Exploring the Social Learning of Taxi Drivers in Latent Vehicle-to-Vehicle Networks

Tong Xu, *Member, IEEE*, Hengshu Zhu, *Member, IEEE*, Hui Xiong, *Senior Member, IEEE*, Hao Zhong, and Enhong Chen, *Senior Member, IEEE*

Abstract—With recent advances in mobile and sensor technologies, a large amount of efforts have been made on developing intelligent applications for taxi drivers, which provide beneficial guidance for improving the profit and work efficiency. However, limited scopes focus on the latent social interactions within cab drivers, and corresponding social learning mechanism to share driving behavior patterns has been largely ignored. To that end, in this paper, we propose a comprehensive study to discover how social learning affects taxi drivers' driving behaviors. To be specific, by leveraging the classic social influence theory, we develop a two-stage framework for quantitatively measuring the latent propagation of driving patterns within taxi drivers. Validations on a real-word data set collected from New York City clearly verify the effectiveness of our proposed framework with better explanation of future taxi driving pattern evolution, which prove the hypothesis that social factors indeed improve the predictability of taxi driving behaviors, and further reveal some interesting rules on social learning mechanism.

Index Terms—Behavior analysis, mobile data mining, social network

1 Introduction

Recent years have witnessed the growing interests on data-driven technologies for developing new paradigms of taxi business. Let's take a look at the dramatic expansion of urban areas and population, for instance, the population of Beijing increases nearly 35 percent during the last decade, while on the contrary, the amount of taxis keeps almost the same. This phenomenon not only raises the novel business mode of taxi services like Uber and Didi in China, but also urges the demand of more efficient and intelligent taxi services, which cannot be solved by simply increasing the amount of cabs or drivers.

At the same time, thanks to the rapid development of wireless sensor technologies in mobile environments, such as GPS, Wi-Fi and RFID, the abundant real-time trajectories could be promptly collected [2] to support the deep analysis of taxi trajectory records. Along this line, a variety of intelligent services can be enabled for extracting effective transportation patterns, e.g., the fastest driving route like [3] and [4], sequence of pick-up points [5] or passengers within the shortest driving distance [6]. Usually, these techniques will lead to the improvement of work efficiency and profit of taxi drivers. However, in some cases, they might be inadequate with ignoring the subtle differences between two types of taxi services. Generally, in most regions of America and Europe, taxis pick up passengers via appointments, which could be easily controlled by the centralized command center. But, in Asia like China or Japan, or huge cities in USA like New York, taxis wander along the streets to search and pick up the next passengers. In these cases, driving routes could be more random and personalized, thus intelligent services may fail to control the situation.

Moreover, prior arts may suffer some defects as follows. First, the case-by-case recommendations are sensitive to the current context, thus frequent update is required, which results in heavy burden of computation. Second, it will be difficult to distribute the cabs for keeping regional balances. Last but not least, predictability of taxi route might be limited due to *personalized habits*. Different from the algorithms with unified optimization task, e.g., maximal benefit or shortest distance, taxi drivers, especially those experienced ones, usually hold their own driving habits. For example, Fig. 1 illustrates the driving routes of one taxi driver, which is a snapshot extracted from a visualization app for New York City taxi services.[1] Interestingly, we find this driver tend to drive just around the central park. In these cases, if recommender system designs faraway routes without considering personalized habits, the drivers may tend to refuse even with higher benefits.

To formulate the personalized driving habits, intuitively, we would like to reveal how these habits emerge and evolve. Generally, on the one hand, those experienced drivers, who are sensitive to the routes and rules, could effectively summarize the patterns and regulate the routes by

- *T. Xu and E. Chen are with the Anhui Province Key Laboratory of Big Data Analysis and Application, School of Computer Science, University of Science and Technology of China, Hefei 230027, China. E-mail: {tongxu, cheneh}@ustc.edu.cn.*
- *H. Zhu is with the Baidu Talent Intelligence Center, Baidu Inc., Beijing 100085, China. E-mail: zhuhengshu@baidu.com.*
- *H. Xiong and H. Zhong are with the Management Science and Information Systems Department, Rutgers Business School, Rutgers University, Newark, NJ 07102. E-mail: {hxiong, h.zhong31}@rutgers.edu.*

Manuscript received 16 Nov. 2017; revised 17 Mar. 2019; accepted 29 Apr. 2019. Date of publication 7 May 2019; date of current version 1 July 2020. (Corresponding author: Enhong Chen.)
Digital Object Identifier no. 10.1109/TMC.2019.2915228

1. http://chriswhong.github.io/nyctaxi/

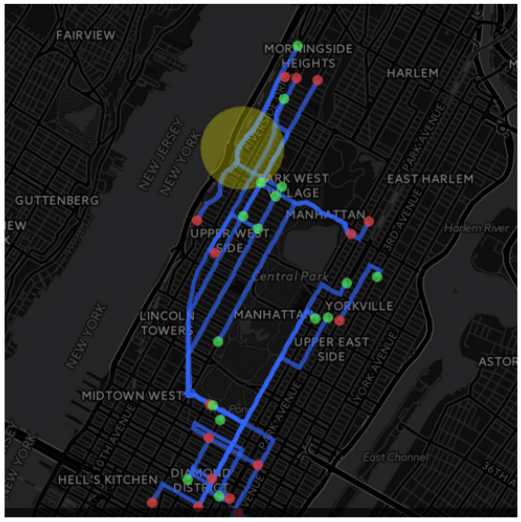


Fig. 1. A toy example of personalized driving behavior patterns.

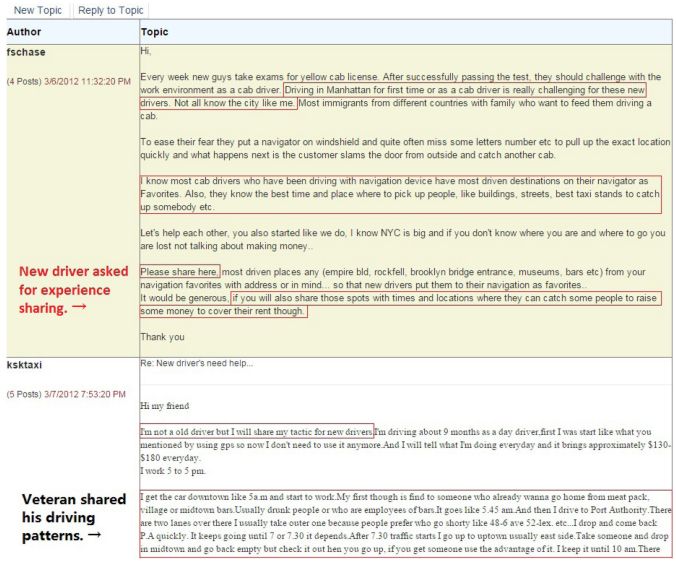


Fig. 2. Example of social sharing with detailed driving patterns.

themselves. On the other hand, for those inexperienced ones or newcomers, thanks to rapid development of social network services (SNS), not only offline gatherings (e.g., refueling or lunching), but also online communications (e.g., forums, or SNS platforms) are now available, where driving experience could be shared within drivers, namely the "*social learning*" mechanism of driving behaviors. A motivating example demonstrated in Fig. 2 may intuitively illustrate this phenomenon, in which questions from a new taxi driver in New York City have been replied by experienced drivers with comprehensive suggestions, e.g., some detailed time and trajectories, or even specific locations to pick up certain type of passengers for more benefits. Based on these suggestions, newcomers could now adjust their driving patterns to improve work efficiency.

If we review the above example in the "social" perspective, taxi drivers here could be treated as "*mobile social agent*" in the latent vehicle-to-vehicle social network, and the driving patterns are "*propagated*" within them. Correspondingly, with simulating the function of "*social propagation*" in the driving pattern evolution, the predictability of taxi routes may be improved, especially for those who wander through the street and randomly pick up passengers. Along this line, some social-oriented taxi services, e.g., "tutors" or pattern recommendations for taxi drivers with "People's Choice" could be further supported.

Unfortunately, due to the constraint of user privacy, there is no exposed signal for social interactions to be observed. To deal with this task, we aim at exploring latent connections among taxi drivers based on the analysis of their driving behaviors. To be specific, we propose a partial-ranking framework, to capture the latent propagation of driving patterns. Specifically, we realize that some common driving patterns are shared within taxi drivers, e.g., some popular routes with high benefits. However, different drivers may hold different habits on these driving patterns, which lead to personalized proportions of patterns. Moreover, preferences may change due to "social influence" from experienced drivers, which lead to the varying proportions of patterns. Thus, we intuitively assume that the increasing proportion of driving patterns may be caused by stronger influence, and vice versa. Validations on a real-word data set clearly validate the effectiveness of our proposed framework with better explanation of future taxi driving pattern evolution, which prove the hypothesis that social factors indeed improve the predictability of taxi driving behaviors, and further reveal some interesting rules on social learning mechanism. To the best of our knowledge, we are among the first ones who discuss the "social learning" mechanism among taxi drivers, and then investigate the impact of social factors for modeling taxi driving patterns evolution.

Overview. The rest of this paper is organized as follows. We summarize related works in Section 2, and then propose our technical solution for analyzing driving pattern propagation with integrating social factors and various constraints in Section 3. Afterwards, we evaluate the performance with extensive validations in Section 4, and conduct more comprehensive discussions on social learning mechanism in Section 5. Finally, in Section 6, we conclude the paper.

2 Related Work

In this paper, we target at revealing the "social factors" within taxi driving behaviors. Thus, three types of prior arts are related to our research, namely traditional studies on taxi business, social-related techniques, as well as researches on urban-based social network.

