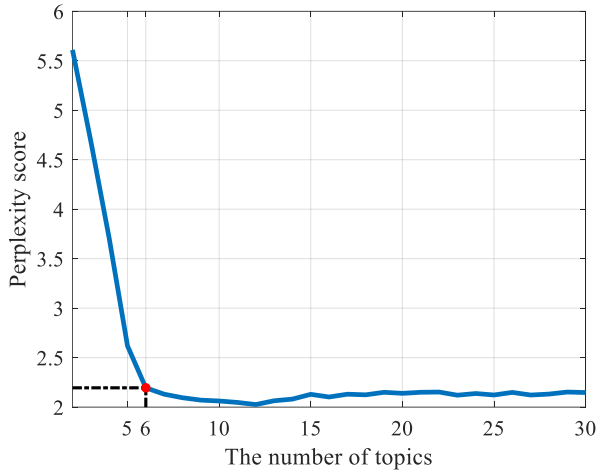


Fig. 6. Perplexity score versus the number of topics. The lower value of perplexity implies the better performance of the model. The knee of the curve is denoted as the red point where the number of topics is 6.

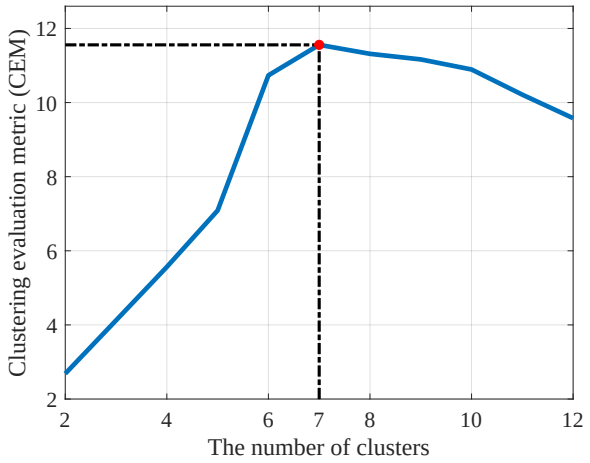


Fig. 7. Clustering evaluation metric versus the number of clusters. The higher the value implies the better the clustering. The optimal clustering is performed with 7 clusters.

3.3 Window Aggregation

To facilitate finding the semantic terms of each window, we take a further step, aggregating similar unique windows in terms of their activity features. Windows from the same cluster have similar activity features, and different clusters represent different activities. For an arbitrary window s, its activity feature is a 6-dimensional vector, i.e., $\theta_s = (\theta_{s,1},\ \theta_{s,2},\ ...,\ \theta_{s,6})$, where $\theta_{s,t}$ is the proportion of activity t for window s.

We perform the Bisecting K-means clustering algorithm on the activity feature vectors of the 9,196,661 unique windows. Compared with other clustering algorithms, such as hierarchical clustering and spectral clustering, only k-means and its variants can handle such large scale windows. Different from basic K-means, Bisecting K-means as a variant that can overcome the problem of getting caught in a local minimum by minimizing the Sum of Squared Errors (SSE) of split clusters [28], [29]. Karypis *et al.* showed that Bisecting K-means outperforms basic K-means in terms of entropy, F measure, and overall similarity [30].

The number of clusters for the Bisecting K-means algorithm needs to be predefined according to the application and determined by measuring the quality of clustering [31]. To evaluate the quality of clustering for different numbers of clusters, K, we defined a clustering evaluation metric (CEM),

$$\text{CEM} = \frac{SP}{CP}, \tag{2}$$

where SP and CP indicate the average separation and compactness of clusters, respectively.

$$\begin{aligned} SP &= \frac{2}{K^2 - K} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} |c_i - c_j|, \\ CP &= \frac{1}{K} \sum_{i=1}^{K} \sum_{\theta \in \Omega_i} \frac{|\theta - c_i|}{|\Omega_i|}, \end{aligned} \tag{3}$$

where Ω_i is the set of members in cluster i, c_i is the centroid of cluster i. SP stands for the average inter-cluster distance. A small value for SP indicates adjacent clusters are similar. Thus, the larger SP, the higher the quality of clustering. On the other hand, CP stands for the average intra-cluster distance. A small value for CP means the less scattering of clusters. As a result, the higher the CEM, the better the quality of clustering. To determine the most appropriate number of clusters, we run the Bisecting K-means algorithm on the Window-Activity matrix and vary the number of clusters ranging from 2 to 12 to compute CEM. Fig. 7 shows how CEM changes with the number of clusters. We then find the optimal number of clusters is 7.

3.4 Activity Identification

After window aggregation, we obtain seven clusters. Windows in the same cluster have similar activities. In this step, we aim to identify each window cluster with semantic terms, i.e., activity labels. Note that activity identification is a very challenging problem. Unlike small scale datasets [32], the large-scale app usage datasets lack meaningful labels mapping user cyber activities to app usage records. Fortunately, we can explore the semantic information of app categories and prior knowledge of activity temporal patterns to identify activity labels. For example, if a user uses food & drink apps during lunchtime, it has a high probability of identifying the activity as exploring food.

In detail, we identify the cyber activity label of a window cluster by considering the following three aspects. 1) The app category configuration in a window cluster. We compute the average proportion of different app categories for each window cluster. According to the calculated proportion, we rank app categories in a window cluster, called *Internal Ranking (IR)*. 2) The window cluster configuration across different app categories. We also rank the window clusters for each app category, called *External Ranking (ER)*. 3) The temporal distribution of windows in each cluster. In Table 3, the top-3 internal app categories for each window cluster are colored by blue from darkest to lightest, while the top window cluster for each app category is colored by red. The seven typical window clusters are identified as follows.

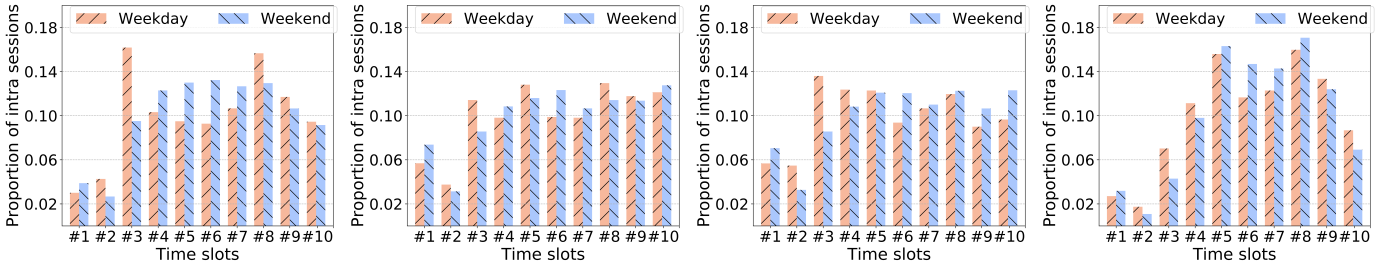
(C1) *Commute and Transportation*. The top-3 app categories in this window cluster are *Music & audio*, *SON & IM*, and *Navigation*. This cluster contains the maximum

1536-1233 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: CAMBRIDGE UNIV. Downloaded on November 16,2020 at 21:46:11 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2020.3021987, IEEE Transactions on Mobile Computing

TABLE 3
Proportion of app categories for window clusters and corresponding internal and external rankings. C: *cluster*, Pro: *proportion*, IR: *internal ranking*.

