Taxi Driving Behavior Analysis in Latent Vehicle-to-Vehicle Networks: A Social Influence Perspective

Tong Xu[1], Hengshu Zhu[2], Xiangyu Zhao[1], Qi Liu[1]
Hao Zhong[3], Enhong Chen[1], Hui Xiong[3]

[1] University of Science and Technology of China
{tongxu, zxy1105}@mail.ustc.edu.cn, {qiliuql, cheneh}@ustc.edu.cn

[2] Baidu Research-Big Data Lab
zhuhengshu@baidu.com

[3] Rutgers, the State University of New Jersey
{h.zhong31, hxiong}@rutgers.edu

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 guide and opportunity to improve the profit and work efficiency. However, limited scopes focus on the latent *social interaction* within cab drivers, and corresponding *social propagation* scheme to share driving behaviors has been largely ignored. To that end, in this paper, we propose a comprehensive study to reveal how the social propagation affects for better prediction of cab drivers' future behaviors. To be specific, we first investigate the correlation between drivers' skills and their mutual interactions in the latent vehicle-to-vehicle network, which intuitively indicates the effects of social influences. Along this line, by leveraging the classic *social influence theory*, we develop a two-stage framework for quantitatively revealing the latent *driving pattern propagation* within taxi drivers. Comprehensive experiments on a real-word data set collected from the New York City clearly validate the effectiveness of our proposed framework on predicting future taxi driving behaviors, which also support the hypothesis that social factors indeed improve the predictability of driving behaviors.

Keywords

Mobile Data Mining, Social Influence, Taxi Trajectories

1. INTRODUCTION

Recent years have witnessed the growing interests on data-driven technologies for developing new paradigms of taxi business. On the one hand, the dramatic expansion of urban areas results in the urge demand of efficient taxi services, which cannot be solved by simply increasing the amount of cabs or drivers. On the other hand, 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 [40]. Therefore, through the analysis of trajectory data from taxi drivers, a variety of intelligent services can be enabled, which will not only lead to the improvement of work efficiency and profit of taxi drivers [10], but also help to strike a balance between the needs of taxi drivers and passengers [33].

In the literature, most of the research efforts on taxi business focus on extracting effective transportation patterns, e.g., the fastest driving route like [32] and [38], sequence of pick-up points [23] or customers within the shortest driving distance [10]. However, although the above works can effectively enhance taxi business, they may still suffer some defects due to the ignorance of drivers' behaviors. First, the case-by-case recommendation are sensitive to the context, thus frequent update is required, which results in heavy burden of computation. Second, as drivers tend to be attracted by popular routes, it will be difficult to distribute the cabs for keeping regional balances. Last but not least, predictability of taxi route is still limited as individual factors of taxi drivers have been largely ignored, i.e., neither the self-learning ability, nor the social sharing scheme are considered in the system design.

Indeed, though related techniques could highly support the intellectual services, cab drivers could also rely on themselves to well fit the taxi business. For the experienced drivers who are sensitive to the route and rules, they could effectively summarize the patterns and regulate the lines. At the same time, for those inexperienced ones, thanks to rapid development of intelligent mobile devices and social network services (SNS), not only offline (e.g., refueling or lunching) but also online (e.g., forum, or twitter) gatherings are now available, where driving experience could be shared within drivers, which results in the "*social propagation*" of driving behaviors. For instance, an motivating example below, which summarizes a real post in a online taxi forum [1], could intuitively illustrate this phenomenon.

A Motivating Story. *A new taxi driver, who was unfamiliar with the traffic in New York City, posted a question*

KDD '16, August 13-17, 2016, San Francisco, CA, USA

DOI: http://dx.doi.org/10.1145/2939672.2939799

[1] http://www.nycitycab.com/forum/Replies.aspx?postid=649

in online forum to ask for sharing driving experience. Then, an experienced driver replied with several detailed suggestions on how to drive in Manhattan to avoid the rush hour or earn more tips. Based on these recommended patterns, the new driver could better adjust his driving trajectory to improve the work efficiency.

The true story above indicates that social interactions existing to influence drivers' driving behaviors. Intuitively, if we treat taxi drivers as "social agent" in the mobile social networks, and simulate how the "social propagation" scheme functions to interpret their future behaviors, taxi route will be more predictable, and further social-oriented taxi services, e.g., social-based "tutor" or pattern recommendations could be effectively conducted. Unfortunately, due to the constraint of user privacy, there is no exposed signal for social interactions to be observed. Thus, techniques to reveal latent connections within taxi drivers are required.

To deal with this task, in this paper, we aim at exploring latent vehicle-to-vehicle networks among taxi drivers based on the analysis of their driving behaviors. To be specific, we propose a two-stage framework to capture the latent propagation of driving behaviors among taxi drivers. We assume that the increasing proportion of certain behavior patterns may be caused by stronger influence, and vice versa. Therefore, sequential driving behaviors with integrating social propagation could be formulated as time series and solved as partial ranking tasks. Based on the framework, we could better model the historical driving behaviors to predict their future trends, and also discover the latent social connections within taxi drivers. To the best of our knowledge, we are the first to investigate the impact of social factors on taxi driving patterns to explain drivers' behaviors.

Finally, extensive experiments are conducted on a real-world data set collected from the New York City. The experimental results clearly verify that our framework can better predict driving behaviors of taxi drivers with dramatic margin outperforming the baselines, which validates the hypothesis that social factors indeed affect the driving pattern decisions, and also demonstrates the capability of social analytics in intelligent taxi services.

Overview. The rest of this paper is organized as follows. Section 2 further illustrates the motivation with some intuitive statistics. In Section 3, we propose the novel framework for driving behavior analysis with integrating social factors. Afterwards, we comprehensively evaluate the performances in Section 4, and then conduct some further discussion in Section 5. Related works are summarized in Section 6. Finally, in Section 7, we conclude the paper.

2. DOES "SOCIAL FACTORS" AFFECT TAXI DRIVERS?

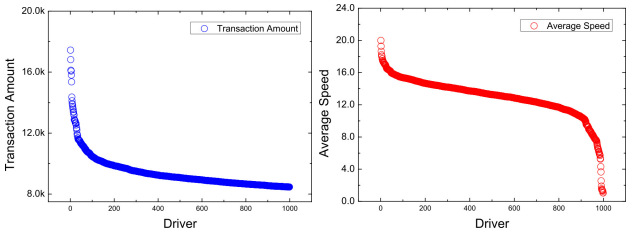
As we propose the idea of driving behavior propagation within cab drivers, in this section, we will intuitively discuss the effects of social propagation with related statistical analysis to further support our motivation.