Indeed, plenty of efforts have been made on the intelligent taxi services, e.g., recommending hotspots to pick up passengers quickly [7], planning practically fastest route [8] or an optimal sequence of pick-up points [6], and even scheduling taxi in a multi-source data fusion perspective [9]. At the same time, for taxi passengers, prior arts may also list locations to easily achieve vacant taxi [7], or support them to share taxi with optimal candidates [10]. Besides, some other prior arts focus on different aspects of taxi analysis. For instance, [11] effectively achieved the optimal route for mobile sequential recommendation to empty taxi cabs, while [12] studied the strategy to pick up passengers with creating decision trees. Moreover, some applications based on taxi driving records

were designed, e.g., isolated trajectories detection [13], traffic jams detection [14], road capacity design [15] and nightly bus routes design [16]. However, few of them studied the taxi driving behaviors in the social perspective, and the latent vehicle-to-vehicle network with experience sharing has been largely ignored.

Another related topic is the social analysis. Since social factors were analyzed for marketing in [17], the "word-of-mouth" effect has become one of the hottest issue in recent years, and several prior arts, e.g., [18] and [19] were proposed to model represent the step-by-step dynamics of influence. Among them, Independent Cascade (IC) model [20] is widely-studied due to intuitive simulation, which motivated several linear approximation like [21], as well as some extended frameworks like [22] for more general application. Along this line, some related works targeted at estimating the influence strength, like [23] discussed several heuristic method to reveal link strength, [24] predicted diffusion probabilities by using the EM algorithm, and [25] investigated the topic-sensitive interactions via re-producing the information propagation process. Besides, more issues were discussed on evolving social network, e.g., [26] studied on how to model the implicit social diffusion with time decay, and [27] attempted to rapidly extract the topic-sensitive subgraphs within evolving graph streams.

With combining social analysis with urban computing issues, prior arts mainly focused on the correlations between social factors in cyber network and human behaviors in physical world. For instance, [28] revealed that 10-30 percent of human mobility could be explained by social factors, and [29] announced that more cohesive communities will be found for offline event-driven social networks. Correspondingly, mobility patterns or even context-aware mobility preference [30], in return, shape and impact social connections like [31] and [32]. As connections revealed, some prior arts attempted to analyze the social effects in offline social network, e.g., [33] discussed about the homogeneity and influence in LBSN, [34] recommended offline geo-friends based on heterogeneous network analysis, [35] discussed the dynamic social influence for event participation, and [36] further formulated this task under conflicting situations. Recently, vehicular social networks has been analyzed in the perspective of Internet of Vehicles (IoVs) to support applications of smart cities like [37] and [38]. However, they mainly focused on global trend with crowd sensing, but not behavior modeling for individual drivers.

Different from these prior arts, to the best of our knowledge, we are among the first ones who analyze the latent vehicle-to-vehicle network within taxi drivers, and then discuss the "social learning" mechanism with investigating the impact on driving behaviors evolution of taxi drivers, which could be novel compared with related works.

3 Technical Solution: Simulating Driving Patterns Propagation

In this section, we will introduce how we reveal and leverage the latent social factors within taxi driving behaviors, in which the technical solution for social influence modeling, optimization task, and extended framework with comprehensive constraints will be explained in detail.

3.1 Preliminary Statement

As mentioned above, we target at discovering the latent social factors within taxi driving behaviors. Thus, in this paper, we mainly focus on the "*social factors*" to explain the evolution of driving patterns. For the other factors, e.g., the profit, traffic status or unexpected incidents that may also influence the routes, we will study parts of them in the following discussions, and conduct more comprehensive researches in future works.

First of all, as mentioned above, we realize that some *common driving patterns* are shared by taxi drivers. Moreover, different drivers may hold different preference on these patterns, which lead to the different proportion of each pattern. To be specific, if K patterns are extracted in total, correspondingly, we could define the K-dimensional vector $\mathbf{s_i}$ to describe the driving behaviors of a taxi driver u_i, in which each $s_{i,k}$ indicates the proportion of kth pattern in u_i's driving behaviors. Definitely, the vector is normalized by $\sum_k s_{i,k} = 1$. Along this line, since driving patterns may evolve as time flies, we further enrich the pattern vector with introducing the timestamp as $\mathbf{s_i^t}$, which indicates the driving behavior of driver u_i in tth round. More details for the driving pattern definitions will be explained in validation part in Section 4.1.

Then, to describe "social learning" mechanism, i.e., social propagation of driving patterns within taxi drivers, we formally define the *vehicle-to-vehicle network* as follow:

Definition 3.1 (Vehicle-to-Vehicle Network). *Similar with social network, a Vehicle-to-Vehicle Network could be formulated as $G = \langle V, E, W \rangle$, where $u_i \in V$ denotes the taxi drivers, and $e_{ij} \in E$ indicates the connection from u_i to u_j. Finally, we have $w_{ij} \in W$ which corresponds to e_{ij} to indicate the edge weight, or the influential strength to propagate the driving patterns.*

What should be noted is that the Vehicle-to-Vehicle Network could be asymmetric and all edges are directional. Afterwards, we have N_i to present the social "neighbors" of driver u_i, and $u_j \in N_i$ if e_{ij} exists, which also means that u_i's driving patterns may be influenced by u_j. In this case, N_i could be treated as "*tutors*" of u_i. Moreover, to integrate drivers' own opinions, i.e., some drivers tend to insist their own patterns. Thus, we treat each driver as *tutor of itself*, i.e., $u_i \in N_i$ with w_{ii} exists. Obviously, higher w_{ii} indicates less propagation and more persistence, and vice versa. The mathematical notations are summarized in Table 1.

Finally, considering that usually no exposed signal for social interactions could be observed, thus, we target at revealing the latent vehicle-to-vehicle connections via modeling the social propagation of driving patterns. Along this line, we define the overall problem as follow:

Definition 3.2 (Overall Problem). *Given the set of drivers $V = \{u_i\}$, as well as their driving behavior records $\mathbf{s_i^t}$ during a period $t = 1, 2 \ldots T$, we target at revealing the vehicle-to-vehicle network $G = \langle V, E, W \rangle$, so that latent social learning will be captured for better understanding their driving behaviors.*

3.2 Loss Function for Partial Ranking

With preliminaries introduced and notations summarized, now we turn to simulate the social propagation of driving patterns. Specifically, as introduced above that driving preferences may change due to "social influence" from other experienced drivers, which lead to the varying proportions

TABLE 1
Mathematical Notations

Symbol	Description
$\mathbf{U} = \{u_i\}$	the set of taxi drivers
w_{ij}	social connection strength from u_i to u_j
$\mathbf{s_i^t}$	pattern frequency vector of u_i in time t
$s_{i,k}^t$	proportion of kth pattern in time t
$p_{i,k}^t$	social influence of kth pattern in time t
N_i	social neighbors of driver u_i
$\mathbf{R}_i^t$	the pattern of u_i that raise in time t
$\mathbf{D}_i^t$	the pattern of u_i that decrease in time t

of patterns. Thus, we attempt to reveal the *social influence* $p_{i,k}^t$ of kth pattern to driver u_i, while the detailed formulation of p^t will be explained in next section.

As we mainly focus on the social factors which affect the evolution of driving patterns, as mentioned above, we intuitively assume that the increasing proportion of driving patterns may be caused by stronger influence, and vice versa. Thus, for a taxi driver u_i, if kth pattern holds an increasing proportion in round $t+1$, we conclude that the social influence $p_{i,k}^t$ could be relatively higher, and vice versa. Following this assumption, intuitively, the increasing/decreasing trend of different driving patterns may reflect the partial ranking of social influence.

To be specific, if we define the set of increasing patterns in time t as $\mathbf{R}_i^t$, i.e., $\forall r \in \mathbf{R}_i^t, \Delta s_{i,r}^{t+1} = s_{i,r}^{t+1} - s_{i,r}^t > 0$. Similarly, $\mathbf{D}_i^t$ presents set of decreased patterns. Then, for each pair of patterns like $\forall \langle r, d \rangle_{i,t}$ where $r \in \mathbf{R}_i^t$ and $d \in \mathbf{D}_i^t$, since increasing proportion may indicate stronger influence, and vice versa, we may obtain pairwise ranking of corresponding social influence, i.e., $p_{i,d}^t < p_{i,r}^t$.

With the assumption above, we realize that the latent social connections could be accurately estimated with optimally reveal the partial ranking of social influence $\{p_{i,k}^t\}$. Thus, the task of revealing latent social connections w_{ij} will be summarized as a partial ranking problem as follows:

Definition 3.3 (Ranking Objective). *Revealing appropriate w_{ij}, so that for $\forall \langle r, d \rangle_{i,t}$ where $r \in \mathbf{R}_i^t$, we will have $p_{i,d}^t < p_{i,r}^t$.*

To deal with this task, we formulate the loss function of pairwise ranking problem as follows:

$$min_w \mathcal{F}(w) = \sum_{i,t} \sum_{r \in \mathbf{R}_i^t, d \in \mathbf{D}_i^t} h\big(p_{i,d}^t - p_{i,r}^t\big), \tag{1}$$

where $h(\gamma_{rd})$ is a loss function to assign a non-negative penalty based on the partial ranking of social influence, in which $\gamma_{rd} = p_{i,d}^t - p_{i,r}^t$. Usually, we have $h(\gamma_{rd}) = 0$ when $p_{i,d}^t \leq p_{i,r}^t$, i.e., correct ranking. When $p_{i,d}^t > p_{i,r}^t$, i.e., wrong ranking, we have $h(\gamma_{rd}) > 0$ as penalty.

