Interactive Visual Analytics for Multi-City Bikeshare Data Analysis

Alicia Bargar, Amrita Gupta [*] , Srishti Gupta, Ding Ma
Georgia Institute of Technology
Atlanta, GA, USA
{abargar3, agupta375, srishtigupta, mading}@gatech.edu

ABSTRACT

Bicycle sharing schemes are gaining traction as an alternative or complementary mode of urban transport. In this paper, we explore the utility of an interactive web-based visual analytics application for comparing usage patterns between different bike sharing programs. We demonstrate the potential for adjustable filters on user demographics and trip characteristics to reveal differences in ridership between cities. We also perform clique-detection and Louvain modularity-based community detection to reveal areas of high connectedness under different contexts. Our work utilizes the ST-DBSCAN algorithm in a novel context to cluster trips as a means of categorizing flow patterns. Finally, using publicly available data from bike share organizations, we conduct some experiments combining the data filters, algorithms and visualization. This preliminary work showcases the value of interactive visual analytics for highlighting notable differences between established bike share systems that may help frame questions for future research or policymaking.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data Mining

Keywords

bike share, community detection, visualization, visual analytics

1. INTRODUCTION

Bike share schemes are an increasingly prevalent mode of intra-city transportation. In April 2013 there were 26 bike-share programs in the US [11], a number expected to double in the subsequent 2 years as many cities are investigating this option for a number of reasons. Bike sharing can provide an alternative to traditional modes of transport or, more likely, a complementary service for solving the "last mile problem" of getting from a public transportation stop to the final destination. Furthermore, bike share systems may help mitigate automobile congestion and reduce pollution, although relatively little research has been done to asses their actual impact in these areas. Benefits to users include potentially reduced commute times by perhaps as much as 10% [9]; and a healthier lifestyle–one recent study investigating the effect of the London cycle hire scheme on the health of over half a million of its users reported a measurable reduction in disability adjusted life years, particularly for male and older participants [17]. These potential benefits have contributed to the recent rise in bike share feasibility studies and policies to promote cycling in the USA.

To ensure the success of such ventures, it is useful to study the operation of existing public bicycle programs to identify features that improve the effectiveness of bike share scheme implementations. For instance, a comparison of bike share schemes in China revealed that government-led investment, enforced bicycle lanes and technologically sophisticated equipment greatly boosted the performance of bike share schemes [18]. Making public bicycles available to non-registered users can increase the number of trips taken and introduce new flow patterns between docking stations [12] compared to a system reserved for use by subscribers only. Studies also suggest numerous station-specific variables (neighboring land use, population density, proximity to transit, distance from the CBD, bike lanes and bicycle infrastructure [6, 7, 15], among others) and city-wide variables (station density, weather, demographics, attitudes towards cycling) that affect bikeshare ridership.

Although bike share usage patterns and trends necessarily vary between cities, there are relatively few comparative studies in the literature. However, jointly analyzing different bikeshare programs may bring to light valuable insights that can be used to improve or expand existing services or to shape planning decisions for a new scheme [19]. Furthermore, the analysis in bikeshare network studies is not accessible to non-technical persons. This paper aims to address these needs through a web-based interactive visual analytics application for simultaneously exploring bikeshare data from multiple US cities. Specifically, we focus on detecting sub-communities in networks from usage patterns subject to filters like date, time, trip duration, and user data. We use maximum clique detection to identify the most interconnected parts of the bikeshare network, as well as modularity-based community detection to find sub-communities. Additionally, our use of ST-DBSCAN to cluster similar trips is a

[*] Corresponding author. Tel: +14108071155
Email address: agupta375@gatech.edu

UrbComp'14 August 24, 2014, New York, NY, USA

novel application of this algorithm. Urban planners may be able to leverage such a tool to improve cycling infrastructure or identify target regions or demographics for expanding a bikeshare system.

