Computers & Graphics 60 (2016) 119–129

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Special Section on SIBGRAPI 2016

Visual analysis of bike-sharing systems

Guilherme N. Oliveira [a], Jose L. Sotomayor [a], Rafael P. Torchelsen [b], Cláudio T. Silva [c], João L.D. Comba [a,*]

[a] Instituto de Informática - UFRGS, Brazil
[b] CDTec - UFPel, Brazil
[c] New York University, USA

ARTICLE INFO

Article history:
Received 1 March 2016
Received in revised form
2 August 2016
Accepted 3 August 2016
Available online 16 August 2016

Keywords:
Bike-sharing systems
Visual analytics

ABSTRACT

Bike-sharing systems are a popular mode of public transportation, increasing in number and size around the world. Public bike-sharing systems attend to the needs of a large number of commuters while synchronizing to the rhythm of big cities. To better understand the usage of such systems, we introduce an interactive visualization system to explore the dynamics of public bike-sharing systems by profiling its historical dataset. By coordinating a pixel-oriented timeline with a map, and introducing a scheme of partial reordering of time series, our design supports the identification of several patterns in temporal and spatial domains. We take New York City's bike-sharing program, Citi Bike, as a use case and implement a prototype to show changes in the system over a period of ten months, ranking stations by different properties, using any time interval in daily and monthly timelines. Different analyses are presented to validate the visualization system as a useful operational tool that can support the staff of bike-sharing programs of big cities in the exploration of such large datasets, in order to understand the commuting dynamics to overcome management problems and provide a better service to commuters.

1. Introduction

Public bike-sharing systems (BSSs) are services of increasing popularity, with many instances running around the world. The system is based on a set of stations located at several spots around the town with bikes available for rent. Commuters can take a bike out of any station, ride it for a limited period and then return it to any station. One problem with this usage scheme is that the operational staff has little control over the distribution of resources (bikes) as commuters are always moving them around. This control ensures that the stations do not get full or empty (an event called *outage*), thus users can get a bike from any station and also leave a bike at any station. Station rebalancing is used to prevent outages. Trucks are used to move bikes from different locations, which raise questions about how to choose the best route that minimizes gas consumption and time. Also, trucks are subject to traffic conditions and popular stations might need to be rebalanced more often than others.

Citi Bike of New York City is an example of a bike-sharing system. It was deployed in May of 2013 and is the largest bike-sharing system in the United States, officially serving 6000 bikes through 330 stations with a total of more than 11,000 docks [1]. Rebalancing efforts in Balancing Bike-Sharing Systems are usually done during the night when the usage frequency is minimal (or there is no service at all). Citi Bike NYC's rebalancing operations are performed during daytime to handle the intense commuting behavior. The recording of the information about each station throughout time (station states) can serve as an indicator of unbalanced stations, as well as circulation habits in the city. The analysis of usage data may lead to strategies for improving rebalancing procedures, plan upgrades in the infrastructure, and help commuters better use the program. Previous work with such data analysis usually focuses on simple scenarios, like small time windows, few variables, only trips or only balance data. In this work, we provide an exploratory view of bike-sharing usage data to understand its underlying dynamics. We provide sample analysis of such data that could be of use to the operational staff of bike-sharing systems trying to improve the quality of service to commuters. The data is composed of the station states (number of bikes and free slots available) recorded at periodic time intervals, as well as commuter's trip information (origin and destination station and timestamps). Fig. 1 shows one of the visual designs proposed to inspect station state data.