To ease the computation, "squared loss function" is widely utilized to estimate the penalty as follow:

$$h(x) = \max\{x, b\}^2. \tag{2}$$

In squared loss function, usually we have a soft margin parameter b to tolerate a tiny error. Here we simply treat $b = 0$ to present no tolerance to the error, thus $h(x)$ of given u_i and t could be re-formulated as

$$\sum_{r \in \mathbf{R}_i^t, d \in \mathbf{D}_i^t} h\big(p_{i,d}^t - p_{i,r}^t\big) = \sum_{r,d: p_{i,d}^t > p_{i,r}^t} \big(p_{i,d}^t - p_{i,r}^t\big)^2. \tag{3}$$

3.3 Social Propagation Simulation and Optimization

Then, we turn to formulate the social influence within taxi drivers. To simulate the propagation process, here we adapt the Steady State Spread (SSS) model [19], which follows the basic format of IC model [20] and could be replaced by other social influence simulation model if needed. Specifically, in the SSS model, all the nodes attempt to influence their neighbors, and then influence will be measured not only by connection strength, but also their current status. Thus, the social influence could be formulated as follow:

$$p_{i,k}^t = 1 - \prod_{j \in N_i} \big(1 - w_{ji} \cdot \delta_{i,j,k}^{t-1}\big), \tag{4}$$

in which $\delta_{i,j,k}^{t-1}$ presents the current status of influential node, i.e., the social "tutor". Here, we design this parameter to present the pattern-sensitive strength of social influence, which is different from the overall strength w_{ji}. Intuitively, when a driver learn from a social "tutor", those patterns hold by the "tutor" will generate social influence on this driver to develop a new driving pattern or enhance existing patterns. And definitely, the more significant difference on proportion between driver and "tutor", namely $s_{j,k}^t - s_{i,k}^t$, lead to the stronger influence on the corresponding pattern. Therefore, the current status of social "tutor" on kth pattern could be utilized to measure the difference of proportion, which could be formulated as the *Sigmoid function* as follow:

$$\delta_{i,j,k}^t = \frac{1}{1 + e^{-\left(s_{j,k}^t - s_{i,k}^t\right)}}. \tag{5}$$

Based on the formulation, $\delta_{i,j,k}^{t-1}$ will be controlled within [0, 1], and the relation between $s_{j,k}^t$ and $s_{i,k}^t$ will affect the influence, i.e., if $s_{j,k}^t > s_{i,k}^t$, we will have $\delta_{i,j,k}^{t-1}$ near 1 to enhance the influence, while for $s_{j,k}^t < s_{i,k}^t$, the pairwise influence will be impaired.

Finally, we could now optimize the loss function Equation (1) to estimate latent social connection strength w_{ij}. To be specific, gradient descent methods will be introduced to achieve the approximated w_{ji} with minimizing $\mathcal{F}(w)$. Specially, with defining $\gamma_{rd} = p_{i,d}^t - p_{i,r}^t$, we have the derivative of $\mathcal{F}(w)$ with respect to w_{ji} as follow:

$$\frac{\partial \mathcal{F}(w)}{\partial w_{ji}} = \sum_t \sum_{r \in \mathbf{R}_i^t, d \in \mathbf{D}_i^t} \frac{\partial h(\gamma_{rd})}{\partial \gamma_{rd}} \left(\frac{\partial p_{i,d}^t}{\partial w_{ji}} - \frac{\partial p_{i,r}^t}{\partial w_{ji}} \right), \tag{6}$$

where $h'(x)$ could be easily achieved as derivation of *square loss function*, while for the social influence part, we have

$$\frac{\partial p_{i,k}^t}{\partial w_{ji}} = \prod_{l \in N_i^t, l \neq j} \big(1 - w_{li} \cdot \delta_{i,l,k}^{t-1}\big) \cdot \delta_{i,j,k}^{t-1}. \tag{7}$$

According to the formulations, finally gradient descent methods could be exploited to deal with the optimization task. The data stream of proposed framework and optimization task is summarized in Algorithm 1.

Algorithm 1. Optimization for Ranking Task.

Input: A set of taxi drivers $\mathbf{U} = \{u_i\}$, corresponding driving transactions records $\mathbf{E}$, and time lag T;
Store: Driving pattern $\mathbf{s}_{i,t}$ for each $u_i \in \mathbf{U}$ in time t;
Output: Social connections between drivers $\{w_{ij}\}$

for $u_i \in \mathbf{U}$, $t = 1, 2, \ldots, T$
  extract $\mathbf{s}_{i,t}$ from $\mathbf{E}$;
  if $t > 1$ **then**
    if $s_{i,k}^{(t-1)} < s_{i,k}^{t}$
      then $\mathbf{R}_i^t = \mathbf{R}_i^t \bigcup k$;
      else $\mathbf{D}_i^t = \mathbf{D}_i^t \bigcup k$;
    end if
  end if
end for
  Iteration = *True*;
while (Iteration)
  Iteration = *False*;
  for $u_i, u_j \in \mathbf{U}$, $t = 2, \ldots, T$
    for $r \in \mathbf{R}_i^t$, $d \in \mathbf{D}_i^t$
      update w_{ij} based on Equation (7);
    update $\langle \mathbf{p}_i, h_{i,0} \rangle$ and $\{w_{ij}\}$ until convergency;
    if w_{ij} changed more than threshold **Iteration** = *True*;
    end if
  end for
end while
return $\{w_{ij}\}$;

3.4 Extended Framework with Constraints

Finally, we turn to extend our framework with comprehensive constraints, thus social impacts will be enhanced by other factors, e.g., location, skills and so on. Along this line, we could also discover which factor may urge the birth of, or hold high correlation with latent social connections.

3.4.1 Extended Loss Function

Specifically, to constrain the social influence and refine the loss function of proposed framework, we intuitively assume that the strength of social connections should be *proportional* to the integrated factors. Thus, considering that different factors may suffer different orders of magnitude, to present the proportional relationship, similar with loss function in Equation (1), we introduce the pairwise ranking method, i.e., higher score (which may indicate stronger social influence) leads to stronger social connection. Then, the loss function will be extended as follow:

$$min_w \mathcal{F}(w) = \sum_{i,t} \sum_{r \in \mathbf{R}_i^t, d \in \mathbf{D}_i^t} h(p_{i,d}^t - p_{i,r}^t) + \lambda \sum_i \sum_{j,k, \forall score_{i,j} < score_{i,k}} \max(w_{ji} - w_{ki}, 0). \tag{8}$$

Here $score_{i,k}$ presents the measure of constraint-oriented factors, while $\max()$ indeed presents the penalty to ensure that ranking relationship between $\langle score_{i,j}, score_{i,k} \rangle$ and $\langle w_{ji}, w_{ki} \rangle$ should be the same. Also, λ here actually means the weight of constraints, and a higher weight leads to a more solid constraint on pairwise ranking. Correspondingly, the gradient function could be extended as follow:

$$\frac{\partial \mathcal{F}(w)}{\partial w_{ji}} = \sum_t \sum_{r \in \mathbf{R}_i^t, d \in \mathbf{D}_i^t} \frac{\partial h(\gamma_{rd})}{\partial \gamma_{rd}} \left(\frac{\partial p_{i,d}^t}{\partial w_{ji}} - \frac{\partial p_{i,r}^t}{\partial w_{ji}} \right) + \sum_{k, \forall k \neq i,j} h(I(score_{i,k} - score_{i,j}) \cdot I(w_{ji} - w_{ki})). \tag{9}$$

As we targets at ensuring the same partial order of ranking, the penalty function $h()$ in the loss function 1 is also borrowed, i.e., if $\langle score_{i,j}, score_{i,k} \rangle$ and $\langle w_{ji}, w_{ki} \rangle$ hold the different pairwise ranking, the product inside $h()$ should be 1, which result in a penalty as 1. On the contrary, the same ranking relationship will lead to no penalty (as 0). Also, $I()$ here means the symbolic function to achieve the sign of difference. With this extended framework, comprehensive constraints, even those without detailed value but only pairwise ranking relationship, could be adopted now to refine the latent social network.

3.4.2 Different Types of Constraints

With optimizing the extended loss function above, we could now reveal the relation between social connection strength and comprehensive factors. To be specific, three types of factor are selected as follows:

1. Counts of co-occurrence

Co-occurrence has been widely studied and utilized in researches on location-based social network, e.g., [39] and [34], as a heuristic method to describe the latent so- cial connections. With GPS records, co-occurrence could be easily estimated, while if there is no GPS coordinates, similarly, we counts the frequency of "co-destinations" to approximate the coefficient.

For instance, if two cabs arrive the same destinations at almost the same time, and stay during a period (e.g., 10 minutes) without picking up another passenger, they may possibly communicate after drop-off, and share driving experience. It is only a rough approximation as no interaction could be ensured. However, usually more co-occurrences may indicate more chances for social learning.

2. Homogeneity between drivers

Homogeneity also means the similarity of driving behaviors vectors. Intuitively, following the traditional assumption of social connection that friends usually hold similar preference, more similar driving behaviors may lead to stronger connection, and then stronger social influence.

3. Levels of different skills

It could be easily understood as top drivers on certain skills could be the more attractive to the rest. To be specific, four skills are considered as follows:

- *Amount of transactions*, which indicates the work effectiveness (i.e., more business).
- *Average driving speed*, which indicates the efficiency (i.e., faster trip).
- *Total income*, which indicates the financial profit (i.e., higher rate of return).
- *Number of followers*, which measures the level of social-oriented skill.

As the framework extended, more constraints could be added in future work, if more comprehensive data sets are available. We will discuss and compare these constraints in validation part in Section 4.5.

TABLE 2
Data Set Description

	Data Statistic
Number of Taxis	14,144
Number of Drivers	43,191
Average Num. of Transactions	3,928.86
Average Num. of Passengers	1.68
Average Trip Time	15.05 min
Average Trip Distance	8.86 miles
Average Trip Fare	$15.39

4 Validations: Predictability of Driving Pattern Evolution

As we target at revealing and leveraging the latent social factors to better explain the evolution of taxi driving behaviors, in this section, we will conduct extensive validations on a real-world data set to verify our hypothesis.

4.1 Data Set Description

First of all, we will briefly introduce the data set we extracted for validation, as well as the details of pre-processing for driving patterns.

4.1.1 Brief Introduction to Data

We conduct our study on a real-world data set collected from the taxi driving transactions in New York City during the whole year of 2013, which is published by NYC Taxi and Limousine Commission (NYC TLC). This is a large-scale data set which totally consists of more than 169 million transaction records of 43,191 drivers in 14,144 cabs. For each transaction, we have the spatial and temporal information for both pick-up and drop-off, as well as fares including tip and toll. The statistical details of data set are shown in Table 2, and the distribution of transaction amount for each driver is shown in Fig. 3. Clearly, we find that transactions amount suffers the long tail effect, thus, to ensure the performance and prevent the interference of data sparsity, we have to remove some drivers based on their transaction amounts.