2. RELATED WORK

Cycling is seen as a cost-effective, eco-friendly and healthy mode of urban transport, and as such numerous research efforts have sought to determine factors that may promote its adoption by urban populations. Much of this work was conducted through surveys, e.g. Dill and Carr used census data across 50 US cities to statistically analyze correlations between number of bicycle commuters and city density or average population age [4]. With the advent of modern bicycle sharing systems, however, automated data collection at docking stations has made possible the application of quantitative assessments of bicycle-based transport. Earlier studies on bike share systems were largely concerned with characterizing system behavior by extracting spatiotemporal activity patterns from station occupancy data. Many of these studies began by clustering docking stations into groups based on their temporal occupancy profiles. Froehlich, Neumann and Oliver were some of the first to analyze bike share usage and sought to use their results to infer the underlying human mobility patterns of the city [8]. Recognizing the potential of such research to improve the performance of existing bike share programs, Kaltenbrunner et al. conducted a similar study to use the mined activity cycles of stations to produce short-term predictions of station occupancy [10]. Vogel et al. also clustered stations by pickup and return activity profiles and attempted to account for the behavior of the resulting clusters by examining the stations' surroundings [14]. Indeed, a number of research efforts have attempted to determine factors that affect docking station activity ([15, 7]), revealing complex dependencies on station density, population density, proximity to public transport, altitude, neighboring businesses and job density.

In contrast to these studies, which focus on a single bicycle sharing system, O'Brien et al. analyzed 38 systems from around the world to develop a classification of bike share systems based on temporal usage patterns [13]. This work also included a hierarchical clustering of global bike shares based on system properties (e.g. number of stations, system compactness ratio, number of weekday and weekend usage peaks, etc.). This study highlighted the value of drawing comparisons between bike sharing systems to gain insights into the effects of city- and system-specific parameters.

The literature discussed so far has mostly conducted analyses using station occupancy data. However, this precludes the examination of links between specific pairs of stations, which could afford more precise measurements of bicycle flows. Borgnat et al. were able to acquire trip-level data and conducted a broad exploratory study of Lyon's bike share system features [5]. On a system-wide scale, they modeled the effects of the program's popularity, weather, holidays and day of the week on the overall number of rentals. In addition, the availability of trip data enabled them to perform a hierarchical modularity-based community detection and K-means clustering to group edges based on their weights at certain high-activity times. The availability of trip data, therefore, invites the use of graph theoretic algorithms to explore network properties and to broaden the notion of "groups" of stations. In our work, we have implemented clique-detection and community detection for this purpose, as well as a novel application of ST-DBSCAN [3] to cluster similar trips.

In recent years, a number of bike sharing organizations have made station and trip data publicly available as part of data visualization challenges. This demonstrates a growing demand from bike sharing providers for lucid presentations of data analysis results, a problem which has received relatively little attention thus far from the information visualization and visual analytics communities. Even simply visualizing trips between stations can be challenging, since the large number of possible links between nodes and the non-uniform geographical distribution of stations can cause visual clutter and misleading impressions due to visual salience effects. Wood et al. have developed an interactive visualization of trips taken in the London bicycle sharing system [16] that attempts to minimize the consequences of visual clutter caused by displaying trips as curved lines drawn in order of increasing frequency, so that links obscured by overlap are the least common. The authors have recently used the visualization to examine specific questions about the usage of the London bikeshare system, such as journeys made by commuters [2] or by male versus female subscribers [1]. Hence it is clear that interactivity and flexible querying can facilitate the exploration of a variety of usage-related questions.

Our work contributes to this body of literature by bringing together numerous aspects of these studies to create a visual analytics tool targeted at bikeshare scheme managers and researchers alike. It is our belief that an interactive application offering a variety of data filters will allow users of the interface to quickly ask questions of interest. We also augment the visual presentation of the filterd data with the option to perform community-detection, clique-detection and trip clustering. Lastly, the application has a side-by-side layout to facilitate comparisons between multiple cities.

3. DATA ACQUISITION

A number of bicycle sharing organizations have made some of their recorded system data publicly available as part of data visualization competitions. We obtained station and trip data from three such organizations: Hubway based in Boston, Capital Bikeshare in Washington DC, and Divvy from Chicago. The data comprised station information and trip logs. The former contained station ID's, names, coordinates and capacities while the logs recorded trip origin and destination stations, bike checkout and return timestamps, trip duration and user information. In Chicago, registered users accounted for 53% of trips in Chicago, 64% of recorded trips in Boston and 80% in Washington DC. The data were cleaned and entered into SQLite databases.

4. ALGORITHMS

With filters in place for date, time of day, trip duration, user type, age and gender, it is possible to select a subset of the trip data and build a directed graph or network, where the nodes are the docking stations, edges are trips between stations, and edge weights correspond to the number of trips that took place between pairs of stations. Subsequently, we are able to apply various graph theoretic algorithms to this network to investigate its connectivity properties.