The main contributions introduced in this work for the analysis of this data can be summarized as follows:

* Corresponding author.
E-mail address: comba@inf.ufrgs.br (J.L.D. Comba).

http://dx.doi.org/10.1016/j.cag.2016.08.005
0097-8493/

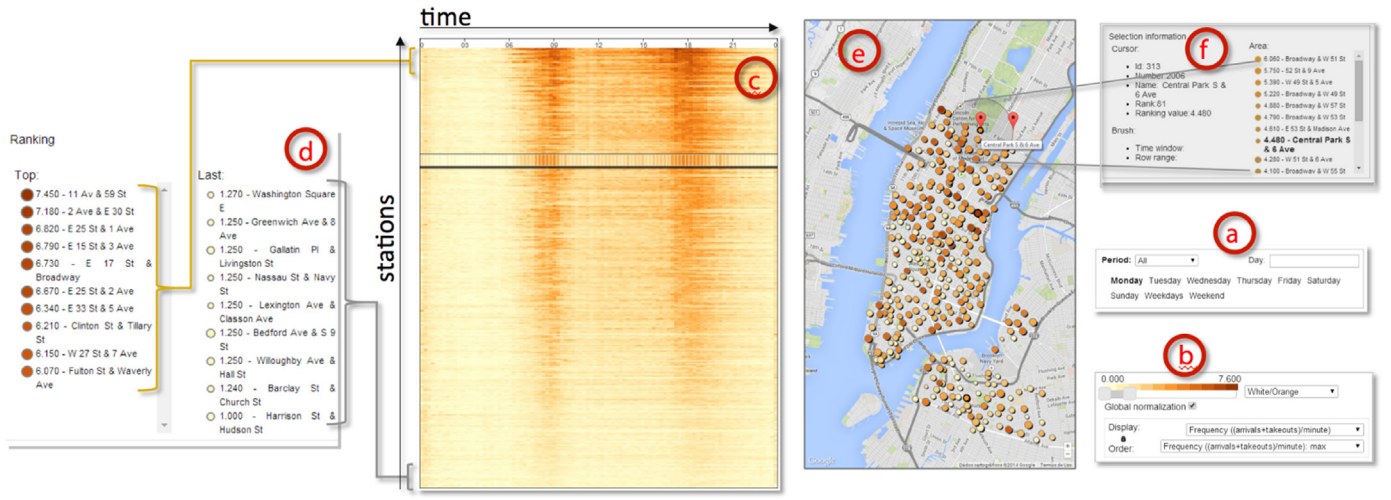

Fig. 1. Average frequency (number of commuters leaving or taking a bike from a station per minute) on Mondays in New York City. (a) Station data can be inspected for a given day, days of the week, or other periods of the year. Each station is presented as a row in the station state matrix (c) and as a circle on the map (e). In the station state matrix, each cell is color-coded based on the value of a selected variable, sampled at a 15-min interval in a day-long timeline. On the map, circles represent station locations with area proportional to the total capacity (maximum number of bikes supported). Rows (stations) can be ordered by any variable (b). A selected ordering is used in the ranking of the top and last ten stations (d), which are shown in the map using colored circles. The timeline color changes indicate the behavior on working days, where frequency peaks around 9am and 6pm. Ordering stations by maximum frequency show the 11 Av & 59 street station at the top with a frequency of at least 7.4 usages per minute. Selecting the Central Park South & 6 Avenue station shows its row and surrounding stations (f). (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

- encoding of the system usage data in pixel-oriented visualization designs that help understand the dynamics of station states and trip circulation patterns;
- visual designs that support flexible partial reordering of the pixel-oriented representation using an interactive brush that selects time intervals and station groups;
- several analysis scenarios conducted on real data from Citi Bike illustrate the capabilities of the proposed solutions to reveal relevant spatial and temporal patterns.

2. Related work

Several works analyze the behavior of bike-sharing systems. In this section, we report the related work and discuss how they differ from our proposal. We organize this material in three sections: rebalancing systems for finding optimal routes and loading operations, statistical tools for prediction and planning purposes, and visual analytics tools for a similar type of data.