	Commute & transportation (C1)		Entertainment (C2)		Shopping (C3)		Socializing (C4)		Reading & checking (C5)		Life & health (C6)		Exploring food (C7)	
App categories	**Pro**	**IR**	**Pro**	**IR**	**Pro**	**IR**	**Pro**	**IR**	**Pro**	**IR**	**Pro**	**IR**	**Pro**	**IR**
Game	0.0109	9	0.3435	2	0.0083	16	0.0093	5	0.0127	18	0.0076	10	0.0997	2
Finance	0.0063	13	0.0017	21	0.0127	12	0.0056	10	0.0087	19	0.0074	11	0.0254	6
Stock	0.0027	18	0.0059	12	0.0130	11	0.0024	14	0.0642	5	0.0028	17	0.0047	17
Shopping	0.0127	8	0.0077	11	0.1223	3	0.0141	3	0.0197	10	0.0171	5	0.0191	7
Parent & child	0.0009	24	0.0107	8	0.0093	15	0.0008	23	0.0049	24	0.0015	20	0.0049	16
Education	0.0026	19	0.0005	24	0.0048	22	0.0022	15	0.0151	13	0.0020	19	0.0015	23
Weather	0.0022	20	0.0017	22	0.0049	20	0.0016	18	0.0084	20	0.0014	21	0.0023	22
Travel	0.0016	22	0.0023	19	0.0049	21	0.0009	21	0.0133	17	0.0009	24	0.0038	18
Navigation	0.1812	3	0.0039	15	0.0387	5	0.0085	6	0.0715	4	0.0138	7	0.0346	5
Transportation	0.1322	4	0.0129	7	0.0269	7	0.0036	12	0.0199	9	0.0049	14	0.0148	8
SON & IM	0.2226	2	0.0551	5	0.4335	1	0.8294	1	0.1410	2	0.1141	2	0.0976	3
Food & drink	0.0480	5	0.0729	3	0.1374	2	0.0765	2	0.0432	6	0.0224	4	0.5825	1
Photography	0.0062	14	0.0041	14	0.0071	18	0.0016	16	0.0064	23	0.0010	23	0.0416	4
Lifestyle	0.0080	12	0.0032	17	0.0484	4	0.0060	7	0.0150	14	0.6962	1	0.0113	9
Health & fitness	0.0031	17	0.0098	9	0.0050	19	0.0016	19	0.0140	16	0.0478	3	0.0052	14
Sports	0.0017	21	0.0026	18	0.0073	17	0.0009	20	0.0406	7	0.0142	6	0.0027	21
News	0.0056	15	0.0052	13	0.0237	8	0.0056	9	0.3280	1	0.0038	16	0.0051	15
Reading	0.0046	16	0.0082	10	0.0098	14	0.0016	17	0.0878	3	0.0064	13	0.0071	13
Media & video	0.0103	10	0.3632	1	0.0342	6	0.0052	11	0.0236	8	0.0069	12	0.0076	12
Music & audio	0.2693	1	0.0579	4	0.0138	10	0.0134	4	0.0173	11	0.0090	9	0.0103	11
Business	0.0302	6	0.0208	6	0.0108	13	0.0028	13	0.0149	15	0.0104	8	0.0103	10
House & home	0.0097	11	0.0035	16	0.0040	24	0.0008	22	0.0076	21	0.0024	18	0.0034	19
Car	0.0013	23	0.0010	23	0.0043	23	0.0005	24	0.0151	12	0.0014	22	0.0010	24
Tools & others	0.0260	7	0.0017	20	0.0149	9	0.0058	8	0.0071	22	0.0046	15	0.0033	20



(a) Commute and transportation (b) Entertainment (c) Life and health (d) Exploring food

Fig. 8. The temporal distribution of windows in clusters (C1) *Commute and Transportation*, (C2) *Entertainment*, (C6) *Life and health*, and (C7) *Exploring food*.

number of *Transportation* apps as well. We also investigate the distribution of these windows in the temporal domain. As shown in Fig. 8(a), on weekdays, the windows in this cluster mainly happen in the time slots 3 and 8, i.e., morning rush hour and evening rush hour. The specific time slot division scheme is presented in section 3.1. On weekdays, the number of windows occurring in the evening is higher than that in the afternoon, while it is contrary to weekends. This phenomenon corresponds to the fact that on weekdays people, especially white-collar workers, are 'bound' in workplaces during the afternoon and free from work on weekends. According to the app category features of this cluster, we also infer that people prefer to listen to music and check emails during commute time. It is not surprising that navigation apps are highly used during transportation activities as well.

(C2) *Entertainment*. This cluster contains typical entertainment activities with the highest both internal ranking and external ranking of *Media & video* and *Game*, as shown in Table 3. Fig. 8(b) shows the temporal features of the windows in this cluster. On weekdays, there are three time slots of the highest proportion of windows, namely time slots 3 (morning rush hour), 5 (noon break), and 8 (evening rush hour). These three time slots are the main 'free time' for people on weekdays. On the other hand, on weekends, the time slots in the evening account for the highest proportion. Especially for the time slot 1, from 0.00 to 5.00, the proportion on weekends is much larger than weekdays, because people do not need to get up early and go to work on weekends.

(C3) *Shopping*. This cluster mainly has shopping activities with the most *shopping* apps. There is no significant difference in its temporal features on weekdays and weekends. Like the general pattern of user behavior, the proportion increases during the day and decreases during the night. In the windows of this cluster, the socializing apps, i.e., *SON & IM* category, account for the highest proportion in the internal ranking. We infer that people usually browse the products in online shops and share them with friends via socializing apps to ask for suggestions and comments.

(C4) *Socializing*. This cluster is identified as the social activity since not only the *SON & IM* category has the highest

1536-1233 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: CAMBRIDGE UNIV. Downloaded on November 16,2020 at 21:46:11 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2020.3021987, IEEE Transactions on Mobile Computing

internal ranking and external ranking but also the other app categories are of lower proportions compared with other clusters (see Table 3). In other words, the windows in this cluster are 'pure' social states. Its temporal features are similar to the *Shopping* activity, peaking during the daytime while reducing over the night period.

(C5) *Reading and checking*. The most characteristic app categories in this window cluster are *News*, *Reading*, *Sport*, *Weather*, *Stock*, and *Education*, with a significant higher proportion than other window clusters. These app categories are all about reading and checking activities. It seems that people are habituated to using these kinds of apps in the same period. Besides, its temporal features are similar to the *Shopping* activity as well.

(C6) *Life and health*. In this window cluster, the typical app category features are *Lifestyle* and *Health & fitness*, which are with both high external ranking and internal ranking as presented in Table 3. The temporal distribution of this cluster's windows is shown in Fig. 8(c). It is clear that, on weekdays, the windows are concentrated in the morning, especially in the time slot 3, from 7.00 to 9.00. In contrast, on weekends, time slot 10, from 21.00 to 24.00, accounts for the most significant part of windows. Since apps of *Health & fitness* category are mainly used to record users' exercises such as jogging, swimming, and keep-fit, we infer that people prefer to take workouts in the morning on weekdays while in the evening on weekends.

(C7) *Exploring food*. Table 3 shows that the windows in this cluster are of the maximum proportion of *Food & drink* app usage both in external ranking and internal ranking. We then look into its temporal features, as shown in Fig. 8(d). We observe that both on weekdays and weekends, time slot 5 (lunchtime, from 11.00 to 13.00) and time slot 8 (dinner time, from 17.00 to 19.00) take up the largest part of windows. Also, *SON & IM* and *Game* are with a high ranking in this cluster. We guess this is because people share restaurant localization with friends via social apps and like playing mobile games while waiting for dishes. We also notice *Finance* in this cluster is with a significantly higher proportion than other clusters, which suggests that people pay their orders by electronic payment methods like Paypal.

4 Discovery of Activity Patterns

Activity patterns are of significant value for both individuals and society. For individuals, service providers can provide personalized service by exploring users' lifestyles, habits, occupations, and socio-economic status from their activity patterns. For society, the government can understand people's living status and detect disrupting trends from activity patterns and then make policies to improve people's well-being. Specifically, in this work, we sequence users' cyber activities[1] identified from their app usage traces and apply sequence analysis methods to extract activity patterns.

4.1 Similarity Measurement of Activity Sequences

After identifying users' activities, the next goal is to use the collected one-week individuals' behavior to determine daily cyber activity patterns. To do so, we treat users' app usage traces of one day as an incidence of sequential activities. Apart from the discovered seven activities, we add an *Unknown* label for those silent time slots during which no app usage records are observed and denote it as $C8$. For each user, each day app usage traces reveal a cyber activity sequence of length ten that contains combinations of the eight general activities including *Commute and transportation*, *Entertainment*, *Shopping*, *Socializing*, *Reading and checking*, *Life and health*, *Exploring food*, and *Unknown*. Hence, each user's activity sequence can be expressed as,

$$A_u = \{[a_1^{d_1}, a_2^{d_1}, ..., a_{10}^{d_1}], ..., [a_1^{d_7}, a_2^{d_7}, ..., a_{10}^{d_7}]\}, \quad (4)$$

where A_u stands for the activity sequence of user u and $a_m^{d_n}$ denotes the activity label of window m in the n-th day, $a \in \{C1, C2, C3, C4, C5, C6, C7, C8\}$.

