Context Aware Flow Prediction of Bike Sharing Systems

Yang Zhou
Department of Computer Science and Engineering
University of North Texas
Denton, Texas, USA
Email: yangzhou2@my.unt.edu

Yan Huang
Department of Computer Science and Engineering
University of North Texas
Denton, Texas, USA
Email: yan.huang@unt.edu

***Abstract*—The number of bike-sharing services worldwide has grown over 1800% from 2014 to 2018. Accurate flow prediction of bike-sharing systems is critical for increasing user satisfaction, improving the efficiency, and planning new stations. Current flow prediction methods leave out important contextual features. In this paper, we propose a context-aware framework to predict bike flows for both existing stations and new stations. The model incorporates spatiotemporal, network, and environment contexts in a synergistic manner. The proposed model is evaluated on a bike-sharing system in New York City (NYC). Our model reduced error rate by 9% and root mean squared logarithmic error by 8% compared with the state-of-the-art neural network based method using context features. We also extend our model to a clustered bike flow prediction model, which is the focus of most of the current literature. Our model outperformed the state-of-the-art clustered model by 23% and 25% under anomalous situations in error rate and root mean squared logarithmic error respectively.**

***Keywords*-Bike-sharing system; traffic prediction; bike flows; point of interest; contextual feature;**

I. Introduction

Bikes are compact in size, easy to park, and environment-friendly. Due to these characteristics, bikes are popular in many large cities. Bike-sharing system started in 1965 in Amsterdam with 300 bikes. As of 2015, there are more than 200 cities around the world having a bike-sharing system [17]. Example bike-sharing systems include Velo'v in Lyon (4,000 bikes at 334 stations, started in May 2005), Velib in Paris (2,000 bikes at 1,500 stations, started in 2007), Bixi in Montreal (5,000 bikes at 400 stations, started in 2009), and Ecobi in Mexico (1,000 bikes at 85 stations, started in Feb 2010). The bike-sharing system in Hangzhou, China, started in 2008 is considered as the largest today. The system operates with 50,000 bikes at 2,000 stations. One station is available for every 100 meters in the city [11].

Bikes provide a convenient way for people to go to places and can help to reduce traffic congestion. Timothy et al. [13] found that the availability of a bike-sharing system reduces traffic congestion by 2 to 3% within a neighborhood in Washington, DC, area. Bike-sharing programs also contribute to environment preservation. Montreal's Bixi proudly states that its program has saved over 3,000,000 pounds of greenhouse gases since inception in May, 2009. Lyon states that its Velo'v program has saved the equivalent of 18,600,000 pounds of CO2 pollution from the atmosphere since 2005 [9].

In a typical bike-sharing system, users can easily borrow (i.e. check-out) a bike from nearby stations by swiping their ID card and return (i.e. check-in) the bike to a station close to their destination [23]. In such a system, flow prediction is one of the most important issues [8], [19]. Accurate flow prediction can help the bike-sharing systems to operate more efficiently and enhance the resource utilization. Flow prediction can help avoid overflow where a station is full and users cannot return their bike there. Overflow forces a user to return to another station or park the bike privately overnight, which incurs additional time and monetary costs. Flow prediction can also avoid underflow where a station is running out of available bikes and a user could not rent a bike as the result. Furthermore, flow prediction is useful when a new station is planned. Knowing number of bikes to be borrowed and returned helps the relevant agency to select the right location for the new station and instrument the station with proper number of bikes.

Current works of flow prediction are mainly in cluster level and perform poorly on predicting flows of a single station. Furthermore, they only consider time and environment factors and rely heavily on historical flow data. We argue that the surrounding points of interests, distances to other stations, and structure of a station in relation to others play important roles in flow prediction. We consider several contextual features:

1) Spatiotemporal Context: People have purposes for their trips, e.g., work, shopping, and school. So, the point of interest (POI) such as work place, shopping mall, school has strong influence on bike flows. For instance, during morning rush hours, most of the bikes are used to go to work and during evening, bikes are often used to go back home as well as recreation places.

2) Network Context: This context includes a bike station's degree and distance from other stations. The degree of a bike station is the number of other stations that the station has flow to/from during a time period. Distance between stations influence bike flow as people need to borrow from and return to stations.

978-1-5386-5035-6/18/$31.00 ©2018 IEEE

3) Environment Context: This context models environment impact on flow and includes factors such as temperature and weather, e.g. sunny, rainy, windy, and snowy [24].

In this paper, we propose a context-based attractiveness model to predict bike flows and make the following contributions:

- We collect the point of interest(POI) around each bike station and categorize them into different groups. POIs with similar properties and functions are classified in the same group;
- We give definitions of spatiotemporal, network context, and environment contexts. We build an attractiveness model based on historical bike flow data, POIs, degree and distance;
- We conduct experiments on real datasets to show the context-based attractiveness model can predict check-out/in data of each station accurately. Our model outperforms the state-of-the-art neural network based method using context features [18] (9% reduced rate on error and 8% reduced root mean squared logarithmic error) on real data from NYC.
- A Bipartite Station Clustering algorithm is implemented, which is from the former work of Li [15]. We combine bipartite station clustering algorithm with our attractiveness model to predict check-out/in data in cluster level. Our model outperforms the former work on both normal situations and (2% reduced rate on error and 4% reduced root mean squared logarithmic error) and anomalous situations (a 23% reduced rate on error and 25% reduced root mean squared logarithmic error).

II. Related Work

Bike-sharing systems have experienced significant growth worldwide and are quickly gaining popularity. This growth has created challenges. Recent studies on bike-sharing system are summarized into three categories: flow prediction, flow pattern discovery, and system balance.

Flow prediction: Flow prediction is an important topic. Li et al. [15] proposed a hierarchical prediction model to predict the number of bikes that will be rented from/returned to each station cluster in a future period so that reallocation can be executed in advance. They proposed a bipartite clustering algorithm to cluster bike stations into groups, formulating a two-level hierarchy of stations. The total number of bikes that will be rented in a city is predicted by a Gradient Boosting Regression Tree (GBRT). Then a multi-similarity-based inference model is proposed to predict the rented proportion across clusters and the inter-cluster transition, based on which the number of bikes rented from/ returned to each cluster can be easily inferred. Vogel et al. [26], [25] used time series analysis to forecast bike demand in Vienna based on bike data from Dublin. Both of these works can only predict bicycle flows at cluster level and can not predict the flow of each station. Yoon et al. [27] proposed a modified ARIMA model, considering spatial interaction and temporal factors, to predict the available bike/docks at each station.