2.1 Data Set Description

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 provided by NYC Taxi and Limousine Commission (NYC TLC). This is a large-scale data set that 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 are shown in Table 1.

Table 1: 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



(a) (b)

Figure 1: The distribution of transaction amount and average driving speed of taxi drivers.

As mentioned above that usually there is no explicit signal of social interactions. In the following sections, we will conduct our two-stage framework to reveal the latent social connections via simulating the driving behavior propagation. However, for the intuitive statistics here, we choose the heuristic method which is widely used in *"ephemeral social networks"* study such as [31] and [39], i.e., co-occurrences lead connections, and the frequency indicates the strength. Along this line, if two cabs stop at the same place for a relatively long time (e.g., more than 10 minutes), we guess that a short-term communication may occur, which probably results in social interactions. Therefore, to be the same with the prior arts like [39], an undirected edge will be built between two drivers if long-time co-parking happened, and the frequency will be counted as the link strength.

2.2 Social Effects on Taxi Driver Skill

With the above *"ephemeral vehicle-to-vehicle networks"*, we could now discuss about the social effects on taxi drivers. In details, two questions remain to be answered: 1) if drivers with similar skills hold stronger connections, and 2) if drivers with better skills hold stronger influences. To measure cor-

Table 2: Correlations between social link strength and driving skill of taxi drivers.

Term	Transactions		Speed		Income	
3-4 Driver	T10	R10	T10	R10	T10	R10
Correlation	0.093	0.018	0.177	0.011	0.094	0.021
Significant*	7	0	9	3	7	2
Positive	0	5	0	2	1	3

*P-Value<0.05

Table 3: 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 k-th pattern in time t
$p_{i,k}^t$	social influence of k-th 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

relations of driving skills, here we choose three evaluation metrics, i.e., the *transaction amount*, the *average driving speed* and the *total income*, to study on this issue. Particularly, we choose these three metrics since they can be regarded as the representative symbols of work effectiveness (i.e., more business), efficiency (i.e., faster trip) and profit (higher benefits). Specifically, the distributions of further two metrics are shown in Figure 1, while the curve for total income is quite similar with the first one.

For better understand the correlation between skill level and social influence, here for each evaluation metric, the top 10 drivers are ranked, compared with other 10 randomly selected drivers. Afterwards, for each driver, we list his/her "friends" with two values: *connection strength* and *skill similarity* (absolute value of difference), and then measure correlation between them. Also, we have *P-Value* presents the significance of correlation. Besides, those with positive coefficients, i.e., strong social connection leads to higher difference, are individually listed as they may violate the assumption that social effects encourage similar driving patterns.

The statistical results are shown in Table 2, in which "T10" presents the top 10 drivers and "R10" indicates the random ones for comparison. For each group, we list average correlation, count of significant correlations, and count of positive correlations. Expectedly, we find that for all the three metrics, the top-ranked drivers reflect more significant correlation, and further almost no positive correlation occur. As random selected drivers reflect no significant correlation between skill level and social connections, "homogeneity" may fail to explain this phenomenon. Thus, we may draw the conclusion that skilled drivers may act as the influential nodes to share expertise, and stronger connections lead to better skills via "social learning" scheme, which further supports our motivation.

3. SIMULATING DRIVING BEHAVIORS WITH SOCIAL PROPAGATION

As significance of social effects have been revealed, in this section, we will introduce our two-stage framework to capture the driving behavior propagation within taxi drivers, in which the technical solution for social influence modeling and optimization task will be explained in detail.

3.1 Preliminary and Problem Statement

In this paper, we focus on the social interaction which affects the driving behaviors. Thus, some other factors, e.g., the profit, traffic flow or festivals that may also influence the routes will be temporally ignored, and more complicated framework will be studied in future work.

Though we intuitively discover the social factors with introducing heuristic "ephemeral vehicle-to-vehicle networks", however, it could hardly summarize the complete social interactions due to the following reasons. First, co-occurrences do not necessarily means face-to-face communication, thus they are not adequate clues to reveal latent social connections. Second, occasional co-occurrences are indeed events of small probability ($\ll 1\%$), while majority of interactions might be neglected. To deal with these challenges, in this section, we target at formulating social propagation of driving behaviors within taxi drivers, and then the social interactions could be revealed conversely.

Firstly, to describe the social effects on driving behaviors, we define the vector $\mathbf{s_i}$ to present the driving behaviors of a certain taxi driver u_i, in which each element corresponds to a kind of driving patterns. Specially, each pattern here will be described as a triple contains the information about pick-up area, drop-off area and pick-up time period, such as (*Central Park, JFK Airport, 10:00AM-11:00AM*). The details of extracting these patterns will be introduced in experimental part. Then, we have $s_{i,k}$ to indicate the proportion of k-th pattern in u_i's driving behaviors, and definitely, the vector is normalized by $\sum_k s_{i,k} = 1$.

Further, to integrate social factors, we assume that the driving behaviors could be represented as temporal consequences, and in each round, the effects of social propagation will be reflected by the fluctuation in next round. Thus, timestamp information should be introduced into the driving behavior vectors as $\mathbf{s_i^t}$, which indicates the driving behavior of driver u_i in t-th round. And intuitively, if we could accurately reveal the latent social connections, we will precisely predict drivers' behavior in the future, i.e., $\mathbf{\Delta s_i^{t+1}}$ will be well estimated.

At the same time, to describe the social factors within taxi drivers, i.e., to draw the latent vehicle-to-vehicle network, similar with traditional social network, we introduce e_{ij} to present the edge from u_i to u_j, and correspondingly, w_{ij} indicates the edge weight, or the influential strength, which will be estimated during the training stage. What should be noted is that the social network here is asymmetric and all edges are directional. Afterwards, we have N_i to present the entire social "neighbors" of driver u_i. Furthermore, different from the traditional social propagation problems, to integrate drivers' own opinion into this framework, here we treat each driver as *neighbor of itself*, i.e., $u_i \in N_i$, and then introduce w_{ii} to measure how the drivers insist their own habits. Obviously, higher w_{ii} indicates less propagation and more succession, and vice versa.