To quantify the degree of similarity among activity sequences, we apply the string metric. Each user's activity sequence for one day is regarded as a string, which is a combination of eight kinds of characters, i.e., activities. Particularly, in our work, we use the Levenshtein distance metric [33] which has been widely used in the social analysis [34], information theory [35], linguistics [36], and computer science [37]. In [38], William W. C. *et al.* compared different string metrics and showed the Levenshtein distance is better than others. The Levenshtein distance is also be referred to as edit distance. Given two sequences, the Levenshtein distance is the minimum number of single-character edits, including insertions, deletions, and substitutions, required to transform one sequence into another.

Compared with another commonly used distance metric, i.e., Hamming distance, Levenshtein distance is more suitable for our measurement to capture the sequential patterns. Hamming distance does not use insertions and deletions [39]. It only uses substitutions and is only possible to compare when activities occur, while Levenshtein distance takes insertions and deletions into account and can capture the order in which user activities are organized over time. We then justify that with a practical example. We are given three sequences: 'C8C1C5C4C1C2', 'C2C8C1C5C4C1' and 'C8C2C2C2C2C2', where C denotes the discovered activities which are presented in Table 3. Since our goal is to investigate how cyber activities are sequenced throughout a day, i.e., sequential features, the distance between 'C8C1C5C4C1C2' and 'C2C8C1C5C4C1' should be smaller than the distance between 'C8C1C5C4C1C2' and 'C8C2C2C2C2C2'. Note that, for 'C8C1C5C4C1C2' and 'C2C8C1C5C4C1', there is only one time slot shift. The Hamming distance and Levenshtein distance between 'C8C1C5C4C1C2' and 'C2C8C1C5C4C1' are 6 and 2 respectively, while both Hamming distance and Levenshtein distance between 'C8C1C5C4C1C2' and 'C8C2C2C2C2C2' are 4. Hence, we apply Levenshtein distance to measure the similarity of activity sequences. Particularly, the two edits to change 'C8C1C5C4C1C2' into 'C2C8C1C5C4C1' for Levenshtein metric are shown as follows,

1) C8C1C5C4C1C2 ⇒ C8C1C5C4C1 (deletion of 'C2'),
2) C8C1C5C4C1 ⇒ C2C8C1C5C4C1 (insertion of 'C2' at the beginning).

1. Apart from discovered seven activities, we add an Unknown label to denote silent time slots.

1536-1233 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: CAMBRIDGE UNIV. Downloaded on November 16,2020 at 21:46:11 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2020.3021987, IEEE Transactions on Mobile Computing

4.2 Individuals' Activity Analysis

We first investigate the similarity of different days to examine the regularity of activities in days' scale. As shown in Table 4, we calculate the average Levenshtein distance between two days in pairwise. Given day i and day j, their distance is computed as,

$$\frac{1}{U}\sum_{u=1}^{U} \text{lev}(\mathbf{A}_u^{d_i}, \mathbf{A}_u^{d_j}), \tag{5}$$

where U is the number of unique users and $\mathbf{A}_u^{d_i}$ denotes the activity sequence of the i-th day for user u.

We notice that there is an apparent difference between weekdays' sequences and weekends' sequences because the average distance between any two weekdays is less than that between any weekday and weekend. Besides, we discover an interesting appearance that the activity sequence of a weekday is more similar to yesterday's sequence. This implies that *people intentionally or unintentionally obey yesterday's activity sequence, and there should be a daily pattern of activities for individuals.*

We further look into the coherence of each user's activity sequences on weekdays. As shown in Fig. 9, we compute the average intra-distance for each user,

$$\frac{2}{D^2 - D}\sum_{i=1}^{D-1}\sum_{j=i+1}^{D} \text{lev}(\mathbf{A}^{d_i}, \mathbf{A}^{d_j}), \tag{6}$$

where D is the number of weekdays, here $D = 5$, and $\mathbf{A}^{d_i}$ stands for $[a_1^{d_i}, a_2^{d_i}, ..., a_{10}^{d_i}]$ namely the activity sequence of day i. We observe that nearly 74% of users have the average intra-distances less than 5, implying that they repeat at least half of their daily activities in cyberspace. Therefore, a large part of individuals' daily lives follows a regular daily activity pattern. We give the definition of the daily activity pattern of an individual as follows.

Definition 1. *Daily activity pattern of an individual. Given an individual's activity sequences on weekdays,* $\mathbf{A}^{d_1}$, $\mathbf{A}^{d_2}$, ..., $\mathbf{A}^{d_5}$, *the daily activity pattern of the individual,* $\mathbf{A}$, *has the minimum sum-distance between all pairs of* $\mathbf{A}$ *and* $\mathbf{A}^{d_i}$. *Mathematically, denoting* $\mathbf{A} = [\mathbf{a_1}, \mathbf{a_2}, ..., \mathbf{a_{10}}]$, *then*

$$\mathbf{A} \leftarrow \arg \min_{\mathbf{a}\in\{\mathbf{C1},\mathbf{C2},...,\mathbf{C8}\}} \sum_{\mathbf{i=1}}^{\mathbf{5}} \mathbf{lev}(\mathbf{A^{d_i}}, \mathbf{A}). \tag{7}$$

4.3 Identifying Common Activity Patterns

We have examined the existence of daily activity patterns in cyberspace for individuals. Next, we investigate whether there are common activity patterns for millions of users. To do this, we first quantify the distance between each pair of daily activity patterns for all users. Once the distance matrix is calculated, we apply the agglomerative hierarchical algorithm to identify homogeneous clusters of daily patterns. In terms of existing studies [30] conducted on labeled datasets, the agglomerative hierarchical algorithm usually has a better performance compared with bisecting K-means. Also, there are 654,092 users, which satisfies the scalability limitation of hierarchical clustering algorithms. Hence, instead of bisecting K-means, we employ the agglomerative hierarchical algorithm for common pattern discovery. In detail, the agglomerative hierarchical algorithm is a 'bottom-up' approach in which each object starts in its own cluster, and the pairs of clusters are merged as one moves up the hierarchy [40].

TABLE 4
Average Levenshtein distance between arbitrary two days. We round numbers to two decimals.

	MON	TUE	WED	THU	FRI	SAT	SUN
MON	/	4.26	4.37	4.42	4.58	5.26	5.18
TUE	/	/	4.37	4.48	4.59	5.29	5.21
WED	/	/	/	4.28	4.49	5.07	5.01
THU	/	/	/	/	4.47	5.27	5.29
FRI	/	/	/	/	/	5.27	5.31
SAT	/	/	/	/	/	/	4.70

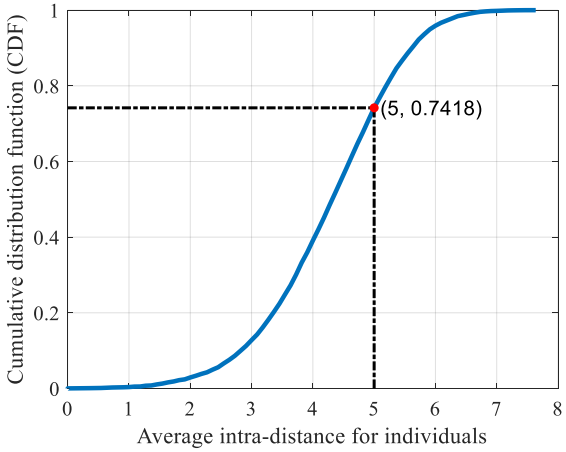


Fig. 9. Cumulative distribution of users with respect to the intra-distance. The average intra-distances of nearly 74% users are less than 5.