4.1.2 Data Set Pre-Processing

Then, we turn to introduce the details of data pre-processing, i.e., the extractions of driving patterns. Recently, prior arts may enrich the mobile patterns with specific track and contextual information [40]. However, due to the limitation of the NYC TLC data set, the GPS coordinates are not included, thus patterns considering specific track could not be achieved. Instead, we define a driving pattern as a *triple* which contains the *pick-up area*, *drop-off area* and the *pick-up time*, e.g., we have (World Trade Center, Wall Street, 7:00 AM-9:00 AM) as a pattern. The rest information, e.g., the driving speed or fare will be treated as features, and discussed later in Section 5.

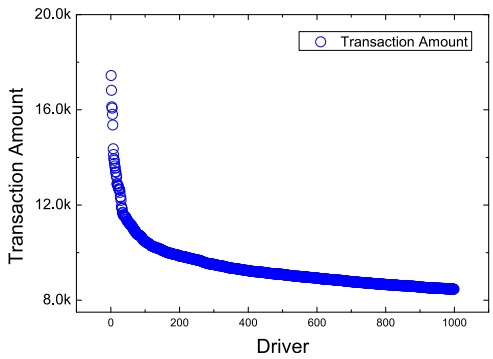


Fig. 3. The overall distribution of transaction amount for each driver.

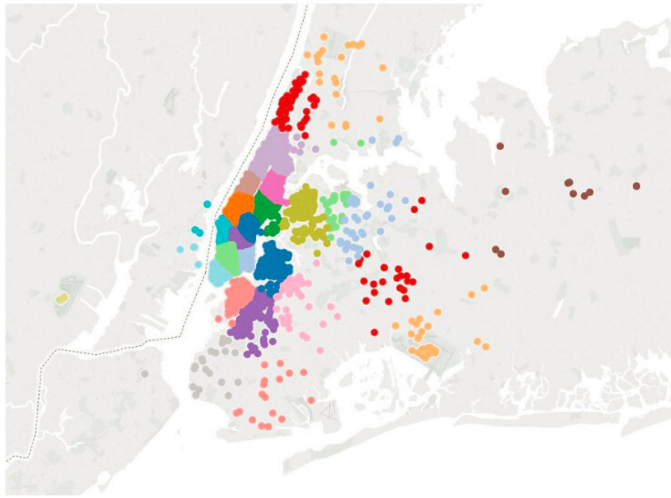


Fig. 4. Clustering results of New York City Locations.

For the pick-up / drop-off area, we first collected all the pick-up and drop-off locations in the historical transaction records. Along this line, we conducted a bottom-up *hierarchical clustering* with minimum variance criterion until only 30 clusters were kept. The clustering result is shown in Fig. 4, and the sensitiveness of cluster amounts will be discussed in Section 4.6 as a parameter.

According to the statistical analysis on pattern extraction, we realize that the distribution of patterns could be imbalanced, i.e., most of the taxi transactions happened in a few hot routes. For instance, as shown in Fig. 5 which summarizes the most frequent patterns appear around 8:00 AM as the peak of the morning rush hour, we can see many passengers take taxi from WTC station to the downtown for work, or return home from JFK Airport after the long-time international flight. On the contrary, in Fig. 6, which indicates the most frequent patterns around 6:00 PM, passengers return home in Queens and Brooklyn after one day's work. Along this line, we generalized the patterns with divided period as long as every 2 hours, e.g., 7:00 AM to 9:00 AM, and then 24 hours lead to 12 intervals. As mentioned above, due to the imbalance distribution of transaction amount with respect to different areas, only the most frequent patterns are considered in order to reduce the

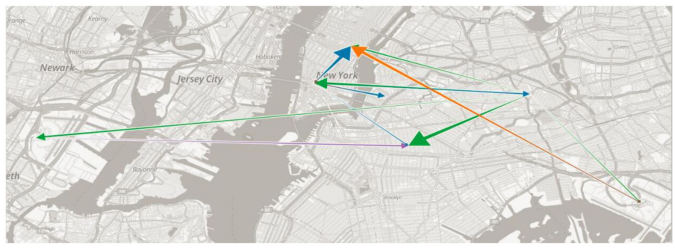


Fig. 5. Frequent patterns in NYC taxi driving around 8:00 AM.

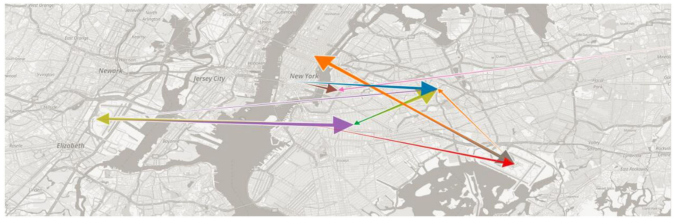


Fig. 6. Frequent patterns in NYC taxi driving around 18:00 PM.

interference of data sparsity. The sensitiveness of pattern amount will be also discussed in Section 4.6.

4.2 Validation Settings

As data set described above, in this section, we will introduce the validation settings, including the two-stage framework, selected baselines, and the evaluating metrics.

4.2.1 Two-Stage Framework for Validation

According to our intuitive assumption in Section 3.2, if we optimally reveal the partial ranking of social influence $\{p_{i,k}^t\}$, we will accurately reveal the latent social connections among taxi drivers. Correspondingly, more precise estimation of latent social connections will lead to better prediction of drivers future behaviors. Along this line, we design the validation as a *two-stage framework* as follows:

Training Stage. Given a group of taxi drivers $V = \{u_i\}$ and their driving pattern vectors $\mathbf{s_i^t}$ during the period $t = 1, 2, \ldots, T$, in the training stage, we aim at inferring the latent vehicle-to-vehicle network $G = \langle V, E, W \rangle$, which achieve the best explanation for the partial ranking of driving pattern evolution $\mathbf{\Delta s_i^{T+1}}$.

Test Stage. After obtaining the latent vehicle-to-vehicle network $G = \langle V, E, W \rangle$, in the test stage, given the taxi drivers group $V = \{u_i\}$ with their pattern vectors $\mathbf{s_i^t}$ during the $\mathbf{p}$-*time lag* as $t = T - p + 1, \ldots, T - 1, T$, we aim at predicting the driving behavior vector fluctuation $\mathbf{\Delta s_i^{T+1}}$ with accurate sign and ranking.

With the definitions above, latent social connections revealed in the first stage will be leveraged for predicting drivers pattern evolution, and the *performance of prediction* task will be measured to validate the effectiveness of our framework. The framework is summarized in Fig. 7, which illustrates the three steps (two steps for two-stage framework, as well as one step for data pre-processing) of our validations, where *blue arrows* indicate the work flow, and *red arrows* mean the data flow.

4.2.2 Tasks and Evaluating Metrics

As the precise estimation of driving patterns evolution could be a tough task, we design two tasks to measure the performance in different perspective, i.e., the *binary classification* to distinguish the trend (incresing / decresing) of pattern evolution, and then *ranking* the patterns with respect to their increments. For each task, related metrics will be selected to measure the performance.

For the binary classification task, typically, we select the common used *Precision* and *Recall* rates for validation.

For the ranking task, similar with the state-of-the-art learning to rank problems, Normalized Discounted cumulative

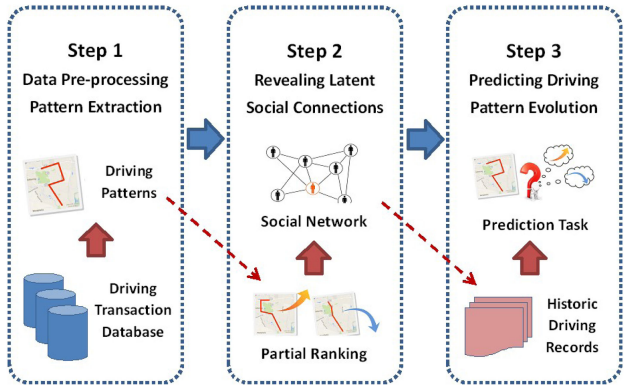


Fig. 7. Framework designed for the validations with three steps.

gain (*NDCG*) and Mean Average Precision (*MAP*) are selected. Specially, we get NDCG following the equation $\frac{DCG}{iDCG}$, in which $iDCG$ presents the *ideal results* of DCG with all rank- ings are correctly estimated, and DCG will be calculated based on the formulation as below:

$$DCG = \sum_i \frac{2^{r_i} - 1}{\log(1 + i)}, \tag{10}$$

where r_i denotes the relevance of result, which is set as reversal order of correct ranking in our validation. Furthermore, when calculating *MAP*s, we treat the top 10 patterns in ground truths as "*expected results*", and the score will be calculated based on their ranks in the result list.

4.2.3 Selected Baselines

Generally, in this paper, we attempt to discover the evolution of taxi driving patterns. To study the effects of "social learning" mechanism, we select baseline which mainly focus on the temporal evolution, or time-series analysis. Thus, here we exploit three baselines as follows:

1. *Personalized Average (Ave).* As the basic time-series analytical tool, we follow the simple assumption that driving patterns may evolve just around the average value. Thus, this baseline uses the average value of previous p intervals to predict the evolution of pattern frequency in next round.
2. *Overall Popularity (Pop).* Another heuristic assumption is that drivers will follow the overall popularity to update their own patterns. Based on this assumption, we intuitively rank the overall popularity for ranking task. For the binary classification, we compare the ranking of last time interval to distinguish the increasing / decreasing trend.
3. *Vector Autoregression (VAR)* [41]. Classical econometric model to capture the linear interdependencies among multiple time series, which suits modeling the auto regression for more than one evolving variable. As we analyze the evolution of multiple patterns simultaneously, it will be proper to utilize the VAR model. What should be noted is that there will be one personalized VAR model trained for one driver, and the estimation results will also be normalized.