Along this line, finally, we could now formally define our research problem as "*social-driven behavior prediction*" as a typical time series analysis as follows.

Problem Statement. *Given the target group of taxi drivers* $\mathbf{U}$*, and for each* $u_i \in \mathbf{U}$*, we have corresponding pattern vectors* $\mathbf{s_i^t}$ *during the time period* $t = 1, 2, ..., T$*. The problem of driving behavior prediction is to accurately reveal weighted social connections* w_{ij} *within each pair of drivers* u_i *and* u_j*, then the fluctuation of driving behavior vector* $\mathbf{\Delta s_i^{T+1}}$ *in* $T+1$ *could be estimated.*

Technical details to solve the problem will be introduced in following subsections, and the mathematical notations used throughout this paper are summarized in Table 3.

3.2 Loss Function for Behavior Prediction

With problem defined and notations summarized, now we turn to formulate the behavior prediction task in detail. Firstly, to describe the social propagation, for each pattern, we have the accumulated **social influence** $p_{i,k}^t$ of k-th pattern to driver u_i, while the technical solution will be explained in next subsection.

As we assume that fluctuation of driving behavior vectors is due to the social propagation within cab drivers, specially, for a certain cab driver i_i, if k-th pattern is holding 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, as it could be difficult to exactly estimate the value of $s_{i,k}^{T+1}$, here to ease the modeling, we target at predicting whether $s_{i,k}^{T+1}$ will increase or decrease in the $T+1$ round, and how the increment $\Delta s_{i,k}^{T+1}$ ranks among all the patterns, which may reflect the rank of corresponding social influence.

Therefore, the problem defined above could be intuitively transformed as a partial ranking problem, i.e., rank $\Delta s_{i,k}^{t+1}$ and label the top results in the list as increased. Here, we define the set of patterns whose frequency are raised 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 ones. Then, for each pair $\forall \langle r, d \rangle_{i,t}$ where $r \in \mathbf{R}_i^t$ and $d \in \mathbf{D}_i^t$, the fluctuation is due to the pairwise ranking of corresponding social influence, i.e., $p_{i,d}^t < p_{i,r}^t$.

With the assumption above, finally, we realize that if we could accurately reveal the latent social interactions, we could achieve the optimal ranking results of social influence $\mathbf{p_i^t}$. Correspondingly, optimization of the ranking task will leads to the solution of social connections. Thus, the task of learning latent social connections w_{ij} will be summarized as a pairwise ranking problem as follows:

Ranking Objective. *Finding 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:

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

where $h(\gamma_{rd})$ is a loss function to assign a non-negative penalty according to the difference of social influence $\gamma_{rd} = p_{i,d}^t - p_{i,r}^t$. Usually, we have the penalty $h(\gamma_{rd}) = 0$ when $p_{i,d}^t \leq p_{i,r}^t$. While for $p_{i,d}^t > p_{i,r}^t$, we have $h(\gamma_{rd}) > 0$ as loss. To ease the computation, here we utilize the squared loss function as follow:

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

In this framework, we have a soft margin parameter b to tolerate a small error. To ease the computation, here we simply treat $b = 0$, and $h(x)$ could be rewritten as

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

3.3 Social Propagation Simulation

As ranking objective is proposed, now we turn to formulate the *social influence* within taxi drivers. Traditionally, to simulate the propagation, or so-called "social influence" process, classic models like Independent Cascade or Linear Threshold model [16] might be selected, in which each nodes usually has only two statuses, i.e., activated or inactivated. However, in our task, the proportion of a certain driving pattern may never jump from 0 to 1 sharply. Thus, here we adapt the Steady State Spread (SSS) model [1] to simulate the social propagation, in which each node holds its own *activating probability*. Then, in each round, all the nodes (but not only activated nodes in IC model) attempt influence their neighbors, and then influence will be measured not only by strength of social connections, but also their activating probabilities. With adapting the formulation by introducing pattern proportion $s_{I,k}^t$ instead of activating probability, we have the equation as follow:

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

For $\delta_{i,j,k}^{t-1}$, we design it to present pairwise influential probability sensitive to the pattern, which is different with the overall social connection strength w_{ji}. Similar with the Sigmoid function, we have the formulation as follow:

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

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.

3.4 Optimization Task

As all the formulations established, we could now optimize the loss function Equation 1 to estimate latent social factors w_{ij}. To be specific, we approach the latent social factors w_{ji} by first deriving the gradient of $\mathcal{F}(w)$ with respect to w_{ji}, and then use a gradient based optimization method to find proper $\{w\}$ that minimize $\mathcal{F}(w)$. Specially, as defining $\gamma_{rd} = p_{i,d}^t - p_{i,r}^t$, we have the derivative 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 function, while for the social influence model, we have:

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

According to the formulations, finally gradient descent methods could be exploited to deal with the optimization task.

3.5 Two-stage Framework

Based on the formulations above, we could now formally define our two-stage framework as follows.

Training Stage. Given a group of taxi drivers $\mathbf{U} = \{u_i\}$ as well as their pattern vectors $\mathbf{s_i^t}$ during the time period $t = 1, 2, ..., T$, in the training stage, we aim at inferring the latent social connections $\{w_{ij}\}$ within drivers, which achieve the best explanation for the ranking of driving behavior vector fluctuation $\mathbf{\Delta s_i^{T+1}}$.

Figure 2: Two-stage framework.

Test Stage. After obtaining the latent connections $\{w_{ij}\}$ in the training stage, in the test stage, given the taxi drivers group $\mathbf{U} = \{u_i\}$ with their pattern vectors $\mathbf{s_i^t}$ during the certain **p-time lag** as $t = T - p + 1, ..., T - 1, T$, we aim at predicting the driving behavior vector fluctuation $\mathbf{\Delta s_i^{T+1}}$ with accurate sign and ranking.

With these definitions, for the training stage, we target at optimizing the Equation 1 to achieve the latent social connections w_{ij}. For the test stage, we could directly calculate $p_{i,k}^t$ according to Equation 4 for each pattern of driver u_i, and then rank the propagation to estimate the ranking of driving pattern proportions, or classify the increment sign. According to the framework, social interactions will be discovered and social propagation scheme will be simulated, which will be beneficial for further discussion on intellectual taxi services.