Flow pattern discovery: Understanding the pattern of a bike-sharing system helps to know the mobility of a city. Coffey et al. [7] combine social networks with bike-sharing system. They model bike trip counts and social media check-ins into topics (mobility purposes) using Latent Dirichlet Allocation. They then try to incorporate topics detected from social media to help predict the bike flow using a multi-layer perception neural network. The flow estimation is on a large level such as a city and could not drill down to a station. Froehich et al. [12] and Kaltenbrunner et al. [14] provide spatiotemporal analysis of Barcelona's bike station usage pattern by clustering techniques. Borgnat et al. [2], [3] detected the mobility pattern of Velo'v by interpreting the system as a dynamical network and analyzing how the flows are distributed spatially along the network. Works in [10] proposed a model to analyze the patterns in different areas of a city by different functions, considering the latent factors of each station.

Bike-sharing system balance: The design of bike-sharing system needs to know the demands from users. Sayarshed et al. [22], Chemla et al. [4], [6] and Benchimol et al. [1] present an optimization formulation to design a bike-sharing system for travel inside small communities, or as a means to extend public transport for access and egress trips.

The mathematical model [16] attempts to optimize a bike-sharing system by determining the minimum required bike fleet size that minimizes simultaneously unmet demand, unutilized bikes, and the need to transport empty bikes between rental stations to meet demand.

Lin et al. [17] proposed a model which attempts to determine the number and locations of bike stations, the network structure of bike paths connected between the stations, and the travel paths for users between each pair of origins and destinations.

However, previous flow prediction methods cannot be adopted to our work directly. Our work focuses on the check-out/in prediction of each station. In addition, previous works predict bike flow from system level or cluster level and cannot work for single station. Our work does not use the number of available bikes/docks at each station, but those in [5], [27], [10] do. People always have a purpose for a trip. They will have a particular destination such as work place, shopping mall and so on. It means, points of interest can represent their travel purposes. However, previous works ignored this important factor. In our work, a context aware flow prediction model is proposed, which uses POIs as an important factor. Our work is a new aspect for flow prediction of bike-sharing system.

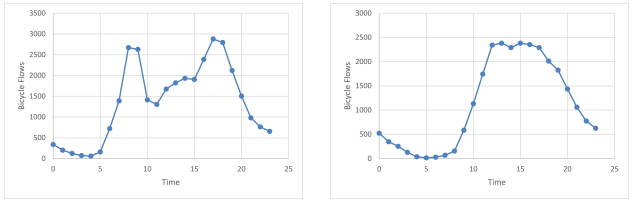
III. Overview

In this section,we give a detailed description of contextual features which will influence the bike-sharing system. To

make our model efficient and effective, we analyze the importance of different features.

A. Spatiotemporal Context

1) Points of interest: To the prosperity of New York City, there are so many points of interest (e.g., shopping malls, schools, residential, work places, churches) located in the city. Understanding the purpose of the trip by each person who participates in bike-sharing system will help us to analyze the flow prediction. So, we need to extract the POI around each bike station. There are a lot of POIs in New York City. The POIs are very close to each other. We set a radius around each station and then collect the POIs within the radius. Based on our experiment, choosing a radius as 50 meters is proper. In our work, we get 80 different types of POI totally. Some of the POIs are very similar to each other. Therefore, we group 80 POIs in further step. By grouping POIs, we get 10 groups at last. Table I gives the groups and POIs in detail.



(a) Weekday (b) Weekend

Figure 1. Average bike flows

2) Time: The bike-sharing system will be affected by time. As bike-sharing system is a dynamic system, bike flows will change over time. The time period is set as one hour. We focus on bike flows of different hours within one day as well as the difference between weekday and weekend. Fig.1(a) and Fig.1(b) show the average bike flows of weekdays and weekends of NYC in August, 2014, during different hours of a day. From Fig.1(a) and Fig.1(b), we can see that bike flow is different during each time period(one hour). There are more bike flows in the day time. Weekday and weekend also have different bike flow patterns. Fig.1(a) consists of morning rush hours and evening rush hours. Morning rush hours are around 8 am and evening rush hours around 6 pm. However, Fig.1(b) only has day rush hours, which are around 12 pm. Weekdays also have higher bike flows than weekend. The reason is that, on weekday many people will go to work around 8 am and after work around 6 pm by using bikes as their transportation vehicle.

Fig.2 shows the bike flows of each station during different time periods.

B. Network context

1) Degree: In the bike-sharing system, the degree of a bike station is the number of connections with other stations during a time period. Larger degree of a station means more stations have connection with this station. Large degree of a station also means more bike flows related to this station. With larger degree, the flow of bike station is more stable. So degree has influence on the flow prediction.

2) Distance: In the flow prediction, we need to consider distance. People will not choose a bike as a vehicle if the destination is too far away. They may use car or other transportation methods. In the system, if there are two stations with same POIs, people will choose a station which has a smaller distance from them as their destination. It saves time and is convenient.

C. Environment context

Environment context contains temperature and weather, e.g. sunny, rainy, windy, and snowy. Environment context separates bike-sharing system into normal and anomalous parts. In the normal part, weather is sunny and temperature changes slowly. However, in anomalous part, it will be atrocious weather and temperature will change suddenly. So different environment context affect flow prediction. In the experiment, environment context is used to get the anomalous situation.

IV. Context Based Model

In a specific time period, we combine spatiotemporal, network context and historical bike flows of stations to make a prediction. During different time of one day, there are different number of bikes flow into the stations. For example, there are more bike flows in daytime than midnight. So the flow correlates with time. In our experiment, we choose one hour as time period. T = $<t_1, t_2 ... t_i>$, t_i is the time period of [i-1,i), $(1 \leq i \leq 24)$. There are 344 bike stations in our New York city data set. We use s_n $(1 \leq n \leq 344)$ to represent station n, and d_{nm} is the distance from station s_n to s_m, $(1 \leq n, m \leq 344)$. Table II shows all the definitions of parameters we will use in our model.

