Data-Driven Utilization-Aware Trip Advisor for Bike-sharing Systems

Ji Hu
Zhejiang University
huji@zju.edu.cn

Zidong Yang
Zhejiang University
zdyang@zju.edu.cn

Yuanchao Shu
Microsoft Research Asia
Yuanchao.Shu@microsoft.com

Peng Cheng
Zhejiang University
pcheng@iipc.zju.edu.cn

Jiming Chen
Zhejiang University
cjm@zju.edu.cn

***Abstract*—Rapid development of bike-sharing systems has brought people enormous convenience during the past decade. On the other hand, high transport flexibility comes with dynamic distribution of shared bikes, leading to an unbalanced bike usage and growing maintenance cost. In this paper, we consider to rebalance bicycle utilization by means of directing users to different stations. For the first time, we devise a trip advisor that recommends bike check-in and check-out stations with joint consideration of service quality and bicycle utilization. From historical data, we firstly identify that biased bike usage is rooted from circumscribed bicycle circulation among few active stations. Therefore, with defined station activeness, we optimize the bike circulation by leading users to shift bikes between highly active stations and inactive ones. We extensively evaluate the performance of our design through real-world datasets. Evaluation results show that the percentage of frequent used bikes decreases by 33.6% on usage number and 28.6% on usage time.**

I. Introduction

With the development of the economy, pollution and destruction caused by human activities to natural environment was becoming more and more serious in recent years, and therefore sustainable development has become a consensus of the international community [1, 2]. In this circumstance, bike-sharing systems (BSS) are developed as a replacement for short vehicle journeys due to its low pollution, low energy consumption and high flexibility. In addition to the reduce of need for personal vehicle trips, public bike-sharing systems can not only help extend the reach of transit and walking trips, providing people with a healthy transportation option, but also trigger greater interest in cycling, and increase cycling ridership. By the end of 2016, over 1,100 cities actively operate automated bike-sharing systems deploying an estimate of 2,000,000 public bicycles worldwide [3].

With bike-sharing systems, a user can easily rent a bike by a smart card at a nearby station, use it for a short journey, and return it at another station. Despite high convenience and flexibility, a notable problem in bike-sharing systems is unbalanced bike usage, which means a small part of bikes are used much more frequently than others. Bikes that are used too much are vulnerable and hence increase repair bills and lead to potential service denied. In 2012, the very first bicycle from Hangzhou bike-sharing system became a permanent exhibit in the Low-Carbon Technologies Museum in China. This bicycle is reported to be rented for over 6,000 times and ridden for more than 20,000 kilometers in 3 years. Similarly, the most tireless bicycle from 2016 has been rented for 5,616 times, over 15 times on average each day. According to Hangzhou public bike-sharing company, the average life of their bicycles is less than 4 years due to longtime high load operation and lack of timely renewal and maintenance. On the contrary, average life of private bicycles is 10 years and above. Meanwhile, the cost of repair and labor accounts for a large proportion in overall operating expenses. In 2012, the repair cost of Hangzhou bike-sharing system was near 6 million yuan [4]. In Washington, D.C., the annual maintenance cost was $200 to $300 per bike in the year of 2012 [5]. The bike shops in New York completed 5,604 bike repairs in April 2017 with a total number of 9,367 bikes in the system [6].

Intuitively, operators can balance bike usage by leading users to use those unpopular bikes based on usage counts of each bike. However, leading users to rent a specific bike is not practical. Based on our analysis on real bike-sharing dataset from Hangzhou, we observe that bikes located in some stations are much more likely to be used and moved to another active station. Hence, by introducing the station property of activeness, we transform the original problem of picking bikes to recommending check-in and check-out stations. By using the proposed trip advisor, we aim to guide users to ride bicycles between stations with different levels of activeness, therefore avoiding circumscribed circulation among active stations. For users, an advisor can not only help them choose stations with adequate bicycles, but also ensure a higher success rate when returning bikes. Also, different incentive mechanisms can be leveraged to better prompt the balancing process.

In this paper, we propose a trip advisor that recommends the optimal pair of stations to rent and return bikes. Through guiding the actions of users, it can help balance bike usage, reduce operation cost and enhance user experience. Firstly, to make sure users can find bikes and available lockers, success rates of rental and return should be predicted for each station. Different from traditional demand prediction methods, we present probabilistic forecast methods on a minute timescale instead of predicting the exact stock number on sub-hour granularity. Secondly, in order to balance bike usage through station recommendation, a station property must be associated with bike usage frequency. We define activeness for each station by exploiting the idea of PageRank. These two parts constitute the core content of the trip advisor framework.

2374-8486/17 $31.00 © 2017 IEEE
DOI 10.1109/ICDM.2017.26

Table I
PRIMARY FIELDS IN THE BIKE-SHARING DATASET.

user_id	rent_netid	tran_date	tran_time
8601940	9926	20150601	070641

return_netid	return_date	return_time	bike_id
9205	20150601	071635	1708133

In summary, in this paper we propose a novel utilization-aware trip advisor to lead users to help balancing bike usage without compromising the quality of service. We highlight our key contributions as follows:

- We propose a probabilistic forecast method which adopts Monte Carlo simulation and random forest model to improve prediction accuracy.
- We introduce the concept of activeness to link bike usage frequency to station property which utilizes the topological characteristics of bike sharing network and the relative check out amount of each station. Meanwhile, we dynamic update the activeness to take the effect of the advisor on system into account.
- We present a novel framework to balance bike usage with the help of users and validate our proposed method with real-world human mobility datasets.

II. PRELIMINARY ANALYSIS

In this section, we first present some statistics and preliminary mobility analysis derived from the bike-sharing dataset from Hangzhou City in China. Inspired by insights obtained from the study we propose our utilization-aware trip advisor.

A. Dataset Description

The Chinese city of Hangzhou has the world's largest public BSS with more than 3300 stations and over 84,000 shared bicycles [7]. Since deployed in May 2008, thousands of bicycles have been rented for more than 700 million times. The concept of public bicycles has since spread to 30 other provinces in China and around 175 cities nationwide.

The system is classified as a third-generation bike-sharing program due to its IT-based system, automated check-in and check-out, and distinguishable bicycles and docking stations [8]. The system automatically collects user ID, bicycle ID, check-in and check-out time etc. every time users rent or return bikes. The dataset used in this paper was collected in June 2015, which contains 58,647 bikes and 3,329 stations. Each bike-sharing trip contains an origin and a destination with information of locations and timestamps. The primary fields of the dataset are shown in Table I.