To determine the most appropriate number of clusters, i.e., patterns, we apply the dendrogram to evaluate the agglomerative hierarchical clustering algorithm, as shown in Fig. 10. The dendrogram is a branching diagram representing the hierarchy of clusters based on the degree of similarity [41]. As Fig. 10 shows, the distance from the root to a subtree indicates the similarity of subtrees. Highly similar nodes or subtrees have joining points farther from the root. We know how the nodes are combined into larger parent clusters from the dendrogram, i.e., the detailed clustering process. In Fig. 10, we find five is the most appropriate number of clusters, where the clusters are of high intra-cluster and low inter-cluster distances. The five clusters are boxed using orange lines.

4.4 Pattern Annotation

Given the clustering results, we then annotate each cluster of daily activity patterns with semantic terms, which will contribute to understanding the hidden image of these patterns. We first visualize them by randomly selecting fifty users for each cluster and show how their cyber activities are sequenced. As shown in Fig. 11, the x-axis refers to the windows, and the y-axis indicates the random fifty users. Each bin refers to the activity label of that window for that user, and we use different colors to distinguish different activities. The distribution of all users among clusters is shown in Table 5.

1536-1233 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: CAMBRIDGE UNIV. Downloaded on November 16,2020 at 21:46:11 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2020.3021987, IEEE Transactions on Mobile Computing

TABLE 5
The number and proportion of users for each cluster.

	Cluster A	Cluster B	Cluster C	Cluster D	Cluster E
Number of users	55,513	120,495	49,374	191,291	236,419
Proportion of users	8.50%	18.45%	7.56%	29.29%	36.20%

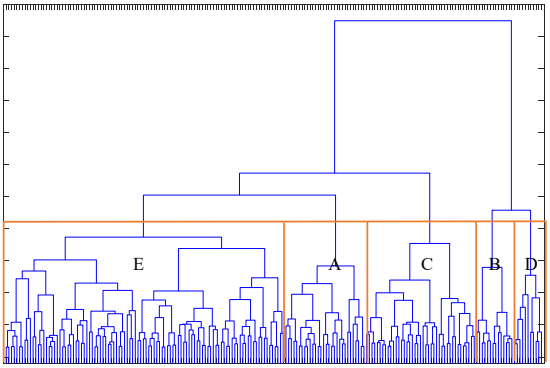


Fig. 10. Dendrogram of the hierarchical clustering where highly similar nodes or subtrees have joining points farther from the root. Five is the most appropriate number of clusters, where the clusters are of high intra-cluster and low inter-cluster distances.

Afternoon reading (Cluster A). The users in this cluster are mostly involved in *Reading and Checking* during the day, as shown in Fig. 11(a). The users, on average, start to use apps from time slot 4, 9.00 to 11.00, and become to be inactive after time slot 8, around 19.00. Hence, we annotate this cluster as afternoon reading to reflect the main active periods and activity of this group. Although *Reading and Checking* activity dominates during time slots 4 to 7, there are still many users like *Shopping* during these hours. Generally, both two activities are leisure activities. We still notice that there are several *Commuting and Transportation* activities in this cluster. However, unlike Cluster D, *Commuting and Transportation* activities in this cluster are randomly distributed over time slots. Therefore, the users in this cluster do not have regular commute schedules, e.g., on and off work. Moreover, by considering the dominating pattern of reading and shopping activities, we infer the users in this group are senior citizens. Recalling the statistics in Table 5, they represent 8.5% of the total users, which is similar to the proportion of senior citizens, 10.1%, in Shanghai.

Nightly entertainment (Cluster B). Fig. 11(b) visualizes the daily patterns of this cluster. The users are engaged in *Entertainment* activities during evening and night, from 17.00 to 24.00. Hence, we annotate this cluster as nightly entertainment. Compared with cluster A, C, and E, the users have fewer activities during the usual active hours, i.e., from 7.00 to 15.00. Also, the users in this cluster are mostly nocturnal, and they are more than 18% of all the users. We infer that they are likely the younger generation. Due to the daytime classes, they only have free time in the evening and night, which may be why their app usage is so sparse during the day time. Besides, the daily patterns show a high number of users in this cluster sleep late, still active in the time slot 1, from 0.00 to 5.00. Most of their *Entertainment* activities last more than 6 hours, which indicates that the younger generation is addicted to the *Entertainment* activity, e.g., mobile games, and it is harmful to their health.

Pervasive socializing (Cluster C). As shown in Fig. 11(c), the users in this cluster are engaged in the *Socializing* activity from 7.00 till 24.00. Compared with the patterns of other clusters, the patterns of this cluster are more regular concerning the active time of users and the duration of the dominating activity. This unusual pattern of social activities of nearly 7.5% users can be explained as follows. With modern social networking apps, social activities are not limited only to known friends and families. Peoples are making friends and communicating with people from different social classes via social apps. The social platforms are not only for interaction but also for various businesses, such as advertising and self-media. For example, on WeChat, people advertise and sell their products via instant messages, video calls, group chats, and WeChat Moments. This method [42] utilizes the business relationship and friendship to maintain the customer relationship and is called the WeChat business. Therefore, we suspect that the users in this cluster work in call centers, customer services, or they are bloggers, cyberspace writers, and online shop owners.

Commuting (Cluster D). The patterns in this cluster shown in Fig. 11(d) are typical commuting patterns. The *Commuting and Transportation* activities are sequenced regularly and mainly occur during rush hours. Due to the regular commute patterns, we infer most people in this cluster are involved in white-collar jobs. Meanwhile, we find an important phenomenon. In the morning, nearly 90% of *Commute and Transportation* activities happen in the morning rush hour. However, these activities are spread over multiple time slots around the evening rush hour. This phenomenon implies that many workers cannot knock off on time, and even working overtime becomes a habitual pattern for them. Like cluster B, most of the users are with limited activities from 9.00 to 17.00, as they are busy at work and do not have time to use smartphones.

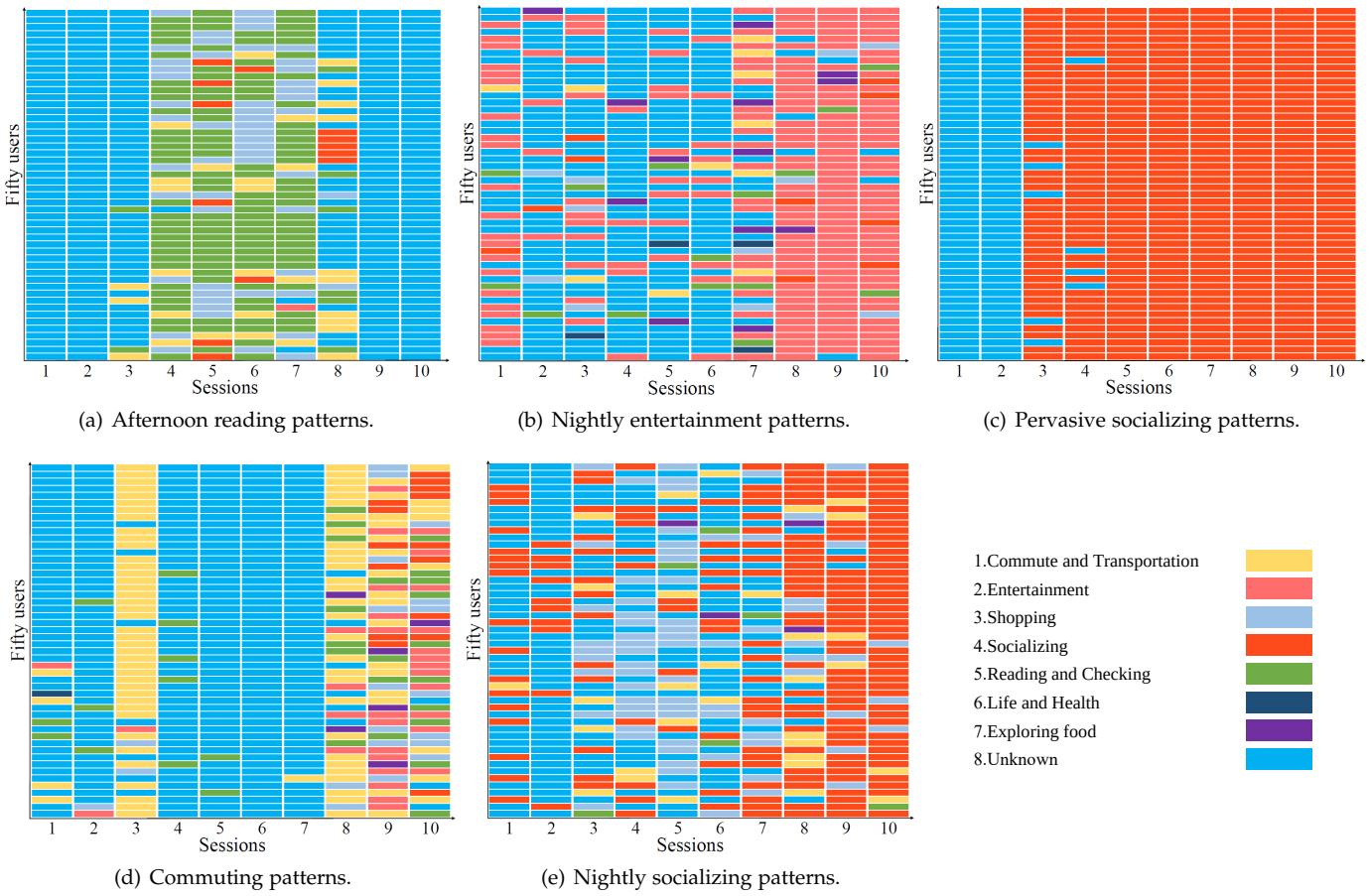
Nightly socializing (Cluster E). The patterns of this cluster are shown in Fig. 11(e). It is the largest cluster having 36.2% of users and the most diverse patterns as well. The dominating activity is *Socializing*, which mostly happens after the evening, i.e., after time slot 6. The users in this cluster are mostly active during the day and engaged in other activities, such as *Shopping*. This implies their time is more flexible compared to other cluster users. The lack of sufficient commuting suggests that people are mostly staying at or near home, involved in household work during the day time, and socializing in the evening. Hence, we infer the users in this cluster should be self-employed.