Table II
Definition of Parameters

Parameters	Definition
R	Total number of training records
N	Total number of stations
s_m	Station m
d_{nm}	Distance between s_n and s_m
$\overline{d_m^{t_i}}$	Average distance between other stations and s_m during t_i
t_i	Time period of [i-1,i) $(1 \leq i \leq 24)$
α, β	Tuning parameter
x_{nm}	Status of connection between s_n and s_m
$k_m^{t_i}$	Degree of station s_m during t_i
$\overline{k_m^{t_i}}$	Average degree of station s_m during t_i
$w_{nm}^{t_i}$	Bicycle flows from s_n to s_m during t_i
$f_m^{t_i}$	Total flows of s_m during t_i
$fc_m^{t_i}$	Flow centrality of s_m during t_i
$at_m^{t_i}$	Attractiveness of s_m during t_i
x_i	Feature group number $(1 \leq i \leq 10)$

Table I
GROUPS AND POINT OF INTERESTS

Group	Point of Interests
store	bicycle_store, book_store, clothing_store, convenience_store,grocery_or_supermarket, electronics_store, furniture_store, hardware_store, shoe_store, car_dealer, car_rental, home_goods_store, jewelry_store, liquor_store, pet_store, car_repair
government place	courthouse, embassy, local_government_office, police, postoffice
recreation	art_gallery, bar, cafe, casino, movie_theater, spa, night_club
finance	accounting, atm, bank, finance
public place	church, cemetery, library, gym, park, museum, school, university, parking, synagogue, funeral_home, subway_station, mosque, train_station, campground
food	bakery, food, meal_takeaway, restaurant
healthy	dentist, doctor, health, pharmacy
living	beauty_salon, hair_care, laundry, locksmith, storage, lodging, neighborhood
work type	lawyer, florist, painter, plumber, place_of_worship, travel_agency, real_estate_agency, insurance_agency, general_contractor, veterinary_care
other	establishment, natural_feature, point_of_interest, sublocality_level_1, route

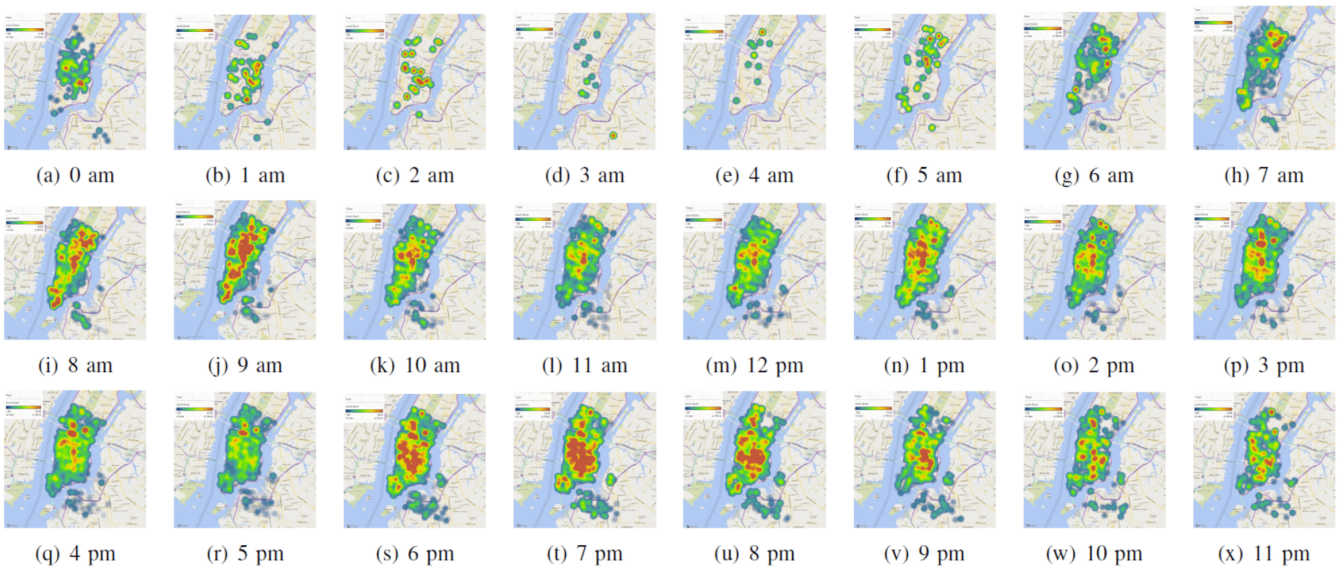


Figure 2. The Heat Map of Check-in Data of Each Hour During One Day.

Definition 1: (**Surrounding Point of Interest**) A station will be surrounded by several typical POIs. For example, shopping mall, supermarket, hospital, park, residential, highway, bus-stop and so on. We use Google Map API to get the features in New York City. Some of the POIs are similar. So we classify the POIs into different 10 categories. The categorized groups are listed in Table I. We choose the POIs within 50 meters, which is the radius around the station.

Definition 2: (**Flow**) The flow of a station is the total number of bikes ended up in the station during a time period. Within t_i, $w_{nm}^{t_i}$ is the number of bikes from station s_n to station s_m. So during this time period, the flow of station s_m is defined as:

$$f_m^{t_i} = \sum_{n=1}^{N} w_{nm}^{t_i}$$

If no bike is taken from station s_n to station s_m, $w_{nm}^{t_i} = 0$. Larger number of $f_m^{t_i}$ means more bikes from other stations are flowing into station s_m.