B. Station Distribution

Bike stations in Hangzhou are located within the urban area spanning over 600 square kilometers; the average distance to the closest neighboring station being 300 meters [8]. Figure 1 shows the probability distribution function (PDF) of the number of stations within a certain range of one station. From this figure, we notice that half the stations have more than 3 neighbors within the range of 300 meters, and typically a station may have 8 neighbors within the range of 500 meters.

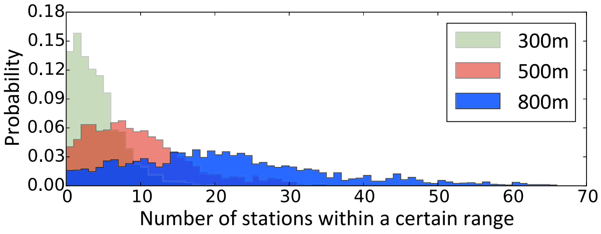


Figure 1. Station distribution.

This provides a reference to the range settings when designing the trip advisor. If we only consider stations within a very small range, there will be few stations to be selected. Otherwise, the number of candidate stations will increase significantly but users will suffer from extra walking distance. Here, we set the range threshold to 500 meters which provides 8 stations in expectation.

C. Station Diversity

After we are sure that there are enough stations to be selected near the origin and destination, we need to find out whether the stock levels of those stations are quite different from their neighbors. If the stock levels are almost the same, there is no need to predict the stock level of each station. The success rate of rental and return would be exactly the same for all the candidate stations.

Figure 2 shows the cumulative distribution function (CDF) of the number of unbalanced stations around each station in June 2015. For each station, if the difference in stock level between it and a station located within 500 meters exceeds 50%, it is considered as an unbalanced event. If the accumulated time of unbalanced events is longer than h hours in a month, the unbalanced station number increases by 1. Here, let h be 120, 180 and 240. From Figure 2, we notice that when h is set to 180, there are more than 61% of stations have at least 1 station nearby that is distinct from them in stock level. When h gets smaller, the percentage of stations that have at least 1 unbalanced station nearby is obviously increased. When h equals to 240, the corresponding percentage is 42%. According to the above analysis results, the stock level of stations within a small range could be quite different from each other, which means that it's necessary to predict the stock level and ensure that users can rent or return bikes successfully.

D. Unbalanced Bike Usage

After the analysis of station distribution and station unbalance, the most essential issue is bike usage unbalance. Because historical records contain the ID of bikes, we can extract the usage characteristics by summing up the number of occurrences and trip durations of each bike. The preliminary results are depicted in Figure 3. As shown in Figure 3(a), 57% of bikes are used for less than 150 times in a month, less than 5 times per day on average. However, about 10% of bikes

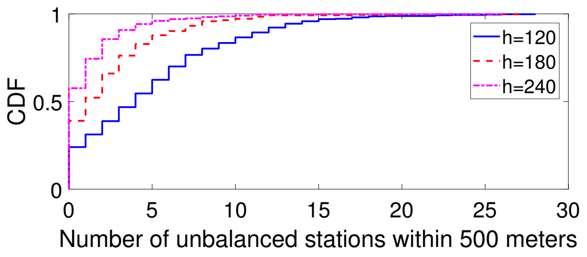
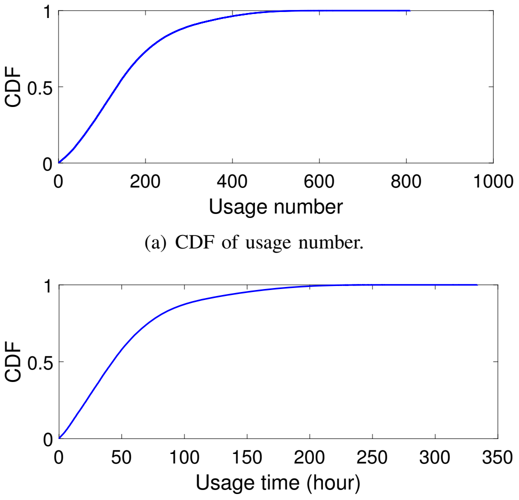


Figure 2. Station unbalance.

are used more than 310 times in a month, which is twice as frequent as less used bikes. From Figure 3(b), we can see that the usage time of 64.5% bikes is less than 57 hours in a month while that of 10% bikes is over 115 hours. These statistics clearly indicate that the usage of bikes is unbalanced and a small part of bikes have much higher usage frequency and longer usage time than others, which is the leading cause of bike damage [9].



(a) CDF of usage number.

(b) CDF of usage time.

Figure 3. Usage unbalance.

Further, we describe the usage characteristic by using the idea of the Lorenz curve. The Lorenz curve plots the percentage of total income earned by various portions of the population when the population is ordered by the size of their incomes [10]. In Figure 4, the vertical axis represents the cumulative percentage of bikes (in ascending order of usage number/time), while the horizontal axis shows cumulative percentage of bike usage number/time. We find that 60% less used bikes only contribute about 30% usage time and 33% usage number. Thus, it can be concluded that bike usage unbalanced problem does exist, and we need to design a trip advisor to guide users to help balancing bike usage.

E. Insight

In this part, we offer some insights into explaining the observed bike usage unbalance problem. A direct and effective way to explore the reasons is to identify those most frequently used bikes and observe their mobility patterns. From the historical check in and check out records, we have calculated the usage number and usage time of bikes and the results are shown in Table II. We found that the top 3 bikes on usage number is consistent with those on usage time. The most frequently used bike numbered 687,500 has been rented 809 times in a month with a total time of 333.67 hours.

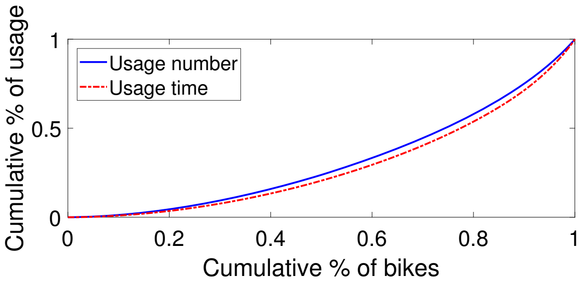


Figure 4. Cumulative contribution rate of usage.