4.5 Verification through Controlled Study

To validate the detected activity patterns, we make a controlled study and collect a small-scale app usage dataset from 100 users with occupation information. The small-scale app usage dataset is over one week, i.e., seven days and users are in the age range from 25 to 55. Users' occupations

1536-1233 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: CAMBRIDGE UNIV. Downloaded on November 16,2020 at 21:46:11 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2020.3021987, IEEE Transactions on Mobile Computing



(a) Afternoon reading patterns. (b) Nightly entertainment patterns. (c) Pervasive socializing patterns.

(d) Commuting patterns. (e) Nightly socializing patterns.

Fig. 11. Visualization of daily activity patterns by randomly selecting fifty users in each cluster. Each row represent the activity patterns for one user. The windows correspond to the time segments illustrated in Fig. 3.

TABLE 6
The average distance matrix between common daily activity patterns and occupations, where the corresponding occupations and clusters are highlighted.

Distance: Occupation \ Cluster	Afternoon Reading (Cluster A)	Nightly entertainment (Cluster B)	Pervasive socializing (Cluster C)	Commuting (Cluster D)	Nightly socializing (Cluster E)
White-collar workers	6.6904	7.2043	7.4817	5.5983	6.6183
Socializers	7.5787	8.2933	3.5213	7.9347	5.7453
Freelancers	7.1092	7.2277	5.4939	6.7462	5.0953

are in three categories, i.e., white-collar workers, socializers, and freelancers. Specifically, the white-collar workers include clerks, engineers, teachers, editors, lawyers, and others. The socializers consist of bloggers, advertisers, cyberspace writers, online shop owners, and others.

Applying our proposed cyber activity discovery approach, we obtain the activity sequences of that 100 users. We then compute the average distance matrix between common daily activity patterns and occupations, as shown in Table 6. Mathematically, the average distance is calculated as,

$$\frac{\sum_{u\in\mathbf{U}}\sum_{\hat{u}\in\hat{\mathbf{U}}_\mathbf{i}} \mathrm{lev}(A_u, A_{\hat{u}})}{|\mathbf{U}|\cdot|\hat{\mathbf{U}}_\mathbf{i}|}, \tag{8}$$

where A denotes the activity sequence, $\mathbf{U}$ is the set of users for each occupation category, and $\hat{\mathbf{U}}_\mathbf{i}$ is the set of users in cluster i.

From Table 6, we find that white-collar workers, socializers, and freelancers have the lowest average distance with cluster D, C, and E, respectively. Since white-collar workers go work and back home in the working days, it is reasonable that their activity sequences belong to the commuting pattern. Similarly, socializers are usually engaged in social-economic activities for a whole day and of the pervasive socializing pattern. As for freelancers, they have more freedom to arrange their time so that their activity pattern is the most diverse, similar to the nightly socializing pattern. Besides, due to the age bias of the small-scale dataset, we do not find the groups of small distances with cluster A (Afternoon reading) and cluster B (Nightly entertainment). This reflects the correctness of our former inferences to some extent as well.

In summary, these results are consistent with our annotation of the discovered activity pattern. These results also demonstrate the potential of our model to leverage the

1536-1233 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: CAMBRIDGE UNIV. Downloaded on November 16,2020 at 21:46:11 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2020.3021987, IEEE Transactions on Mobile Computing

features of app usage in cyber activity profiling and social status inferring.

5 Discussion

We now revisit the research questions presented in the introductory part and present answers based on our methodology and empirical observations.

- What activities can be discovered from app usage data? *We have discovered a set of seven dominant activities and provided these seven activities with meaningful labels based on their temporal patterns and the semantic information of app categories. The discovered activities are commute and transportation, entertainment, shopping, socializing, reading and checking, life and health, and exploring food options.*
- What common patterns we share with others in our daily cyber activities? *We have examined the existence of daily activity patterns for individuals and found a set of five common regular activity patterns for different groups of people. More specifically, the afternoon reading pattern for senior citizens, the nightly entertainment pattern for the younger generation, the pervasive socializing pattern for socially active people, the commuting pattern for white-collar workers, and the nightly socializing for freelancers.*

In this study, we answered these two fundamental research questions. We also discuss how the discovery of such patterns may help to understand human behavior and further improve the quality of user experience and contribute to the well-being of people. In this section, based on our findings of both activities and daily patterns, we will discuss the implications for policymakers, researchers, service providers, and app developers.

5.1 Implications for Policymakers

Mobile app usage data and the insights on human behavior are relevant and vital for policymakers. The data analysis can provide valuable feedback regarding the activities and well-being of citizens. The insights are expected to contribute to better decisions on multiple scales: from cities and urban planning to the level of individual companies and environments. Traditional methods, like survey and questionnaire techniques require significant personnel resourcing and cannot address the need for timely data and insights. Thus, app data analysis provides an alternative methodology that can be near real-time and low cost.

Through smartphone app data analysis, social issues can be detected and traced at an early stage. For example, in March 2019, an 'anti-996' protest was launched via GitHub[2], followed by over 250 thousand people against working overtime.

We have analyzed working patterns and examined signs of overtime with our smartphone app dataset. We have detected that approx. 35% of workers tend to stay late at the office following a habitual pattern of working overtime. Similarly, we find that a significant proportion of the younger generation has addicted to the *Entertainment* activity. The observations indicate that policymakers and employers can use smartphone data for providing early advice and counseling when working overtime or becoming addicted to games is becoming an addiction.

5.2 Implications for Researchers

The identified seven activities and five common daily cyber activity patterns indicate that app usage records reflect user activities in real life. The usage records provide a rich basis for research on multiple levels of abstraction from the analysis of specific apps and app categories to user activities and habits.

We show that demographics have a significant impact on users' app usage and activities. This may result in biased results and may explain the inability to replicate results across studies, as mentioned in [43]. For example: since our small-scale validation dataset does not cover senior citizens and teenagers, we could not discover the groups associated with the afternoon reading and nightly entertainment patterns that are presented in the large-scale dataset. This issue can be mitigated by understanding the demographics aspects of the study at hand and taking this information into account when designing the data gathering and analysis.