2.1. Balancing Bike-Sharing Systems

The balancing of BSSs discusses strategies to find optimal routes that visit unbalanced stations while performing rebalancing operations. Several papers address the optimization issues derived from this problem, and our system could be used to visualize the different results obtained in these algorithms. Rainer-Harbach et al. [2] generate candidate routes for visiting unbalanced stations and optimal loading operations (load or unload) to be performed along the route. Papazek et al. [3] also describe a hybrid heuristic to find efficient vehicle routes, showing how it scales with benchmark instances derived from real data from Vienna. Raidl et al. [4] calculate optimal loading operations but use graph maximum flow algorithms. Schuijbroek et al. [5] propose a heuristic to find loading operations and near-optimal vehicle routes to rebalance the inventory. Urli et al. [6] address the problem of instance generation for benchmarking proposed approaches to the optimization problem of balancing BSSs. They describe a process to generate input instances based on data from Citi Bike and rely on very simple box plots to display the data. They display individual daily plots for each station, which generates a great number of plots to evaluate. Our system could be used in conjunction with this data to more easily identify sinks and sources stations, as well as rebalancing procedures that occurred.

2.2. Statistical tools for prediction and planning

Statistical tools allow analyzing system dynamics with minimal use of visualizations. Our individual visual designs or complete tool could be used to improve these statistical tools. Guenther et al. [7] focus on the forecasting of future bicycle migration trends. An analysis using data from the Barcelona's Shared Bicycling program is described by Froehlich et al. [8] to gain an understanding of human behavior and city dynamics. This proposal relies on analytics tools to analyze 45 days of data. On the other hand, our work makes extensive use of visualization and interaction to support the exploration of a 10-month long dataset, which can be easily extended to all data generated by Citi Bike. A follow-up work by Froehlich et al. [9] compares experimental results from four predictive models of near-term station usage, considering impact factors such as time of day and station activity in the prediction capabilities of the algorithms. In [10], a study is given that relies on statistical modeling and data mining to model the evolution of the dynamics of movement. Visual designs as we proposed in this work can help in finding interesting patterns of trip circulations.

2.3. Visual analytics

Visualization and analytics tools are used in the analysis of bike-sharing dynamics. Maps displaying the operational status of the system are given in the WorldMap system [11]. Such a visualization is useful for understanding the current status of the stations, but inadequate to present the evolution of the operational status in time such as we propose in this work, which helps to identify bottleneck stations, among other problems. Several works discuss the visual interpretation of bike data and their impact on city life. Zaltz et al. [12] use visualization techniques, statistics, spatial, and network analysis tools to explore bike-sharing system usage in five different cities. A Voronoi diagram is created using

Station State	
Property	Value
Timestamp	2014-07-22T13:30:05.196Z
Number	72
Address	W 52 St & 11 Ave
Latitude	40767272
Longitude	-73993928
Bikes	0
Free Slots	35

Trip	
Property	Value
Duration (s)	1547
Start time	2013-07-01 00:00:02
End time	2013-07-01 00:25:49
Start Station #	388
End Station #	459
Bike #	19816
User Type	Annual subscriber
User Gender	2 (female)
User Year of Birth	1980

Fig. 2. Citi Bike data: station state and trip sample records.

station locations to partition the city into regions [13], which are then color-coded based on trip data (e.g. trip duration, age, user profile, the number of bikes, etc). Kaufman [14] shows an analysis of the popularity of Citi Bike stations between genders. Beecham et al. [15] use coordinated views to understand cycling behaviors in London.

The visualization of several trajectories allows understanding the flow of bikes in the city, which is represented by lines with a different thickness. A design study using visualization techniques for bike-sharing systems is described by Wood et al. [16]. The bike-sharing system is represented as a connected graph, where a single resource (bike) circulates over edges to nodes with different capacities. A visual analysis of London's bicycle system is described in [17]. Two views show trips and station balance using a restructuring scheme that avoids cluttering, while bike trips are displayed on the map. Citi Bike trip data is used by O'Brien [18] to estimate bike trajectories since trip data only contains origin and destination. Ferzoco [19] shows trip circulations in NYC for a two-day period. Our system is more flexible since it can be used to inspect trip data in any period of time with different possibilities of ranking this data. Hurter et al. [20] visualize changes in dynamic graphs and paths using edge bundling. This approach could be integrated into our visualization of trips to reduce cluttering. Taxi trips are used by Ferreira et al. [21] to understand the city life dynamics. To derive traffic information and visualize how traffic jams propagate in the city, Wang et al. [22] fit trajectories given by taxis with GPS to the road network. Guo et al. [23] use coordinated views that are used to explore spatiotemporal data with a reordering matrix representation of time series connected to a map. While these works target a single question, variable, data facet (balance or trips), or very limited periods, in our work, we aimed for flexibility to support a more detailed understanding of BSS's usage. In this sense, our work is closer to the one by Wood et al. [17]. We add to the design by Guo et al. [23] a partial reordering to explore a long history of system's activity in both stations' balance and trips, and to identify patterns, outliers, raise questions and confirm prior theories.