Table II
MOST FREQUENTLY USED BIKES.

bike_id	687500	683676	687119
usage num	809	783	780
bike_id	1502964	688515	1500966
usage num	630	616	608
bike_id	687500	687119	683676
usage time (h)	333.67	319.16	314.56
bike_id	1501877	1502628	1502407
usage time (h)	259.63	258.10	257.88

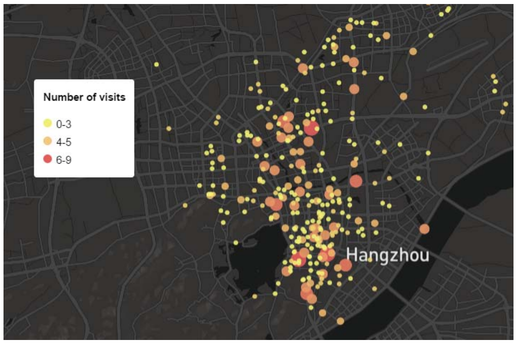


Figure 5. Geographical distribution of stations that that top 3 frequently used bikes have been visited.

It's possible that the usage frequency of each bike has close relation with the stations it has been visited. Thus, the stations where top 3 frequently used bikes have been checked out are found and the amounts of visits are counted. The geographical distribution of those stations is depicted in Figure 5. From this figure, we notice that the number of visits in main urban area is much higher. The purpose of rental in main urban areas could be going to work or school or even buying breakfast. The significant features of this kind of rental are short trip, high efficiency and quick turnover. In this case, bikes are usually rented from one station and then quickly returned to another station. After being returned, bikes are likely to be checked out again and flow to the next station quickly. Such preliminary results demonstrate that the main reason for unbalanced bike usage is the continuous circulation of bikes among active stations. On the other side, bike utilization can be balanced by introducing flows between active stations and inactive stations. How to define the activeness of stations will be elaborated in the section below.

III. Methodology overview

In this section, we first formulate the problem of station recommendation, and then show the details of the proposed trip advisor framework.

A. Problem Definition

Considering a bike-sharing system consisting of stations, bikes and users, the inputs of trip advisor are user requests including origin location l_o, destination location l_d and leaving time t_l. The user requests are stochastic and can occur at every station at any time. Let $S_o = s_{o1}, s_{o2}, ..., s_{on}$ be a set of stations in R meters zone around the origin and $S_d = s_{d1}, s_{d2}, ..., s_{dn}$ be a set of stations near the destination. Each station has its location (e.g., latitude and longitude) and stock level r_i with sub-hour granularity, where $i \in S_o, S_d$. Based on user inputs and current status of the system, the output of trip advisor is a pair of optimal stations (s_i^*, s_j^*) for users to rent and then return a bike, where $s_i^* \in S_o$ and $s_j^* \in S_d$. The problem is dynamic because decisions can be adapted over the planning horizon. In decision making process, the first step is to filter the stations in S_o and S_d based on success rate of rental and return. Hence, we will obtain a middle variable $S_o^{'}$ and $S_d^{'}$ representing candidate stations after probabilistic forecasts. The important notations used in this paper are listed in Table III.

Table III
Symbols and definitions.

l_o, l_d	location of origin/destination
t_l	leaving time
S_o, S_d	stations near the origin/destination
R	range
r_i	stock level of station $i \in S_o, S_d$
$S_o^{'}$, $S_d^{'}$	candidate stations after probabilistic forecasts

B. General Framework

Before leaving, users can send a query including their origin, destination and leaving time to the trip advisor and then get the recommended stations for rental and return. The key problem is how to guide the users to balance bike usage through station recommendation while not affecting the user experience. In this section, we will introduce the framework of our method, as shown in Figure 6. The framework is comprised of two major components: probabilistic forecasts and activeness calculation.

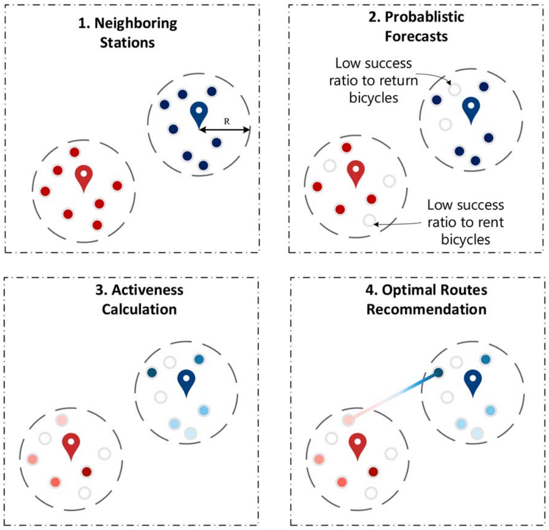

1. Neighboring Stations
R
2. Probablistic Forecasts
Low success ratio to return bicycles
Low success ratio to rent bicycles
3. Activeness Calculation
4. Optimal Routes Recommendation

Figure 6. Framework of the trip advisor.

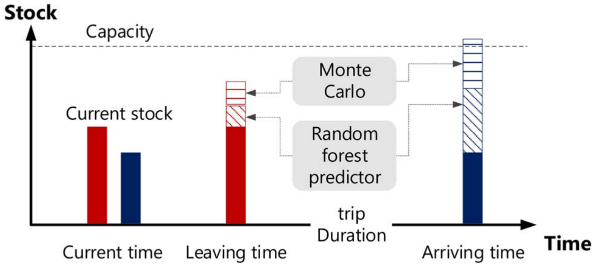


Figure 7. The idea of probabilistic forecasts.

1) Probabilistic Forecasts: In order to encourage users to use the advisor and continue to help balancing bike usage, we need to firstly make sure that users can rent or return bikes successfully. Therefore, the first component, probabilistic forecasts, is designed to solve the no-service problem and guarantee the higher success rate for rental and return when users arrive at the stations. No-service means the situations in which a user can't find available bikes to rent, and those in which he/she finds there's no parking spot to return. This problem is mainly caused by the asymmetric and fluctuating user demand among the stations. For users, they may know where the nearest station is, but what they really want to know is the probability of successfully renting or returning bikes when he/she arrives there. To obtain the success rate at a precise moment, simply predicting the forthcoming user demand on half-hour granularity is not enough to meet the above requirement. The component of probabilistic forecasts is needed to predict the stock level on a minute timescale and further derive success rate through the Monte Carlo method.

