We Know How You Live: Exploring the Spectrum of Urban Lifestyles

Nicholas Jing Yuan[†], Fuzheng Zhang[*†], Defu Lian[*†], Kai Zheng[‡], Siyu Yu[§], Xing Xie[†]
[†]Microsoft Research Asia
[*]School of Computer Science and Technology, University of Science and Technology of China
[‡]School of Information Technology & Electrical Engineering , The University of Queensland
[§]Department of Sociology, University of California, Berkeley
{nicholas.yuan, v-fuz, v-delian, xing.xie} AT microsoft.com,
kevinz AT itee.uq.edu.au, syyu AT berkeley.edu

ABSTRACT

An incisive understanding of human lifestyles is not only essential to many scientific disciplines, but also has a profound business impact for targeted marketing. In this paper, we present **LifeSpec**, a *computational* framework for exploring and hierarchically categorizing urban lifestyles. Specifically, we have developed an algorithm to connect multiple social network accounts of millions of individuals and collect their *publicly available* heterogeneous behavioral data as well as social links. In addition, a nonparametric Bayesian approach is developed to model the lifestyle spectrum of a group of individuals. To demonstrate the effectiveness of LifeSpec, we conducted extensive experiments and case studies, with a large dataset we collected covering 1 million individuals from 493 cities. Our results suggest that LifeSpec offers a powerful paradigm for 1) revealing an individual's lifestyle from multiple dimensions, and 2) uncovering lifestyle commonalities and variations of a group with various demographic attributes, such as vocation, education, gender, sexual orientation, and place of residence. The proposed method provides emerging implications for personalized recommendation and targeted advertising.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Data mining; J.4 [**Social and Behavioral Science**]: Sociology

Keywords

LifeSpec, lifestyles, living patterns, human behavior

1. INTRODUCTION

Understanding human lifestyles is essential to many scientific disciplines as varied as sociology [4], biomedicine [1], and economics [37]. In marketing research, understanding individual lifestyles is particularly crucial since consumer lifestyles are strong indicators of their buying behaviors [24]. Hence, if you know more about your consumer's lifestyles, you can reach your targets faster, and more effectively.

In the past, it was quite costly (in both time and money) for social scientists to investigate human lifestyles, since such studies depended heavily on large-scale demographic data, e.g., by surveying thousands of participants. The National Census dataset might be a good resource for studying individual lifestyles. However, the typical time cycle between two consecutive censuses is extremely long (10 years for both the U.S and China) and the data at the individual level is usually not available to the public (refer to the 72-years rule[1]). In short, both the time lag and data granularity limit the effectiveness and efficiency of traditional survey-based approaches for understanding dynamic urban lifestyles[24]. In addition, all the survey-based approaches rely on retrospective self-reports and thus are vulnerable to memory error, not to mention the well-known experimenter effects[31].

That is now changing. The emerging era of "big data" provides unprecedented (in terms of both breadth and depth) potential for us to uncovering the underlying patterns of our everyday lives. Imagine a typical day in your life: You are awakened on a Friday morning by the alarm clock, a bit earlier than usual due to an early meeting on that day (the alarm is synced with your online calendar). You rush to take a taxicab to your company (the GPS trajectories are logged by the transportation center) and arrive on time. After a boring meeting that takes up the whole morning, you decide to have a good lunch. So you search on Yelp for a high-scoring restaurant, and you check-in (the process of announcing your arrival at a place and sharing it on a social network) at the restaurant on Foursquare in order to get a discount. After work, you book yourself a ticket for a movie at night that has a high score at IMDB. It ends up being a fantastic Friday evening–not just because of the movie, but because you come across a wonderful woman/man at the theater after posting to Twitter about the movie and catching her attention when she clicked the "Who's Nearby" button.

As can be seen in the above example, many of us already live in an online world. During the past few years, mobile devices, ubiquitous sensing technologies, and various kinds of social networks have proliferated tremendously, which has turned out to be the most important catalyst for bridging our offline world with the online world. The meetings we attend, the restaurants we go to, the movies we see, the people we meet–everything we do during a day–will eventually produce behavioral data stored somewhere in the "Cloud." Intentionally or not, this data mirrors our daily lives, just as the digital "*footprints*" we leave in the online world. Some of these footprints reveal our movements in the physical world, such

COSN'13, October 7–8, 2013, Boston, Massachusetts, USA.
Copyright 2013 ACM 978-1-4503-2084-9/13/10 ...$15.00.
http://dx.doi.org/10.1145/2512938.2512945.

[1] http://1.usa.gov/ZS5T0s

as check-ins and cell-phone traces. Moreover, a footprint can also be generated without the necessity of mobility. A diverse range of data falls into this scope, e.g., posting a tweet, sharing a link to a song, purchasing a book online, rating a movie and so forth. If we understand footprints as the linkages between *human* and *entities* (locations, music, videos, etc.), another kind of link–social link–is the connection between humans that are impatiently migrating from offline to online, even more rapidly than digital footprints.

Given the overwhelming heterogeneous behavioral data of individuals, it is tempting to think that exploring their lifestyles through such data should be easy. This is, however, still not the case. The major challenges of this work are:

- How to *connect* multiple network accounts of a user and *collect* users' *publicly available*[2] footprints (as many as possible) residing in different online networks?
- How to *computationally* model the lifestyle of an individual and a group of individuals, by integrating users' heterogeneous behavioral data?

To address the above challenges, in this paper, we propose a data-driven framework termed **LifeSpec**, to explore urban lifestyles with users' heterogeneous behavioral data and social links. Our exploration ranges from the **specification** of an individual's daily lifestyle as shown in Figure 1, to the lifestyle **spectrum** (will be explicitly defined later) within a group of individuals. This model is flexible enough to deal with groups with various sizes, e.g., a group can either be as small as hundreds of students in a university, or as large as the whole population in a megacity.

To the best of our knowledge, LifeSpec is the first attempt to investigate and model human lifestyles in a *computational* way, based on *millions* of people's *heterogeneous* behavioral data. Our main contributions are summarized as follows:

- We have developed an algorithm for connecting multiple network accounts of users, based on our key observations of users' *self-disclosure* behavior and social "hub sites." In turn, we built a data platform which successfully crawled a large dataset covering 997,500 users (identified to be unique) from 493 cities including their profiles, footprints, and social links. (Section 3)
- We have derived a nonparametric Bayesian approach to computationally model the lifestyle spectrum for a group of individuals, as well as the lifestyle of an individual. This method provides an automatic and data-driven way of generating a hierarchical lifestyle segmentation for a group of individuals. (Section 4)
- We present in-depth analytics on the collected behavioral dataset. Based on this dataset, we conducted extensive experiments and user studies to validate the effectiveness and flexibility of LifeSpec for different groups of users, considering a variety of demographic attributes, such as *vocation*, *education*, *gender*, *sexual orientation*, and *place of residence*. (Section 5)

2. RELATED WORK

Lifestyle Research in Social Science. Human lifestyles have long been studied in social science [16]. In 1967, Ansbacher [3] provided a historical and systematic review of lifestyle research in social science literatures, in which they recognized the similarities among different individuals' lifestyles and suggested the existence of lifestyle *typologies*. Furthermore, they discussed three different levels of aggregation of lifestyles, including "an *individual*," "a *group*," and a "*(generic) class or category*." While there is a broad range of lifestyle research focusing on the US, European, and India markets [26, 36, 23], we have found there is little research on systematically studying the lifestyles in contemporary China. The targeted population in this work, people who were mostly born in the 1980s and 1990s, have experienced profound changes during their lives. As the most lucrative target for the market, their lifestyles significantly differ from their previous generations. Discovering the dynamics and variations of their lifestyles will have a far-reaching impact for both marketers and governments.

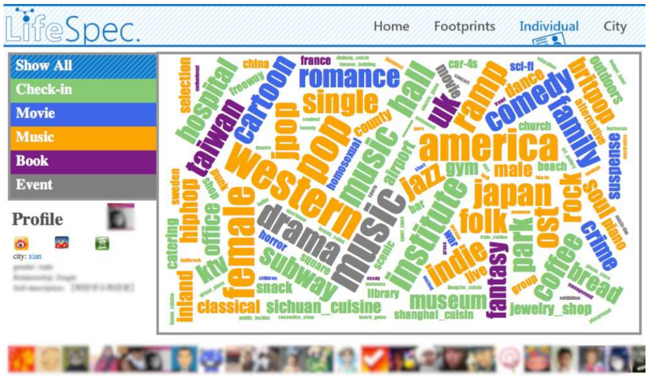


Figure 1: A screenshot of the individual view in the LifeSpec system, where the categories and frequency of footprints are indicated with different colors and sizes. A user can also switch to the city view for exploring the lifestyle spectrum (as detailed later) of a city by clicking either the user's place of residence or the menu above.

Cross-domain User Linking. Linking users from different domains is crucial for targeted advertising and personalization. IP and/or Cookie-based user identification and personalization approaches have been used for years [17]. Some recent methods addressed this problem with various new solutions [25, 21, 14]. These approaches have shown a powerful ability to identify the same user from different networks. However, in most existing methods, a user's different accounts are identified and linked in a *passive* way. In other words, a user might have no intention of being linked from different networks or websites. As a result, these approaches may sometimes raise privacy concerns among users and thus become controversial. In addition, these methods are successful in a probabilistic sense, which means that an error rate (i.e., mis-linking) is inevitable. Our method, however, is based on users' *explicit* and *active* self-disclosure (detailed later) of the connections between their different accounts. Instead of *inferring* any links between their accounts–we *discover* their linked accounts.