3. Materials and methods

The proposed method was designed to conform to requirements that guided our data analysis and visual representations.

3.1. Desiderata

The goals in the analysis of bike-sharing systems were used to formulate tasks that should be supported by visualization and interaction. The result desiderata to be satisfied by the method described in Section 4 are presented in the requirements below:

- **R1**: Identify stations that become bottlenecks in the system, frequently getting outages (empty or full of bikes). This can help design changes to improve the resilience of the system, providing better service for commuters.
- **R2**: Verify the influence of city life changes and events in the behavior of bike-sharers. As the popularity of bike-sharing programs increases, its dynamics becomes a relevant indication of changes in the city life routine.
- **R3**: Understand how the distribution of station roles, into source/provider and sink/receiver, changes through the day. This division of tasks is recurrent in bike-sharing systems, being usually a good indicator of commercial and residential areas, and aspect of primary importance when designing balancing solutions in BSS's.
- **R4**: Compare the dynamics of the system at different periods since its deployment. As Citi Bike was deployed only recently, the city is still adapting to it and vice versa. Looking into how the usage dynamics changed over its first year may give rise to valuable insights on what to expect when the system expands to cover new areas and for the years to come. Also, it can show how the cycling behavior changes through the seasons as the weather varies.

3.2. Citi Bike data

The state of all stations in the Citi Bike system can be queried by fetching the JSON feed [24]. The feed has current balance for all stations and is updated every time the balance of one station changes. The feed consists of a list of state entries, one for each station. Each entry has the station id, name (address), number of bikes available, number of free parking slots, latitude, longitude, and timestamp of the last change. We tracked changes in the JSON feed for eight months, at an interval of 30 s, since June of 2013, for a total of more than ten million updates about the state of 330 stations. Fig. 2 illustrates both station state and trip data.

Changes in the station states happen when a commuter is either parking or taking one bike out of the station. A sequence of these events leads to a time-series sampled at irregular intervals. The usage activity for one station on a given day is a sequence of n events. A bike return is detected when the number of bikes increases after an event. Otherwise, it is called a bike rental. The *capacity* of each station is defined as the total number of slots. The *station balance* is calculated as the ratio between the number of bikes at the station and its capacity. Due to visualization constraints that we address in the next subsection, we reduced the size of the time series data. We applied a piecewise aggregation to resample the series into regular 15-min intervals. This interval was satisfactory for a 24-h period and the station rate of usage.

Trip data was aggregated using a 1-h interval. We computed aggregated measures for trips that began or ended in each interval: balance, capacity, in/out difference, number of cyclic trips, number of incoming trips, number of outgoing trips, outage state (empty, full or no outage), number of incoming origins and outgoing destinations, number of trips, trip duration and trip distance. Trips are divided into three classes: outgoing, incoming, or cyclic. Outgoing and incoming trips have different origin and destination stations while in cyclic trips they are the same. For each interval, we counted each class separately and the difference between incoming and outgoing stations. Outages can be identified by the average balance (0 defines an empty outage, 1 a full outage). We use a threshold to allow values closer to 0 and 1 to indicate outages. For example, if a threshold is 0.1, average balance above 0.9 is considered a full outage, and below 0.1 as an empty outage. Since trip distances are not provided, we estimated them by the shortest path between the origin and the destination station given by Google Directions using cycling routes. For cyclic trips, we estimated the distance by multiplying the duration by the average speed of 2.7 m/s. Our Citi Bike dataset has station state information for each day from June 2013 to March 2014, leading to about 270 time series per