Based on the definitions above, in training stage, latent social connections will be revealed to support the social-driven behavior prediction task in test stage. The data flow of two-stage framework is summarized in Figure 2.

4. EXPERIMENTS: SOCIAL INFLUENCE INDEED AFFECTS

To verify our hypothesis that social factors may affect the driving behaviors of taxi drivers, which result in the change of pattern frequency, in this section, we conduct extensive experiments on a real-world data set. And further, some related discussion will be conducted.

4.1 Experimental Setup

Here, we introduce the data set pre-processing and baseline algorithms for evaluation.

4.1.1 Data Set Pre-processing

Similar with the statistical analysis in pre-study, our experiments were conducted on the real-world taxi data set collected from New York City. To generalize the driving patterns, we first clustered all the pick-up and drop-off locations in the historical transaction records. Specially, we conducted a bottom-up hierarchial with minimum variance criterion until only 30 clusters were left (indeed, the number of zones doesn't disturb the results if around 20-50). The

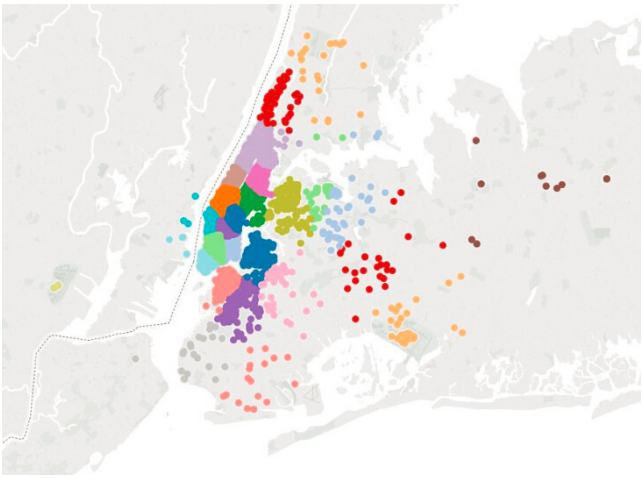


Figure 3: Clustering results of New York City Locations.

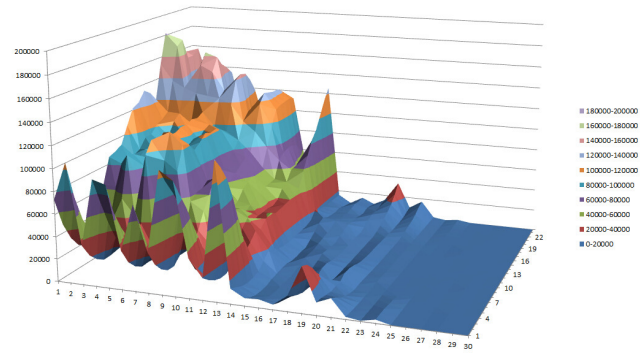


Figure 4: Pattern frequency in each hour, in which X-axis presents the label of area, and Y-axis presents the hour.

clustering results are shown in Figure 3. Interestingly, we find some landmarks have been distinguished, such as the JFK Airport (i.e., the orange part in the southeastern corner) and EWR Airport (i.e., the brownish-green part in the west), as well as some key regions in Manhattan, e.g., the red region around the Central Park, and the pale green in Lower Manhattan representing the Wall Street.

After location clustering, we counted pattern frequency in each hour with respect to different pick-up locations, which is shown in Figure 4. We could find except for the check out time around 18:00 PM, in Manhattan downtown (e.g., area No. 8 and 13), the peak time usually happen around midnight when passengers finish their night life. But for the suburb area (e.g., area No. 3 and 6), the peak time happen around 8:00 AM for those who take taxi for work. Also, only a few areas attract most of the taxi transactions, while some other areas, especially for area No. 25-30, may attract even completely no transactions.

Furthermore, we generalized the patterns by dividing the time period into every 2 hours, e.g., 7:00 AM to 9:00 AM, then 24 hours lead to 12 intervals. Figure 5 shows the most frequent patterns appear around 8:00 AM, which is the peak of the morning rush hour. We can see many passengers take taxi from WTC station to the downtown for work, or return

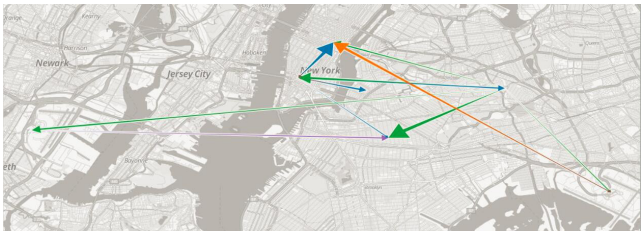


Figure 5: Frequent patterns in NYC taxi driving around 8:00 AM

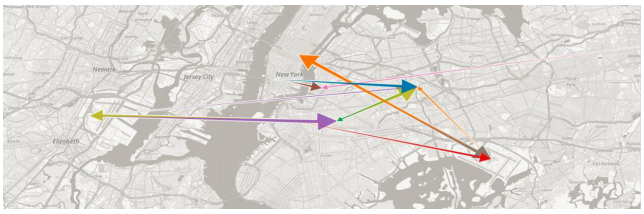


Figure 6: Frequent patterns in NYC taxi driving around 6:00 PM

home from JFK Airport after the long-time international flight. While on the contrary, in Figure 6, which indicates the most frequent patterns around 6:00 PM, passengers return home in Queens and Brooklyn after one day's work. As Figure 4 shows, 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 interference of data sparsity. The sensitiveness of pattern amount will be discussed in Section 4.3.

4.1.2 *Evaluating Metrics*

In this paper, we target at discussing the social factors within taxi drivers which may affect their driving behaviors. As mentioned above, we assume that accurate estimation of social propagation leads to better explanation of future driving behaviors. Thus, here we indirectly validate the prediction results of driving pattern fluctuation to measure the performance of our framework.

As mentioned in Section 3.5, to predict the pattern change, we indeed have two tasks, i.e., the **binary classification** to distinguish the sign (positive / negative) of pattern increment, 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, **NDCG** and **MAP** are selected. Specially, we get NDCG following the equation $\frac{DCG}{iDCG}$, in which $iDCG$ presents the ideal results and DCG will be calculated based on the formulation as below:

$$DCG = \sum_i \frac{2^{r_i} - 1}{\log(1+i)}, \tag{8}$$

where r_i denotes the relevance of result, which is set as reversal order in our experiments. Furthermore, when calculating **MAP**, 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.