Bridging Online and Offline. Recent studies converge to suggest that the barrier between individuals' online and offline lives is tremendously blurred. Such trend is still being accelerated by the advance in ubiquitous sensing, social networks, and big data [11, 12, 39, 2, 29]. For example, Cranshaw et al. [10] showed that human mobility patterns have strong connections with the structure of their underlying social network. Recently, Kosinski et al. [19] reported that the "Like" behavior in Facebook can be used to predict users' psychological traits with a surprisingly high accuracy. However, we have found that there is still a lack of research into human lifestyles that leverages the emerging heterogeneous behavioral data mirroring users' offline behavior from multiple dimen-

[2] In accordance to the United States Code of Federal Regulations defining Human Subjects Research (45 CFR 46), our collection of publicly available data would be considered exempt from Institutional Review Board (IRB) review. We anonymized the IDs of each user on each website when collecting the data, and removed all the identities before the experiments.

sions. Our work takes one more step forward towards the goal of bridging the offline world and the online world.

3. COLLECTING AND CONNECTING

This section first presents some observations that enables us to connect multiple accounts of a user from heterogeneous online networks (Section 3.1). Next, we detail the methodology for connecting users' accounts while collecting the data (Section 3.2).

3.1 Observations

Many people have multiple social network (or website) accounts, e.g., one person is likely to have both a Twitter account and a Foursquare account. There might be a number of reasons behind this, while the most obvious one is that users sign up for different social networks to fulfill different needs. It is because of the heterogeneity of online networks that we have heterogeneous behavioral data, i.e., footprints. A critical challenge for collecting users' cross-domain footprints is the identification of users' multiple social network accounts. Unlike existing approaches that leverage machine learning or rule-based methods to "infer" the connection between a single user's different accounts (which is not 100% accurate and may also be deemed to be a serious invasion of users' privacy), we identify the connection with strong "evidence" that users *actively* and *explicitly* disclose. Specifically, our method is based on the following observations:

O1. Hub site. Some websites function as a *hub* to serve for users' other social networks or websites. Here, we call a website "hub site" if it satisfies the following conditions: 1) It supports indexing users by entities or categories, e.g., places (by Foursquare), and songs (by Last.fm); 2) Users can sync their contents to other networks. For example, Foursquare can usually sync users' check-ins to Facebook and Twitter, once authenticated by users themselves. Other examples of hub sites include About.me and Klout.

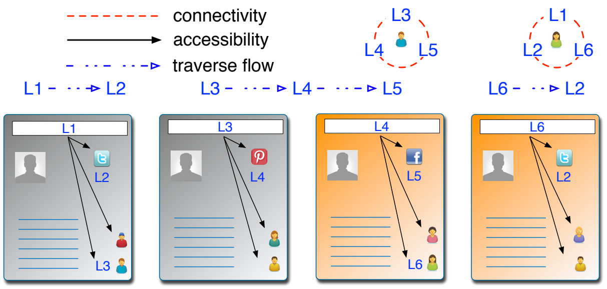
O2. Self-disclosure. Many users explicitly display their other network accounts on their profile pages of one/multiple social networks. For example, many users' Facebook ID and Twitter ID are easy to identify from their Foursqaure profiles. This is also true for Jiepang (known as China's Foursquare).

3.2 Methods

Assume we have a set of known social networking sites Ψ (e.g., $\Psi = \{\text{Facebook, Twitter, Foursqure}\}$) and a hub site $h \in \Psi$. Based on observation O1, we can obtain a set of profile pages in h by querying h with a list of entities. Given the profile pages in h as seeds, we have developed an algorithm (Algorithm 1) to efficiently connect users' accounts on Ψ, as well as to collect their publicly available footprints and social links from different social networks. Note that in this paper, we consider a friendship relationship (either directed or undirected) identified from any network as a social link.

Accessibility and Connectivity. Given a user u's profile page p belonging to a website in Ψ (where p can be accessed by a URL), let Q_p be all the profile page URLs displayed on p, including both her own profile URLs (note p's URL is also in Q_p) and her friends' profile URLs (i.e., social links). We say q is **directly accessible** from p if $q \in Q_p$. If there is a *path* that starts from p, and reaches another profile page p' by traversing only through *directly accessible* URLs (one by one), we say p' is **accessible** from p, denoted as $p \rhd p'$.

Furthermore, let Q be the set of all profile pages in Ψ. For each website $W \in \Psi$, by inspecting the structure of a user's profile page, we can identify whether a user discloses her accounts (profile page URLs) of other websites in Ψ based on observation O1 and O2 (since profile pages in W follow a certain HTML template). For any $q_1, q_2 \in Q$, we say q_1 and q_2 are **directly connected** (or with a *direct connectivity* relation) if there exist a profile page p and a user u, such that $q_1, q_2 \in Q_p^u$, where $Q_p^u \subset Q_p$ denotes u's self-disclosed profile URLs on p. Then we define the **connectivity** relation between any pair of profile pages in Q as the *transitive closure* [32] of the *direct connectivity* relation.



(a) L1 on website A (b) L3 on website A (c) L4 on website B (d) L6 on website B

Figure 2: Accessibility relationships and connectivity relationships among users' different profile pages.

To explain the connectivity relation in a more natural way, we can regard each profile page in Q as a node in an undirected graph G. For any two nodes $q_1, q_2 \in Q$, there is an undirected edge between q_1 and q_2 if q_1 and q_2 are *directly connected* in some profile page p of user u. Thus, the connectivity relationship defined above is equivalent to the common sense of connectivity in an undirected graph, i.e., there exist a path connecting q_1 and q_2 in G. Since the connectivity relation is reflexive, symmetric and transitive, it is an *equivalence* relation. Hence our problem is formulated as:

Given a set of profile pages h, find ***all equivalence classes*** *U from all* ***accessible*** *profile pages of h.*

Actually, when all nodes (profile pages) in G are known in advance, this problem can be efficiently solved using the Union-Find algorithm [33]. In our situation, however, both the nodes and edges are unknown. To address this issue, we propose the ICONNECT algorithm (Algorithm 1) to keep track of all the equivalence classes (unique users) in real time as we discover new nodes (accessible profile pages).

Specifically, as formally presented in Algorithm 1, a queue P_U stores all unvisited profile pages, and each iteration starts by traversing from a popped profile page p (line 6), which contains a user's *directly connected* profile URLs (e.g., URL $L2$ in Fig.2a). Here, we traverse only through these *directly connected* and *directly accessible* profile pages to obtain *connected* profile pages D of this user until we encounter previously visited profile pages P_O (line 8). Since all these profile pages are *connected* with each other, we merge all the original profile pages in P_O as one user (line 9) using the Union-Find structure with almost constant time [34]. For example, $L1, L2, L6$ and $L3, L4, L5$ are merged into two equivalence classes respectively, as shown in Fig.2. Meanwhile, we collect all the footprints and social links obtained from p. In particular, if we discover new hub sites (e.g., another check-in service provider) from the footprints (typically indicated by the patterns in the URL, such as "I'm AT"), we can further add them into the hub site set (line 18), and enqueue discovered new users (line 15 and line 17). Iteratively, we crawl more profiles, footprints, and social links.

Here, to prevent potential confusion between connectivity and accessibility, we note that in Algorithm 1, accessibility relation is merely leveraged for discovering new profile pages (i.e., nodes in the undirected graph G), while we traverse from one profile page

Algorithm 1: IdentifyConnectedUsers (ICONNECT)

```
   Input: seed hub site $h$
   Output: users $U$, where each user $u \in U$ is a set of profile pages;
           footprints $\mathcal{F}$; social links $\mathcal{L}$, hub sites $H$
1  $H \leftarrow \varnothing$; $H$.add($h$);
2  $\mathcal{F} \leftarrow \varnothing$; $\mathcal{L} \leftarrow \varnothing$;
3  A queue storing unvisited profile pages $P_U \leftarrow h$.getAllUsers();
4  A set of visited profile pages $P_V \leftarrow \varnothing$;
5  while $P_U$.length$> 0$ do
6      $p \leftarrow P_U$.dequeue();
7      $D \leftarrow \varnothing, P_O \leftarrow \varnothing$;
8      $(D, P_O) \leftarrow$ VisitProfile($p, P_V, D, P_O, \mathcal{F}, \mathcal{L}$);
9      Merge $\{u \in U | P_O \cap u \neq \varnothing\}$ to $u'$;
10     $P_V \leftarrow P_V \cup D$;   $P_U \leftarrow P_U \setminus D$;
11     if $u' = \varnothing$ then $u' \leftarrow U$.createUser();
12     foreach $d \in D$ do
13         $u'$.add($d$);
14         $H_0 \leftarrow$ hub sites obtained from $\mathcal{F}(d)$;
15         $P_U$.enqueue($\mathcal{L}(p)$.getProfiles()$\setminus(P_V \cup P_U)$);
16         foreach $h' \in H_0 \setminus H$ do
17             $P_U$.enqueue($h'$.getAllUsers()$\setminus(P_V \cup P_U)$);
18         $H \leftarrow H_0 \cup H$;
19 return $U, \mathcal{F}, \mathcal{L}, H$
```

to another only if they are *directly connected* (i.e., an edge in G), which means that we can maintain equivalence classes (connected users) in real time by traversing each profile page once. The correctness and complexity of this algorithm are given in Theorem 1.

THEOREM 1. *Algorithm 1 keeps track of all connected users in currently visited profile pages after each loop (line 18) by accessing each profile page once, and finally finds all connected users from all accessible profile pages $h_{\triangleright}$ of hub site h, with time complexity $O(|h_{\triangleright}|\alpha(|h_{\triangleright}|))$, where α is the inverse Ackermann function*[3] *(the proof is left to Appendix A).*

Another benefit of this algorithm is that we can stop it at any time, while still guarantee having numerous unique users (instead of one single user's different profile pages). Specifically, at any time, we can obtain a set of users $U = \{u_i\}_{i=1}^{N}$ with disjoint profile pages on heterogeneous networks. Let U^W denote the users who have profile pages on website W. Assuming a single user does not have two profile pages on W (which is usually true), theoretically we have in total $|U^W|$ different users. Note that choosing a different W may lead to a different number of users as well as their footprints and social links. In our method, we have chosen a W which maximizes the number of footprints. Hence, the total number of footprints is

$$\max_{W} \left| \bigcup_{u \in U^W} \bigcup_{p \in u} \mathcal{F}(p) \right|. \qquad (1)$$

Based on this algorithm, we developed a data platform, to incrementally and continuously identify new users, connect their accounts on different networks, and collect their heterogeneous behavioral data as well as social links, which in turn supports our further exploration of their lifestyles.