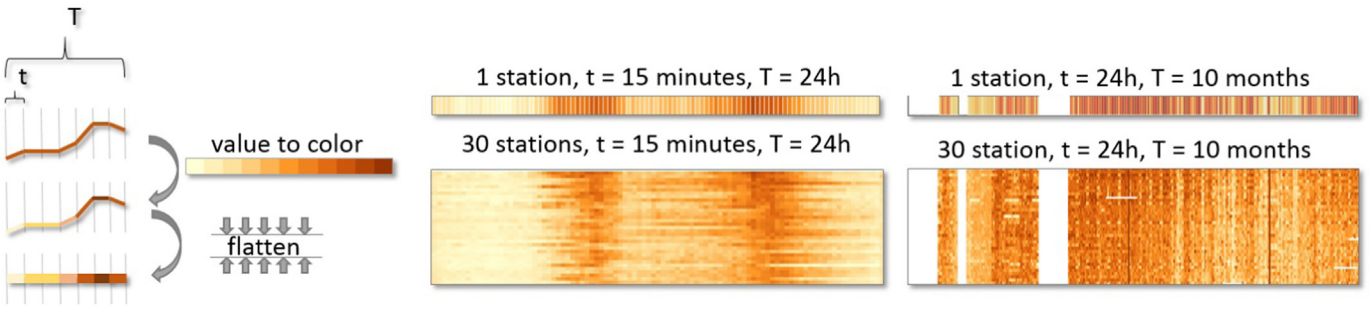

Fig. 3. Station linear representation. (left) Representation of station states into color-mapped rows. (middle) Single day timeline for one station (top) and 30 stations (bottom). (right) Ten month timeline for one station (top), and 30 stations (bottom). The white vertical stripes correspond to missing data for part of June, beginning of July, and part of September. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

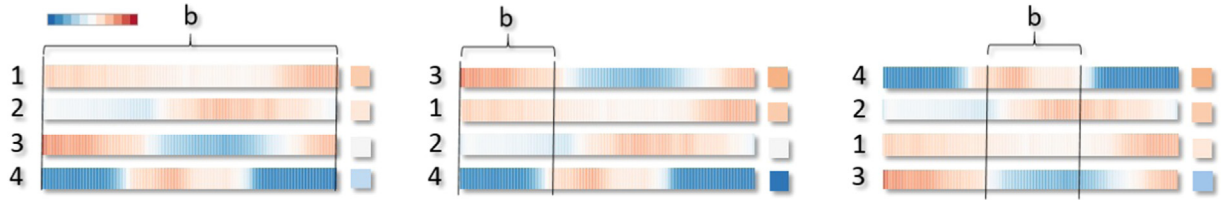

Fig. 4. Partial reordering. (left) The reduce operator (arithmetic average in this example) is applied on the entire time span to create the initial ordering for stations 1–4. (middle) Using the timeline brush (*b*) to change the domain of the reduce operator, thus changing the ordering. (right) Stations reordered by their average balance during working hours. Stations 4 and 2 had more bikes than free slots while 1 was usually evenly balanced, and 3 had a shortage of bikes.

station and 8000 series in total. We created daily and monthly aggregations (from June 2013 to March 2014), as well as the summer, fall, and winter seasons within this period.

4. Visualization designs

In this section, we describe visualization designs with interactive capabilities to explore data from bike-sharing systems.

4.1. Partially re-ordered station state matrix

Station state data might give insights about which stations are often used during a given period of time, and how stations relate to one another. This natural comparison among stations suggests that station states be displayed in order of some attribute (number of bikes available, balance, etc). However, this order changes during the time series, and different rankings can be observed with different time intervals. We use the linear representation of time series from the work of Guo et al. [23] as it makes the best use of screen space up to our knowledge, being able to represent hundreds of series on the same screen with no overlap. In order to find spatial patterns related to the different properties of the stations over time, with adjustable temporal resolution, we improve this representation with an interactive reordering scheme using brushing, as explained below.