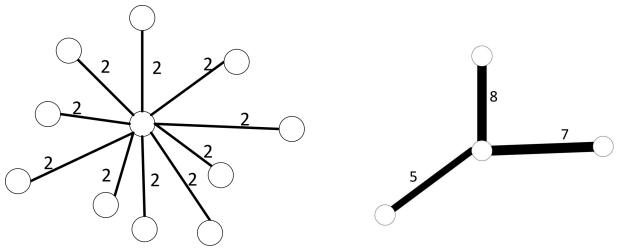
Definition 3: (**Degree**) The degree of a bike station s_m is the total number of stations that connected with s_m during time t_i. The measurement can be formalized as:

$$k_m^{t_i} = \sum_{n=1}^{N} x_{nm}^{t_i}$$

We define m as focal node and n as all other nodes. If during t_i, s_n connected with s_m, $x_{nm}^{t_i}$ is defined as 1. Otherwise it is 0. In our model, based on historical data, the degree of each station is the average degree. For instance, the average degree of s_m during t_i can be defined as:

$$\overline{k_m^{t_i}} = \sum_{r=1}^{R} k_m^{(r)t_i}/R$$

Definition 4: (**Flow Centrality**) Degree and flow can be both indicators of the level of involvement of a station in the surrounding bike-sharing network [21]. Because if a station connects with more stations, it will have higher chance to get flows. If stations have the same flows, degree still has influence on bike-sharing system. Fig.3(a) and Fig.3(b) show the stations have the same flows. But, the degree of Fig.3(a) is larger than Fig.3(b). Fig.3(a) and Fig.3(b) are two different patterns. Fig.3(a) is more stable than Fig.3(b). If there is an anomaly, the flow of Fig.3(b) will change more than Fig.3(a). Incorporating flow and degree are important when deciding flow centrality. In order to combine both degree and flow, we use a tuning parameter α, which determines the relative importance of the degree to flow. So the flow centrality is the product of degree that a focal station connects to and the average flow to these stations is adjusted by the tuning parameter. The formal definition is defined as:



(a) Degree=10, Flow=20 (b) Degree=3, Flow=20

Figure 3. Relation between Degree and Flow

$$fc_m^{t_i} = \overline{k_m^{t_i}} \times (\frac{f_m^{t_i}}{k_m^{t_i}})^{\alpha} = (\overline{k_m^{t_i}})^{1-\alpha} \times (f_m^{t_i})^{\alpha}$$

α is a positive number and can be decided by the research setting and data.

Definition 5: (**Distance**) The distance between s_n and s_m is defined as d_{nm}. The distance between each two stations is calculated by the lat/long information of the stations. The distance of s_m during time t_i of record r ($1 \leq r \leq R$) can be defined as:

$$d_m^{(r)t_i} = \sum_{n=1}^{M} d_{nm}^{(r)t_i}/M$$

M is the total number of stations that connected with station s_m during time period t_i of records r ($1 \leq r \leq R$).

The average distance of s_m during t_i can be defined as:

$$\overline{d_m^{t_i}} = \sum_{r=1}^{R} d_m^{(r)t_i}/R$$

Distance has an influence on flow prediction. For instance, if a user want to go to Walmart, the user will choose the one around him instead of riding a bike across New York City to find another one. In our model we use average distance for each station.

Definition 6: (**Attractiveness**) If a user would like to travel a longer distance to the end station s_m, that means s_m appeals to the user. If station s_m has large flows and degrees, it also means s_m is attractive to other stations. So attractiveness of a station can be decided by three factors: average distance, flow and degree. During t_i, the attractiveness of s_m is defined as:

$$at_m^{t_i} = (\overline{k_m^{t_i}})^{(1-\alpha)\times(1-\beta)} \times (f_m^{t_i})^{(\alpha)\times(1-\beta)} \times (\frac{\overline{d_m^{t_i}}}{f_m^{t_i}})^{\beta}$$

Based on POIs, we combine our attractiveness and multiple linear method together to form the prediction model:

$$at_m^{t_i} = b_0 + b_1 x_1 + b_2 x_2 + ... + b_{10} x_{10} \quad (1)$$

First step, we calculate all the POIs in our system. Second step, around each station s_n, we count the number of POIs. If the POI did not appear in the area, we assign the number of this type of POI 0. Otherwise, we assign it to its corresponding values. Third step, we use the equation (1) to calculate b_i ($0 \leq i \leq 10$) and then get the final results of the predicted flows.

V. Experiments

A. Datasets

We use the data of Citi Bike system[1], which was in NYC, from 1^{st} Apr. to 20^{th} Oct. in 2014 as presented in Table III. There are 5,855,763 records. The records contain: duration time, start/end time, station ID and station lat/long. The training data is from 1^{st} Apr. to 10^{th} Sep. The testing data is from 11^{th} to 30^{th} Sep. The validation data is from 1^{th} to 20^{th} Oct, which is used to get the tuning parameters. There are totally 344 stations in NYC. Fig.4 shows the details. For the bike data, the records of duration time more than 3 hours are considered as anomalous data and not taken into consideration, because few people will ride a bicycle for a trip more than three hours. The records with duration time more than 3 hours appear due to the reason that some people may go to the destination and not put the bike in a dock or they forget to swiping their cards. As a result, we can not know the purposes of these kind of trips. So before we use the dataset, we need to remove records which duration time is more than 3 hours.

[1]https://www.citibikenyc.com/system-data

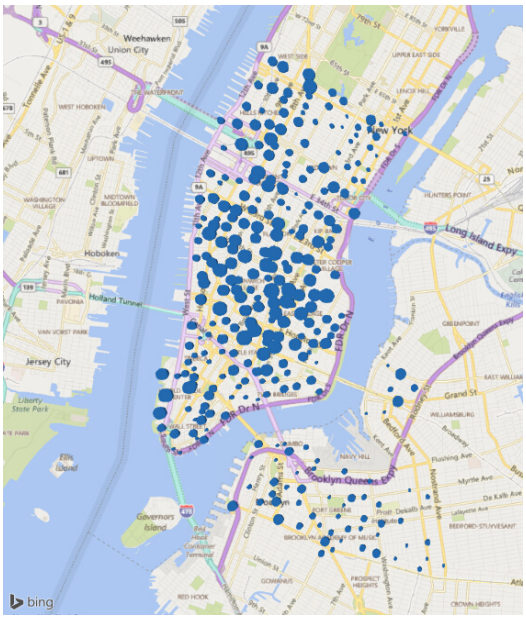


Figure 4. Locations of all Bicycle Docks in New York City, Each Circle Represents One Dock, the Size of the Area is Proportional to Total Check in Data on 8 am August.1.2014

Table III
DETAILS OF THE DATASETS IN 2014 NYC

Data Sources		NYC
Time Span		1^{st},Apr-20^{th},Oct
Bike Data	# Stations	344
	# Bikes	6,800+
	# Records	5,855,763