4. MODELING LIFESTYLES AND LIFESTYLE SPECTRUM

This section first explicitly clarifies some related concepts in our model (Section 4.1), then tackles the challenge of integrating heterogeneous footprints and social links to learn the lifestyle spectrum of a group of individuals (Section 4.2).

Procedure VisitProfile($p, P_V, D, P_O, \mathcal{F}, \mathcal{L}$)

```
Input: profile page $p$, visited pages $P_V$, newly discovered pages $D$,
       previously visited pages $P_O \subset P_V$, footprints $\mathcal{F}$, social links$\mathcal{L}$
Output: $D$, $P_O$
$D$.add(p);
$\mathcal{F}(p) \leftarrow$ footprints obtained from the account of $p$;
$\mathcal{L}(p) \leftarrow$ social links obtained from the account of $p$;
$P' \leftarrow$ other profile pages displayed on $p$ or $\mathcal{F}(p)$, directly connected
  with $p$;
foreach $p' \in P'$ and $p' \notin D$ do
    if $p' \in P_V$ then $P_O$.add($p'$);
    else VisitProfile ($p', P_V, D, P_O, \mathcal{F}, \mathcal{L}$);
return $(D, P_O)$
```

4.1 Preliminary

A **footprint** f of an individual u is a combination of domain-specific tokens or tags, which discriminatingly describe the behavior of individual u on a certain domain at a particular time. The granularity of a footprint can vary according to different demands and the data format obtained from the data providers. For example, a mobility-related footprint can be represented with a timestamp and a geo-coordinates, or a POI category, such as "shopping mall" and "office"; a movie-related footprint can be a movie's exact name, or the category of that movie, e.g., "drama and romance".

A **living pattern** S is a combination of frequently co-occurring footprints. For example, an individual may often listen to British pop music and read sci-fi fictions. Note that different individuals may share some typical living patterns and different living patterns may also share some common footprints.

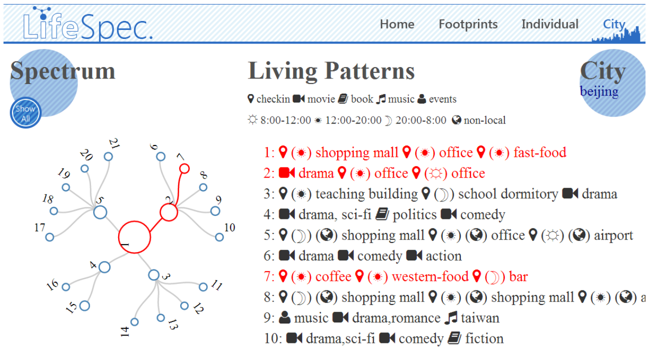
Given a group of individuals, A **lifestyle spectrum** $\mathcal{T}$ is a tree-structured hierarchy summarizing the living patterns of these individuals, where each node of $\mathcal{T}$ is a living pattern. The higher nodes in this tree stand for more commonly shared living patterns and the lower nodes are variations. Thus, a specific **lifestyle** l is a path from the root to a leaf in $\mathcal{T}$, i.e., a sequence of living patterns, ordered by the degree of commonality.

For example, Fig.3 shows the lifestyle spectrum of 100,000 Beijing citizens (sampled from the data we collected). As is shown, Node 1 is the most common living pattern (each living pattern is represented with the top 3 frequent footprints). Node 2, 3, and 4 are various living patterns pertaining to subgroups, e.g., Node 2 is a typical living pattern for an office worker and Node 3 is common for students. The path connecting $1 \rightarrow 2 \rightarrow 7$ forms a typical lifestyle for urban-dwelling office workers who love coffee, western-food and frequent bars at night. Here, the size of a node in the spectrum indicates the number of people who own such a living pattern (the larger, the more). As a result, we can easily target a subgroup of individuals with a certain lifestyle (path) in $\mathcal{T}$. Since a large group (e.g., a megacity) usually contain millions or even tens of millions of citizens, many people may share similar lifestyles, compared with a flat model or simple enumerations, the hierarchical topology of lifestyle spectrum inherently captures the similarity and difference between members in a group.

4.2 Learning the Lifestyle Spectrum

Given millions of footprints and social links of a group of individuals, we leverage topic modeling to learn their lifestyles and lifestyle spectrum, where a "topic" is a distribution of words in a document [6]. As shown in Fig. 4, this is established by building an analogue from the lifestyle spectrum to a hierarchical topic struc-

[3] A quasi-constant function, which grows incredibly slowly [34].



Living Patterns

📍 checkin 🎥 movie 📖 book ♫ music 👤 events

☼ 8:00-12:00 ✹ 12:00-20:00 ☽ 20:00-8:00 🌐 non-local

1: 📍 (✹) shopping mall 📍 (✹) office 📍 (✹) fast-food
2: 🎥 drama 📍 (✹) office 📍 (☼) office
3: 📍 (✹) teaching building 📍 (☽) school dormitory 🎥 drama
4: 🎥 drama, sci-fi 📖 politics 🎥 comedy
5: 📍 (☽) (🌐) shopping mall 📍 (✹) (🌐) office 📍 (☼) (🌐) airport
6: 🎥 drama 🎥 comedy 🎥 action
7: 📍 (✹) coffee 📍 (✹) western-food 📍 (☽) bar
8: 📍 (☽) (🌐) shopping mall 📍 (✹) (🌐) shopping mall 📍 (✹) (🌐) a
9: 👤 music 🎥 drama,romance ♫ taiwan
10: 🎥 drama,sci-fi 🎥 comedy 📖 fiction

Figure 3: A screenshot of LifeSpec showing the lifestyle spectrum and living patterns (partially presented) of Beijing citizens.

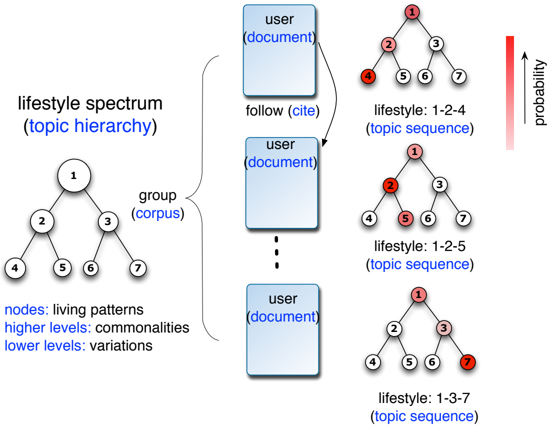


Figure 4: From lifestyle spectrum to a topic hierarchy.

ture as follows: Given a group of users as a *corpus*, we regard all the users in this group as *documents*, where the *words* in each document are an individual's footprints. Just as a topic is described using a collection of words, a living pattern is represented by a set of footprints (recall the definition of living pattern), so it can be considered as a latent *topic* in a document. Here, the lifestyle spectrum is a *hierarchical topic tree* in which each node is a topic. Each document can exhibit multiple topics, which are derived as a path (containing a set of nodes) from the topic tree. In this tree, more commonly shared topics are near the root and more specified topics are close to the leaves.

Blei et al. [7] proposed the hierarchical Latent Dirichlet Allocation (hLDA) for the above model. This model is powerful in the sense that it allows both the parameters and the structure of the model to be automatically adapted as more data is observed. For example, this model can support arbitrary branches and depth of the tree-structured spectrum. This is achieved with the aid of the "nested-Chinese restaurant process" (nCRP), which is widely used in Bayesian nonparametric statistics (refer to [15] for details of this process).

Nevertheless, hLDA still does not take the full advantage of our data, since another important signal that we have captured–the social graph–has not been considered. Actually, social psychologists have found that "perceived and real activity similarity would be equally good predictors of liking." [38] Inspired by this theory, we model a social link between two individuals as a function of their similarity with respect to their living patterns. Given the living patterns $\mathbf{x}_u = \{x_{u,1}, x_{u,2} \ldots, x_{u,n}\}$ of individual u, we calculate the empirical living pattern distribution of u by $\bar{\mathbf{x}}_u = \frac{1}{n}\sum_i x_{u,i}$ (where each $x_{u,i}$ is a distribution over the vocabulary of footprints). For any pair of individuals u, u', the probability that u *follows* u' on a social network is given by

$$\exp(\zeta^\top(\frac{\bar{\mathbf{x}}_u \circ \bar{\mathbf{x}}'_u}{|\bar{\mathbf{x}}_u|}) + \upsilon), \tag{2}$$

where $\circ$ is the Hadamard product and ζ, υ are parameters that need to be learned from the data.

Intuitively, we can deem social links between individuals as citations between documents, e.g., a paper usually cite papers that are similar or related. Several approaches have been developed to deal with the citation relationships between documents, such as the Relational Topic Model (RTM) [8, 9], however, they are for flat topic models instead of a hierarchical topic structure.