4.1 Maximal Clique Detection

Maximal clique detection was chosen to determine the largest interconnected part of the network given some subset of the trip data. The weighted directed graph structure is converted into a weighted undirected graph and edges whose weight falls below a minimum traffic threshold are disregarded in order to focus on high-frequency connections. We then find the maximum maximal clique for the high traffic graph. This computation may be useful for identifying a core group of stations serving a tight-knit subcommunity or a self-contained region within the city.

4.2 Louvain Modularity Optimization

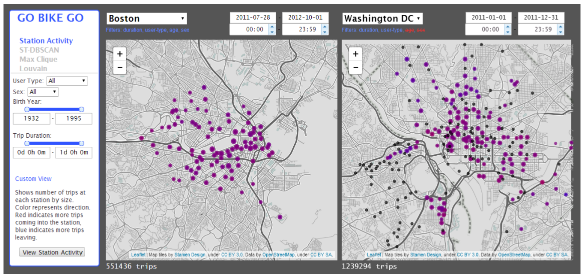
A greedy modularity optimization method (the Louvain algorithm [4]) was used to perform community detection, in order to find groups of stations that do not necessarily form perfect cliques but are still highly connected. First, the Louvain method looks for "small" communities by optimizing modularity locally. Then it aggregates nodes belonging to the same community and builds a new network whose nodes are the communities. These steps are repeated iteratively until a maximum of modularity is attained and a hierarchy of communities is produced. This approach was selected in order to find groups of stations that are more connected than average, possibly signifying the presence of hidden subcommunities.

4.3 ST-DBSCAN

We chose ST-DBSCAN [3] to cluster similar trips due to its ability to incorporate temporal and other non-spatial features of data into Density-Based Spatial Clustering. The core of the DBSCAN algorithm is to define density using neighbors. In our implementation, ST-DBSCAN utilizes both spatial (geo-location of trips) and temporal (start and end time of trips) information to find "similar" trips before performing clustering. The purpose of this approach is to extract coherent flow patterns by grouping trips between one set of neighboring stations and another, occurring at roughly the same time of day.

5. VISUALIZATION

We created a visual tool for comparing patterns in bike usage across different cities. To achieve this, we designed our application to display two maps side-by-side with a variety of filters allowing the user to adjust which aspects of the programs they wish to compare. Stations are designated by circles whose size is proportional to the number of journeys that start or end at that station, relative to the total number of trips shown. The color of the circle encodes the ratio of incoming to outgoing trips, with "sink" type stations colored more red and "source" type stations colored more blue. Clicking on a station reveals data and edges starting and ending there. We have separate panels for our three algorithms and a station activity viewer panel.



GO BIKE GO

Boston: 2011-07-28 00:00 - 2012-10-01 23:59

Filters: duration, user-type, age, sex

Washington DC: 2011-01-01 00:00 - 2011-12-31 23:59

Filters: duration, user-type, ~~age~~, ~~sex~~

User Type: All

Sex: All

Birth Year: 1932 - 1995

Trip Duration: 0d 0h 0m - 1d 0h 0m

Shows number of trips at each station by size. Color represents direction. Red indicates more trips coming into the station, blue indicates more trips leaving.

551436 trips

1239294 trips

Station Activity
ST-DBSCAN
Max Clique
Louvain
Custom View
View Station Activity
+
−
Leaflet | Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under CC BY SA.

Figure 1: Visualization tool. Left column contains tabs for switching between visualization components and filters.

The maps and station markers are implemented in Leaflet, an open-source Javascript library for map creation. Individual map tiles are provided by Stamen Design (free for non-commercial users). D3's scaling features are used to determine color and width of individual paths. The filtering widgets and logic are achieved through a mixture of JQuery and JQuery UI. Specifically, JQuery UI provides a calendar widget and range scaler that aids the user in selecting date and time intervals. Once the filtering parameters have been selected through the widgets, they are sent via JQuery AJAX calls to the server where they are passed to the relevant Python scripts. The results are sent back in JSON format to the client for visualization.