B. Baselines

The method of our work is to predict check-out and check-in data of NYC bike-sharing system. Because check-out and check-in data can be predicted with the same method, we just predicted the check-out flow in our experiments. The bike-sharing system will be affected by temperatures and weather conditions. We choose the sunny days and normal temperatures as our all hour data. Other weather types (e.g., cloudy, rainy, snowy) and extreme temperatures are considered as anomalous hours, which describes as follows:

1) Anomalous periods: An anomaly is a period which satisfies one of the following two conditions: 1) The entire bike flow in this time period is much different from the pattern; 2) The check-out/in across clusters in this time period is much different from the pattern, e.g., Fig.5 shows the entire bike flow in every hour from 15^{th}, Sep. to 24^{th}, Sep. in NYC. Points in rectangles are anomalous periods, because their entire bike flow deviates from the pattern significantly. We can also detect the second situation in a similar way. There are several factors that can cause anomalies, such as weather, temperature and so on. *1) Weather:* Anomalies will be caused by extreme weather types (e.g., cloudy, rainy, snowy). For instance, people prefer to ride a bike on sunny day instead of rainy, cloudy or snowy day. From New York Weather data set [2], (a record of weather data set contains basic temperature information, weather types (e.g., sunny, rainy, snowy and cloudy)), we use the data of 1^{st} Apr. to 20^{th} Oct. in 2014. We pick up all the sunny, cloudy, snowy and rainy days. Fig.6(a) to Fig.6(d) shows the average bike flows under different weather types. From Fig.6(a) to Fig.6(d) we can see that weather types (e.g., cloudy, rainy, snowy) will cause anomalies.

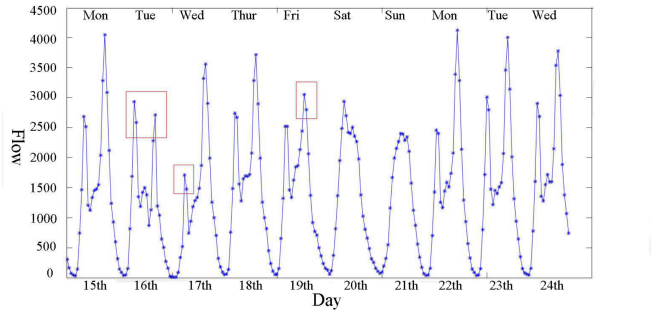
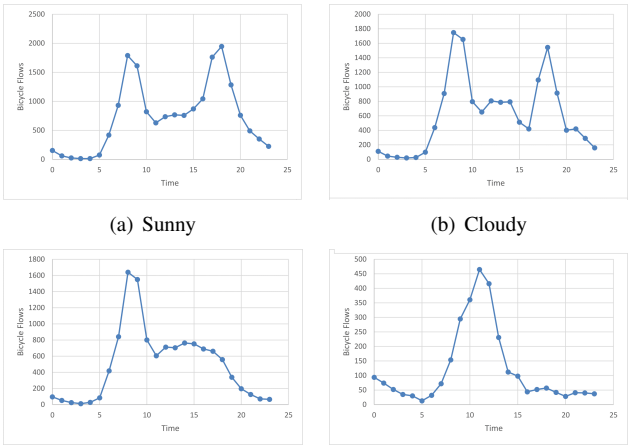


Figure 5. Anomanous Periods



(a) Sunny (b) Cloudy

(c) Rainy (d) Snowy

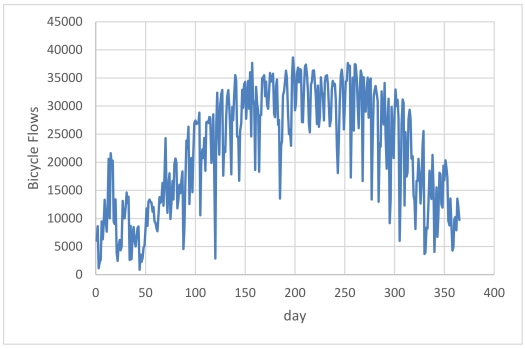
Figure 6. Influence of Weather Types

2) Temperature: If the temperature changes suddenly, it will cause anomalous as well. Less people will use the bike because of the sudden drop in temperature and it is freezing cold. Fig.7(a) and Fig.7(b) show the relation between bike flows and temperatures. We use the whole year of 2014 data set to show the relations. From Fig.7(a) and Fig.7(b) we can see that, if the temperature changed suddenly, the bike flow will change as well, it will cause the anomalies.

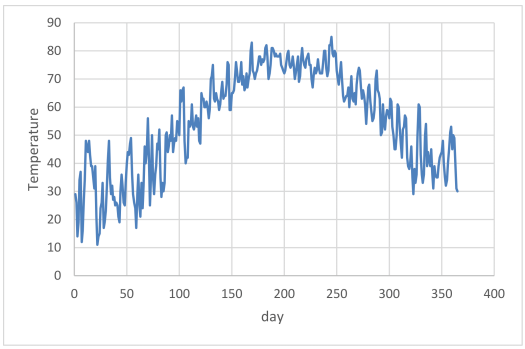
Because the anomalous periods in the testing data are small and the deviations are obvious, we can detect them

[2]http://www.weather.gov/okx/

manually. After finding the anomalous periods, we combine the anomalous periods to our attractiveness method and conduct the experiments.



(a) Bike flows of 2014 Year in New York



(b) Temperature of 2014 Year in New York

Figure 7. Relationship between Bike Flows and Temperatures in 2014.

Based on these features, we combine attractiveness and multiple linear regression method to form the prediction model. In order to confirm the efficiency and effectiveness of our model, we conduct experiments to compare our methods with six baselines.

FL: During each time period, check-out and check-in data relate to bike flows directly. So based on POIs, we combine flow and multiple linear regression model as a method to predict the check-out/in.

FLC: According to the definition of flow centrality, check-out/in also relate to degree factor. We form the flow centrality method, which combines degree and flow factor as a whole and uses feature based regression model to make a prediction. We use a tuning parameter to control the weight of these two factors.