To integrate the signals from footprints and social links for learning the lifestyle spectrum, we propose a hybrid model, termed Relational Hierarchical Latent Dirichlet Allocation (RH-LDA), which is a generalization of both hLDA and RTM. Specifically, the generative process of RH-LDA is as follows:

1. For each node k in the spectrum tree $\mathcal{T}$
   (a) Draw a living pattern $\beta_k \sim$ Dirichlet(η).
2. For each individual $d \in \{1, 2, \ldots, D\}$
   (a) Draw a path $\mathbf{c}_d$ ∼nCRP(γ).
   (b) Draw a distribution over levels in the tree, $\theta_d|(m, \pi) \sim$ GEM(m, π)[27].
   (c) For each footprint,
      i. Choose level $Z_{d,n}|\theta_D \sim$ Discrete(θ).
      ii. Choose footprint $W_{d,n}|\{z_{d,n}, \mathbf{c}_d, \beta\} \sim$ Discrete(β_{c_d}), which is parameterized by the living pattern in position $z_{d,n}$ on the path $\mathbf{c}_d$.
3. For each pair of individuals d, d',
   (a) Draw binary link with a probability given by Equation (2).

Note that Step 3 differs from the approach proposed in [9], where we extend the original RTM to directed graph by Eq. 2 (social links are usually directed). The inference of this model is implemented using collapsed Gibbs Sampling (refer to Appendix B) and the Metropolis-Hasting(MH) algorithm.

5. EXPERIMENTS

In this section, we first describe and analyze the data we collected using Algorithm 1 (Section 5.1). Later, based on this huge dataset, we conduct experiments to explore lifestyle spectrums of different groups in terms of various demographic attributes (Section 5.2), as well as a user study on the lifestyle spectrum of different cities (Section 5.3).

5.1 Data Description and Analytics

We chose Jiepang (China's Foursquare) as a hub site. We first crawled the city list and retrieved all the Points of Interest (POI) for each city. Then, for each POI, we obtained all the users who have checked-in at that POI. Using Algorithm 1 and choosing Jiepang as the W which maximizes Eq. 1, we eventually obtained a collection of 997,500 unique users from 439 cities all over China (note some users do not indicate their places of residence on their profile pages), with their 53 million footprints, as well as 3,094,965 social links. Among these users, 99.1% users have at least 2 network accounts and 33.7% users have at least 3 network accounts. We crawled users' *publicly available* profiles, footprints, and social links from the following 4 social networking sites: Jiepang,

Table 1: Summarization of collected footprints for different cities (partially presented due to page limit).

	city	Shanghai	Beijing	Guangzhou	Tianjin	Hangzhou	Hongkong	Xiamen	Suzhou	Nanjing	Chengdu	Wuhan	Xian
	users	417,681	162,764	53,089	15,490	34,322	12,599	10,123	19,673	21,558	23,372	20,975	15,261
Footprints	check-in	25,178,189	5,898,447	1,092,138	392,943	619,219	424,650	369,231	560,274	414,202	327,634	321,646	229,678
	movie	1,661,214	1,466,479	171,789	118,775	238,721	57,003	70,172	89,706	174,664	191,042	166,337	123,223
	music	766,165	737,254	85,953	60,658	103,936	30,313	29,716	39,701	82,513	88,426	76,316	62,876
	book	402,318	387,138	51,913	28,188	57,835	18,117	18,516	19,521	44,345	42,241	44,804	28,435
	event	609,076	803,158	101,246	52,133	78,587	18,277	20,889	27,400	46,788	66,640	44,764	72,902
	total	28,616,962	9,292,476	1,503,039	652,697	1,098,298	548,360	508,524	736,602	762,512	715,983	653,867	517,114

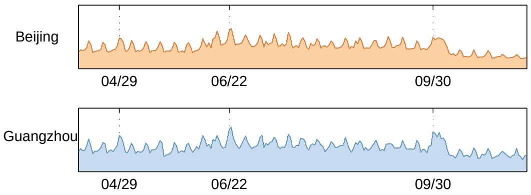
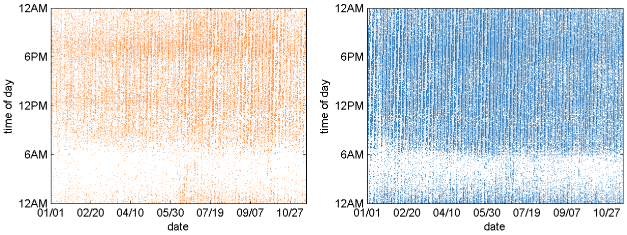


Figure 5: Daily trends of check-ins in different cities.



(a) Beijing

(b) Guangzhou

Figure 6: Diurnal distribution users' check-ins.

Sina Weibo (China's Twitter), Douban (an interest-based social network), Dianping (China's Yelp).

Check-ins. We collected in total 39,358,679 check-ins, where each check-in was represented with a timestamp, a latitude, a longitude, and a POI category. Since these physical footprints are extremely crucial for understanding individual lifestyles, we analyzed them in several dimensions.

For example, Fig. 5 shows the daily trends (during Apr. 9, 2012 to Nov. 15, 2012) of the total number of check-ins with respect to two major cities in north and south China respectively (Beijing and Guangzhou). As is shown, the number of check-ins periodically rises on weekends and falls on weekdays. Here, the dates that we have labeled on the x-axis are important public holidays (at least three days) in China, including Labor Day (3 days), Dragon Boat Day (3 days), and National Day (8 days). On these days, the number of check-ins is relatively higher than normal weekends, especially on National Day, which has a clearly longer peak than other holidays. This is mainly because that a lot of people travel with their families and friends during this long holiday.

We further examined the diurnal distribution of people's check-in behavior by sampling the 1000 individuals in different cities and analyzing their diurnal check-in distributions, as shown in Fig. 6. Unlike the aggregated daily trends, the diurnal distribution varies in different cities. For example, Fig. 6a and Fig. 6b reveal that people in Guangzhou check in earlier than people in Beijing in the morning (7AM), and later in the evening (12AM). This result conforms very well to a recent survey [4] with 1 million respondents performed by Chinese Medical Doctor Association, which shows that the average bedtime of Guangzhou citizens is 23:08pm and 22:15pm for Beijing.

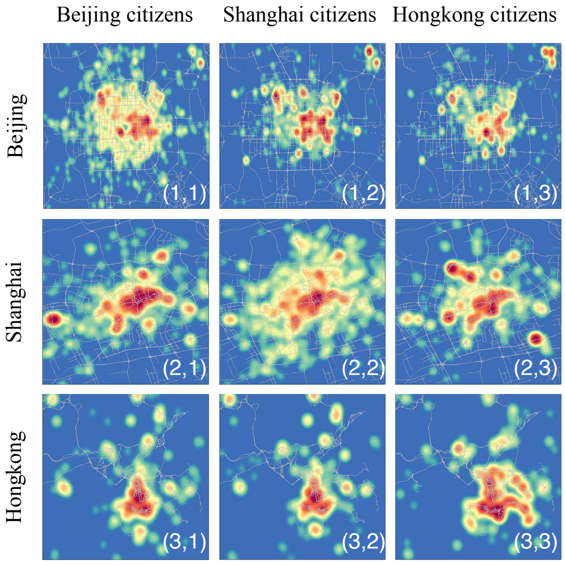
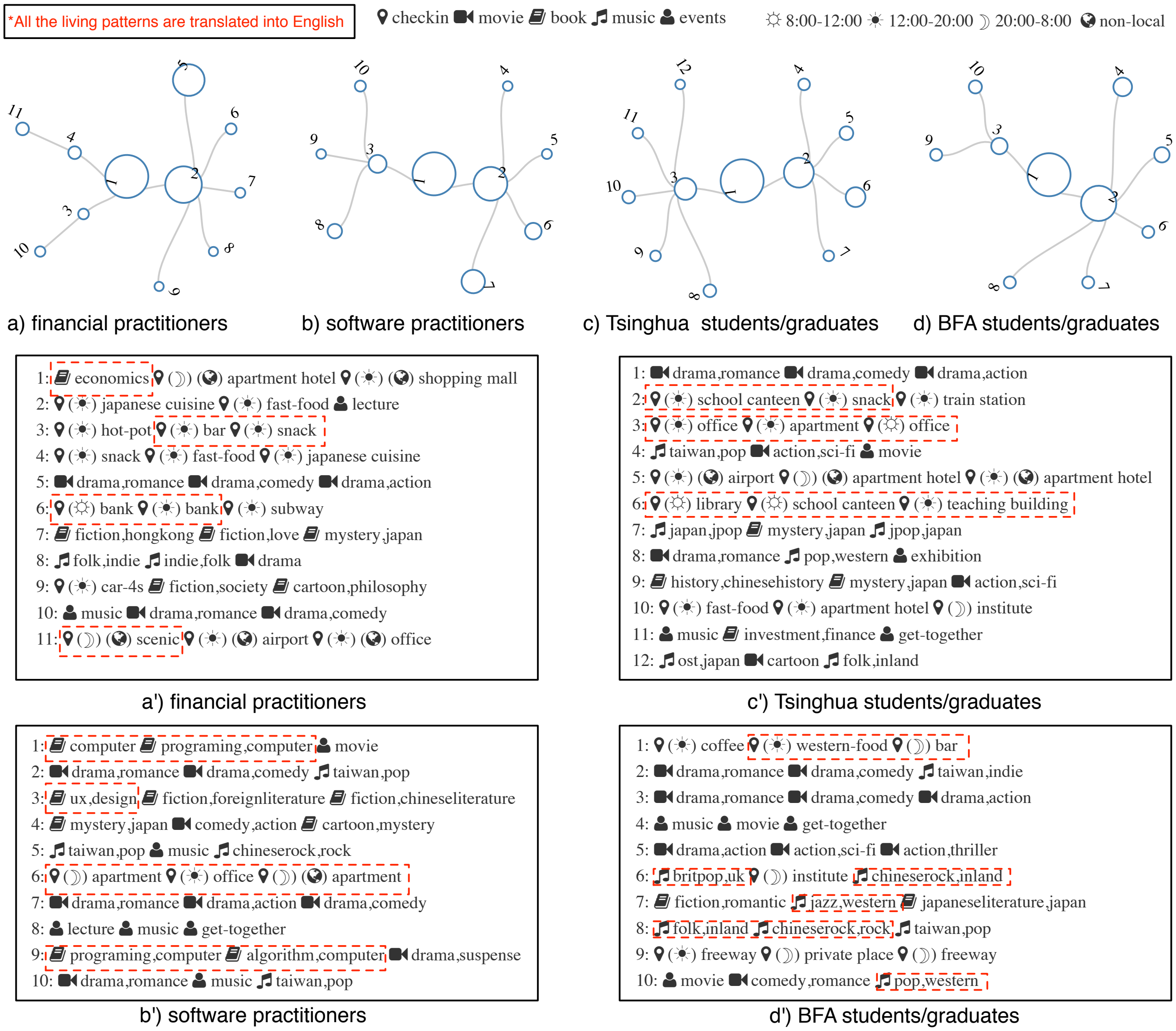


Figure 7: Check-in density distribution of 3 cities showing where people check-in in each other's cities. The diagonal subplots (local citizens) show significantly higher diversity than other subplots (travelers).