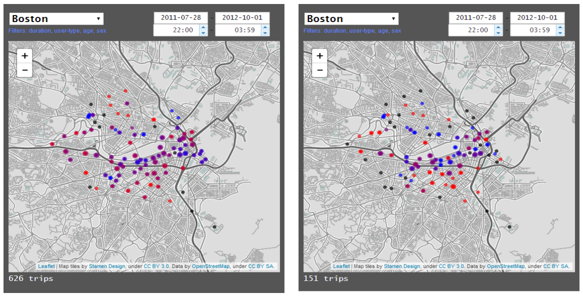
6. EXPERIMENTS AND DISCUSSION

When designing our visual tool, we decided to focus on providing visually simplistic representations of our algorithms with minimal interpretation of the results. Our intention is to allow people with deeper understanding of the cities we consider to employ our methods as a tool while attempting to limit potential for misinterpretation. To this end, our visualizations are based in primary shapes, colors, and sizes, with text limited to station details available on mouse click. The four components of our tool has a separate visualization: one for each algorithm and one for exploratory analysis, in keeping with the focus of the individual component. Each visualization is cleared before applying a visualization corresponding to a different component. We tested our visualizations on data provided by bike sharing systems in Boston, Washington DC, and Chicago.

6.1 Exploratory Analysis: late night riders in Boston

We developed an exploratory analysis component to our tool to give users the ability to find or consider patterns in the data using our filtering and visualization capabilities. In this component, a circle represents each station on the map. Its radius corresponds to the number of trips at this station. We use logarithmic scaling over the total number of trips to determine the exact radius size. The color indicates the ratio of incoming vs outgoing trips per station as a gradient. A station with all incoming trips will be red, whereas a station with all outgoing trips will be blue. Black is assigned to all trips without stations.

Here we examine trips made late at night using Boston's bike sharing program to demonstrate how the ability to apply filters to the data can help identify system usage trends among subpopulations. This particular case may be of real significance to program directors looking to promote the sys-

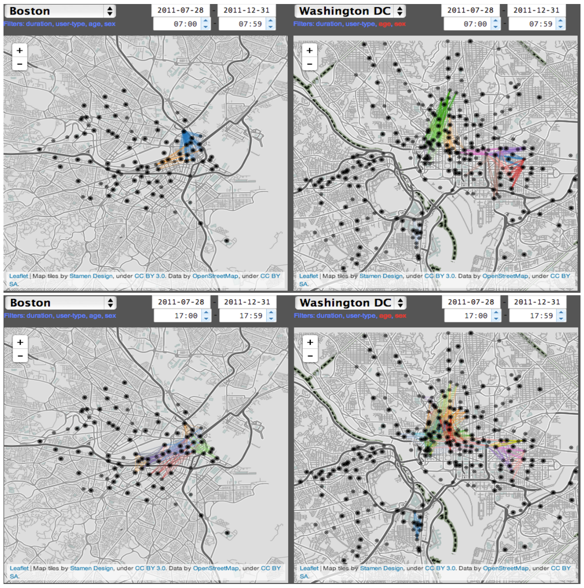


Boston
Filters: duration, user-type, age, sex
2011-07-28 - 2012-10-01
22:00 - 03:59
626 trips

Boston
Filters: duration, user-type, age, sex
2011-07-28 - 2012-10-01
22:00 - 03:59
151 trips

+
−
Leaflet | Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under CC BY SA.

Figure 2: Trips taken by male (left) and female (right) cyclists in Boston between the hours of 10pm and 4am.



Boston
Filters: duration, user-type, age, sex
2011-07-28 - 2011-12-31
07:00 - 07:59

Washington DC
Filters: duration, user-type, age, sex
2011-07-28 - 2011-12-31
07:00 - 07:59

Boston
Filters: duration, user-type, age, sex
2011-07-28 - 2011-12-31
17:00 - 17:59

Washington DC
Filters: duration, user-type, age, sex
2011-07-28 - 2011-12-31
17:00 - 17:59

+
−
Leaflet | Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under CC BY SA.

Figure 3: ST-DBSCAN applied to Boston and Washington DC trips between 7 and 8 am (top) and 5 and 6pm (bottom).

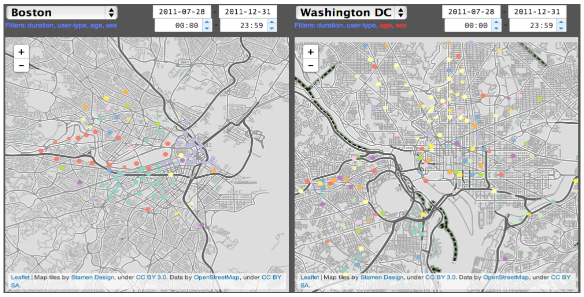
tem as a safe, reliable option for late-night travel. First, the time filters are used to select trips made between 10pm and 4am. Next, the user filter is applied to select for trips taken by registered users, for whom demographic information is available. Finally, we can specify "male" or "female" in the sex filter to determine whether there are differences in usage patterns, potentially due to safety concerns late at night.