The process is illustrated in Figure 7. At the beginning, the stock levels of candidate stations near the origin/destination are known. The forecasts consist of two parts. The first part is coarse-grained prediction using random forest model, the second part is fine-grained prediction based on Monte Carlo method.

Here, we take predicting return success rate at arriving time as an example to elaborate on the details. Let $[t]$ represent the rounded time of t to the nearest 30 minutes before. At the rounded current time $[t_{now}]$, we already know the stock status r_i of station i within R meters of the destination. Firstly, we predict the base check in and check out demand at each station with sub-hour granularity by using random forest model. Random forests are an ensemble learning method for regression, that operate by constructing a multitude of decision trees with different samples and different initial variables. The final output is the mean prediction of the individual trees. We apply the random forest theory to model and predict the users behaviors with a joint consideration of time factors, meteorology and real-time bike availability[11]. Let $CI_i(t)$ and $CO_i(t)$ be the predicted check in and check out number of station i within a temporal window $(t, t+T)$, where $i \in S_d$ and $T = 30min$. The coarse-grained prediction of stock level at the rounded arriving time $[t_a]$ is as follows:

$$Stock_i([t_a]) = r_i + \sum_{t=[t_{now}]}^{[t_a]-T} (CI_i(t) - CO_i(t)) \qquad (1)$$

Then, to get more accurate stock number, we adopt the Monte Carlo method to simulate the bike rental and return process at the temporal window $([t_a], t_a)$. The general method of Monte Carlo is to obtain numerical results through repeated random sampling. We assume that the number of bikes rented or returned in the predicted time window follows a Poisson distribution. Given the station i with the predicted bike check in and check out number $CI_i([t_a])$ and $CO_i([t_a])$ in the time window $([t_a], [t_a] + T)$, we divide time delta into T small consecutive time intervals $\delta t = 1min$. The number of bikes returned to this station in each δt, noted as x, follows a Poisson distribution with mean parameter $\lambda = CI_i([t_a])/T$:

$$P(x = k) = \frac{e^{-\lambda}\lambda^k}{k!}, k = 0, 1, 2, ... \qquad (2)$$

For each simulation, we generate a stochastic sequence Q_{+i} from the return distribution to simulate the bike return events of each station. Similarly, we generate a stochastic sequence Q_{-i} for the bike rental events. Afterward, we randomly arrange the return and rental events based on the two sequences and update the stock number over time. If the stock number exceeds the capacity of the station, we mark it as an over-demand station and stop the process.

We repeat the simulation for M times to count the over-demand occurrences U. In the end, we estimate the probability of successfully returning bikes at arriving time as the rate:

$$p = 1 - \frac{U}{M}. \qquad (3)$$

The success rate for bike rental at leaving time can be calculated in a similar manner.

In summary, the main idea of probabilistic forecasts is to simulate the probabilistic process of check in and check out and derive the probability of success-of-service across a sufficiently large number of simulations. We choose the stations as candidate stations $S_o^{'}$, $S_d^{'}$ on the basis of whether its success rate is larger than a threshold P, which is set as 0.8 in our work.

2) Activeness Calculation: For the candidate stations $S_o^{'}$, $S_d^{'}$, we need to further decide which is the best pair of stations to recommend. Our ultimate goal is to balance bike usage and extend their lifespan, but we can only lead users to a station instead of recommending a specific bike. Therefore, we have to concern about how to link up the bike usage characteristic with a certain property of the station, such as activeness.

According to the previous analysis, active stations are characterized by the following properties: (1) Bikes returned to this station are easily checked out and flow to many other stations; (2) The stations that those bikes flowed to are also very active. These properties remind us of the way to measure a web page's importance. PageRank is an algorithm used by Google Search to rank websites in their search engine results [12]. According to Google: PageRank works by evaluating the quality and quantity of links to a web page to determine a relative score of that page's importance. The idea that PageRank brought up is that more important websites are likely to receive more links from other websites.

In bike-sharing systems, activeness can be defined to measure the active level of bike usage for each station based on the idea of PageRank. We begin by picturing the station network as a directed graph, with nodes represented by stations and edges represented by the bike flow (rent to return) between them. The underlying assumption is that more active stations in the network are likely to send more links to other stations. This makes sense because bikes do tend to be checked out extensively to many other stations at active stations and the bike usage in stations with more links out are usually more frequent. But this is only a start, the bikes must continue to flow to active stations so they can enter a high-speed circulation and be repeatedly used. This leads to the next assumption that stations that are themselves active weigh more heavily and help to make the stations that link to them active. If bikes rent from one station to stations with lower activeness, the bikes are likely to stay there and it will take a long time for them to be checked out again. Therefore, this station may have low activeness as well. Finally, the activeness of station i is given as

$$A(i) = \frac{1-\alpha}{N} + \alpha \sum_{j \in out(i)} \frac{n(i,j)A(j)}{n_{in}(j)} \qquad (4)$$

where

- $A(i)$ is the activeness of station i,
- α is a damping factor which can be set between 0 and 1,
- N is the number of stations,

- $n(i,j)$ is the number of bikes rent from i and return to j,
- $n_{in}(j)$ is total number of bikes return to j and
- $out(i)$ is the set of stations that have bikes rent from i.

So we can see that the activeness of station i is recursively defined by the activeness of those stations which are linked to by station i. If station i links to a lot of stations, the common belief is that station i is active. The activeness of station j which station i links to does not influence the activeness of station i uniformly. Within this algorithm, the activeness of a station j is always weighted by $n(i,j)/n_{in}(j)$. This means that the more return bikes station j has, the less will station i benefit from the link to station j. In addition, if a node has no ingoing edges, it cannot transfer its activeness to any other stations. Therefore, a damping factor is added for giving each node a probability that a bike can be returned to this station from any other station, each station has $1/N$ probability to be the source.

In the above formula, flow patterns in the station network is the main consideration, but the rental scale of each station has to be concerned as well. Stations with large amount of rentals will certainly affect the mobility of more bikes. Bikes in those stations are usually easier to spread to more stations which is an expression of high activeness. So we adopt the normalized relative check out number to indicate the rental scale and suppose that stations with large rental scale are more active. Therefore, we rewrite the activeness of station i as following:

$$\begin{aligned} A(i) &= (1-\alpha)r_i + \alpha \sum_{j \in out(i)} \frac{n_{in}(i,j)A(j)}{n(j)} \\ r_i &= \frac{n_{out}(i)/c(i)}{\sum_{j=1}^{N} n_{out}(j)/c(j)} \end{aligned} \tag{5}$$

where $c(i)$ is the capacity of station i and $n_{out}(i)$ is the absolute check out number of station i. In this way, bikes are more likely to come from stations with higher relative check out number. By introducing this prior distribution, this method provides a more comprehensive measure of the activeness of stations.