HA: The check-out/in can be predicted by the average value of historical check-out/in in the corresponding periods, e.g., for 6:00 am-7:00 am on Monday, its corresponding periods are all historical time intervals from 6:00 am-7:00 am on weekdays.

ARMA: Similar to method of HA, in experiments with ARMA, we just differentiate the hour of the day and the day of the week as well, e.g., for 6:00 am-7:00 am on Monday, its corresponding periods are all the the historical time intervals from 6:00 am-7:00 am on Mondays. ARMA is a common tool for understanding and predicting future values in a time series.

ANN: We use Liu's [18] work, as another baseline. In Liu's work, they extracted 10 features from the station network, bicycle trajectories, taxi trajectories and POIs for bike flow prediction problem. The 10 features include: walking distance from each bike to its nearest parking lot (PL), the walking distance to the nearest subway entrance (SE), the taxi pick-up densities (TP), the number of faster bicycle routes (FR), number of entertainment (PO_EN), number of restaurants (PO_R), number of shopping centers (PO_S), number of educations (PO_E), total number of docks of a station (ND), and nearby station score (NSS). After they extracted features, they proposed an artificial neural network (ANN) to predict the bike flows of stations.

Clustering: We implement the station clustering method of Li [15]. We combine our attractiveness method and bipartite clustering algorithm together to do the experiment. By this way, we can compare the result with Li's work. Li's work can only predict the flow of each cluster. For each station, it can not work.

In our experiments, we cluster all the stations of the bikesharing system in NYC into 23 clusters respectively, by using the bipartite clustering algorithm. As shown in Fig.8, points with the same color denote stations pertaining to the same cluster. After clustering the stations into 23 clusters, we combine clusters with our attractiveness method and calculate the error of each cluster.

C. Metric

We use Error Rate*(ER)* and Root Mean Squared Logarithmic Error*(RMLSE)* [20] as metrics to measure the results, which are formally defined as follows:

$$ER(t) = \frac{\sum_{i=1}^{m} |\overline{X_{i,t}} - X_{i,t}|}{\sum_{i=1}^{m} X_{i,t}}$$

$$RMLSE(t) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(log(\overline{X_{i,t}}+1) - log(X_{i,t}+1))^2}$$

where, $X_{i,t}$ is the ground truth of the check-out/in of station i during time period t while $\overline{X_{i,t}}$ is the corresponding prediction value, m is the total number of stations.

D. Result

Results of tuning parameters: Fig.9(a) to Fig.9(d) show the relation of tuning parameters with Error Rate (ER) and Root Mean Squared Logarithmic Error (RMLSE). To make our model work well, we need to choose proper tuning parameter α and β. Degree and flow are two important factors to decide the flow centrality. We use tuning parameter α to indicate the weight of degree and (1-α) represents

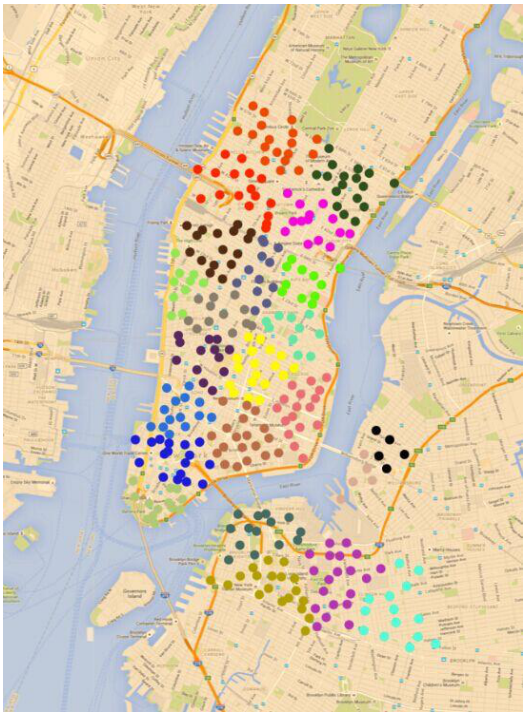
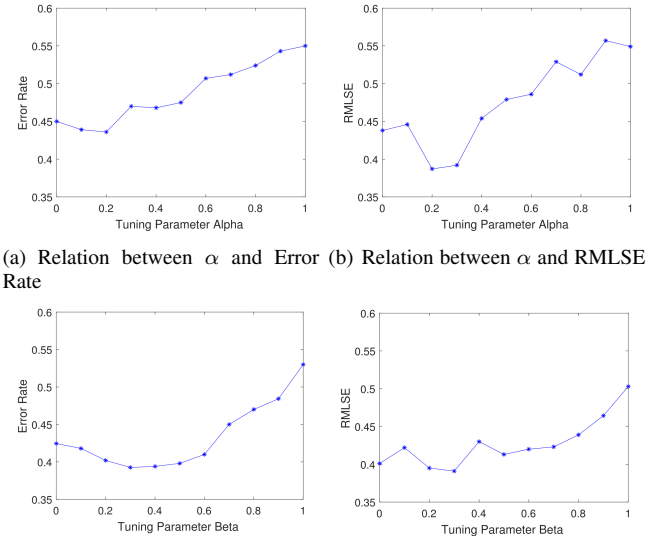


Figure 8. Bipartite Stations Clustering



(a) Relation between α and Error (b) Relation between α and RMLSE Rate

(c) Relation between β and Error (d) Relation between β and RMLSE Rate

Figure 9. Relation between Tuning Parameters and Error

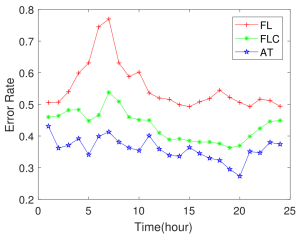
the weight of flow. From Fig.9(a) we can see that, when α=0.2, the ER has the smallest value 0.42. Fig.9(b) shows the relation between tuning parameter α and RMLSE. From the figure, we can find that, when α is 0.2, it can achieve the smallest RMLSE value, which is 0.413. So in our experiment we choose α=0.2. Based on flow centrality we get the attractiveness model. In attractiveness model, distance is an important factor, so we import tuning parameter β to illustrate the relationship between distance and flow centrality. Because we get the optimal tuning parameter α in flow centrality, in this step we just assign α=0.2. Fig.9(c) shows the relation between β and ER. We can see that when β=0.3, the ER has the smallest value 0.39. We also analyze the relation between β and RMLSE, which shows in Fig.9(d). From the figure, we can also find that, when we choose β as 0.3, the curve can get the smallest value. So in our attractiveness model, we choose tuning parameter α=0.2 and tuning parameter β=0.3. The final ER of our attractiveness model is 0.39, and the RMLSE is 0.385.