Figure 2 shows that female cyclists have more unused stations (black circles) at this time and a more distinctive flow pattern compared to male cyclists (more red and blue nodes, fewer mixed nodes). Female riders also take far fewer trips: 151 trips compared to the 626 trips taken by male cyclists. Identifying why the male and female cycling patterns differ so extensively may be of interest to program directors looking to increase their system's accessibility to female riders.

6.2 ST-DBSCAN

Figure 3 shows the outcome of ST-DBSCAN's trip clustering applied to Boston and Washington DC trips in the morning and the early evening. The resulting color-coded trip clusters allow the user to identify patterns in trip flows and to make sense of different 'types' of trips that occur. In both cities, there is a greater variety of flows in the evening compared to the morning. Further, flows associated with a specific time of day are easily distinguishable, such as the flows between Boston's North Station and the financial district (blue cluster at 7-8am) or between downtown Washington DC and the residential neighborhoods to the north (green cluster at 7-8am).

6.3 Louvain



Boston
Filters: duration, user-type, age, sex
2011-07-28 - 2011-12-31
00:00 - 23:59

Washington DC
Filters: duration, user-type, age, sex
2011-07-28 - 2011-12-31
00:00 - 23:59

+
−
Leaflet | Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under CC BY SA.

Figure 4: Results of Louvain algorithm on trips in Boston and Washington DC.

We represent station clusters provided by Louvain with matching node color indicating common cluster membership. This algorithm allows users to easily see the geographical spread of stations within a highly-connected community. Figure 4 illustrates the results of Louvain applied to all trips taken in Boston and Washington DC. In this example, we note that the Washington DC clusters are far more spread out than they are in Boston, suggesting differences in cycling infrastructure or layout between the two cities that is not otherwise immediately clear from the data alone.

6.4 Max Clique

We distinguish the stations belonging to the maximum clique by coloring them a bright coral and leaving the other stations a pale gray color. Max clique helps us distinguish the areas of maximum traffic at any given time, with the map overlay providing us with additional contextual information. For example, our application of max clique to Boston and Washington DC for the hours 7-8am and 5-6 pm indicate differing areas of maximum trip activity for the morning and evening commutes (Figure 5). We note that there appears to be a greater number and diversity of stations included in the evening cliques, which indicates either a growing number of users and/or common routes for the evening rush hour.

7. CONCLUSIONS

We aimed to provide an accessible interface for examining bike share programs across cities through the use of various algorithms for clustering data, namely ST-DBSCAN, Louvain modularity optimization and max clique detection. By coupling these back-end algorithms with visualizations on a city map overlay, their results become accessible to potential users seeking to identify system usage patterns. Furthermore, by including a variety of filtering capabilities for exploratory visual analysis, users of this tool can narrow the data they examine to answer more specific questions or discover behaviors unique to certain time periods or subpopulations.

We believe that this marks a first step towards developing a tool that provides visual data analytics for bike share

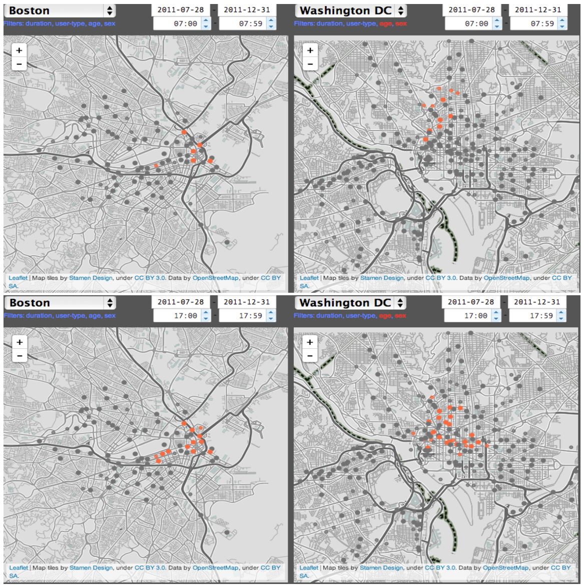

Boston
2011-07-28
2011-12-31
Filters: duration, user-type, age, sex
07:00
07:59
Washington DC
2011-07-28
2011-12-31
Filters: duration, user-type, age, sex
07:00
07:59
Boston
2011-07-28
2011-12-31
Filters: duration, user-type, age, sex
17:00
17:59
Washington DC
2011-07-28
2011-12-31
Filters: duration, user-type, age, sex
17:00
17:59
+
−
Leaflet | Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under CC BY SA.