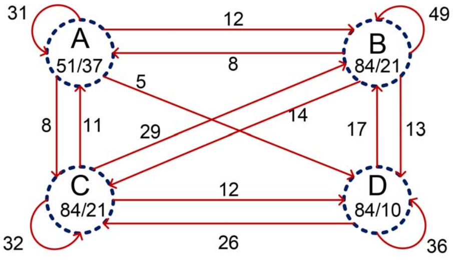


Figure 8. An example of BSS network.

Then, we use a simple example to better explain the process of activeness calculation. As shown in Figure 8, we regard a small network consisting of just 4 stations A,B,C and D referencing each other. When bikes move from station A to B, we add a directed edge between node A and node B in the graph. The weight of each edge represents the amount of bikes. For instance, there are 12 bikes rented from A and then returned to B. The relative check out numbers are noted in the center of the circles. In our model, each station should transfer its activeness to the station that links to it. Let T denote the transition matrix of the graph and Q denote normalized relative check out numbers of the stations, we get the following form of the new transition matrix M by:

$$M = (1-\alpha)Q + \alpha T \tag{6}$$

Suppose that initially the importance is uniformly distributed among the 4 nodes, each getting $1/4$. Denote by v the activeness vector of stations, we have the following equation:

$$v_{i+1} = Mv_i, i = 0, 1, 2, ... \tag{7}$$

where $v_0 = [1/4, 1/4, 1/4, 1/4]^T$. We can iterate the process until the sequences of v_0, v_1, ..., v_i tends to the equilibrium value v^* which is the activeness of our station graph. The damping factor α is to balance the influence of network topology and check out amount. The exact value of the damping factor α admittedly has effects on the final results. The activeness of stations under different α is shown in Figure 9. From this figure, we notice that the most active station is D and the activeness of A increases as α gets larger which means more emphasis on network topology.

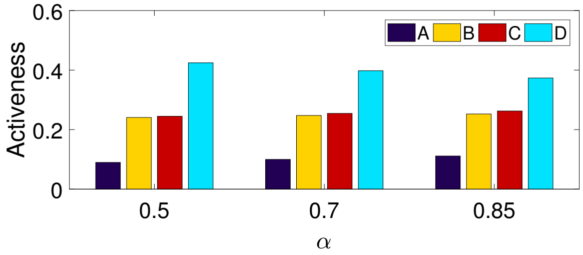


Figure 9. Activeness of stations in the example.

Finally, to obtain the optimal pair of stations (s_i^*, s_j^*), we select stations according to the following equation:

$$(s_i^*, s_j^*) = \arg\max |A(s_i) - A(s_j)| \tag{8}$$

where $s_i \in S_o^{'}, s_j \in S_d^{'}$. If users strictly follow the advisor, the activeness of stations could have a distinct change due to the altered user behaviors. Taking into account this counteraction of the advisor to the network, we update the activeness each hour using the check in and check out records within the last hour.

IV. Evaluation

In this section, we empirically evaluate the performance of our proposed method. We conduct experiments on dataset of Hangzhou bike-sharing system in June 2015. There are 10,190,841 records, which contains 58,647 bikes and 3,329 stations. The data format is presented in Table I. The records that check out and check in at the same station with a trip duration less than 2 minutes are considered as noise data and removed from the original records.

A. Probabilistic Forecasts

In our experiments, we use the results of probabilistic forecasts as a condition for filtering stations, so we evaluate the probabilistic forecasts step as a classification problem and the metrics is as follows:

Precision and Recall: Given the results of whether stations will be over-demand, precision and recall are defined as:

$$Precision = \frac{|N_{pre-od}| \cap |N_{real-od}|}{|N_{pre-od}|} \tag{9}$$

$$Recall = \frac{|N_{pre-od}| \cap |N_{real-od}|}{|N_{real-od}|} \tag{10}$$

where N_{pre-od} represents the number of events that are predicted to be over-demand, and N_{pre-od} represents the number of events that are really over-demand.

F-measure: F-measure is a weighted average of the precision and recall. We use F_β which weighs precision higher than recall by setting $\beta = 0.5$:

$$F_\beta = (1+\beta)^2 \frac{Precision \cdot Recall}{\beta^2 Precision + Recall} \tag{11}$$

We compare our proposed probabilistic forecasts method with the following three algorithms:

- **Historical average (HA)** predicts the usage demand by averaging the historical values for the same day and time [13]. For instance, the check out number of Monday 08:00 a.m. equals to the average of check out numbers of Monday 08:00 a.m. in the history and check out number of 08:00 a.m. last day.
- **Auto-Regressive and Moving Average (ARMA)** belongs to time series analysis methods and has been applied in demand prediction in [14]. It captures the temporal patterns of rental and return by leveraging check in/out information of the most recent p time windows.
- **Random forest (RF)** is the basic model where fine-grained prediction is not considered. Therefore, this method directly gives prediction of stock number instead of probabilistic results for each station.

For the experiment setup, we divide the historical records into two parts: the first 20 days for training and last 10 days for testing. We extract over-demand events by comparing the predicted stock with the threshold β multiplying the capacity. β equals to 0.2 for check out prediction and 0.8 for check in prediction.

The results are shown in Figure 10. As one can see from Figure 10, the precision of RF_MT method is as much as 0.826, 25.9% more than the HA method. ARMA and RF methods have relative higher precision but the recall of ARMA is only 0.55, which is the lowest among the three methods. On the other hand, we observe that the recall of HA is significantly larger than other methods. This is because HA method tends to predict more over-demand events, which makes most of the real over-demand events can be predicted successfully. Due to this characteristic, HA method are low in precision. Among all the approaches, RF_MT method demonstrates the best performance both in terms of precision and F-score.