Fig. 7 plots the density distribution of check-ins posted by Beijing, Shanghai and Hongkong citizens when they are in each other's city, e.g., grid (1,2) in the 3 × 3 grids tells us where Shanghai citizens check-in at Beijing. This figure clearly indicates that the check-ins of local citizens are much more diverse than those of travelers (as expected), and people in different cities have differentiated preferences when traveling in other cities. e.g., the hot spots in grid (2,1) and grid (2,3) are quite different.

Movies/Music/Books/Events. We collected in total 82,451 movies, 477,712 songs , 406,564 books, and 407,950 social events/gatherings. All of the above entities have their taxonomies, which were also crawled and leveraged for constructing the footprints. Note these are the number of entities, not footprints. In terms of footprints, we crawled 6,241,036 movie footprints, 3,075,305 music footprints, 1,560,206 book footprints, and 2,596,252 event footprints. Table 1 summarizes the different kinds of footprints for different cities.

[4] http://bit.ly/14cvnem


*All the living patterns are translated into English
📍 checkin 🎥 movie 📖 book ♫ music 👤 events
☼ 8:00-12:00 ☀ 12:00-20:00 ☽ 20:00-8:00 🌐 non-local

a) financial practitioners b) software practitioners c) Tsinghua students/graduates d) BFA students/graduates

1: 📖 economics 📍 (☽) (🌐) apartment hotel 📍 (☀) (🌐) shopping mall
2: 📍 (☀) japanese cuisine 📍 (☀) fast-food 👤 lecture
3: 📍 (☀) hot-pot 📍 (☀) bar 📍 (☀) snack
4: 📍 (☀) snack 📍 (☀) fast-food 📍 (☀) japanese cuisine
5: 🎥 drama,romance 🎥 drama,comedy 🎥 drama,action
6: 📍 (☼) bank 📍 (☀) bank 📍 (☀) subway
7: 📖 fiction,hongkong 📖 fiction,love 📖 mystery,japan
8: ♫ folk,indie ♫ indie,folk 🎥 drama
9: 📍 (☀) car-4s 📖 fiction,society 📖 cartoon,philosophy
10: 👤 music 🎥 drama,romance 🎥 drama,comedy
11: 📍 (☽) (🌐) scenic 📍 (☀) (🌐) airport 📍 (☀) (🌐) office

a') financial practitioners

1: 🎥 drama,romance 🎥 drama,comedy 🎥 drama,action
2: 📍 (☀) school canteen 📍 (☀) snack 📍 (☀) train station
3: 📍 (☀) office 📍 (☀) apartment 📍 (☼) office
4: ♫ taiwan,pop 🎥 action,sci-fi 👤 movie
5: 📍 (☀) (🌐) airport 📍 (☽) (🌐) apartment hotel 📍 (☀) (🌐) apartment hotel
6: 📍 (☼) library 📍 (☼) school canteen 📍 (☀) teaching building
7: ♫ japan,jpop 📖 mystery,japan ♫ jpop,japan
8: 🎥 drama,romance ♫ pop,western 👤 exhibition
9: 📖 history,chinesehistory 📖 mystery,japan 🎥 action,sci-fi
10: 📍 (☀) fast-food 📍 (☀) apartment hotel 📍 (☽) institute
11: 👤 music 📖 investment,finance 👤 get-together
12: ♫ ost,japan 🎥 cartoon ♫ folk,inland

c') Tsinghua students/graduates

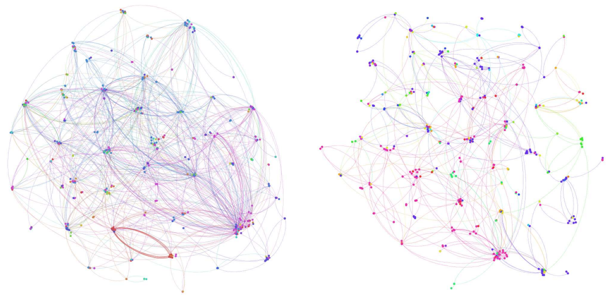
1: 📖 computer 📖 programing,computer 👤 movie
2: 🎥 drama,romance 🎥 drama,comedy ♫ taiwan,pop
3: 📖 ux,design 📖 fiction,foreignliterature 📖 fiction,chineseliterature
4: 📖 mystery,japan 🎥 comedy,action 📖 cartoon,mystery
5: ♫ taiwan,pop 👤 music ♫ chineserock,rock
6: 📍 (☽) apartment 📍 (☀) office 📍 (☽) (🌐) apartment
7: 🎥 drama,romance 🎥 drama,action 🎥 drama,comedy
8: 👤 lecture 👤 music 👤 get-together
9: 📖 programing,computer 📖 algorithm,computer 🎥 drama,suspense
10: 🎥 drama,romance 👤 music ♫ taiwan,pop

b') software practitioners

1: 📍 (☀) coffee 📍 (☀) western-food 📍 (☽) bar
2: 🎥 drama,romance 🎥 drama,comedy ♫ taiwan,indie
3: 🎥 drama,romance 🎥 drama,comedy 🎥 drama,action
4: 👤 music 👤 movie 👤 get-together
5: 🎥 drama,action 🎥 action,sci-fi 🎥 action,thriller
6: ♫ britpop,uk 📍 (☽) institute ♫ chineserock,inland
7: 📖 fiction,romantic ♫ jazz,western 📖 japaneseliterature,japan
8: ♫ folk,inland ♫ chineserock,rock ♫ taiwan,pop
9: 📍 (☀) freeway 📍 (☽) private place 📍 (☽) freeway
10: 👤 movie 🎥 comedy,romance ♫ pop,western

d') BFA students/graduates

Figure 8: Lifestyle spectrums of different groups in terms of vocation and education.



(a) homosexual men (b) homosexual women

Figure 9: Social graphs of homosexual men and women in our dataset, where different colors indicate different places of residence.

5.2 Results on Different Demographic Groups

We extract profiles from users' multiple networks, such as gender, place of residence, sexual orientation, education, and vocation. We note that all these networks have privacy options for users to hide their profiles from others, and we only crawled *public profiles*. Based on which, we segmented users into different groups and generated the lifestyle spectrums for each group. Due to space limit, we only present part of our findings and results here. For each lifestyle spectrum, we show the top-3 levels, where each living pattern is presented by the top-3 frequent footprints.

Vocation and Education. Fig. 8a,b present the lifestyle spectrums for two vocational groups: financial practitioners and software practitioners, both containing 1,000 samples. As is shown, the most common living pattern for these financial practitioners is reading economics books and checking-in at apartment hotel (indicating that they are often on business trips). However, for software practitioners, reading programming books is their most-typical living pattern, which makes perfect sense. Node 6 (N6 for short, similarly hereinafter) in Fig. 8a targets a subgroup of individuals who are probably working for banks. Compared with software practitioners, the result suggests that these financial practitioners live a chicer life, e.g., more often they show at bars and scenic places (N3, N11 of Fig. 8a), while the software engineers are still coding or reading programming books at apartment or office (N6, N9 of Fig. 8b).

Fig. 8c and Fig. 8d are lifestyle spectrums of graduates and students from two universities: Tsinghua University (known as one of the best science and engineering university in China) and Beijing Film Academy (BFA), which graduated many famous alumni in filming industry such as Yimou Zhang and Kaige Chen. The results (generated with 300 samples for each group) show that their lifestyles widely differ from each other, e.g., N1 in Fig 8c reveal that most students/graduates in BFA go to bar and western food restaurants frequently and their music tastes are more diverse than Tsinghua students/graduates. In addition, students and graduates from Tsinghua are automatically categorized into two subgroups: Living patterns rooted at N2 reveal many characteristics of a student such as "teaching-building"; while the living pattern of N3 is commonly shared by working people.

Gender and sexual orientation. Gender and sexual orientation differences were intensely studied in sociology and social psychology[28]. Based on self-identified sexual orientation provided in users' *public* profiles, we randomly sampled 500 homosexual men and women to generate the lifestyle spectrums, as depicted in Fig. 10a,b. In 2006, China was estimated to have 5–9 million homosexual men among 452 million adult males (aged 15 to 64)[13]. To be comparable, we sampled 45,000 heterosexual men and women (for each), and generated their lifestyle spectrums. As a result, some characteristic living patterns for homosexual men/women are prominently revealed, e.g., watching homosexual movies, gym, and reading tanbi books (describing the love between boys). For homosexual women, however, the living pattern "watching homosexual movie" is not as dominant as for homosexual men. Another remarkable signal implied by the result is that a certain number of homosexual men are students (N10 in Fig. 10a). Note that social links are also leveraged in our model for learning lifestyle spectrums. Fig. 9 is a visualization of their social graphs (using OpenOrd layout [22] for edge-cutting and community clustering), where we removed 21 isolated nodes for homosexual men and 27 for homosexual women. It's clear that the gay community has a much stronger social connection than the lesbian community.