5.3 Implications for Service Providers

Smartphone app data analysis is expected to provide valuable insights for mobile service providers. They can configure and optimize their services in a dynamic manner according to the discovered patterns and behavior. For example, the usage patterns show when the high bandwidth apps are in use of the day, allowing the provider to optimize service delivery and reduce operating costs predictively. The data is also very important for generating recommendations to users based on their observed app usage.

As another example, the nightly entertainment pattern implies that low latency network service is essential in the evening and night since game and media & video apps dominate this pattern. However, in terms of our findings in section 4, there are nearly 42% of users in cluster B addicted to mobile games, still involved in entertainment activity after midnight. A recent report [44] suggests that gaming and smartphone addiction have significant detrimental effects on physical and mental well-being. The app data analysis can highlight these problems through early detection.

5.4 Implications for App Developers

Our work also provides useful information for app developers. For example, during the step of activity identification, we found several highly related app pairs, such as *Navigation* and *Transportation*, *Game* and *Media & video*, *Shopping* and *SON & IM*, *Food & drink* and *Photography*. These observations help application developers in making their apps more intuitive and user-friendly by grouping and merging functions of highly related applications. For example, WeChat, one of the most popular apps in China, already supports such grouped functions through mini-programs[3]. Mini-programs are embedded in WeChat, and they enable the user to access other apps' services effectively.

2. https://996.icu/#/en_US

3. https://wechatwiki.com/wechat-resources/wechat-mini-program-epic-tutorial-guide/

1536-1233 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: CAMBRIDGE UNIV. Downloaded on November 16,2020 at 21:46:11 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2020.3021987, IEEE Transactions on Mobile Computing

5.5 Study Limitations

We have examined the smartphone app usage data gathered from the urban area. However, the activity patterns might be different for suburban and rural areas. The presented work is based on mobile app usage data; thus, it is only possible to detect cyber activities performed while using and carrying smartphones. It is also possible that the app usage records may not reflect the actual engagement of users' activities, as some apps may be executed automatically in the background. For example, an email app may download emails even when the user is not using that app. The basic model can be improved by introducing a user attention mechanism.

6 Related Work

6.1 App Usage Patterns

A lot of previous works or analyses focused on how individuals use their smartphones and their applications. For example, Falaki *et al.* [45] used detailed traces from 255 users to characterized user activities and found that users interact with their smartphones between 10 to 200 times per day on average and use 10-90 applications. Xu *et al.* [46] investigated the diverse usage patterns of smartphone apps via network measurements from a tier-1 cellular network provider in the US. They found that some apps have a high likelihood of co-occurrence across smartphones, that is, when a user uses one app, he or she is also likely to use another one. In our work, we also found similar results. For example, recalling *Commute and transportation* activity presented in section 3, we found that people listen to music and check emails along with the transportation apps while commuting. We also discovered that people are habituated to using news, reading, and weather apps during the same period in *Reading and checking* activity. Yang *et al.* [47] collected continuous cellular traffic over a week to characterize user behavior on mobile Internet. They showed that a user visits seven applications over a week and five categories of applications within a day on average. Canneyt *et al.* [5] collected a sample of Flurry data which consists of events from 600 million daily unique users and covers users from 221 countries. They showed how application usage behavior is disrupted through major political, social, and sports events.

Some existing studies clustered users into several particular groups and provide comprehensive descriptions for these groups. Jones *et al.* [2] analyzed users' application revisitation patterns based on three months of application launch logs from 165 users and identified three distinct user clusters, checkers, waiters, and responsives. Checkers refer to the users exhibit brief revisit patterns of fast revisitation (less than one hour). Waiters stand for the users who show longer revisit patterns, which are uniformly distributed between short-medium re-visitations (between 1min and 4hrs) and long re-visitations (from 2hrs to 3days). Responsives are the users who sometimes exhibit brief and occasionally long revisit patterns. Zhao *et al.* [4] analyzed one month of application usage from 106,762 users and discovered 382 distinct types of users based on their usage behaviors in app-category granularity. They also gave a meaningful label to the users in each cluster, such as Night communicators, Evening learners, and Financial users. In [48], Katevas *et al.* collected daily mobile phone activity data from 340 users and revealed five smartphone use profiles, i.e., limited use, business use, power use, and personality-& externally induced problematic use.

Unlike grouping users of similar application usage habits, Welke *et al.* [49] showed the significant diversity of application usage among users. They demonstrated that it is possible to differentiate users according to their application usage and found out that 500 most frequent applications are sufficient to identify 99.67% of the users. Further, Tu *et al.* [3] quantified the uniqueness of individual app usage and showed that the fingerprints of mobile app usage are highly unique. They also found user demographics, users' online, and offline behavior all influence the uniqueness level.

Since different demographic attributes can lead to differences in app usage behavior, many studies have sought to study the relationship between user personality traits and their app usage traces. For example, Seneviratne *et al.* [50] collected an app usage dataset from over 200 users and exploited linear support-vector machine to predict users' gender based on the apps used by users. Further, Malmi *et al.* [51] conducted a similar study on a more extensive app usage dataset covering 3,760 mobile users and demonstrated that, apart from gender, the app usage traces can be used to predict income as well. Zhao *et al.* [52] collected an app usage dataset from 15,000 mobile users. They extracted topic features from app descriptions and then applied the topic features of used apps to infer users' gender.

6.2 App Usage Prediction and Recommendation

Some scholars also worked on modeling mobile app usage and tried to predict which applications will be launched and recommend accordingly. Given the app usage sequences, it is often assumed that Markovian property stands. Zou *et al.* [53] proposed using Markovian models to learn the app usage sequences. They compared first and second-order Markov models with the weighted linear combination of these two models. The results showed that the combined model is the best model with an accuracy of 85% in predicting the top-5 apps each time of the sequence. Besides, Natarajan *et al.* [54] proposed a cluster-level Markov model to make personalized app usage prediction to a user. It first clusters users based on their usage patterns and then computes personalized PageRank for users corresponding cluster Markov graph.

Apart from only using app usage sequences, some works computed app usage features separately for different temporal contexts. In [55], Verkasalo et al. studied how mobile services are used in different contexts and found that time of day is the most useful context for app usage prediction. Liao *et al.* [56] proposed a temporal-based apps predictor to dynamically predict the apps that are most likely to be used. They extracted three Apps usage features, i.e., global, temporal, and periodical, from the Apps usage trace. Later they dynamically derive an app usage probability model from estimating the current usage probability of each app in each feature. Other studies [57], [58] applied similar approaches, i.e., taking time-based features into account

1536-1233 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: CAMBRIDGE UNIV. Downloaded on November 16,2020 at 21:46:11 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2020.3021987, IEEE Transactions on Mobile Computing

to predict app usage dynamically. Like time features, the activity features derived in our work can be used as context, which helps to predict app usage.

With a large pool of apps currently available and the fast proliferation rate of new apps, app recommendation as an important topic also attracted many researchers. Shi *et al.* [59] proposed a similarity-based recommendation algorithm that measured the similarity of apps based on the usage patterns found among a group of users. Yan *et al.* [60] presented the AppJoy system to make personalized recommendations by picking the apps with similar usage patterns to a user's installed apps. Bae *et al.* [61] made app recommendations based on co-occurrence in usage behavior. Meanwhile, in our work, we also find several high related app pairs, like *Navigation* and *transportation*, *Game* and *Media & video*, *Shopping* and *SON & IM*, *Food & drink*, and *Photography*, which can provide prior knowledge for app recommendation systems. Besides, the contextual information, e.g., users' mobility status, location, and time of the day, is also useful for app recommendations. Davidsson *et al.* [62] combined both the context and user feedback to develop the app recommendation system. Kaji *et al.* [63] developed an app, AppLocky, which requests users to select their current context for context-aware app recommendations. The activity labels derived in our work can also be applied in recommendation systems as a kind of contextual information.