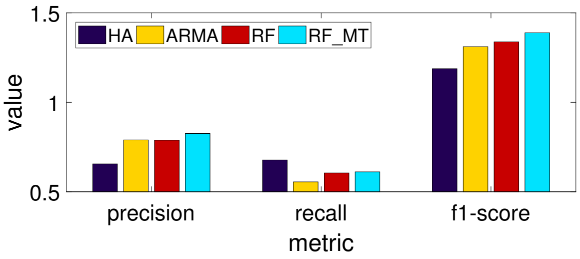
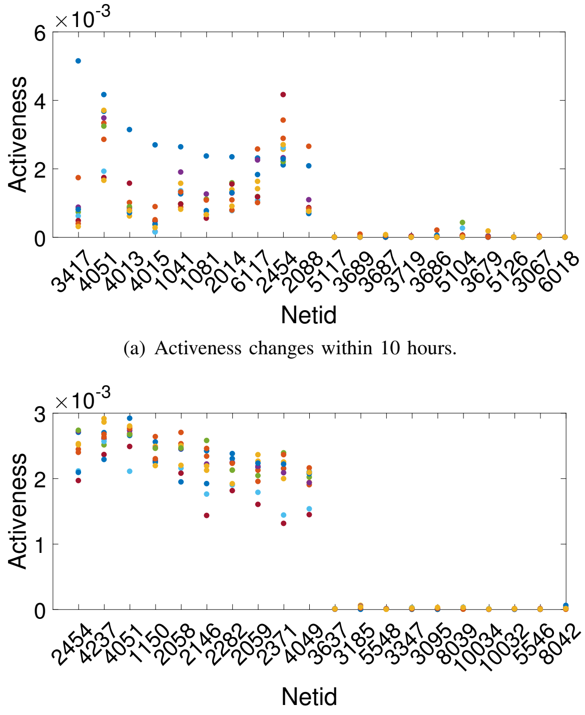


Figure 10. Precision, recall and F-measure for probabilistic forecasts.



(a) Activeness changes within 10 hours.

(b) Activeness changes within 10 days.

Figure 11. Activeness changes with the time.

B. Activeness Changes

In the simulation, we notice that the activeness of stations has different characteristics under different time granularities. The results are shown in Figure 11. Figure 11(a) reflects the activeness changes of Top 10 active and inactive stations within 10 hours. Different colors represent different hours/days. Since check out number in one hour is uncertain and random, the activeness of active stations fluctuates wildly. Meanwhile, the difference between active and inactive station looks rather small due to the short time interval. Figure 11(b) reflects the activeness changes of Top 10 active and inactive stations within 10 days. It shows relatively smooth changes of activeness for active stations and there are deep gaps between active and inactive ones. In the simulation, we update the activeness of stations for each hour because the activeness changes can be more obvious among hours especially when only small part of users follow the advisor.

C. Bike Usage Distribution

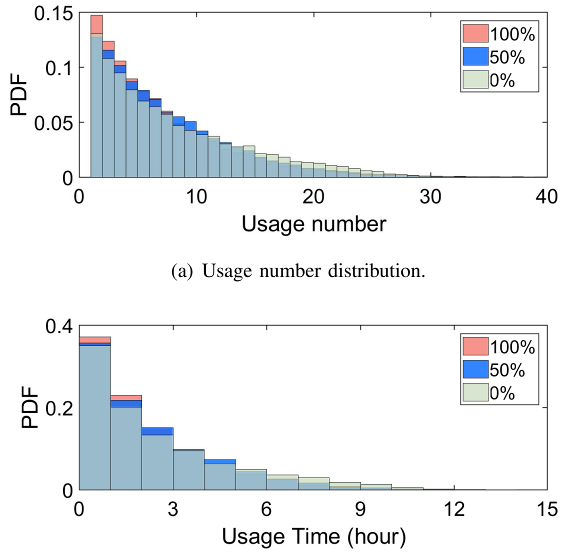
To study the model performance on bike usage distribution, we adopt PDFs of both usage number and usage time of bikes as performance metrics. In addition, we also use average (AVG) and standard deviation (STD) of usage number and usage time for evaluation. As shown in Figure 12 and Table IV, we compare situations when different proportions of the users, with 100, 50 and 0 percent, respectively, follow the advisor. We have two observations. Firstly, we can see from Figure 12(a) that compared with 0%, the percentage of less used bikes whose usage number belongs to [0,5] increases by 14.8% and the percentage of frequent used bikes whose usage number belongs to [15,40] decreases by 33.6% when the user proportion is 100%. We find out that the average usage number per day for each bike decreases from 7.656 to 6.901 when 50% of the users listen to the advisor. When the percentage rises to 100%, the average usage number is 6.625 which is down by 13.5%. The reason is that the advisor tends to use bikes that are rarely or never used more frequently. Since the total user demand stays the same with the original records, the more bikes are used, the smaller the average usage number will be. Secondly, the average usage time per day becomes more balanced as shown in Figure 12(b), especially for the bikes with usage time larger than 6 hours per day. The percentage of frequent used bikes whose usage number belongs to [6,15] decreases by 28.6% when the user proportion is 100%. The standard deviation of usage time for 100% and 50% proportion of users are 1.99 and 2.04 while that of the historical records is 2.39. These results prove that the proposed method can help to balance both bike usage number and usage time. In addition, with the proportion of users grows, the effect of usage balancing gets better.

Table IV
AVG AND STD USAGE UNDER DIFFERENT PROPORTIONS OF THE USERS.

User proportion	AVG of usage number	STD of usage number	AVG of usage time	STD of usage time
100%	6.56	5.60	2.11	1.99
50%	6.83	5.45	2.22	2.04
0%	7.57	6.16	2.50	2.39

D. Impact of Range Settings

Experimental results for the advisor derived in this paper show high performance, demonstrating the potential of the approach. To better understand the performance of the proposed method, we further conduct an evaluation by varying the range parameter in the model. The range R is the distance allowed between stations and the origin/destination, which is set from 500m to 1000m and 200m. Here, we assume that all the users follow the advisor. The bike usage distribution under different range settings are shown in Figure 13. When the range is set to 200m, usage number between 5 and 15 per day take the large proportion compared with other settings which has benefit effect on usage balancing. However, there are only few stations to be chosen when $R = 200m$ and the simulator failed to offer a suggestion for more than 15,000 time per day. When the range is set to be 1000m, the experiment results have been improved, but too large range settings will cause added walking distance of users and seriously impact user experience.



(a) Usage number distribution.

(b) Usage time distribution.

Figure 12. Usage distribution under different proportions of the users.