Results of baseline: Fig.10(a) and Fig.10(b) show the ER and RMLSE of FL, FLC and AT. From the results we can see that, for all time periods, both of the ER and the RMLSE obtained from our proposed method are much lower than all the baselines with a significant margin. The ER and RMLSE of FL is larger than FLC in each hour, and the average ER of FL is 0.55, and average RMLSE of FL is 0.543. The average ER of FLC is 0.43, and average RMLSE of FLC is 0.418. Both of the ER and RMLSE of FLC is smaller than FL, which means bike flow not only relates to flow, it also relates to degree. Our attractiveness model(AT) can get the smallest average ER of 0.35, and average RMLSE of 0.359. Because distance will affect the bike flow, and by choosing the proper tuning parameters for degree and distance, our model can achieve a better result than FL and FLC.

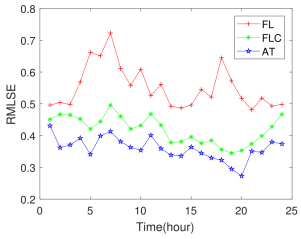
Fig.10(c) and Fig.10(d) show the performance comparison of ARMA, HA, ANN, and AT. As can be seen, our proposed method AT lower the ER and RMLSE of different time periods, which means our method is better than these three baselines. We can also find that, the ER and RMLSE of ARMA and HA are very close. However, in most situations, ARMA has a smaller ER and RMLSE than HA. Because it is more accurate if we predict the bike flow of Monday by using the historical data of Mondays than weekdays. Among the three baselines, ANN model performs better than ARMA and HA, because neither of ARMA and HA models have taken any external features. They only use the historical data to make a prediction, which is simple and not accuracy. In contrast, ANN model incorporates the context features from the station network, bicycle trajectories, taxi trajectories and POIs, which are related to bike flows. However, our AT model is still better than ANN model, because we incorporate more detailed POI context features than ANN, and we also incorporate network context, which is very important in bike flow prediction system.

In summary, compared to ANN model, which achieved the best performance of the baselines, our AT model can reduce average ER and RMLSE by 0.088 and 0.82 respectively for all the predicted hours. The detailed average ER and RMLSE values of these methods are shown in Table IV.

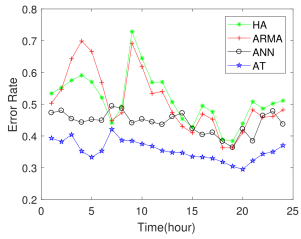
Results of clustering: Table V shows the average ER and



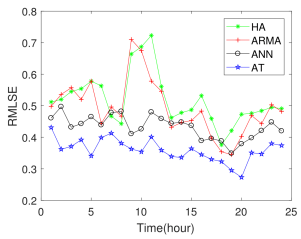
(a) Comparison of FL,FLC and AT with Error Rate



(b) Comparison of FL,FLC and AT with RMLSE



(c) Comparison of ARMA, HA, ANN and AT with Error Rate



(d) Comparison of ARMA, HA, ANN and AT with RMLSE

Figure 10. Results of baselines

Table IV
PREDICTION ERROR OF BASELINES

Method	All Hours	Anomalous Hours	All Hours	Anomalous Hours
	ER		RMLSE	
HA	0.514	1.641	0.519	1.731
ARMA	0.503	1.743	0.492	1.862
ANN	0.442	1.216	0.433	1.139
AT	**0.354**	**0.582**	**0.351**	**0.573**

RMLSE of bipartite cluster based hierarchical model(HP-MSI), which is the result of Li's work, and the bipartite cluster based attractiveness model(AT-BC), which proposed by us. By comparing the results of these two models, we can see that our model can reduce ER by 0.021, and reduce RMLSE by 0.043 for all hours, which is significant. Because the bike-sharing system is sensitive, it can fluctuate obviously. In Li's work, they only consider the time and weather conditions, however, there are more factors that will influence the bike flow, such as surrounding POIs, which is used in our model. Because people always have a purpose for a trip, for instance most people will go to work during the morning hours from 7 am to 9 am, that means more bike flows will head to work places during that time. Moreover, for the geography information, Li's work just simply considers the closest bike station situation, however, network context is also a key factor for bike flow prediction as we explained in the method part. From the results we can also see that, AT-BC has a lower ER and RMLSE than AT. The reason is that, AT is used to predict the flow of each station. However, AT-BC is used to predict the flow of a cluster, which usually contains multiple stations. Based on Li's work, we know that, the flow of a cluster level is much more stable than a single station. The larger size of the cluster, the more stable on the prediction of flow.

Table V
PREDICTION ERROR OF BIKESHARING SYSTEM

Method	All Hours	Anomalous Hours	All Hours	Anomalous Hours
	ER		RMLSE	
HP-MSI	0.282	0.503	0.291	0.535
AT-BC	**0.261**	**0.274**	**0.248**	**0.287**

Results of anomalies: Table IV shows AT has a ER value of 0.582 and RMLSE value of 0.573, which are much lower than HA, ARMA, and ANN under anomalous hours. It means AT can achieve a better result than HA, ARMA, and ANN under anomalous hours. Table V shows the average prediction error of cluster methods under anomalous hours. We can see that the ER and RMLSE of our AT-BC model is 0.274 and 0.287. Compared with Li's work, our AT-BC model can reduce ER by 0.229 and reduce RMLSE by 0.248, which means under the anomalous hours, AT-BC can predict bike flow much better than HP-MSI. The results from both Table IV and Table V indicate that our proposed models can achieve better results than baselines under anomalous hours, because our models consider more important factors, which will influence the bike flow prediction. With these key factors, our models can be more stable than baselines under anomalous situations. Table IV and Table V also show that, all hours can have a lower ER and RMLSE than anomalous hours both in baselines and our works. The reason is that, with anomalous hours, it is much more difficult to predict bike flows and the number of the flows can change randomly under different situations.