7 Conclusions

Although we are overwhelmed by the chaotic flow of everyday obligations, we have patterns in our digital activities that characterize our daily lives individually and collectively. In this paper, we designed a probabilistic topic model based activity detection framework for discovering daily activity patterns across mobile app usage data. By applying our framework on a large-scale and real-world dataset collected from Shanghai, one of the world's largest cities, we identified seven typical activities, i.e., commuting and transportation, entertainment, shopping, socializing, reading and checking, life and health, and exploring food. From users' activity sequences, we examined the regularity of individuals' daily activities and successfully extracted five common patterns among millions of people, including afternoon reading, nightly entertainment, pervasive socializing, commuting, and nightly socializing. We also showed that demographics have an important impact on users' daily lives. Finally, we demonstrated how our findings could be used by policymakers, government, researchers, service providers, and app developers. In this work, the patterns discovered are based on the dataset collected from only one city, i.e., Shanghai, which might not represent the whole community. Therefore, verifying the findings and comparing the differences across different cities is an exciting future direction.

References

[1] Tong Li, Mingyang Zhang, Hancheng Cao, Yong Li, Sasu Tarkoma, and Pan Hui. what apps did you use?: Understanding the long-term evolution of mobile app usage. In *Proceedings of The Web Conference 2020*, pages 66–76, 2020.

[2] Simon L. Jones, Denzil Ferreira, Simo Hosio, Jorge Goncalves, and Vassilis Kostakos. Revisitation analysis of smartphone app use. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, pages 1197–1208, New York, NY, USA, 2015. ACM.

[3] Zhen Tu, Runtong Li, Yong Li, Gang Wang, Di Wu, Pan Hui, Li Su, and Depeng Jin. Your apps give you away: Distinguishing mobile users by their app usage fingerprints. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3):138, 2018.

[4] Sha Zhao, Julian Ramos, Jianrong Tao, Ziwen Jiang, Shijian Li, Zhaohui Wu, Gang Pan, and Anind K. Dey. Discovering different kinds of smartphone users through their application usage behaviors. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '16, pages 498–509, New York, NY, USA, 2016. ACM.

[5] Steven Van Canneyt, Marc Bron, Andy Haines, and Mounia Lalmas. Describing patterns and disruptions in large scale mobile app usage data. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, pages 1579–1584, Republic and Canton of Geneva, Switzerland, 2017. International World Wide Web Conferences Steering Committee.

[6] Adam J Oliner, Anand P Iyer, Ion Stoica, Eemil Lagerspetz, and Sasu Tarkoma. Carat: Collaborative energy diagnosis for mobile devices. In *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*, page 10. ACM, 2013.

[7] Hongyi Yao, Gyan Ranjan, Alok Tongaonkar, Yong Liao, and Zhuoqing Morley Mao. Samples: Self adaptive mining of persistent lexical snippets for classifying mobile application traffic. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, MobiCom '15, pages 439–451, New York, NY, USA, 2015. ACM.

[8] Qinglong Wang, Amir Yahyavi, Bettina Kemme, and Wenbo He. I know what you did on your smartphone: Inferring app usage over encrypted data traffic. In *2015 IEEE Conference on Communications and Network Security (CNS)*, pages 433–441. IEEE, 2015.

[9] Xin Wang, Shuhui Chen, and Jinshu Su. Real network traffic collection and deep learning for mobile app identification. *Wireless Communications and Mobile Computing*, 2020.

[10] Jieba. Jieba chinese text segmentation. https://github.com/fxsjy/jieba, 2017.

[11] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[12] Marc Langheinrich. A privacy awareness system for ubiquitous computing environments. In *Proceedings of the 4th International Conference on Ubiquitous Computing*, UbiComp '02, pages 237–245, Berlin, Heidelberg, 2002. Springer-Verlag.

[13] Huandong Wang, Yong Li, Sihan Zeng, Gang Wang, Pengyu Zhang, Pan Hui, and Depeng Jin. Modeling spatio-temporal app usage for a large user population. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(1):27, 2019.

[14] George Hripcsak, David J Albers, and Adler Perotte. Parameterizing time in electronic health record studies. *Journal of the American Medical Informatics Association*, 22(4):794–804, 2015.

[15] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494. AUAI Press, 2004.

[16] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[17] David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[18] Nasser M. Nasrabadi. Pattern recognition and machine learning. *Journal of Electronic Imaging*, 16, 2007.

[19] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. *An Introduction to Variational Methods for Graphical Models*, pages 105–161. Springer Netherlands, Dordrecht, 1998.

[20] Matthew Hoffman, Francis R. Bach, and David M. Blei. Online learning for latent dirichlet allocation. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 856–864. Curran Associates, Inc., 2010.

[21] Matthew D Hoffman, David M Blei, Chong Wang, and John

1536-1233 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: CAMBRIDGE UNIV. Downloaded on November 16,2020 at 21:46:11 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2020.3021987, IEEE Transactions on Mobile Computing

Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

[22] Liangjie Hong and Brian D Davison. Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics*, pages 80–88, 2010.

[23] Xuerui Wang and Andrew McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433. ACM, 2006.

[24] Thomas Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 289–296. Morgan Kaufmann Publishers Inc., 1999.

[25] L. R. Bahl, F. Jelinek, and R. L. Mercer. A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-5(2):179–190, March 1983.

[26] Peter F Brown, Vincent J Della Pietra, Robert L Mercer, Stephen A Della Pietra, and Jennifer C Lai. An estimate of an upper bound for the entropy of english. *Computational Linguistics*, 18(1):31–40, 1992.

[27] Graham Neubig. Neural machine translation and sequence-to-sequence models: A tutorial. *arXiv preprint arXiv:1703.01619*, 2017.

[28] Peter Harrington. Machine learning in action. *Shelter Island, NY: Manning Publications Co*, 2012.

[29] Anil K. Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651 – 666, 2010. Award winning papers from the 19th International Conference on Pattern Recognition (ICPR).

[30] George Karypis, Vipin Kumar, and Michael Steinbach. A comparison of document clustering techniques. In *KDD workshop on text mining*, volume 400, pages 525–526. Boston, 2000.

[31] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. On clustering validation techniques. *Journal of intelligent information systems*, 17(2-3):107–145, 2001.

[32] Tâm Huynh, Mario Fritz, and Bernt Schiele. Discovery of activity patterns using topic models. In *UbiComp*, volume 8, pages 10–19, 2008.

[33] Li Yujian and Liu Bo. A normalized levenshtein distance metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1091–1095, June 2007.

[34] David Stark and Balázs Vedres. Social sequence analysis. *The Emergence of Organizations and Markets*, pages 347–64, 2012.

[35] Albertus SJ Helberg and Hendrik C Ferreira. On multiple insertion/deletion correcting codes. *IEEE Transactions on Information Theory*, 48(1):305–308, Jan 2002.

[36] M. Serva and F. Petroni. Indo-european languages tree by levenshtein distance. *EPL (Europhysics Letters)*, 81(6):68005, 2008.

[37] Robert A. Wagner and Michael J. Fischer. The string-to-string correction problem. *J. ACM*, 21(1):168–173, January 1974.

[38] William Cohen, Pradeep Ravikumar, and Stephen Fienberg. A comparison of string metrics for matching names and records. In *Kdd workshop on data cleaning and object consolidation*, volume 3, pages 73–78, 2003.

[39] Richard W Hamming. Error detecting and error correcting codes. *Bell System technical journal*, 29(2):147–160, 1950.

[40] Fionn Murtagh. A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal*, 26(4):354–359, 1983.

[41] Jinwook Seo and B. Shneiderman. Interactively exploring hierarchical clustering results [gene identification]. *Computer*, 35(7):80–86, July 2002.

[42] Shuai Yang, Sixing Chen, and Bin Li. The role of business and friendships on wechat business: An emerging business model in china. *Journal of Global Marketing*, 29(4):174–187, 2016.

[43] Karen Church, Denzil Ferreira, Nikola Banovic, and Kent Lyons. Understanding the challenges of mobile phone usage data. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '15, pages 504–514, New York, NY, USA, 2015. ACM.

[44] Chun-Hao Liu, Sheng-Hsuan Lin, Yuan-Chien Pan, and Yu-Hsuan Lin. Smartphone gaming and frequent use pattern associated with smartphone addiction. *Medicine*, 95(28), 2016.