Table 4: Overall Performance

	SPC	Ave	Pop	VAR
NDCG	0.3502	0.1603	0.2211	**0.3619**
Improve (%)	-	+118.46	+58.39	-3.23
P-Value	-	0.000	0.000	0.755
MAP@10	**0.2128**	0.0254	0.1042	0.2018
Improve (%)	-	+737.79	+104.22	+5.45
P-Value	-	0.000	0.000	0.472
Precision	**0.1579**	0.0134	0.0474	0.0192
Improve (%)	-	+1078.35	+233.12	+722.39
P-Value	-	0.000	0.000	0.000
Recall	**0.6892**	0.0298	0.4151	0.0875
Improve (%)	-	+2212.75	+66.03	+687.66
P-Value	-	0.000	0.000	0.000

4.1.3 *Selected Baselines*

As mentioned in preliminary, the prediction task could be generally regarded as a time-series analysis problem. Thus, here we exploit three related baselines in our experiments.

1) Personalized Average (Ave). As the basic time-series analytical tool, we follow the simple assumption that drivers' pattern ratio will only fluctuate around the average value of each dimension. Thus, this baseline uses the average value of previous p intervals (i.e., the selected time lag), to predict the change of pattern frequency in next time interval.

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 within time lag for ranking task. For the binary classification, we compare the ranking with last time lag, and the raising ones will be labeled as positive, while falling ones as negative.

3) Vector Autoregression (VAR) [20]. Vector AutoregRession (VAR) is a classical econometric model to capture the linear interdependencies among multiple time series, which suits modeling the autoregression for more than one evolving variable. As we extracted the driving records into pattern vectors, it will be appropriate to analyze the time-series with VAR model. What should be noted is that there will be one VAR model trained for a certain driver, and the estimation will also be normalized.

In summary, we select two baselines, i.e., personalized average and VAR model that are based on time-series estimation, while one more baseline according to the overall popularity, which follows the similar assumption of our framework that taxi drivers tend to follow suggestions from external information source. With these baselines, comprehensive analysis comparing different assumptions will be achieved.

4.2 Overall Results

Here, we show the overall prediction performance of our approach **SPC** (**S**ocial-aware **P**attern-**C**hange prediction) and other 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 variance of 6th month), while the parameter sensitiveness will be discussed later. Similarly, to ensure the data quality

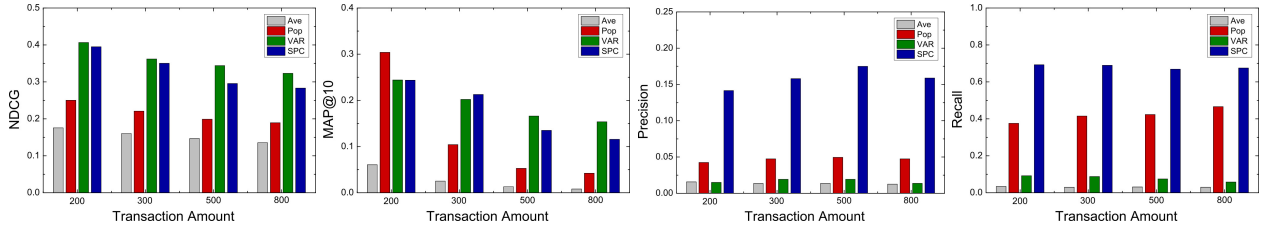


Figure 7: The verification on robustness with different set of patterns in terms of different metrics, a) NDCG, b) MAP@10, c) Precision, d) Recall.

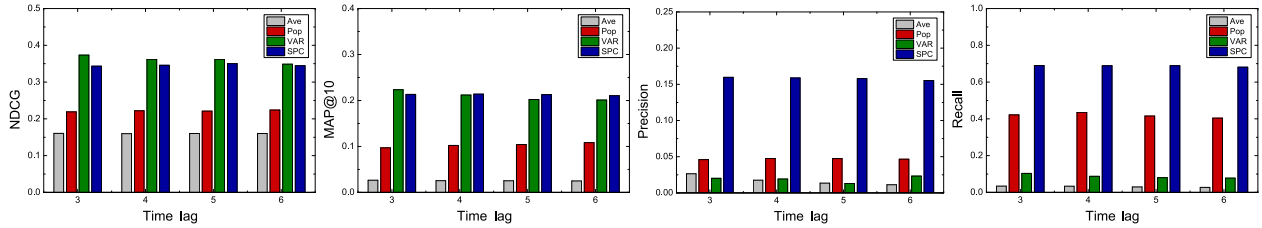


Figure 8: The verification on robustness with different time lag in terms of different metrics, a) NDCG, b) MAP@10, c) Precision, d) Recall.

with reducing sparsity, we selected top 100 taxi drivers with the most records in our experiments. Besides, we set the top 20% of results in ranking list as positive, while the detailed ROC curve will be studied at the end of this section.

The overall results are shown in Table 4. According to the results, we realize that behavior patterns of cab drivers could be largely random, as all the performance are relatively poor. However, we can find that except for the comparison with VAR on ranking problem, our approach outperforms the other baselines with dramatic margin, even 20 times better in some experiments. The performance highly supports our assumption that with introducing the latent social factors, drivers' behaviors could be better explained. The conclusion could also be partially supported by the comparison between overall popularity and personalized average, as the former one beats the later on all the metrics, and achieves a relatively good result in terms of MAP, which indicates that, especially for the most popular patterns, taxi drivers will be glad to follow the trend.

Another interesting finding is that for VAR model, it performs truly great in ranking task, but terribly fails for binary classification. With deep looking of the output of VAR, 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 those drivers tried for once but never reappear. With considering auto correlation, VAR model indeed "refused" to change. Combined with the terrible performance on binary classification, we could conclude that the ranking list of VAR might be meaningless as it fails to reveal the real pattern but only maintains the outmoded ones.

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 do not *"teach the model"* how to predict the change, but intuitively "simulate the social propagation scheme", which is finally proved as effective. Clearly, except for those intellectual services, taxi drivers themselves could be the **"best learner"**.