The gender difference between heterosexual men and women is also significant, e.g., men watch more sci-fi and action movies while women prefer romance movies (N1 of Fig. 10c,d). N2 and N4 of Fig. 10c may refer to two kinds of men: men who spend more time on career vs. men who spend more time with family. For women, regardless of which subgroup they belong to, they all love shopping (N1 of Fig. 10d). A majority group of females like Taiwan or western pop-style music (N2 of Fig. 10d). Many typical living patterns for Chinese women are also brought to light, such as shoe-store, hot-pot, snack, bread and KTV (N3,4,7,11,13 of Fig. 10d).

Place of Residence. We generated the lifestyle spectrum of 15 cities in China using 10,000 sampled citizens for each city, e.g., Fig. 3 shows the result of Beijing (results of other cities are not visualized here due to space limit). Meanwhile, we calculated the similarity of lifestyle spectrums between different cities based on Hausdorff distance [30], which is commonly used as a similarity measure between two sets.

Specifically, a lifestyle spectrum $\mathcal{T}$ can be represented by a set containing all the lifestyles (i.e., paths), as modeled in Section 4.1. Consider a lifestyle l as a point in a m-dimension space, where m is the number of unique footprints. For each footprint f that occurs in some living pattern $S \in l$, we assign the dimension in l that corresponds to f with the proportion of people who exhibit lifestyle l. Then the distance between any two lifestyles can be calculated using the Euclidean distance. Hence, the Hausdorff distance between $\mathcal{T}_1$ and $\mathcal{T}_2$ is defined as

$$d_H(\mathcal{T}_1, \mathcal{T}_2) = \max(\max_{l \in \mathcal{T}_1} \min_{l' \in \mathcal{T}_2} d(l, l'), \max_{l' \in \mathcal{T}_2} \min_{l \in \mathcal{T}_1} d(l, l')).$$

Fig. 11 shows the similarity matrix of these cities, where green cubes indicate a smaller distance (i.e., similar) and red cubes stand for a larger distance. Many similarities found in this matrix can be explained by geographical proximity and culture homology. For example, Chengdu and Chonqing have similar lifestyles, which is widely known to the public since they are geographically close to each other and they share the root of Ba-Shu Culture[5]. Another example is Shanghai and Tianjin, the two major seaports in China, which were both colonized in the 19th century, and for both of them, there is a "mother" river flowing through their hinterlands (Haihe River[6] and Huangpu River[7]). It has been found long ago that a river plays a key-role in a city's economy, civilization and culture development [18].

Table 2: Average recognition ratio

Method	RTM	hLDA	RH-LDA
Check-in	0.361	0.500	0.667
+Movie	0.389	0.556	0.694
++Music	0.444	0.583	0.722
+++Book	0.472	0.639	0.806
++++Events	0.472	0.667	0.833

5.3 User Study

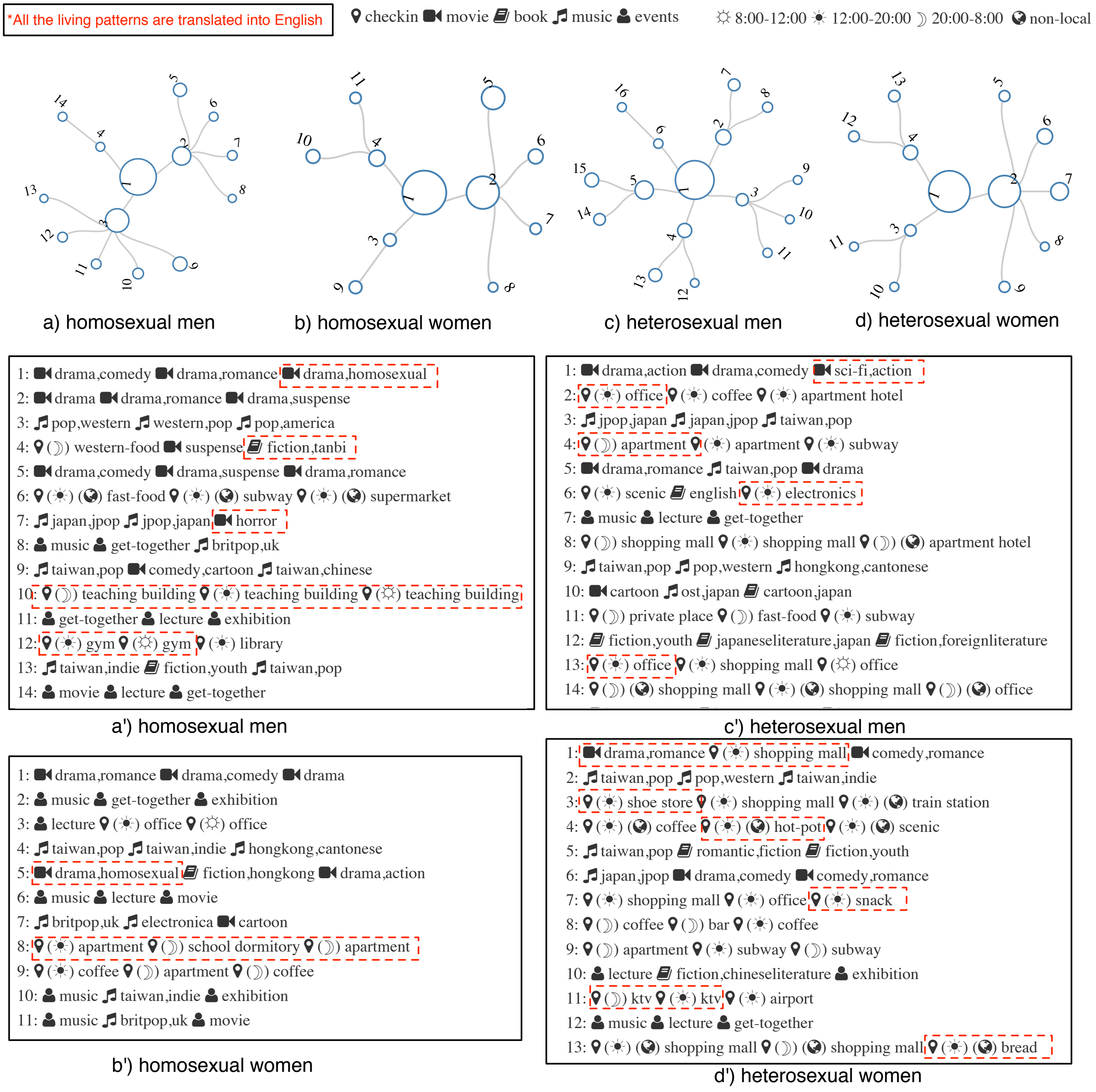
In order to further validate our method in the field, we performed a user study, which focused on evaluating the lifestyle spectrum of a city. Our study is guided by the following questions: 1) Whether our model can capture the characteristics of lifestyles in a city, thus reveal the intergroup difference between different cities? 2) Whether our method can reveal diverse lifestyles of a city, thus can uncover the intragroup variations?

Participants. We recruited 36 participants (aged 21–45, 20 males and 16 females) who reside in 6 cities, including Beijing (BJ), Shanghai (SH), Guangzhou (GZ), Hangzhou (HZ), Chengdu (CD), and Xiamen (XM) (6 participants for each city), and all of them have lived in their cities for more than 8 years.

Baselines. In terms of learning the lifestyle spectrum, we compared our model (RH-LDA) with two baselines: 1) The hLDA model, which only considers words of a document (i.e., footprints). 2) The RTM model, which only leverages social links and does not support a hierarchical structure of topics. Thus the lifestyle spectrum generated by RTM is a list of topics instead of a tree-structure. In addition, RTM requires the number of topics n to be fixed beforehand. To make RTM comparable with hLDA and RH-LDA, we fixed n to be identical to the number of living patterns generated by hLDA and RH-LDA respectively, and chose the best performance between them as the result of RTM. All these models were inferred using collapsed Gibbs sampling, and the hyper parameters were learned through Metropolitan Hasting [7].

Intergroup Difference. Using each of the 3 methods, we generated lifestyle spectrums of these cities and visualized them to the participants, without telling them which city each spectrum stands

[5] http://bit.ly/15B2Qvm
[6] http://bit.ly/ZSHO9T
[7] http://bit.ly/10tpl5U


*All the living patterns are translated into English
📍 checkin 🎥 movie 📖 book ♫ music 👤 events
☼ 8:00-12:00 ☀ 12:00-20:00 ☽ 20:00-8:00 🌐 non-local
1 2 3 4 5 6 7 8 9 10 11 12 13 14
1 2 3 4 5 6 7 8 9 10 11
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
1 2 3 4 5 6 7 8 9 10 11 12 13

a) homosexual men b) homosexual women c) heterosexual men d) heterosexual women

1: 🎥 drama,comedy 🎥 drama,romance 🎥 drama,homosexual
2: 🎥 drama 🎥 drama,romance 🎥 drama,suspense
3: ♫ pop,western ♫ western,pop ♫ pop,america
4: 📍 (☽) western-food 🎥 suspense 📖 fiction,tanbi
5: 🎥 drama,comedy 🎥 drama,suspense 🎥 drama,romance
6: 📍 (☀) (🌐) fast-food 📍 (☀) (🌐) subway 📍 (☀) (🌐) supermarket
7: ♫ japan,jpop ♫ jpop,japan 🎥 horror
8: 👤 music 👤 get-together ♫ britpop,uk
9: ♫ taiwan,pop 🎥 comedy,cartoon ♫ taiwan,chinese
10: 📍 (☽) teaching building 📍 (☀) teaching building 📍 (☼) teaching building
11: 👤 get-together 👤 lecture 👤 exhibition
12: 📍 (☀) gym 📍 (☼) gym 📍 (☀) library
13: ♫ taiwan,indie 📖 fiction,youth ♫ taiwan,pop
14: 👤 movie 👤 lecture 👤 get-together