In summary, two baselines, i.e., personalized average and VAR model are selected based on time-series estimation,

TABLE 3
Overall Performance (SPC Indicates Our Method)

	SPC	Ave	Pop	VAR
NDCG	0.3502	0.1603	0.2211	**0.3619**
Improve (%)	-	+118.46	+58.39	-3.23
p-Value	-	<0.001	<0.001	0.755
MAP@10	**0.2128**	0.0254	0.1042	0.2018
Improve (%)	-	+737.79	+104.22	+5.45
p-Value	-	<0.001	<0.001	0.472
Precision	**0.1579**	0.0134	0.0474	0.0192
Improve (%)	-	+1078.35	+233.12	+722.39
p-Value	-	<0.001	<0.001	<0.001
Recall	**0.6892**	0.0298	0.4151	0.0875
Improve (%)	-	+2212.75	+66.03	+687.66
p-Value	-	<0.001	<0.001	<0.001

while one more baseline is chosen to reflect the overall popularity, which follows the similar assumption of our framework that taxi drivers tend to follow suggestions from external information source. Along this line, comprehensive analysis with different assumptions will be achieved.

4.3 Overall Results

As all the validation settings are introduced, we now show the overall prediction performance of our *SPC* (**S**ocial-aware **P**attern-**C**hange prediction) approach and baselines. To be specific, the top 300 patterns were studied and the time lag was set as 5 months (i.e., we have transactions in 5 months as training data to predict the evolution of 6th month). The parameter sensitiveness of 300 patterns will be discussed later. Similarly, to ensure the data quality with reducing sparsity, we selected top 100 taxi drivers with the most records in our validations, which could be extended if needed. Besides, we label the top 20 percent of results in ranking list as increasing for the binary classification task, while the detailed P-R curve will be studied at the end of this section.

The overall results are shown in Table 3, in which p-Values are listed to measure how our SPC method significant outperformed the baselines. Specifically, if p-Value is less than 0.05, the advantage of our method could be significant. According to the results, we realize that driving patterns of taxi drivers could be largely random, as all the performance are relatively poor. However, generally our approach outperforms the other baselines in most cases with dramatic margin, even 20 times better. These results highly support our assumption that social factors may better explain evolution of driving behaviors. The conclusion could also be partially supported by the comparison between overall popularity and personalized average, which indicate that taxi drivers will be glad to follow the social trend.

Another interesting finding is that for VAR model, it performs truly great in ranking task, but terribly fails in binary classification task. With deeply checking of the VAR output, we realize that usually VAR predicts the proportion as 0 or negative, not only for those patterns that the drivers never try (i.e., no training data), but also for the patterns that drivers tried for once but never reappear. Combined with the terrible performance on binary classification, we conclude that the ranking list of VAR might be meaningless as it fails to reveal the real pattern but only maintains the outmoded ones. In other words, due to the features of auto-regression, VAR model tend to "refused" the change, which impair its performance.

TABLE 4
Performance with Various Intervals

Interval Length (Months)		1	2	3	4
NDCG	*M*	0.3499	0.3442	0.3421	0.3393
	SD	0.0116	0.0181	0.0194	0.0213
MAP10	*M*	0.2127	0.2104	0.2101	0.2082
	SD	0.0113	0.0148	0.0165	0.0184
Precision	*M*	0.1579	0.1552	0.1543	0.1565
	SD	0.0092	0.0138	0.0151	0.0166
Recall	*M*	0.6870	0.6794	0.6723	0.6679
	SD	0.0262	0.0381	0.0371	0.0420

According to the results, we may finally draw the conclusion that the heuristic methods might not be appropriate to estimate driving patterns of taxi drivers if without considering additional factors, like financial benefits or running speed. This phenomenon might further explains why our model could outperforms the baselines, as we don't try to "*teach the model*" how to predict the evolution, but intuitively "*simulate the social learning mechanism*", which is finally proved as effective. Clearly, except for those intellectual services, taxi drivers themselves could be the "*best learner*".

4.4 Results for Prediction with Various Intervals

Then, we target at discovering the effectiveness of our SPC approach with different *time intervals*. Basically, we conduct validations on training data to predict drivers' behavior in following period, e.g., we train the model with driving records within the period [*January, June*] to predict the driving pattern evolution in *July*. However, since till now we only validate the performance on the next month, we would like to discover whether SPC is adequate for further prediction with a longer time interval, i.e., following several months. Under this validation, the robustness of our framework could be somehow verified.

To that end, we design an additional validation, in which the *time lag* is set as 4, and different time intervals are chosen as 1-4. For instance, given the interval set as 4 (months), the SPC model based on records during [*January, April*] will be used to reveal the patterns in *May*, *June*, *July* and *August*, separately. For each time interval, we conducted 5 sets of validations to achieve the average performance.

The results are shown in Table 4 with mean value (short as *M*) and standard deviation (short as *SD*). Generally, we realize that even with longer interval like 4 months, our SPC approach still performs well, better than most of the baselines shown in Table 3. These results may also proves that social connections might keep relatively stable within a short period, like sereval months. At the same time, according to the results, we find that the performance generally becomes worse with longer intervals, and standard deviations usally keep incresing which means that results tend to be more unstable. This phenomenon could be reasonable as parts of connections may change as time goes by, thus outdated social factors may mislead the prediction. In summary, balance

TABLE 5
Performance with Comprehensive Constraints

	NDCG	MAP@10	Precision	Recall
Original	0.3502	0.2128	**0.1579**	**0.6892**
+Co-occurrence	0.3577	**0.2180**	0.1564	0.6840
+Homogeneity	**0.3626**	0.2013	0.1372	0.6007
+Skill-Trans	0.3232	0.1718	0.1297	0.5659
+Skill-Speed	0.3077	0.1689	0.1274	0.5520
+Skill-Income	0.3175	0.1596	0.1201	0.5244
+Skill-Social	0.3214	0.1666	0.1200	0.5259

could be carefully achieved between prediction accuracy and forecasting advance.

4.5 Results with Comprehensive Constraints

After that, we turn to analyze the performance with constraints. The comparisons are shown in Table 5, in which the parameters are kept the same with overall performance in the former section, and λ is intuitively set as 1. Further, to discuss the sensitiveness of λ, we also conduct two more sets of validations on different λ as 0.1 and 10, which are shown in Fig. 8. Here we only list the trend of homogeneity constraint and skill constraint with income level, with NDCG (MAP could be similar), Precision and Recall metrics, while the rest of constraints share the similar trend.

Generally, we realize that for the results with pairwise constraints, e.g, co-occurrence and homogeneity outperform the original ones. Considering that when $\lambda = 0.1$, the performance will be even further improved, we conclude that the pairwise constraints could indeed refine the trained social connection strength. Since both co-occurrence and homogeneity may indicate higher probability of direct interactions, and more interactions definitely mean stronger social influence, they could be reasonable to enhance the social-based prediction.

On the contrary, the global constraints, namely the skill levels may impair the performance instead. Though it might be true that a few top drivers who are expert in certain skills indeed attract followers with strong influence, for the majority of drivers, their connections might not highly correlate with their skill levels. In summary, pairwise constraints with direct influence could be a better choice to refine the proposed framework.

Besides, we realize that the performance generally degrades with increasing λ. As we proved that constraints could indeed improve the performance, we may conclude that an appropriate value of λ should be carefully selected, as a too much larger λ may overly highlight the constraints and conceal the original loss function, which disturb the optimization task of partial ranking.

4.6 Parameter Robustness and P-R Curve

As the performance has been extensively validated, in this part, we will discuss the sensitiveness of parameters. To be specific, three parameters are concerned, i.e., the number of clustering, the amount of pattern and transactions, as well as the time lag. Besides, the P-R curve will be also disucssed.

4.6.1 Number of Clustering

First, we conduct validations on our SPC approach with 10, 20, 30, 40 or 50 clusters respectively, compared with all the

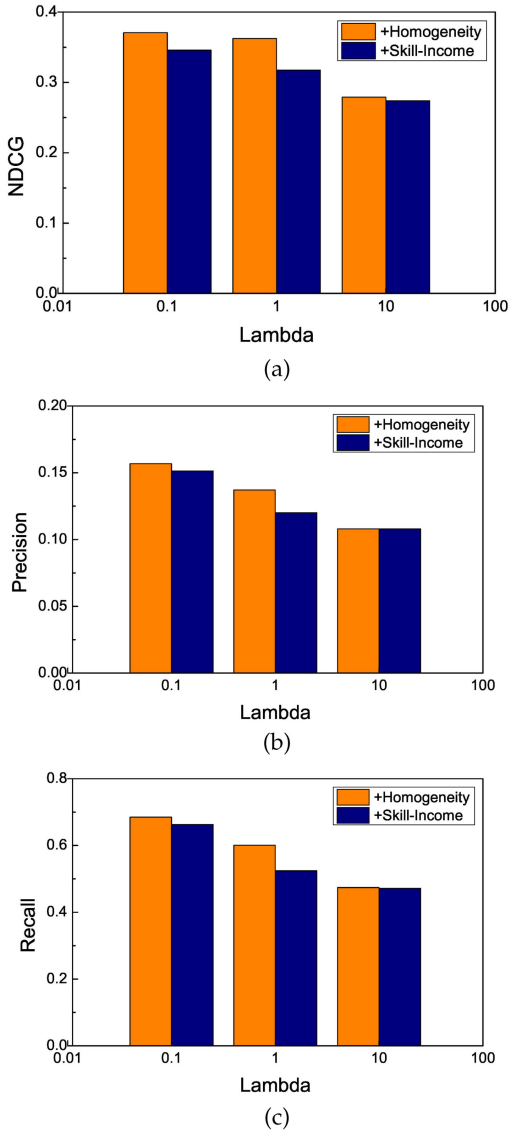


Fig. 8. The performance with constraints on different λ , (a) NDCG, (b)Precision, and (c) Recall.

baselines. The results of NDCG metric are shown in Fig. 9, and the rest three metrics share the similar trend. According to the results, we realize that the performance keeps almost stable, and 30 clusters, which are chosen for overall test, even performs the worst.