4.3 Parameter Robustness

As the overall performance has been validated, in this subsection, we evaluate the sensitiveness of parameters. In this task, two parameters are concerned, i.e., the amount of pattern, as well as the time lag.

For the amount of pattern, we conducted our experiments on four sets of patterns, which contains 200, 300, 500 and 800 patterns separately. The results are shown in Figure 7. We realize that for the ranking task, the metrics become worse with more patterns that may be due to the following two reasons: 1) more patterns results in more sparse data set, especially when those unpopular ones are studied; 2) more patterns raise the difficulty of ranking. Interestingly, we find that for the overall popularity method, the performance dramatically deteriorates when patterns increase from 200 to 300, which may further prove our conclusion that drivers will be glad to follow the popular trend, but for the unpopular ones, it will be difficult to predict. On the other hand, though data sparsity will also disturb the binary classification, the precision and recall rates keep relatively stable when patterns increase.

For the time lag, similarly, four sets of experiments are conducted with lag as 3, 4, 5 and 6. The results are shown in Figure 8. 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 are based on time-series estimation, when time lag increases, the performance seems slightly worsen. It is inter-

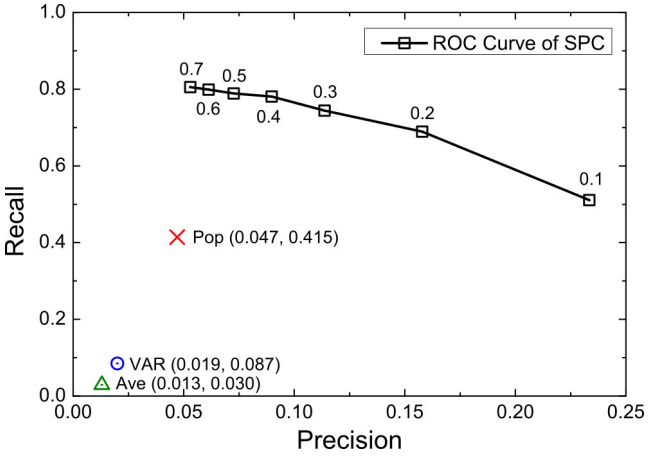


Figure 9: ROC Curve of our framework with different positive ratio.

esting as for time-series analysis, longer lag is usually better to catch the trend, but for taxi driving patterns, as proved above that it might be even a random event, thus time-dependent rules may suffer the over-fitting problem.

4.4 ROC Curve for SPC

Finally, we discuss about the ROC curve of our approach. In former experiments, we treated the top 20% of ranking list as positive. Here we conducted experiments with the same parameters of overall result, while the threshold fluctuates from 10% to 70% with interval as 10%. The ROC curve is shown in Figure 9, in which numbers near the line present the percentage of positive labels. From the figure, we can clearly find that for any threshold, 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 threshold, which indicates that social influence from other taxi drivers or overall popularity may explain at most 80% for pattern selected, while for 20% left, they may be caused by other factors, like municipal engineering or festivals.

5. DISCUSSION: POTENTIAL RULES

In this section, we will further conduct discussions on the estimated social connections for revealing the potential rules of how taxi drivers connect and how driving patterns spread.

5.1 Social Skills: Quantity v.s. Quality

Firstly, we attempt to discover whether the driving skill levels will affect the social connections. What should be noted is that all the connections here are indeed in-edge. Similar with the statistics in pre-study, *average speed* and *total income* will be selected to measure the skills. Generally, with some interesting rules found by statistical analysis, we realize that though the driving skills may not affect how many connections you have, they can determine who you will connect to or study from.

When checking the correlation between the amount of connections and two skills, we find it as completely insignificant by statistical analysis, with P-Value even larger than 0.5 for all the tests. It may indicate that better skills do not necessarily mean more in-edge connections. Clearly, "number" of "tutors" is not a critical factor.

However, since we treat the drivers themselves as one of their connections, when we check the correlation between connection amount and their own position in the ranking list of connection strength. Indeed, it surprisingly results in significantly negative correlation (with the Pearson's coefficient as -0.17), meaning that for those who have better skills are even more willing to learn from the others and update the patterns, while for the ordinary ones, they tend to insist past patterns and refuse to change.

And finally, we check the correlation between driver's skill level, as well as their difference (of rank) with "authorities", i.e., "tutors" with top 5 strongest connection strength. We find a significant correlation with Pearson's coefficient as 0.323 and P-Value 0.001, which indicates that all the drivers tend to learn from those top drivers with best skills, and those with worse driving skills will be more dependent on shared experience from top drivers.

Table 5: Social Metrics for Pattern Spread Graph

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

5.2 Pattern with Profit: Effectiveness v.s. Efficiency

Then, we discuss about the spread trend of driving patterns. Similar with traditional social spread analysis, we build one spread graph for each pattern, and for each driver u_i, if the proportion of k-th pattern is increased in time t compared with $t-1$, we defined it as u_i is **activated** in time t. Further, if u_i is activated in t and one of the in-edge connection u_j was activated in $t-1$, it is defined as u_j successfully **activates** u_i in time t and there exists an directed edge from u_j to u_i in k-th spread graph. Based on these definitions, we could draw the pattern spread track similar with traditional social influence analysis.

Then, we count four metrics to measure the graph structure of pattern spread, i.e., the *amount of edges*, which indicates the pattern's popularity; the *longest path*, which indicates the depth of spread; the *maximal out-degree*, which indicates the width of spread; and finally the *density*, which indicates the frequency of spread.

We conduct our study with the same setting of overall experiments in Section 4.2, i.e., top 300 patterns with time lag as 5 for 100 drivers. Table 5 shows the average statistics of this four metrics. With these four metrics, we could now evaluate the correlation between social spread and pattern profit. Specially, to measure the profit, we select the average speed, the average distance and total fare.

The correlation statistics are shown in Table 6. Interestingly, we find that for almost all the pairs, no matter significant or not, the co-efficient is negative. In other words, higher metrics lead to lower social attraction. Indeed, this phenomenon could be reasonable. Take the "distance" metric as an example since all the correlations are significant, we can conclude that long-range trip might not be popular among taxi drivers. Similar conclusion could be drawn for

Table 6: Correlation within Pattern Spread and Profit 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

the "fare" metric, it seems that experienced drivers are not so willing to share patterns with high rewards.