a') homosexual men

1: 🎥 drama,romance 🎥 drama,comedy 🎥 drama
2: 👤 music 👤 get-together 👤 exhibition
3: 👤 lecture 📍 (☀) office 📍 (☼) office
4: ♫ taiwan,pop ♫ taiwan,indie ♫ hongkong,cantonese
5: 🎥 drama,homosexual 📖 fiction,hongkong 🎥 drama,action
6: 👤 music 👤 lecture 👤 movie
7: ♫ britpop,uk ♫ electronica 🎥 cartoon
8: 📍 (☀) apartment 📍 (☽) school dormitory 📍 (☽) apartment
9: 📍 (☀) coffee 📍 (☽) apartment 📍 (☽) coffee
10: 👤 music ♫ taiwan,indie 👤 exhibition
11: 👤 music ♫ britpop,uk 👤 movie

b') homosexual women

1: 🎥 drama,action 🎥 drama,comedy 🎥 sci-fi,action
2: 📍 (☀) office 📍 (☀) coffee 📍 (☀) apartment hotel
3: ♫ jpop,japan ♫ japan,jpop ♫ taiwan,pop
4: 📍 (☽) apartment 📍 (☀) apartment 📍 (☀) subway
5: 🎥 drama,romance ♫ taiwan,pop 🎥 drama
6: 📍 (☀) scenic 📖 english 📍 (☀) electronics
7: 👤 music 👤 lecture 👤 get-together
8: 📍 (☽) shopping mall 📍 (☀) shopping mall 📍 (☽) (🌐) apartment hotel
9: ♫ taiwan,pop ♫ pop,western ♫ hongkong,cantonese
10: 🎥 cartoon ♫ ost,japan 📖 cartoon,japan
11: 📍 (☽) private place 📍 (☽) fast-food 📍 (☀) subway
12: 📖 fiction,youth 📖 japaneseliterature,japan 📖 fiction,foreignliterature
13: 📍 (☀) office 📍 (☀) shopping mall 📍 (☼) office
14: 📍 (☽) (🌐) shopping mall 📍 (☀) (🌐) shopping mall 📍 (☽) (🌐) office

c') heterosexual men

1: 🎥 drama,romance 📍 (☀) shopping mall 🎥 comedy,romance
2: ♫ taiwan,pop ♫ pop,western ♫ taiwan,indie
3: 📍 (☀) shoe store 📍 (☀) shopping mall 📍 (☀) (🌐) train station
4: 📍 (☀) (🌐) coffee 📍 (☀) (🌐) hot-pot 📍 (☀) (🌐) scenic
5: ♫ taiwan,pop 📖 romantic,fiction 📖 fiction,youth
6: ♫ japan,jpop 🎥 drama,comedy 🎥 comedy,romance
7: 📍 (☀) shopping mall 📍 (☀) office 📍 (☀) snack
8: 📍 (☽) coffee 📍 (☽) bar 📍 (☀) coffee
9: 📍 (☽) apartment 📍 (☀) subway 📍 (☽) subway
10: 👤 lecture 📖 fiction,chineseliterature 👤 exhibition
11: 📍 (☽) ktv 📍 (☀) ktv 📍 (☀) airport
12: 👤 music 👤 lecture 👤 get-together
13: 📍 (☀) (🌐) shopping mall 📍 (☽) (🌐) shopping mall 📍 (☀) (🌐) bread

d') heterosexual women

Figure 10: Lifestyle spectrum of hetero- and homosexual men and women.

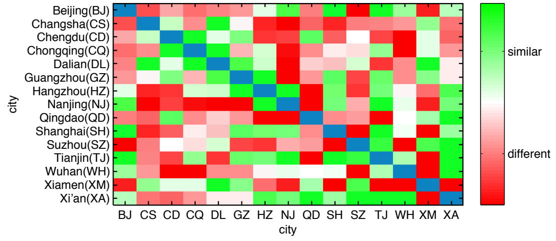


Figure 11: Similarity matrix w.r.t 15 cities.

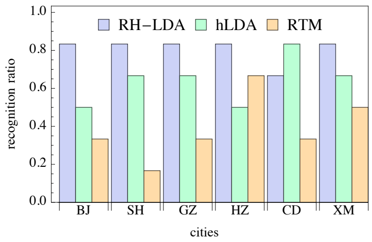


Figure 12: Recognition ratio.

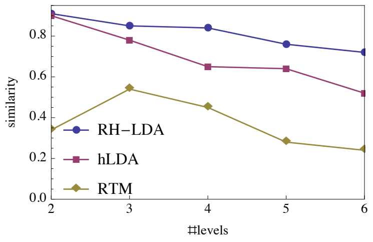


Figure 13: Jaccard Similarity.

for. Then each participant picked one among the 6 different spectrums, which she believed is most likely to be her place of residence (we had told them to consider not only their own lifestyles but the variations of lifestyles according to their knowledge of their cities). Later, we calculated the recognition ratio (RR) (the number of users who successfully identified the spectrum of their cities divided by the number of users) for each model. Fig.12 shows the recognition ratio for each city when we leveraged all kinds of footprints. We further studied how well these models can perform if we reduce the heterogeneity of footprints. Table 2 summarizes the average recognition ratio for all the considered cities, given different types of footprints including checkins, movies, songs, books, and offline events. Clearly, the performance of these models were all improved by increasing the diversity of footprints. However, for each setting, RH-LDA achieves the highest recognition ratio, which suggests that our model is more effective than competing methods in terms of the ability to summarize the lifestyles of a city and uncover the intergroup differences.

Intragroup Variation. For studying the effectiveness of our model in terms of capturing lifestyle variations within a group, we asked the participants for their online network accounts, and collected their footprints. Among them, 23 participants own multiple social network accounts. We used the trained model for each city to infer these participants' lifestyles beforehand. Then we compared the generated lifestyle (a path on the spectrum for RH-LDA and hLDA, and a set of living patterns for RTM) with their own lifestyles chosen by themselves for each model pertaining to their own cities. We then calculated their similarity by extracting the top-5 footprints for each related living pattern of these two sequences (lifestyles), and compared their average Jaccard Similarity (JS). Furthermore, we studied the effect induced by the number of levels in the spectrum (the number of topics for RTM is determined as before), as shown in Fig.13. As a result, our model outperforms the competitors significantly, however, when the number of levels increases, the similarity between the inferred lifestyle and the user's perceived lifestyle decreases for all the models. In particular, when the number of level goes from 3 to 4, the performance of the other methods declines precipitously, nevertheless, our method is relatively more stable and robust. According to users' feedback, when the number of levels becomes too large, it's not easy for them to choose the most relevant lifestyles.

6. DISCUSSIONS

• **Limitations.** While showing the potential to leverage massive behavioral data for learning lifestyles, we are aware that this method has several limitations. First, human lifestyle is complicated. It is still possible that the actual lifestyle of a person deviates from what reflected in the behavioral data. Second, the targeted population mainly consists of young people who use online social networks extensively, which may bias the lifestyle spectrum. However, this is actually induced more by the limitation of the data rather than the model. Note that conventional social studies might also suffer from sampling bias [5], with much smaller scale and coverage (e.g., dozens of college students) compared to the dataset used in this work. We believe by employing more types of footprints that mirror users' offline behaviors, e.g., credit card transactions and public transit records, this framework can cover a broader demographic and attain a more faithful understanding of their lifestyles.

• **Privacy.** We re-emphasize that in this work we only collected users' *publicly available* data (i.e., visible to everyone on the web) including profiles, social links, and footprints. Besides, the connections between users' different accounts are identified from their self-disclosed contents (refer to Section 3.1). However, we remind that some users may have no intention to (or carelessly) disclose the connections of their different accounts (e.g., by adopting the default privacy options of some websites, or by posting a tweet with location check-in automatically embedded). Thus, we suggest both the users and social networking sites re-consider their privacy policy in terms of the linkage between multiple accounts, which may potentially be exploited by attackers and thus bring privacy risk [20] to the users.

7. CONCLUSION

We have presented LifeSpec, a data-driven framework for exploring and hierarchically summarizing urban lifestyles. In this framework, we have built a data platform to connect users' heterogeneous behavioral data and their social links. Given the behavioral data as digital footprints, we have formally modeled the lifestyle spectrum of a group and generalized a probabilistic model to learn the lifestyle spectrum. We conducted a series of experiments and user studies to validate the usability and flexibility of this framework.

Please note that this framework is not designed to replace traditional methods in lifestyle research. Instead, we believe that these methods can complement each other (actually this work is also a collaboration with sociologists) to enable a better and more comprehensive understanding of human lifestyles, which is not only important for advancing the lifestyle research in social science, but also essential to *personalized* recommendation and *targeted* advertising.

References

[1] A. Agarwal, N. R. Desai, R. Ruffoli, and A. Carpi. Lifestyle and testicular dysfunction: a brief update. *Biomedicine & Pharmacotherapy*, 62(8):550–553, 2008.

[2] M. Allamanis, S. Scellato, and C. Mascolo. Evolution of a location-based online social network: analysis and models. In *Proc. ACM conference on Internet measurement conference*, pages 145–158, 2012.

[3] H. Ansbacher. Life style: a historical and systematic review. *Journal of individual psychology*, 23(2):191, 1967.

[4] M. Benson and K. O'Reilly. Migration and the search for a better way of life: a critical exploration of lifestyle migration. *The Sociological Review*, 57(4):608–625, 2009.

[5] R. A. Berk. An introduction to sample selection bias in sociological data. *American Sociological Review*, pages 386–398, 1983.

[6] D. Blei and J. Lafferty. Topic models. *Text mining: classification, clustering, and applications*, 10:71, 2009.

[7] D. M. Blei, T. L. Griffiths, and M. I. Jordan. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *J. ACM*, 57(2):7:1–7:30, 2010.