[45] Hossein Falaki, Ratul Mahajan, Srikanth Kandula, Dimitrios Lymberopoulos, Ramesh Govindan, and Deborah Estrin. Diversity in smartphone usage. In *Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services*, MobiSys '10, pages 179–194, New York, NY, USA, 2010. ACM.

[46] Qiang Xu, Jeffrey Erman, Alexandre Gerber, Zhuoqing Mao, Jeffrey Pang, and Shobha Venkataraman. Identifying diverse usage behaviors of smartphone apps. In *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*, IMC '11, pages 329–344, New York, NY, USA, 2011. ACM.

[47] Jie Yang, Yuanyuan Qiao, Xinyu Zhang, Haiyang He, Fang Liu, and Gang Cheng. Characterizing user behavior in mobile internet. *IEEE Transactions on Emerging Topics in Computing*, 3(1):95–106, March 2015.

[48] Kleomenis Katevas, Ioannis Arapakis, and Martin Pielot. Typical phone use habits: intense use does not predict negative well-being. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services*, page 11. ACM, 2018.

[49] Pascal Welke, Ionut Andone, Konrad Blaszkiewicz, and Alexander Markowetz. Differentiating smartphone users by app usage. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '16, pages 519–523, New York, NY, USA, 2016. ACM.

[50] Suranga Seneviratne, Aruna Seneviratne, Prasant Mohapatra, and Anirban Mahanti. Your installed apps reveal your gender and more! *SIGMOBILE Mob. Comput. Commun. Rev.*, 18(3):5561, January 2015.

[51] Eric Malmi and Ingmar Weber. You are what apps you use: Demographic prediction based on user's apps. In *International AAAI Conference on Web and Social Media*, ICWSM 16, 2016.

[52] Sha Zhao, Yizhi Xu, Xiaojuan Ma, Ziwen Jiang, Zhiling Luo, Shijian Li, Laurence Tianruo Yang, Anind Dey, and Gang Pan. Gender profiling from a single snapshot of apps installed on a smartphone: An empirical study. *IEEE Transactions on Industrial Informatics*, 16(2):1330–1342, Feb 2020.

[53] Xun Zou, Wangsheng Zhang, Shijian Li, and Gang Pan. Prophet: What app you wish to use next. In *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication*, pages 167–170. ACM, 2013.

[54] Nagarajan Natarajan, Donghyuk Shin, and Inderjit S Dhillon. Which app will you use next?: collaborative filtering with interactional context. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 201–208. ACM, 2013.

[55] Hannu Verkasalo. Contextual patterns in mobile service usage. *Personal Ubiquitous Comput.*, 13(5):331–342, June 2009.

[56] Zhung-Xun Liao, Yi-Chin Pan, Wen-Chih Peng, and Po-Ruey Lei. On mining mobile apps usage behavior for predicting apps usage in smartphones. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 609–618. ACM, 2013.

[57] Tingxin Yan, David Chu, Deepak Ganesan, Aman Kansal, and Jie Liu. Fast app launching for mobile devices using predictive user context. In *Proceedings of the 10th international conference on Mobile systems, applications, and services*, pages 113–126. ACM, 2012.

[58] Chen Sun, Jun Zheng, Huiping Yao, Yang Wang, and D Frank Hsu. Apprush: using dynamic shortcuts to facilitate application launching on mobile devices. *Procedia Computer Science*, 19:445–452, 2013.

[59] Kent Shi and Kamal Ali. Getjar mobile application recommendations with very sparse datasets. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 204–212. ACM, 2012.

[60] Bo Yan and Guanling Chen. Appjoy: personalized mobile application discovery. In *Proceedings of the 9th international conference on Mobile systems, applications, and services*, pages 113–126. ACM, 2011.

[61] Donghwan Bae, Keejun Han, Juneyoung Park, and Mun Y Yi. Apptrends: A graph-based mobile app recommendation system using usage history. In *Big Data and Smart Computing (BigComp), 2015 International Conference on*, pages 210–216. IEEE, 2015.

[62] Christoffer Davidsson and Simon Moritz. Utilizing implicit feedback and context to recommend mobile applications from first use. In *Proceedings of the 2011 Workshop on Context-awareness in Retrieval and Recommendation*, pages 19–22. ACM, 2011.

[63] Katsuhiko Kaji, Motoki Yano, and Nobuo Kawaguchi. App. locky: users' context collecting platform for context-aware application recommendation. In *Proceedings of the 2nd international workshop on Ubiquitous crowdsouring*, pages 29–32. ACM, 2011.

1536-1233 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: CAMBRIDGE UNIV. Downloaded on November 16,2020 at 21:46:11 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2020.3021987, IEEE Transactions on Mobile Computing

Tong Li received the B.S. degree and M.S. degree in communication engineering from Hunan University, China, in 2014 and 2017. At present, he is a dual Ph.D. student at the Hong Kong University of Science and Technology and the University of Helsinki. His research interests include mobile computing, edge network, and mobile big data mining. He is an IEEE student member.

Yong Li (M'09-SM'16) received the B.S. degree in electronics and information engineering from Huazhong University of Science and Technology, Wuhan, China, in 2007 and the Ph.D. degree in electronic engineering from Tsinghua University, Beijing, China, in 2012. He is currently a Faculty Member of the Department of Electronic Engineering, Tsinghua University.

Dr. Li has served as General Chair, TPC Chair, SPC/TPC Member for several international workshops and conferences, and he is on the editorial board of two IEEE journals. His papers have total citations more than 6900. Among them, ten are ESI Highly Cited Papers in Computer Science, and four receive conference Best Paper (run-up) Awards. He received IEEE 2016 ComSoc Asia-Pacific Outstanding Young Researchers, Young Talent Program of China Association for Science and Technology, and the National Youth Talent Support Program.

Mohammad A. Hoque obtained his M.Sc degree in Computer Science and Engineering in 2010, and Ph.D in 2013 from Aalto University. He is a postdoctoral researcher at the University of Helsinki. His research interests include energy efficient mobile computing, data analysis, distributed computing, and resource-aware scheduling.

Tong Xia received the B.S. degree in electrical engineering from School of Electrical Information, Wuhan University, Wuhan, China, in 2017. At present, she is studying for the M.S. degree in big data from Department of Electronic Engineering, Tsinghua University, Beijing, China. Her research interests include human mobility, mobile big data mining, user behavior modelling and urban computing.

Sasu Tarkoma (SM12) received the MSc and PhD degrees in computer science from the Department of Computer Science, University of Helsinki. He is a Professor of Computer Science at the University of Helsinki, and Head of the Department of Computer Science. He has authored 4 textbooks and has published over 200 scientific articles. His research interests are Internet technology, distributed systems, data analytics, and mobile and ubiquitous computing. He is Fellow of IET and EAI. He has nine granted US Patents. His research has received several Best Paper awards and mentions, for example at IEEE PerCom, IEEE ICDCS, ACM CCR, and ACM OSR.

Pan Hui (SM'14-F'18) received his Ph.D. degree from the Computer Laboratory at University of Cambridge, and both his Bachelor and MPhil degrees from the University of Hong Kong.

He is the Nokia Chair Professor in Data Science and Professor of Computer Science at the University of Helsinki. He is also the director of the HKUST-DT Systems and Media Lab at the Hong Kong University of Science and Technology. He was a senior research scientist and then a Distinguished Scientist for Telekom Innovation Laboratories (T-labs) Germany and an adjunct Professor of social computing and networking at Aalto University. His industrial profile also includes his research at Intel Research Cambridge and Thomson Research Paris. He has published more than 300 research papers and with over 17,500 citations. He has 30 granted and filed European and US patents in the areas of augmented reality, data science, and mobile computing. He has been serving on the organising and technical program committee of numerous top international conferences including ACM SIGCOMM, MobiSys, IEEE Infocom, ICNP, SECON, IJCAI, AAAI, ICWSM and WWW. He is an associate editor for the leading journals IEEE Transactions on Mobile Computing and IEEE Transactions on Cloud Computing. He is an IEEE Fellow, an ACM Distinguished Scientist, and a member of the Academia Europaea.

1536-1233 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: CAMBRIDGE UNIV. Downloaded on November 16,2020 at 21:46:11 UTC from IEEE Xplore. Restrictions apply.