6. RELATED WORK

In this paper, we deeply analyze taxi driving patterns with considering latent social factors. Indeed, plenty efforts have been made on understanding behaviors of taxi drivers, and further developing intelligent systems, e.g., recommending hotspots that are more likely to pick up passengers quickly [34], planning practically fastest route to a given destination [33] or route which provides a optimal sequence of pick-up points [10], and even constructing the spatio-temporal profitability map for drivers to select profit locations [22]. On the other hand, for passengers to take taxi, prior arts may also list location to easily achieve vacant taxi [34], or support them to share taxi with optimal candidate [21]. Besides, though not directly generating recommendation, some other works also focus on the taxi analysis. For instance, [17] discussed about the comparison of top and ordinary drivers based on several ranking schemes, while [15] studied the strategy to pick up passengers with creating tree structure with highest probability. For the anomaly detection, historical trajectories maintained for fast distinguishing new trajectories that are isolated [36], and similar way was also utilized to detect the taxi fraud [11] or outlier trajactories [9]. Further, some applications were designed based on the taxi driving analysis, e.g., [12] proposed the method to detect traffic jams by searching the GPS records that are close together, and [4] constructed a model to automatically determine the capacity of each road segment using taxi GPS data. Finally, [5] leveraged taxi GPS traces to suggest nightly bus routes.

Another related topic of this paper is social network analysis, to be specific, the social influence or spread analysis, and the location-based social network analysis. Since social network structures are analyzed for marketing in [7] for the first time, the social-based "word-of-mouth" effect, has become one of the hottest issue in recent years, and several prior arts, e.g., [16], [1], and [19] were proposed following [7] that explicitly represent the step-by-step dynamics of influence. Among them, Independent Cascade (IC) model [16] is treated as one of the most widely-studied models with intuitive simulation and simple computation, which motivated several linear approximation like [29] and [26]. Based on the definition, some related works targeted at estimating the influence probability, i.e., the edge strength, like [14] discussed effectiveness of several heuristic method to reveal link strength, [24] predicted diffusion probabilities by using the EM algorithm, and [35] investigated the relationship between the tie strength and information propagation with several strategies of tie selection. Recently, due to the complexity of evolving social network, more issues were discussed, e.g., [28] studied on how to model the implicit social diffusion with time decay, and [2] discussed about differentiation of social influence and homogeneity, as well as the effects of weak ties in social spread.

For the location based social network analysis, the basic issue is whether human mobility indeed reflects social characteristics. Recent studies showed that human mobility is highly repetitive and non-random [13], so did [6] which revealed that 10%-30% of human movement could be explained by social factors, even more evident on long-ranged travel. Also, as comparing the social structure between online and offline social network, prior arts [18] announced that more cohesive communities will be found for offline event-driven social networks than the ordinary ones. Since social ties lead to similar mobility patterns and frequent physical contacts, mobility patterns, in return, shape and impact social connections like [8] and [25]. For instance, [3] indicated that the place of gather may benefit community detection. Also, prior arts like [37] discussed about the homogeneity and influence in location based network, and [27] further discussed the dynamic social influence within decision-making of event participation. Finally, some related work studied the offline user behaviors in the perspective of *ephemeral social networks*, e.g., [39] developed a factor graph model based framework to infer the likelihood of future encounter, and [31] recommended offline geo-friends based on pattern-based heterogeneous information network analysis.

Different from the prior arts, to the best of our knowledge, we are the first to investigate the impact of latent social factors within taxi drivers, and leverage it to explain drivers' future behaviors. Also, to reveal the social connection strength, we formulate the social influence from difference patterns in the perspective of pairwise ranking optimization, which is novel compared with related works.

7. CONCLUSION AND FUTURE WORK

In this paper, we investigated the latent social factors within the latent vehicle-to-vehicle network with simulating the social-driven behaviors pattern change of taxi drivers. To be specific, we proposed a social-driven two-stage framework to simulate latent social interactions within taxi drivers, which could better explain drivers' future behaviors. A unique characteristic of our framework is that it can transfer the problem of driving behavior prediction into the form of partial ranking with social influence, which can be regarded as a pairwise ranking problem for optimization. Extensive experiments conducted on a large-scale real-world data clearly validate the effectiveness of the proposed framework, which also prove the hypothesis that social factors indeed affect the driving behaviors of taxi drivers.

As we discussed that though social factors could better explain the pattern fluctuation at most for around 80%, there are still some other key factors, e.g., financial profit or traffic environment. In the future, we would like to investigate these factors with more comprehensive prediction framework. Also, social-oriented taxi services, e.g., social-based "tutor" or pattern recommendations will be considered. Finally, we will discover whether similar solutions could be used for other service-oriented professions.

Acknowledgments

This research was partially supported by grants from the National Science Foundation for Distinguished Young Scholars of China (61325010), the Natural Science Foundation of China (61403358), the Natural Science Foundation of China (71329201), the Science and Technology Program for Public Wellbeing (2013GS340302) and the National High Technology Research and Development Program of China (2014AA015203). Also, Qi Liu gratefully thanks the support of CCF-Intel Young Faculty Researcher Program (YFRP). Finally, Tong gratefully thanks for Guangyi Lv's kind help during the paper writing.

8. REFERENCES

[1] C. Aggarwal, A. Khan, and X. Yan, On flow authority discovery in social networks, in *Proc. SDM'11*, pp. 522-533, 2011.

[2] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic, The role of social networks in information diffusion, in *Proc. WWW'12*, pp. 519-528, 2012.

[3] C. Brown, V. Nicosia, S. Scellato, et al., Where online friends meet: social communities in location-based networks, in *Proc. ICWSM'12*, pp. 415-418, 2012.

[4] P. Castro, D. Zhang, and S. Li, Urban traffic modelling and prediction using large scale taxi GPS traces, in *Pervasive Computing*, 2012: 57-72.

[5] C. Chen, D. Zhang, Z. Zhou, et al., B-Planner: Night bus route planning using large-scale taxi GPS traces, in *Proc. IEEE PerCom'13*, pp. 225-233, 2013.

[6] E. Cho, S.A. Myers, and J. Leskovec, Friendship and mobility: user movement in location-based social networks, in *Proc. ACM KDD'11*, pp. 1082-1090, 2011.

[7] P. Domingos, and M. Richardson, Mining the network value of customers, in *Proc. KDD'01*, pp. 57-66, 2001.