Indeed, we select 30 mainly due to it could better distinguish the zones in Manhattan properly, neither roughly nor too much thoroughly divided. For instance, if we have only 10 clusters, there will be a huge block in Manhattan containing more than 80 percent of transactions, which will improve the performance as driving behaviors within one zone could be easily predicted, however, fewer interesting rules will be revealed. In summary, those less clusters may lead to better performance, the results may indeed mislead us with some meaningless patterns. In the future, we will study how to automatically cluster the areas based on their geographic [42] and functional features.

4.6.2 Amount of Patterns

Then, for the amount of pattern, we conducted validations on four sets with different sizes, which contain the top 200,

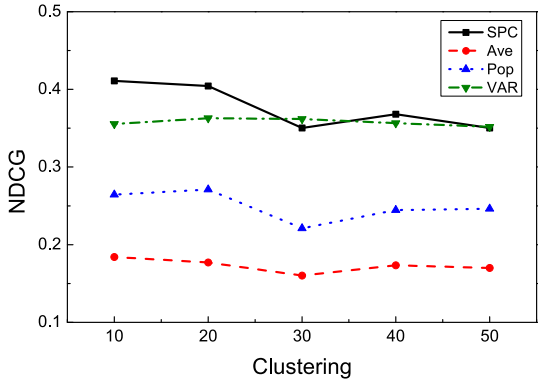


Fig. 9. Performance with different numbers of location clustering.

300, 500 and 800 patterns separately. The results are shown in Fig. 10. We realize that for the ranking task, the metrics become worse with more patterns, which may be due to the data sparsity. Obviously, more patterns lead to sparser data set, which definitely impair the performance, especially when those unpopular patterns are studied.

Interestingly, we find that for the overall popularity method, the performance dramatically deteriorates when the amount of patterns increases from 200 to 300. Usually, drivers will be glad to follow the popular trend. However, if drivers face to more patterns with sparser records, and these patterns could be hardly distinguished via popularity, it will be extremely difficult to predict selection. On the contrary, data sparsity may not severely disturb the binary classification. After all, majority of patterns will be never selected, thus the precision and recall rates keep relatively stable with more patterns.

4.6.3 Amount of Transactions

Third, we would like to reveal how the amount of transactions may affect the social activities of taxi drivers, and then the model performance. As mentioned above, in the overall experiments, we selected the top 100 taxi drivers with the most transactions to prevent interference of data sparsity. For more comprehensive analysis, here we further selected the top 10,000 taxi drivers, and then divided them in order into 3 groups, labeled as "top", "middle" and "bottom", separately. Here "top" means the drivers with most transactions, and "bottom" means the drivers with least transactions. Along this line, for each group, we sampled 100 drivers for our experiments, and repeated the sampling for 5 times to achieve the average performance.

The results are summarized in Table 6. Generally, with less transactions, performance of our SPC framework become worse, and the standard deviation increased, which indicates more unstable results. Under this situation, we may guess that driving behaviors for those inactive drivers (with less transactions) could be less socially motivated, thus the social learning mechanism will be weaken, which impairs the predictability of their driving behaviors and results in worse performance.

TABLE 6
Performance with Different Amount of Transactions

		NDCG	MAP@10	Precision	Recall
Top	*M*	0.3248	0.1583	0.1111	0.6611
	SD	0.0247	0.0177	0.0101	0.0350
Middle	*M*	0.3118	0.1364	0.0870	0.6097
	SD	0.0245	0.0118	0.0106	0.0457
Bottom	*M*	0.3089	0.1420	0.0852	0.6006
	SD	0.0271	0.0267	0.0155	0.0482

Along this line, we would like to know whether the performance of different groups are significant different. Here we took the *NDCG* metric as an example. As samples follow the normal distribution which is ensured by AD Test, we conducted a one-way ANOVA test to check the significance. According to the results, we realize that the differences among these three groups are insignificant (with p-Value as 0.085). Indeed, when reviewing the details, we found that the "top" group performed much better, while value ranges of the rest two groups are almost overlapped. Similar trend also appeared on other metrics, which may indicate that top drivers could be more "socially" active.

4.6.4 Time Lag

Fourth, for the time lag, similarly, we conduct four sets of validations with lag as 3, 4, 5 and 6 (months). The results are shown in Fig. 11. It seems that for our approach as well as overall popularity, the time lag does not reflect significant effect. However, for personalized average and the VAR model, which focus on the time-series estimation, they perform worse with increasing time lags. Usually, longer lag should be beneficial for time-series analysis, as they could better capture the latent trend for more accurate estimation. However, as shown before, the driving patterns of taxi drivers could be largely random, thus time-dependent rules may be misled by the over-fitting problem.

4.6.5 P-R Curve

Finally, we discuss about the P-R curve of our approach. In former validations, we treated the top 20 percent of ranking list as *increasing* patterns. Here we conducted validations

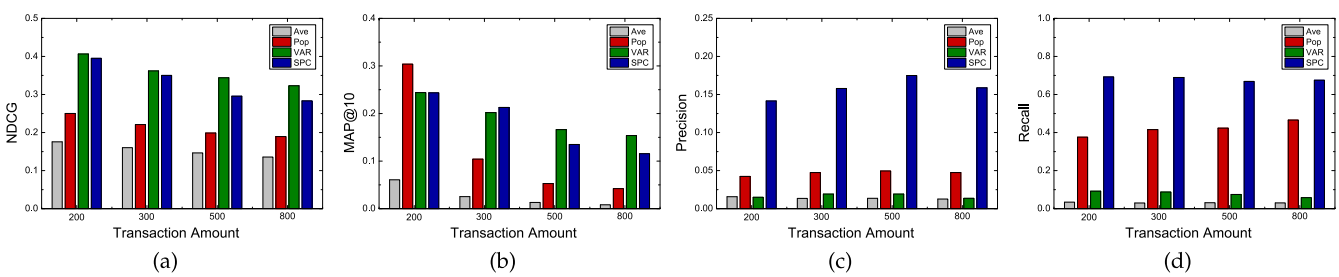


Fig. 10. The verification on robustness with different set of patterns in terms of different metrics, (a) NDCG, (b) MAP@10, (c) Precision, and (d) Recall.

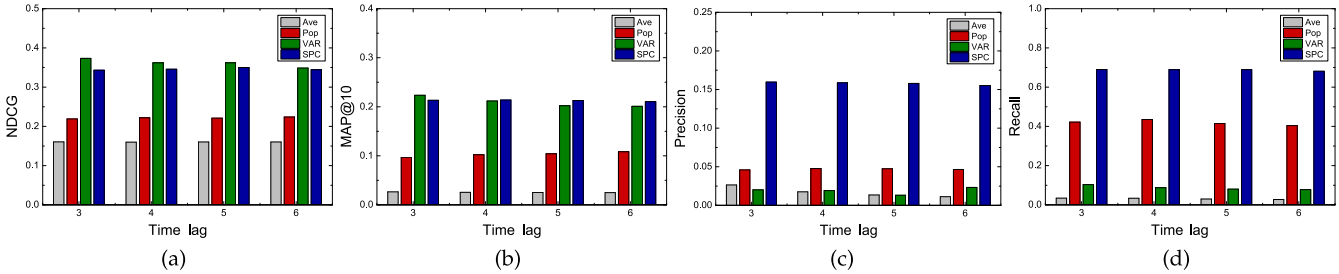


Fig. 11. The verification on robustness with different time lag in terms of different metrics, (a) NDCG, (b) MAP@10, (c) Precision, and (d) Recall.

with different proportions as 10 to 70 percent, while all the rest parameters keep the same. The P-R curve is shown in Fig. 12, in which numbers near the line present the percentage of *increasing* patterns. According to the results, we can clearly find that for any proportion, our framework could outperform all the baselines in binary classification task with significant margin.

However, we also realize that the recall may hardly pass 0.8 no matter how we regulate the proportion of "positive" ("increasing") results. As we attempt to reveal and leverage the latent social factors to explain the evolution of taxi driving behaviors, it seems that the social factors may explain at most 80 percent for driving behaviors. In other words, at least 20 percent of taxi driving behaviors should be explained via other factors, e.g., traffic status, large-scale activities or festivals, which will be discovered in future works.

5 Statistical Analysis: Rules of Social Learning Mechanism

As extensive validations have been conducted to evaluate the effectiveness of our framework, in this section, we will further discover some interesting rules of social learning mechanism with statistical analysis. Specifically, all the statistical analyses are conducted on the drivers with *revealed social network* in validations, and the settings keep the same with the validations.

5.1 Social Skills: Quantity versus Quality

First, we tend to reveal how the drivers select proper social "tutors". To be specific, two impact factors will be considered, namely the number, and experise rank of "tutors".

On the one hand, we attempt to discover the correlation between skill ranking of drivers, and the number of "tutors" they select to follow. All the three skills discussed as "constraints" in Section 3.4 (except the social skill) have been studied. According to the results, we find correlations are completely insignificant, with p-Value even larger than 0.5 for all the tests. It may indicate that finding more tutors are not necessarily lead to the better skill.

On the other hand, we check the correlation between skill ranking of drivers, and the ranking of their top 5 "tutors" with strongest influence. Interestingly, we find a significant correlation with r as 0.323 and p-Value as 0.001. This phenomenon may indicates better skills could be probably related to higher ranking as a "tutor". Along this line, with considering the former statistical analysis, we may draw the conclusion that drivers probably pay more attention on the "quality" but not "quantity", i.e., they prefer to learn from a few experienced experts with better skills, instead of finding many mediocre tutors.

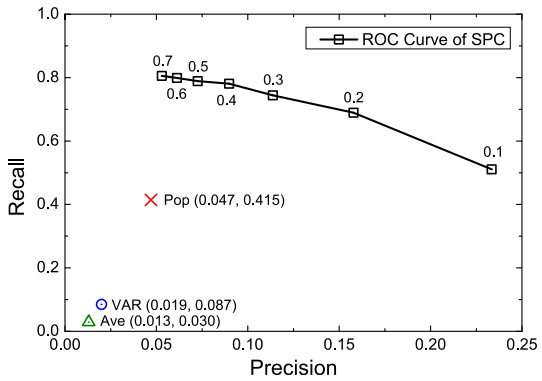


Fig. 12. P-R Curve with different positive ratio.