VI. CONCLUSION

In this paper, we focused on the prediction of bike flows in New York City bike-sharing system. First we analyzed the influence of POI, distance, degree and time on the bike-sharing system, we used Google Map API to get the POIs around each bike station in NYC, and we categorized the POIs into 10 groups. After that, we proposed our POI based attractiveness model to predict check-out/in of each station in a bike-sharing system. We evaluated our model on NYC dataset, the performance of our model significantly beyond the state-of-the-art neural network based method using context features (ER is reduced by 9% for all hours and RMLSE is reduced by 8% for all hours). We implemented the former work of bipartite station cluster algorithm, and combined our model with it. The performance of attractiveness based bipartite station cluster model is better than the former work(ER is reduced by 2% for all hours and 23% for anomalous hours, RMLSE is reduced by 4% for all hours and 24.8% for anomalous hours). The result shows our model is effective and efficient. In the future, we can change the groups of POIs and make them more suitable to our model.

REFERENCES

[1] M. Benchimol, P. Benchimol, B. Chappert, A. De La Taille, F. Laroche, F. Meunier, and L. Robinet. Balancing the stations of a self service bike hire system. *RAIRO-Operations Research*, 45(01):37–61, 2011.

[2] P. Borgnat, P. Abry, P. Flandrin, C. Robardet, J.-B. Rouquier, and E. Fleury. Shared bicycles in a city: A signal processing and data analysis perspective. *Advances in Complex Systems*, 14(03):415–438, 2011.

[3] P. Borgnat, C. Robardet, P. Abry, P. Flandrin, J.-B. Rouquier, and N. Tremblay. A dynamical network view of lyons vélov shared bicycle system. In *Dynamics On and Of Complex Networks, Volume 2*, pages 267–284. Springer, 2013.

[4] D. Chemla, F. Meunier, and R. Wolfler-Calvo. Balancing a bike-sharing system with multiple vehicles. In *Proceedings of Congress annual de la société Française de recherche opérationelle et daidea la décision, ROADEF2011, Saint-Etienne, France*, 2011.

[5] L. Chen, X. Ma, G. Pan, J. Jakubowicz, et al. Understanding bike trip patterns leveraging bike sharing system open data. *Frontiers of computer science*, 11(1):38–48, 2017.

[6] L. Chen, D. Zhang, L. Wang, D. Yang, X. Ma, S. Li, Z. Wu, G. Pan, T.-M.-T. Nguyen, and J. Jakubowicz. Dynamic cluster-based over-demand prediction in bike sharing systems. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 841–852. ACM, 2016.

[7] C. Coffey and A. Pozdnoukhov. Temporal decomposition and semantic enrichment of mobility flows. In *Proceedings of the 6th ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, pages 34–43. ACM, 2013.

[8] A. K. Datta. Predicting bike-share usage patterns with machine learning. 2014.

[9] P. DeMaio. Bike-sharing: History, impacts, models of provision, and future. *Journal of Public Transportation*, 12(4):3, 2009.

[10] C. Etienne and O. Latifa. Model-based count series clustering for bike sharing system usage mining. *ACM Trans. Intell. Syst. Technol*, 5(3):39, 2012.

[11] C. Fricker and N. Gast. Incentives and regulations in bike-sharing systems with stations of finite capacity. *arXiv preprint arXiv:1201.1178*, page 2, 2012.

[12] J. Froehlich, J. Neumann, and N. Oliver. Sensing and predicting the pulse of the city through shared bicycling. In *IJCAI*, volume 9, pages 1420–1426, 2009.

[13] T. Hamilton and C. J. Wichman. Bicycle infrastructure and traffic congestion: Evidence from dc's capital bikeshare. *Resources for the Future Discussion Paper*, pages 15–20, 2015.

[14] 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*, 6(4):455–466, 2010.

[15] Y. Li, Y. Zheng, H. Zhang, and L. Chen. Traffic prediction in a bike-sharing system. 2015.

[16] J.-R. Lin and T.-H. Yang. Strategic design of public bicycle sharing systems with service level constraints. *Transportation research part E: logistics and transportation review*, 47(2):284–294, 2011.

[17] J.-R. Lin, T.-H. Yang, and Y.-C. Chang. A hub location inventory model for bicycle sharing system design: Formulation and solution. *Computers & Industrial Engineering*, 65(1):77–86, 2013.

[18] 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*, pages 883–888. IEEE, 2015.

[19] J. Liu, L. Sun, W. Chen, and H. Xiong. Rebalancing bike sharing systems: A multi-source data smart optimization. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1005–1014. ACM, 2016.

[20] J. Liu, L. Sun, Q. Li, J. Ming, Y. Liu, and H. Xiong. Functional zone based hierarchical demand prediction for bike system expansion. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 957–966. ACM, 2017.

[21] T. Opsahl, F. Agneessens, and J. Skvoretz. Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, 32(3):245–251, 2010.

[22] H. Sayarshad, S. Tavassoli, and F. Zhao. A multi-periodic optimization formulation for bike planning and bike utilization. *Applied Mathematical Modelling*, 36(10):4944–4951, 2012.

[23] S. Shaheen, S. Guzman, and H. Zhang. Bikesharing in europe, the americas, and asia: past, present, and future. *Transportation Research Record: Journal of the Transportation Research Board*, (2143):159–167, 2010.

[24] C. Tian, Y. Huang, Z. Liu, F. Bastani, and R. Jin. Noah: a dynamic ridesharing system. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, pages 985–988. ACM, 2013.

[25] P. Vogel, T. Greiser, and D. C. Mattfeld. Understanding bike-sharing systems using data mining: Exploring activity patterns. *Procedia-Social and Behavioral Sciences*, 20:514–523, 2011.

[26] P. Vogel and D. C. Mattfeld. Strategic and operational planning of bike-sharing systems by data mining–a case study. In *Computational Logistics*, pages 127–141. Springer, 2011.

[27] J. W. Yoon, F. Pinelli, and F. Calabrese. Cityride: a predictive bike sharing journey advisor. In *Mobile Data Management (MDM), 2012 IEEE 13th International Conference on*, pages 306–311. IEEE, 2012.