[8] N. Eagle, A. Pentland, and D. Lazer, Inferring friendship network structure by using mobile phone data, in *PNAS*, 106(36):15274-15278, 2009.

[9] Y. Ge, H. Xiong, Z. Zhou, H. Ozdemir, J. Yu, and K.C. Lee, Top-eye: Top-k evolving trajectory outlier detection, in *Proc. ACM KDD'10*, pp. 899-908, 2010.

[10] Y. Ge, H. Xiong, A. Tuzhilin, K. Xiao, M. Gruteser, and M. Pazzani, An energy-efficient mobile recommender system, in *Proc. ACM KDD'10*, pp. 899-908, 2010.

[11] Y. Ge, H. Xiong, C. Liu, and Z. Zhou, A taxi driving fraud detection system, in *Proc. IEEE ICDM'11*, pp. 181-190, 2011.

[12] F. Giannotti, M. Nanni, D. Pedreschi, et al., Unveiling the complexity of human mobility by querying and mining massive trajectory data, in *VLDB Journal*, 20(5): 695-719, 2011.

[13] M. Gonzalez, C. Hidalgo, and A. Barabasi., Understanding individual human mobility patterns, in *Nature*, 453(7196):779-782, 2008.

[14] A. Goyal, F. Bonchi, and L. Lakshmanan, Learning influence probabilities in social networks, in *Proc. ACM WSDM'10*, pp. 241-250 2010.

[15] H. Hu, Z. Wu, B. Mao, Y. Huang, J. Cao, and J. Pan, Pick-up tree based route recommendation, in *Proc. WAIM'12*, pp. 471-483, 2012.

[16] D. Kempe, J. Kleinberg, and É. Tardos, Maximizing the spread of influence through a social network, in *Proc. ACM KDD'03*, pp. 137–146, 2003.

[17] L. Liu, C. Andris, A. Biderman, and C. Rattt, Uncovering taxi driver's mobility intelligence through his trace, in *IEEE Pervasive Computing*, 2009, pp. 1-17.

[18] 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. ACM KDD'12*, pp. 1032–1040, 2012.

[19] Q. Liu, B. Xiang, E. Chen, H. Xiong, F. Tang, and J. Yu, Influence Maximization over Large-Scale Social Networks: A Linear Approach with Bound, in *Proc. ACM CIKM'14*, pp. 171-180, 2014.

[20] H. Lutkepohl. New introduction to multiple time series analysis. Springer, 2005.

[21] S. Ma, Y. Zheng, and O. Wolfson, T-share: A large-scale dynamic taxi ridesharing service, in *Proc. IEEE ICDE'13*, pp. 410-421, 2013.

[22] J. Powell, Y. Huang, F. Bastani, and M. Ji, Towards reducing taxicab cruising time using spatio-temporal profitability maps, in *Proc. SSTD'11*, pp. 242-260, 2011.

[23] M. Qu, H. Zhu, J. Liu, G. Liu, and H. Xiong, A cost-effective recommender system for taxi drivers, in *Proc. ACM KDD'14*, pp. 45-54, 2014.

[24] K. Saito, R. Nakano, and M. Kimura, Prediction of information diffusion probabilities for independent cascade model, in *Springer KES*, 2008: 67-75.

[25] D. Wang, D. Pedreschi, C. Song, , et al., Human mobility, social ties, and link prediction, in *Proc. ACM KDD'11*, pp. 1100-1108, 2011.

[26] B. Xiang, Q. Liu, E. Chen, H. Xiong, Y. Zheng, and Y. Yang, PageRank with Priors: An Influence Propagation Perspective, in *Proc. IJCAI'13*, pp. 2740-2746, 2013.

[27] 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. SDM'15*, pp. 262-270, 2015.

[28] J. Yang, and J. Leskovec, Modeling information diffusion in implicit networks, in *Proc. IEEE ICDM'10*, pp. 599-608, 2010.

[29] Y. Yang, E. Chen, Q. Liu, B. Xiang, T. Xu, and S. Shad, On Approximation of Real-World Influence Spread, in *Proc. ECML-PKDD'12*, pp. 548-564, 2012.

[30] D. Yang, H. Hung, W. Lee, and W. Chen, Maximizing acceptance probability for active friending in online social networks, in *Proc. ACM KDD'13*, pp. 713-721, 2013.

[31] X. Yu, A. Pan, L.A. Tang, Z. Li, and J. Han, Geo-friends recommendation in gps-based cyber-physical social network, in *Proc. ASONAM'11*, 2011, pp. 361–368.

[32] 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. ACM SIGSPATIAL'10*, pp. 99-108, 2010.

[33] J. Yuan, Y. Zheng, X. Xie, and G. Sun, T-drive: Enhancing driving directions with taxi drivers' intelligence, in *IEEE TKDE*, 25(1):220-232, 2013.

[34] J. Yuan, Y. Zheng, L. Zhang, and X. Xie, T-finder: A recommender system for finding passengers and vacant taxis, in *IEEE TKDE*, 25(10): 2390-2403, 2013.

[35] J. Zhao, J. Wu, X. Feng, H. Xiong, and K. Xu, Information propagation in online social networks: a tie-strength perspective, in *Springer KAIS*, 32(3): 589-608, 2012.

[36] D. Zhang, N. Li, Z. Zhou, C. Chen, L. Sun, and S. Li, iBAT: Detecting anomalous taxi trajectories from GPS traces, in *Proc. UbiComp'11*, pp. 99-108, 2011.

[37] K. Zhang, and K. Pelechrinis, Understanding spatial homophily: the case of peer influence and social selection, in *Proc. WWW'14*, pp. 271-282, 2014.

[38] Y. Zheng, Y. Liu, J. Yuan, and X. Xie, Urban computing with taxicabs, in *Proc. ACM Ubicomp'11*, pp. 89-98, 2011.

[39] H. Zhuang, A. Chin, S. Wu, W. Wang, X. Wang, and J. Tang, Inferring geographic coincidence in ephemeral social networks, in *Proc. ECML-PKDD'12*, 2012, pp. 613–628.

[40] H. Zhu, E. Chen, H. Xiong, K. Yu, H. Cao, and J. Tian, Mining mobile user preferences for personalized context-aware recommendatio, in *ACM TIST*, 5(4): 58, 2015.