5.2 Pattern Learning: Tutors versus Apprentice

Second, we turn to discuss the social learning mechanism, to reveal which type of patterns that the "apprentices" probably tend to learn from "tutors". As we simulate the teaching process as "social propagation" of driving patterns, it is reasonable to extract the propagation graph, and then compare different patterns via graph-based metrics.

Specifically, for each type of driving pattern, we will build one *propagation graph*. For instance, for the kth pattern, if $s^t_{i,k}$ increased compared with $s^{t+1}_{i,k}$ for the first time in time t , we define u_i was *activated* in time t . Along this line, if we have an edge e_{ij} in the latent social network, in which u_i was *activated* in time t and u_j was *activated* in $t+1$, we assume that u_i "successfully propagated" the kth pattern to u_j , then we will add the edge e_{ij} to the *propagation graph* of kth pattern.

Then, four common-used graph metrics are selected to describe the propagation graphs as follows:

- *Amount of edges*, meaning the pattern popularity.
- *Longest path*, meaning the depth of propagation.
- *Max out-degree*, meaning the width of propagation.
- *Density*, meaning the frequency of propagation.

Table 7 summarizes the average statistics of this four metrics. Given the metrics, we could now analyze the correlation between graph metrics and driving-oriented metrics of patterns. In this case study, three metrics, i.e., the average *speed*, *distance and fare* for each transaction are selected to measure the value of each pattern. The correlations are shown in Table 8, in which most cases are uncorrelated.

However, two interesting rules have been captured. First, significantly negative correlation exists between *long-distance trip* and all the graph metrics, which indicates that longer distance may lead to less motivation to learn, as the drivers may

TABLE 7
Metrics for Pattern Propagation Graph

	Statistics
Average Edges	327.9122
Average Longest Path	8.5676
Average Max Out-degree	12.5878
Average Density	0.2244

have to be idling on the way back. Second, significantly negative correlation exists between *density* and *average fares*, which may indicate that drivers are not willing to largely share the patterns with higher benefits, or more willing to share the patterns with lower benefits.

5.3 Skill Update: Tradition versus Innovation

Third, we turn to discuss how the drivers update their driving behaviors. We could like to know whether drivers tend to keep their *traditions*, or evolve with *innovations*.

Considering that we treat drivers as their own "tutors", we first conduct a simple analysis to measure their influence on themselves. Interestingly, as we calculate the Kendall rank correlation coefficient between skill ranking of drivers and their self-influence, i.e., the value of w_{ii}, we realize that they may reflect negative correlation to some extent, e.g., rank correlation coefficient between *speed* and self-influence is around -0.1628. This phenomenon may indicate that better drivers are probably even more willing to learn from the others. In a complementary manner, the ordinary drivers may tend to insist their old patterns and refuse to change.

Along this line, we would like to discover whether the top drivers indeed benefit from the pattern updates. Thus, a more complicated analysis is conducted, in which the top 10 percent drivers are compared with ordinary ones for 1) how many fresh patterns they tried, and 2) how the fresh patterns benefit their effectiveness. The results are shown in Table 9, which list the total amount of fresh patterns, how many patterns hold better/worse skill metrics than average, and the average benefits of skill metrics. Besides, two skill metrics are analyzed to measure the benefits, namely the average *speed* and *income* of patterns.

According to the statistics, the top drivers win out in both amounts of all patterns ($16.1 > 12.8$) and patterns with better metrics ($4.4 > 3.7$, $10.9 > 9.1$), which indicate better innovativeness as they tend to try something new. Also, though the proportion of better patterns could be similar, however, the top drivers gain more benefit via updating patterns, while for the ordinary ones, their average speed even severely decreased.

TABLE 8
Correlation within Pattern Propagation and Skill Metrics

Term	Edge	Long Path	Out Degree	Density
Speed	-0.026	-0.041	-0.040	0.010
p-Value	0.756	0.617	0.631	0.908
Distance	-0.202	-0.182	-0.194	-0.280
p-Value	0.014	0.027	0.018	0.001
Fare	-0.060	-0.044	-0.066	-0.235
p-Value	0.468	0.597	0.426	0.004

TABLE 9
Comparison of Pattern Update

	Top Drivers		Overall	
	Speed	Income	Speed	Income
Total	16.1	16.1	12.8	12.8
+	4.4	10.9	3.7	9.1
-	11.7	5.2	9.1	3.7
Benefit	2.14	0.30	-4.16	0.18

Considering that for taxi drivers who randomly pick up passengers along the street, sometimes they are forced to try new patterns due to the destinations. However, top drivers with experience could better select potential passengers to ensure their working effectiveness. For instance, in the example we mentioned in Introduction part, an experienced drivers pointed out that he liked to pick up "people who have heavy bags", which usually lead to airport meaning higher income. In these cases, the destinations and following patterns could be predicted. In other words, as top drivers could better control their driving patterns, when they are facing new patterns, they could distinguish whether these patterns are "acceptable" or even "rewarding", thus the risk of new patterns may not severely disturb them. On the contrary, for the ordinary drivers, since they are not experienced enough to select proper routes and control their pick-up, their updates may even result in worse performance.

5.4 Social Links: Offline Gathering versus More Channels

Finally, we attempt to measure the similarity between revealed social network in validations, and the heuristic "co-occurrence" as a constraint in Section 3.4, to study whether the latent social connections within taxi drivers are mainly due to the offline gathering. It is true that co-occurrences may not definitely lead to direct interactions, as we mentioned before, thus their social effects may not be convincing enough. However, based on the validation with constraints, "co-occurrence" indeed improve the performance with better ranking metrics. Thus, in this part, we target at discovering the correlation of these two types of social network in different perspectives, and then reveal the potential rules with discussions.

First, we count the overlapping of these two types of social network. Specifically, we find that 36.4 percent social connections revealed in validations have raised "co-occurrence", correspondingly, 47.4 percent pairs of drivers who had "co-occurrence" were captured in the revealed social network as social connections. In other words, as much as half of the heuristic "co-occurrence" may indeed results in actual interactions (Considering the inaccurate approximation of co-occurrence without GPS records, the proportion might be disturbed). However, two thirds of social connections may be due to other factors, e.g., online discussions, instead of face-to-face interactions.

Second, we conduct the correlation analysis between the link strength of revealed social network, and the frequency of co-occurrences. Interestingly, the result presents a significantly weak correlation with r less than 0.1, and p-Value less than 0.01. This result further support our conclusion that co-occurrence may be related to latent social connections, but

clearly, they are not the main reasons. This result could also explain why "co-occurrence" as a constraint could improve the performance, since the correlation indeed exist to further refine the reveal connections.

In summary, though offline gathering with face-to-face interactions may indeed explain parts of the latent social connections within drivers, they must be enhanced by more comprehensive factors to achieve better estimation. We will conduct deep analysis when the data collection is enriched.

6 Conclusion

In this paper, we explored the latent social factors among taxi drivers based on the analysis of their driving behaviors, and then reveal and leverage the social connections to describe the evolution of driving patterns. To be specific, for validating the performance of our approach, we designed a social-driven two-stage framework, which could better explain drivers' future behaviors. A unique characteristic of our framework is that we can deal with the tasks of driving behavior prediction as the problems of partial ranking for optimization, and further enhance the approach with integrating more factors as constraints. Validations on a real-word data set clearly validated the effectiveness of our proposed framework with better explanation of future taxi driving pattern evolution, which proved the hypothesis that social factors indeed improve the predictability of taxi driving behaviors, and further revealed some interesting rules on social learning habits.

Acknowledgments

This is a substantially extended and revised version of [1], which appears in the *Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (KDD'2016). This research was partially supported by grants from the National Natural Science Foundation of China (Grant No. U1605251, 61703386, 91746301, 71531001 and 61836013) and the Anhui Provincial Natural Science Foundation (Grant No. 1708085QF140). This research is also partially funded by Microsoft Research Asia.

References

[1] T. Xu, H. Zhu, X. Zhao, Q. Liu, H. Zhong, E. Chen, and H. Xiong, "Taxi driving behavior analysis in latent vehicle-to-vehicle networks: A social influence perspective," in *Proc. 22nd ACM SIGKDD Conf. Knowl. Discovery Data Mining*, 2016, pp. 1285–1294.

[2] H. Zhu, E. Chen, H. Xiong, K. Yu, H. Cao, and J. Tian, "Mining mobile user preferences for personalized context-aware recommendation," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 4, 2014, Art. no. 58.

[3] J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, G. Sun, and Y. Huang, "T-drive: Driving directions based on taxi trajectories," in *Proc. 18th SIGSPATIAL Int. Conf. Advances Geographic Inf. Syst.*, 2010, pp. 99–108.

[4] Y. Zheng, Y. Liu, J. Yuan, and X. Xie, "Urban computing with taxicabs," in *Proc. 13th Int. Conf. Ubiquitous Comput.*, 2011, pp. 89–98.

[5] M. Qu, H. Zhu, J. Liu, G. Liu, and H. Xiong, "A cost-effective recommender system for taxi drivers," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 45–54.

[6] Y. Ge, H. Xiong, A. Tuzhilin, K. Xiao, M. Gruteser, and M. Pazzani, "An energy-efficient mobile recommender system," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2010, pp. 899–908.

[7] N. J. Yuan, Y. Zheng, L. Zhang, and X. Xie, "T-finder: A recommender system for finding passengers and vacant taxis," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 10, pp. 2390–2403, Oct. 2013.

[8] J. Yuan, Y. Zheng, X. Xie, and G. Sun, "T-drive: Enhancing driving directions with taxi drivers' intelligence," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 1, pp. 220–232, Jan. 2013.

[9] Y. Wang, B. Liang, W. Zheng, L. Huang, and H. Liu, "The development of a smart taxicab scheduling system: A multi-source data fusion perspective," in *Proc. IEEE 16th Int. Conf. Data Mining*, 2016, pp. 1275–1280.

[10] S. Ma, Y. Zheng, and O. Wolfson, "T-share: A large-scale dynamic taxi ridesharing service," in *Proc. IEEE 29th Int. Conf. Data Eng.*, 2013, pp. 410–421.