[8] J. Chang and D. Blei. Relational topic models for document networks. In *Artificial Intelligence and Statistics*, pages 81–88, 2009.

[9] J. Chang and D. M. Blei. Hierarchical relational models for document networks. *The Annals of Applied Statistics*, 4(1): 124–150, 2010.

[10] J. Cranshaw, E. Toch, J. Hong, A. Kittur, and N. Sadeh. Bridging the gap between physical location and online social networks. In *Proc. Ubicomp*, pages 119–128, 2010.

[11] J. Cranshaw, R. Schwartz, J. Hong, and N. Sadeh. The livehoods project: Utilizing social media to understand the dynamics of a city. *Association for the Advancement of Artificial Intelligence*, 2012.

[12] N. Eagle and A. Pentland. Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4): 255–268, 2006.

[13] B. Gill, Y. Huang, and X. Lu. Demography of HIV/AIDS in China. *Center for Strategic International Studies*, 2007.

[14] O. Goga, H. Lei, S. H. K. Parthasarathi, G. Friedland, R. Sommer, and R. Teixeira. Exploiting innocuous activity for correlating users across sites. In *Proc. WWW*, pages 447–458, 2013.

[15] D. Griffiths and M. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. In *Proc. NIPS*, volume 16, page 17, 2003.

[16] R. Havighurst and K. Feigenbaum. Leisure and life-style. *American Journal of Sociology*, pages 396–404, 1959.

[17] A. Juels, M. Jakobsson, and T. Jagatic. Cache cookies for browser authentication. In *IEEE Symposium on Security and Privacy*, pages 5–pp, 2006.

[18] A. Kelman. *A river and its city: The nature of landscape in New Orleans*. University of California Press, 2006.

[19] M. Kosinski, D. Stillwell, and T. Graepel. Private traits and attributes are predictable from digital records of human behavior. *PNAS*, 2013.

[20] B. Krishnamurthy. Privacy and online social networks: Can colorless green ideas sleep furiously? In *IEEE Symposium on Security and Privacy*, 2013.

[21] J. Liu, F. Zhang, X. Song, Y.-I. Song, and C.-Y. Lin. What's in a name? an unsupervised approach to link users across communities. In *Proc. WSDM*, 2013.

[22] S. Martin, W. M. Brown, R. Klavans, and K. W. Boyack. OpenOrd: an open-source toolbox for large graph layout. In *IS&T/SPIE Electronic Imaging*, pages 786806–786806, 2011.

[23] N. Mathur. Shopping malls, credit cards and global brands consumer culture and lifestyle of india's new middle class. *South Asia Research*, 30(3):211–231, 2010.

[24] R. Michman, E. Mazze, and A. Greco. *Lifestyle marketing: reaching the new American consumer*. Praeger Publishers, 2003.

[25] A. Narayanan and V. Shmatikov. De-anonymizing social networks. In *IEEE Symposium on Security and Privacy*, pages 173–187, 2009.

[26] C. Nie and L. Zepeda. Lifestyle segmentation of US food shoppers to examine organic and local food consumption. *Appetite*, 57(1):28–37, 2011.

[27] J. Pitman. Poisson–Dirichlet and GEM invariant distributions for split-and-merge transformations of an interval partition. *Combinatorics, Probability & Computing*, 11(05):501–514, 2002.

[28] S. G. Prus and E. Gee. Gender differences in the influence of economic, lifestyle, and psychosocial factors on later-life health. *Canadian Journal of Public Health*, 94(4):306–309, 2003.

[29] K. K. Rachuri, C. Efstratiou, I. Leontiadis, C. Mascolo, and P. J. Rentfrow. Metis: Exploring mobile phone sensing offloading for efficiently supporting social sensing applications. In *Proc. PerCom*. IEEE, 2013.

[30] R. T. Rockafellar and R. J-B Wets. *Variational analysis*, volume 317. Springer, 2011.

[31] R. Rosenthal. *Experimenter effects in behavioral research*. Halsted Press, 1976.

[32] G. Takeuti, W. M. Zaring, and G. Takeuti. *Introduction to axiomatic set theory*. Springer-Verlag, 1982.

[33] R. E. Tarjan. Efficiency of a good but not linear set union algorithm. *Journal of the ACM*, 22(2):215–225, 1975.

[34] R. E. Tarjan. *Data structures and network algorithms*, volume 14. SIAM, 1983.

[35] R. E. Tarjan and J. Van Leeuwen. Worst-case analysis of set union algorithms. *Journal of the ACM (JACM)*, 31(2): 245–281, 1984.

[36] P. Vyncke. Lifestyle segmentation from attitudes, interests and opinions, to values, aesthetic styles, life visions and media preferences. *European journal of communication*, 17 (4):445–463, 2002.

[37] W. Warren, R. Stevens, and C. McConkey. Using demographic and lifestyle analysis to segment individual investors. *Financial Analysts Journal*, pages 74–77, 1990.

[38] C. Werner and P. Parmelee. Similarity of activity preferences among friends: Those who play together stay together. *Social Psychology Quarterly*, pages 62–66, 1979.

[39] J. Zheng and L. M. Ni. An unsupervised framework for sensing individual and cluster behavior patterns from human mobile data. In *Proc. Ubicomp*, pages 153–162, 2012.

APPENDIX

A. PROOF OF THEOREM 1

PROOF. Since Algorithm 1 only visits each profile page once (line 2-4 of Procedure VisitProfile), and the merge operation (line 9) can be implemented using the Union-Find algorithm [33] (note that the number of connected URLs per profile page is limited by the types of networks), thus Algorithm 1 keeps the complexity of Union-Find with $O(|h_{\triangleright}|)$ find operations and $|h_{\triangleright}|$ elements, i.e., $O(|h_{\triangleright}|\alpha(|h_{\triangleright}|))$ [34, 35]. Therefore, the theorem holds when the following statements are true: 1) After termination, U contains all the profile pages in $h_{\triangleright}$; 2) At any time, there are no two users who have a joint profile page and 3) After each loop (line 18), for any two profile pages $p', p'' \in h_{\triangleright}$, where $p' \in u' \in U$ and $p'' \in u'' \in U$, if $u' \neq u''$ (i.e., p'' and p'' are merged into different classes), then p' and p'' are not connected.

The first statement is true since every visited page is added to a user in U at line 13. The second statement holds because every time, we add all the visited profile pages (line 8 of Algorithm 1) into a single user u', and no profile page is visited more than once. We assume the last statement does not hold, i.e., $u' \neq u''$ and $\exists$ an undirected path $\mathcal{W} = p_0(= p')p_1p_2 \ldots p_n(= p'')$ connecting p' and p'', where each p_ip_{i+1} are *directly connected* $\forall i = 0, 1 \ldots, n-1$ (see Section 3.2). Thus, $\forall i = 0, 1, \ldots, n-1$, $\exists$ a profile page c_i, s.t. the URLs of p_i and p_{i+1} co-occur on c_i, which leads to three cases: 1) $c_i = p_i$, thus $p_i \triangleright p_{i+1}$; 2) $c_i = p_{i+1}$, thus $p_{i+1} \triangleright p_i$; 3) $c_i \triangleright p_i$ and $c_i \triangleright p_{i+1}$. In each of the above cases, Algorithm 1 will add p_i and p_{i+1} into the same equivalence class, i.e., $u' = u''$, which yields a contradiction. Therefore, the theorem holds. □

B. INFERENCE OF RH-LDA

The collapsed Gibbs sampling process is summarized as follows: Given the current state of the sampler,$\{\mathbf{c}_{1:D}^{(t)}, \mathbf{z}_{1:D}^{(t)}\}$, iteratively for each individual $d \in \{1, 2, \ldots, D\}$,

1. Randomly draw $\mathbf{c}_d^{(t+1)}$ from $p(\mathbf{c}_d|\mathbf{w}, \mathbf{c}_{-d}, \mathbf{z}, \eta, \gamma)$, which is exactly the same as given in [7].

2. For each footprint $f \in \{1, 2, \dots, N_d\}$ of u, randomly draw $\mathbf{z}_{n,d}^{(t+1)}$ from

$$\begin{aligned} & p(z_{d,f} = l | \mathbf{z}_{-(d,f)}, \mathbf{c}, \mathbf{w}, m, \pi, \eta, \zeta, \upsilon) \\ \propto & p(z_{d,f} | \mathbf{z}_{d,-f}, m, \pi) p(w_{d,f} | \mathbf{z}, \mathbf{c}, \mathbf{w}_{-(d,f)}, \eta, \zeta, \upsilon) && (3) \\ & \prod_{d' \neq d : y_{d,d'} = 1} \psi_e(y_{d,d'} = 1 | \mathbf{c}_{(d, \mathbf{z}_d)}, \mathbf{c}_{(d', \mathbf{z}'_d)}, \zeta, \upsilon) && (4) \\ & \prod_{d' \neq d : y_{d,d'} = 0} \psi_e(y_{d,d'} = 0 | \mathbf{c}_{(d, \mathbf{z}_d)}, \mathbf{c}_{(d', \mathbf{z}'_d)}, \zeta, \upsilon). && (5) \end{aligned}$$

Eq. (3) is the same as hLDA, given in [7].

Eq. (4)= $\prod_{d' \neq d : y_{d,d'} = 1} \exp(\frac{\eta_k N_{d'}^k}{N_d^2 N_{d'}})$ and

Eq. (5)= $\prod_{d' \neq d : y_{d,d'} = 0} \left(1 - \exp\left(\frac{1}{N_d^2, N_{d'}} \sum_k \left(\eta_k N_d^k N_{d'}^k\right) + \upsilon\right)\right)$,

where $k = \mathbf{c}_{d,l}$ is the assigned living pattern of n, N_d^k is the number of footprints assigned with living pattern k, and N_d denotes the number of footprints in u.

3. Iteratively learn the parameters ζ and υ, using the method provided in the appendix of [9].