Figure 5: Max cliques among Boston and Washington DC stations for 7-8 am (top) and 5-6pm (bottom).

programs. A user study is the next necessary step to improve our designs and test the extent to which the current functionality enables users to explore specific questions of interest. We also hope to develop our understanding of what information is most pertinent for bike share program owners by increasing discussions with these potential users. Beyond design, we believe that continuing to test and compare the potential benefit of other algorithms for this dataset will aid our understanding of what kinds of analyses best facilitate our understanding of the data.

8. ACKNOWLEDGMENTS

The authors would like to thank Duen Horng (Polo) Chau for his guidance throughout the development of this work.

9. REFERENCES

[1] R. Beecham and J. Wood. Exploring gendered cycling behaviours within a large-scale behavioural data-set. *Transportation Planning and Technology*, (ahead-of-print):1–15, 2013.

[2] R. Beecham, J. Wood, and A. Bowerman. Studying commuting behaviours using collaborative visual analytics. *Computers, Environment and Urban Systems*, 2013.

[3] D. Birant and A. Kut. ST-DBSCAN: An algorithm for clustering spatial–temporal data. *Data & Knowledge Engineering*, 60(1):208–221, 2007.

[4] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[5] 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(3):1–24, June 2011.

[6] D. Buck and R. Buehler. Bike lanes and other determinants of capital bikeshare trips. In *91st Transportation Research Board Annual Meeting*, 2012.

[7] A. Faghih-Imani, N. Eluru, A. M. El-Geneidy, M. Rabbat, and U. Haq. How land-use and urban form impact bicycle flows: evidence from the bicycle-sharing system (bixi) in montreal. *Journal of Transport Geography*, 2014.

[8] J. Froehlich, J. Neumann, and N. Oliver. Sensing and predicting the pulse of the city through shared bicycling. In *Proceedings of the 21st International Jont Conference on Artifical Intelligence*, IJCAI'09, pages 1420–1426, San Francisco, CA, USA, 2009. Morgan Kaufmann Publishers Inc.

[9] S. Jäppinen, T. Toivonen, and M. Salonen. Modelling the potential effect of shared bicycles on public transport travel times in greater helsinki: An open data approach. *Applied Geography*, 43:13–24, 2013.

[10] 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. Human Behavior in Ubiquitous Environments: Modeling of Human Mobility Patterns.

[11] J. Larsen. Bike-sharing programs hit the streets in over 500 cities worldwide. *Earth Policy Institute*, 25, 2013.

[12] N. Lathia, S. Ahmed, and L. Capra. Measuring the impact of opening the London shared bicycle scheme to casual users. *Transportation research part C: emerging technologies*, 22:88–102, 2012.

[13] O. O'Brien, J. Cheshire, and M. Batty. Mining bicycle sharing data for generating insights into sustainable transport systems. *Journal of Transport Geography*, 2013.

[14] P. Vogel, T. Greiser, and D. C. Mattfeld. Understanding bike-sharing systems using data mining: Exploring activity patterns. *Procedia - Social and Behavioral Sciences*, 20(0):514 – 523, 2011.

[15] X. Wang, G. Lindsey, J. E. Schoner, and A. Harrison. Modeling bike share station activity: The effects of nearby businesses and jobs on trips to and from stations. *Transportation Research Record*, 43(44):45, 2012.

[16] J. Wood, A. Slingsby, and J. Dykes. Visualizing the dynamics of London's bicycle-hire scheme. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 46(4):239–251, 2011.

[17] J. Woodcock, M. Tainio, J. Cheshire, O. OâĂŹBrien, and A. Goodman. Health effects of the London bicycle sharing system: health impact modelling study. *BMJ: British Medical Journal*, 348, 2014.

[18] L. Zhang, J. Zhang, Z.-y. Duan, and D. Bryde. Sustainable bike-sharing systems: characteristics and commonalities across cases in urban China. *Journal of Cleaner Production*, 2014.

[19] Y. Zheng, L. Capra, O. Wolfson, and H. Yang. Urban computing: concepts, methodologies, and applications. *ACM Transaction on Intelligent Systems and Technology (ACM TIST)*, 2014.