[11] Z. Ye, K. Xiao, Y. Ge, Y. Deng, "Applying Simulated Annealing and Parallel Computing to the Mobile Sequential Recommendation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31,
no. 2, pp. 243–256, 2019.

[12] H. Hu, Z. Wu, B. Mao, Y. Zhuang, J. Cao, and J. Pan, "Pick-up tree based route recommendation from taxi trajectories," in *Proc. Int. Conf. Web-Age Inf. Manage.*, 2012, pp. 471–483.

[13] D. Zhang, N. Li, Z.-H. Zhou, C. Chen, L. Sun, and S. Li, "iBAT: Detecting anomalous taxi trajectories from GPS traces," in *Proc. 13th Int. Conf. Ubiquitous Comput.*, 2011, pp. 99–108.

[14] F. Giannotti, M. Nanni, D. Pedreschi, F. Pinelli, C. Renso, S. Rinzivillo, and R. Trasarti, "Unveiling the complexity of human mobility by querying and mining massive trajectory data," *Int. J. Very Large Data Bases*, vol. 20, no. 5, pp. 695–719, 2011.

[15] P. S. Castro, D. Zhang, and S. Li, "Urban traffic modelling and prediction using large scale taxi GPS traces," in *Proc. Int. Conf. Pervasive Comput.*, 2012, pp. 57–72.

[16] C. Chen, D. Zhang, Z.-H. Zhou, N. Li, T. Atmaca, and S. Li, "B-planner: Night bus route planning using large-scale taxi GPS traces," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun.*, 2013, pp. 225–233.

[17] P. Domingos and M. Richardson, "Mining the network value of customers," in *Proc. 7th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2001, pp. 57–66.

[18] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2003, pp. 137–146.

[19] C. C. Aggarwal, A. Khan, and X. Yan, "On flow authority discovery in social networks," in *Proc. 11th SIAM Int. Conf. Data Mining*, 2011, pp. 522–533.

[20] J. Goldenberg, B. Libai, and E. Muller, "Talk of the network: A complex systems look at the underlying process of word-of-mouth," *Marketing Lett.*, vol. 12, no. 3, pp. 211–223, 2001.

[21] Y. Yang, E. Chen, Q. Liu, B. Xiang, T. Xu, and S. A. Shad, "On approximation of real-world influence spread," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, 2012, pp. 548–564.

[22] T. Jin, T. Xu, H. Zhong, E. Chen, Z. Wang, and Q. Liu, "Maximizing the effect of information adoption: A general framework," in *Proc. SIAM Int. Conf. Data Mining*, 2018, pp. 693–701.

[23] A. Goyal, F. Bonchi, and L. V. Lakshmanan, "Learning influence probabilities in social networks," in *Proc. 3rd ACM Int. Conf. Web Search Data Mining*, 2010, pp. 241–250.

[24] K. Saito, R. Nakano, and M. Kimura, "Prediction of information diffusion probabilities for independent cascade model," in *Proc. Int. Conf. Knowl.-Based Intell. Inf. Eng. Syst.*, 2008, pp. 67–75.

[25] T. Xu, D. Liu, E. Chen, H. Cao, and J. Tian, "Towards annotating media contents through social diffusion analysis," in *Proc. IEEE 12th Int. Conf. Data Mining*, 2012, pp. 1158–1163.

[26] J. Yang and J. Leskovec, "Modeling information diffusion in implicit networks," in *Proc. IEEE Int. Conf. Data Mining*, 2010, pp. 599–608.

[27] T. Xu, H. Zhu, E. Chen, B. Huai, H. Xiong, and J. Tian, "Learning to annotate via social interaction analytics," *Knowl. Inf. Syst.*, vol. 41, no. 2, pp. 251–276, 2014.

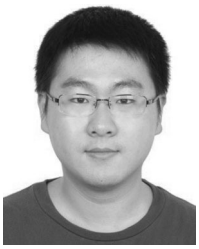
[28] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: User movement in location-based social networks," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2011, pp. 1082–1090.

[29] X. Liu, Q. He, Y. Tian, W.-C. Lee, J. McPherson, and J. Han, "Event-based social networks: Linking the online and offline social worlds," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 1032–1040.

[30] H. Zhu, E. Chen, K. Yu, H. Cao, H. Xiong, and J. Tian, "Mining personal context-aware preferences for mobile users," in *Proc. IEEE 12th Int. Conf. Data Mining*, 2012, pp. 1212–1217.

[31] N. Eagle, A. S. Pentland, and D. Lazer, "Inferring friendship network structure by using mobile phone data," *Proc. Nat. Academy Sci. United States America*, vol. 106, no. 36, pp. 15 274–15 278, 2009.
[32] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi, "Human mobility, social ties, and link prediction," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2011, pp. 1100–1108.
[33] K. Zhang and K. Pelechrinis, "Understanding spatial homophily: The case of peer influence and social selection," in *Proc. 23rd Int. Conf. World Wide Web*, 2014, pp. 271–282.
[34] X. Yu, A. Pan, L.-A. Tang, Z. Li, and J. Han, "Geo-friends recommendation in GPS-based cyber-physical social network," in *Proc. Int. Conf. Advances Social Netw. Anal. Mining*, 2011, pp. 361–368.
[35] T. Xu, H. Zhong, H. Zhu, H. Xiong, E. Chen, and G. Liu, "Exploring the impact of dynamic mutual influence on social event participation," in *Proc. SIAM Int. Conf. Data Mining*, 2015, pp. 262–270.
[36] X. Zhao, T. Xu, Q. Liu, and H. Guo, "Exploring the choice under conflict for social event participation," in *Proc. Int. Conf. Database Syst. Adv. Appl.*, 2016, pp. 396–411.
[37] Z. Ning, F. Xia, N. Ullah, X. Kong, and X. Hu, "Vehicular social networks: Enabling smart mobility," *IEEE Commun. Mag.*, vol. 55, no. 5, pp. 16–55, May 2017.
[38] F. Xia, J. Wang, X. Kong, Z. Wang, J. Li, and C. Liu, "Exploring human mobility patterns in urban scenarios: A trajectory data perspective," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 142–149, Mar. 2018.
[39] H. Zhuang, A. Chin, S. Wu, W. Wang, X. Wang, and J. Tang, "Inferring geographic coincidence in ephemeral social networks," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, 2012, pp. 613–628.
[40] H. Zhu, E. Chen, H. Xiong, H. Cao, and J. Tian, "Mobile app classification with enriched contextual information," *IEEE Trans. Mobile Comput.*, vol. 13, no. 7, pp. 1550–1563, Jul. 2014.
[41] H. Lütkepohl, *New Introduction to Multiple Time Series Analysis*. Berlin, Germany: Springer, 2005.
[42] Y. Fu, H. Xiong, Y. Ge, Z. Yao, Y. Zheng, and Z.-H. Zhou, "Exploiting geographic dependencies for real estate appraisal: A mutual perspective of ranking and clustering," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 1047–1056.

Tong Xu (M'17) received the PhD degree from the University of Science and Technology of China (USTC), Hefei, China, in 2016. He is currently working as an associate researcher of the Anhui Province Key Laboratory of Big Data Analysis and Application, USTC. He has authored more than 30 journal and conference papers in the fields of social network and social media analysis, including the *IEEE Transactions on Knowledge and Data Engineering*, KDD, AAAI, ICDM, SDM, etc. He is a member of the IEEE.



Hengshu Zhu (M'14) received the BE and PhD degrees in computer science from the University of Science and Technology of China (USTC), China, in 2009 and 2014, respectively. He is currently a senior data scientist with Baidu Inc. His general area of research is data mining and machine learning, with a focus on developing advanced data analysis techniques for emerging applied business research. He has published prolifically in refereed journals and conference proceedings, including the *IEEE Transactions on Knowledge and Data Engineering*, the *IEEE Transactions on Mobile Computing*, the *ACM Transactions on Knowledge Discovery from Data*, KDD, IJCAI, and AAAI, etc. He was regularly on the program committees of numerous conferences, and has served as a reviewer for many top journals in relevant fields. He is a member of the IEEE.

Hui Xiong (SM'07) is currently a full professor with Rutgers, the State University of New Jersey, where he received the ICDM-2011 Best Research Paper Award, and the 2017 IEEE ICDM Outstanding Service Award. His general area of research is data and knowledge engineering, with a focus on developing effective and efficient data analysis techniques for emerging data intensive applications. He has published prolifically in refereed journals and conference proceedings (four books, 80+ journal papers, and 100+ conference papers). He is a co-editor-in-chief of the *Encyclopedia of GIS*, an associate editor of the *IEEE Transactions on Knowledge and Data Engineering*, the *IEEE Transactions on Big Data*, the *ACM Transactions on Knowledge Discovery from Data*, and the *ACM Transactions on Management Information Systems*. He has served regularly on the organization and program committees of numerous conferences, including as a program co-chair of the Industrial and Government Track for KDD-2012, a program co-chair for ICDM-2013, a general co-chair for ICDM-2015, and a program co-chair of the Research Track for KDD-2018. For his outstanding contributions to data mining and mobile computing, he was elected an ACM distinguished scientist in 2014. He is a senior member of the IEEE.

Hao Zhong is currently working toward the PhD degree at Rutgers University. His research interests focus on business intelligence and data mining. He has published several papers, e.g., ICDM and ANOR, etc.

Enhong Chen (SM'07) is a professor and vice dean of the School of Computer Science, University of Science and Technology of China. His general area of research includes data mining and machine learning, social network analysis, and recommender systems. He has published more than 100 papers in refereed conferences and journals, including *Nature Communications*, IEEE/ACM Transactions, KDD, NIPS, IJCAI, and AAAI, etc. He was on program committees of numerous conferences including KDD, ICDM, and SDM. He received the Best Application Paper Award of KDD-2008, the Best Research Paper Award of ICDM-2011, and the Best of SDM-2015. His research is supported by the National Science Foundation for Distinguished Young Scholars of China. He is a senior member of the IEEE.