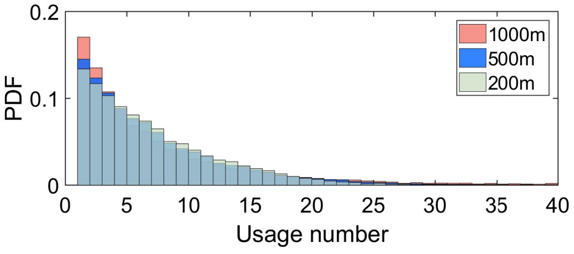
V. Discussion

In this part, we provide some insights into the proposed framework, and provide directions for future work.

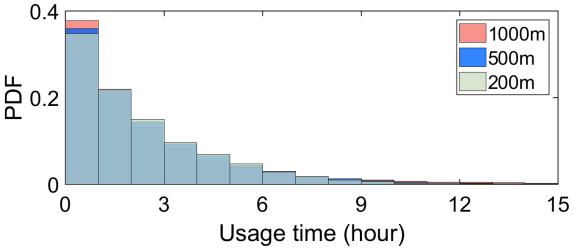
A. Reward Design

Although the advisor can improve the success rate of rental and return in a certain extent, it may also bring additional distance cost to users when realizing the goal of balancing bike usage. For the sake of keeping users' enthusiasm, we can design a reward mechanism to guide the use of shared bikes in the future. For example, the reward can be given by the function $F(d)$ based on the extra distance d that users have to pay. Here,

$$\begin{aligned} &F(d) = k * d, \\ &\quad d = [distance(s_i^*, s_j^*) - \min(distance(s_i, s_j))] \end{aligned} \tag{12}$$



(a) Usage number distribution.



(b) Usage time distribution.

Figure 13. Usage distribution under different range settings.

where $s_i \in O, s_j \in D$. Then, the mechanism transforms the reward of the users into a discount of their public transit cards. Detailed design and evaluation of such reward mechanism is beyond the scope of the paper, and there are many references on this subject [15, 16]. Through this way, users are motivated to help balancing bike usage and it's beneficial to build intelligent and self-sustainable transportation systems.

B. Other Objective Functions

In practical applications, the advisor enables system operators to design other objective functions, thus achieving flexible resource scheduling. For example, we could advise users to rent bikes from active stations and still return them to active stations. Therefore, the aging process of a small part of bikes will be accelerated, allowing the regular upgrades of bikes in the system. Otherwise, it's unacceptable to the normal operation of the systems that a large number of bikes need replacing in the same time.

VI. Related work

Due to the increasing importance and rapid development of bike-sharing systems, a great deal of attention has been focused on a variety of problems that relate to bike-sharing. There are various interesting research questions concerning the establishment, operation and strategic problems of bike-sharing systems [8, 17–22]. For example, Shaheen *et al.* [8, 17, 18] studied the history, business models and the social and environmental benefits of bike-sharing in Europe, the Americas and Asia. Parkes *et al.* [20] explored systems' location, evolution, and their adoption. In addition, a novel use case of the heterogeneous urban open data, namely bike-sharing station placement, was proposed in [21, 22].

Another important research direction concerns user demand prediction. Several papers firstly analyzed user behavior patterns and then proposed predictive models to forecast bike usage demand or stock level of stations in the future period [14, 23–26]. The prediction methods are summarized into two categories: station-centric model and cluster-centric model. The station-centric model predicts demand for each station individually. For instance, Froehlich *et al.* [23] used four basic prediction models to predict available bikes in each station: last value, historical mean, historical trend and Bayesian network. Kaltenbrunner *et al.* [24], Borgnat *et al.* [25] and Vogel *et al.* [14] distinguished typical usage patterns and predicted the hourly user demand in the bike-sharing systems of Barcelona, Lyon and Veinna, respectively, by using time series analysis method. However, these methods show their limitation on prediction performance, especially when predicting the traffic under unusual situations. For cluster-centric model, it usually partitions the stations into clusters and predicts the totoal demand of each cluster [26, 27]. For example, Yexin Li *et al.* [26] proposed a hierarchical prediction model, which contains a bipartite clustering algorithm, a multi-similarity-based inference model, and a check-in inference algorithm, to predict the number of bikes that will be rent from/returned to each cluster, but the geographical granularity of this method is too sparse for trip advisor design.

Based on insights into usage demand analysis, the allocation of resources, bikes and empty places, has to be managed by the operator. To balance the stock level, methodologies in [28–31] tackled the problem of finding truck routes and decided the number of bikes to move between stations that minimizes the distance traveled by trucks. Raviv *et al.* [28] presented two mixed integer linear program formulations to solve the static repositioning problem which assumes that the repositioning is during the night when the usage rate of the system is negligible. Authors in [30] introduced a dynamic public bike-sharing balancing problem when the status of the system is rapidly changing. Redistribution can also be done by users through a crowdsourcing mechanism that incentivizes the users in the bike repositioning process [15, 16]. Similar method has been applied into vehicle sharing systems in [32]. Both dynamic vehicle redistribution and online price incentives were considered in [33]. Different form the above methods, we establish a framework aiming at balancing the usage of bikes instead of the stock level of stations.

VII. Conclusion

In this paper, we propose a novel architecture of a utilization-aware trip advisor which engages users to balance bike usage and prolong the maintenance intervals of bikes. Starting from ensuring users' success rate of rental and return, the advisor is designed to dynamically recommend the optimal stations based on their current activeness of bike usage. We evaluated the proposed system through extensive simulations using historical records from the world's largest bike-sharing system, confirming the effectiveness of our framework.

Acknowledgment

The paper was partially supported by the National Basic Research Program (973 Program) under Grant 2015CB352500, and National Program for Special Support of Top Notch Young Professionals.

References

[1] P. DeMaio, "Bike-sharing: History, Impacts, Models of Provision, and Future," *Journal of Public Transportation*, vol. 12, no. DeMaio 2004, pp. 41–56, 2009. [Online]. Available: http://www.transitinformatics.org/test/nctr/wp-content/uploads/2010/03/JPT12-4DeMaio.pdf

[2] P. Midgley, "Bicycle-sharing schemes: enhancing sustainable mobility in urban areas," *United Nations, Department of Economic and Social Affairs*, pp. 1–12, 2011.

[3] L. MetroBike, "2016 Year-end wrap-up will appear at the end of January," http://bike-sharing.blogspot.com/2017/01/2016-year-end-wrap-up-will-appear-at.html.

[4] Z. Online, "Hangzhou will add three public bicycle maintenance bases," http://zjnews.zjol.com.cn/system/2013/09/23/019608857.shtml.

[5] MSA, "BIKE SHARE PROGRAM," http://www.michigansuburbsalliance.org/wp-content/uploads/MSA-CAP-Strategy-Library_Bike-Share-Program_20211017.pdf.

[6] CitiBike, "Citi Bike Monthly Operating Reports," https://www.citibikenyc.com/system-data/operating-reports.

[7] Wikipedia, "Hangzhou Public Bicycle," https://en.wikipedia.org/wiki/Hangzhou_Public_Bicycle.

[8] S. Shaheen, H. Zhang, E. Martin, and S. Guzman, "China's hangzhou public bicycle: Understanding early adoption and behavioral response to bikesharing," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2247, pp. 33–41, 2011.

[9] dayoo, "Difficulty of renting public bicycles due to many damaged bikes," http://hainan.ifeng.com/news/detail_2014_06/18/2451605_0.shtml.

[10] J. L. Gastwirth, "A general definition of the lorenz curve," *Econometrica: Journal of the Econometric Society*, pp. 1037–1039, 1971.

[11] Z. Yang, J. Hu, Y. Shu, P. Cheng, J. Chen, and T. Moscibroda, "Mobility modeling and prediction in bike-sharing systems," in *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '16. New York, NY, USA: ACM, 2016, pp. 165–178. [Online]. Available: http://doi.acm.org/10.1145/2906388.2906408

[12] S. Brin and L. Page, "Reprint of: The anatomy of a large-scale hypertextual web search engine," *Computer networks*, vol. 56, no. 18, pp. 3825–3833, 2012.

[13] N. Gast, G. Massonnet, D. Reijsbergen, and M. Tribastone, "Probabilistic forecasts of bike-sharing systems for journey planning," in *ACM CIKM*, 2015.

[14] P. Vogel and D. C. Mattfeld, "Strategic and Operational Planning of Bike-Sharing Systems by Data Mining - A Case Study," in *Computational Logistics*, 2011, pp. 127–141.

[15] A. Singla, M. Santoni, G. Bartók, P. Mukerji, M. Meenen, and A. Krause, "Incentivizing users for balancing bike sharing systems." in *AAAI*, 2015, pp. 723–729.

[16] C. Fricker and N. Gast, "Incentives and redistribution in homogeneous bike-sharing systems with stations of finite capacity," *Euro journal on transportation and logistics*, vol. 5, no. 3, pp. 261–291, 2016.

[17] S. a. Shaheen, S. Guzman, and H. Zhang, "Bikesharing in Europe, the Americas, and Asia," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2143, pp. 159–167, 2010.

[18] S. a. Shaheen, A. P. Cohen, and E. W. Martin, "Public Bikesharing in North America: Early Operator Understanding and Emerging Trends," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2387, pp. 83–92, 2013. [Online]. Available: http://trb.metapress.com/openurl.asp?genre=article{&}id=doi:10.3141/2387-10

[19] E. W. Martin and S. A. Shaheen, "Evaluating Public Transit Modal Shift Dynamics in Response to Bikesharing: A Tale of Two U.S. Cities," *Journal of Transport Geography*, vol. 41, pp. 315–324, 2014. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0966692314001409

[20] S. D. Parkes, G. Marsden, S. A. Shaheen, and A. P. Cohen, "Understanding the Diffusion of Public Bikesharing Systems: Evidence from Europe and North America," *Journal of Transport Geography*, vol. 31, pp. 94–103, 2013. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0966692313001130

[21] L. Chen, D. Zhang, G. Pan, X. Ma, D. Yang, K. Kushlev, W. Zhang, and S. Li, "Bike Sharing Station Placement Leveraging Heterogeneous Urban Open Data," in *ACM Ubicomp*, 2015.

[22] J. Liu, Q. Li, M. Qu, W. Chen, J. Yang, H. Xiong, H. Zhong, and Y. Fu, "Station site optimization in bike sharing systems," in *Data Mining (ICDM), 2015 IEEE International Conference on*. IEEE, 2015, pp. 883–888.

[23] J. Froehlich, J. Neumann, and N. Oliver, "Sensing and Predicting the Pulse of the City through Shared Bicycling," in *IJCAI*, 2009.

[24] A. Kaltenbrunner, R. Meza, J. Grivolla, J. Codina, and R. Banchs, "Urban Cycles and Mobility Patterns: Exploring and Predicting Trends in a Bicycle-based Public Transport System," *Pervasive and Mobile Computing*, vol. 6, no. 4, pp. 455–466, 2010.

[25] P. Borgnat, E. Fleury, C. Robardet, and A. Scherrer, "Spatial Analysis of Dynamic Movements of Vélo'v, Lyon's Shared Bicycle Program," in *European Conference on Complex Systems (ECCS)*, 2009.

[26] Y. Li, Y. Zheng, H. Zhang, and L. Chen, "Traffic Prediction in a Bike Sharing System," in *ACM SIGSPATIAL*, 2015.

[27] E. O. Mahony and D. B. Shmoys, "Data Analysis and Optimization for (Citi) Bike Sharing," in *AAAI*, 2015.

[28] T. Raviv, M. Tzur, and I. Forma, "Static Repositioning in a Bike-sharing System: Models and Solution Approaches," *EURO Journal on Transportation and Logistics*, vol. 2, no. 3, pp. 187–229, 2013. [Online]. Available: http://dx.doi.org/10.1007/s13676-012-0017-6

[29] J. Shu, M. C. Chou, Q. Liu, C.-P. Teo, and I.-L. Wang, "Models for Effective Deployment and Redistribution of Bicycles Within Public Bicycle-Sharing Systems," *Operations Research*, vol. 61, no. 6, pp. 1346–1359, 2013.

[30] Contardo, Claudio, C. Morency, and L.-M. Rousseau, "Balancing a Dynamic Public Bike-sharing System," Tech. Rep., 2012.

[31] J. Schuijbroek, R. Hampshire, and W.-J. van Hoeve, "Inventory Rebalancing and Vehicle Routing in Bike Sharing Systems," Tech. Rep., 2013.

[32] A. Waserhole and V. Jost, "Pricing in vehicle sharing systems: Optimization in queuing networks with product forms," *EURO Journal on Transportation and Logistics*, vol. 5, no. 3, pp. 293–320, 2016.

[33] J. Pfrommer, J. Warrington, G. Schildbach, and M. Morari, "Dynamic vehicle redistribution and online price incentives in shared mobility systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 4, pp. 1567–1578, 2014.