4.1.1. Station state timeline matrix

Each station state timeline is shown in individual horizontal rows to avoid overlap. Data for all stations create a station state timeline matrix, with rows identifying stations, and columns associated to subsequent instants of time. Matrix cells are color-coded based on a given variable. The column range of this matrix has two resolutions: a 24-h timeline, with samples of data aggregated over a 15-min period, and a 10-month timeline, with samples aggregated for each day (Fig. 3).

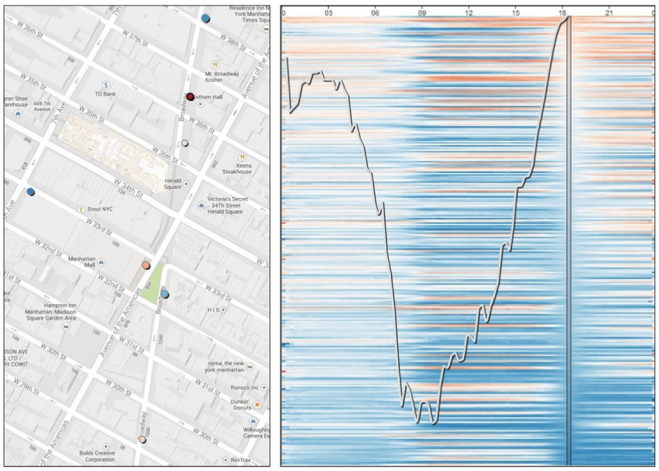

Fig. 5. History of ranking positions on Mondays for the Broadway and West 37 Street Station (red circle on the map) ordered by maximum balance. The history is shown as a line graph connecting ranking positions of the station for a sliding window of 15 min from midnight to 6pm. The ranking for this station changes from lower balance values at 9am to higher values at 6pm. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

4.1.2. Partial reordering using brushes

The rows of the station state matrix are ordered by their content. In the left of Fig. 4, the rows are ordered according to the average value of the entire timeline. If a particular time interval is important, it is possible to reorder the rows considering only the data within this interval. In the middle and right parts of Fig. 4, the rows are ordered using other time intervals, defined by a timeline brush over the columns of the matrix. The timeline brush can also be animated as a sliding window to display how the station ranking changes over time. The ranking history of each station is shown over the matrix as a line graph when the station is selected (Fig. 5). Additionally, a ranking brush is used to define an interval along the rows of the matrix. This allows constraining the analysis to a subset of the ranking, important when displaying station information on the map. We keep a ranking

panel with a list of the top and bottom 10 stations in the ranking brush interval.

4.1.3. Coordinated view of map and station state matrix

We integrated the partially re-ordered station state matrix with a map showing station locations. Fig. 1 gives an overview of the components that create this coordinate view. The station state matrix in Fig. 1(c) is the major interaction and informative component. The average series can be displayed for different periods (e.g. specific day or month, days of the week, weekdays, weekends, season, etc). The selected variable, ordering scheme, color mapping, and normalization can be changed in the panel of Fig. 1(b). Selected variables include: *balance, bikes available, free slots, frequency, station capacity, bike return, bike return frequency, bike rental* and *bike rental frequency*. The *station capacity* is not always constant as expected. The sum of *bikes available* and *free slots* does not always add to the same value at different timestamps. Keeping track of this value over time may reveal unexpected changes that might indicate problems in a station. Ordering options correspond to the same variables available for color coding in the cells, reduced using one of the four reduction operators (*max, min, mean* and *range*), and time-invariant properties of the stations (*id, name, latitude* and *longitude*). The map in Fig. 1(e) shows each station as a circle whose color is associated with the index of the row of the respective station in the matrix, while the area encodes the capacity of the stations. The selection of one station on the map highlights its row using a lens metaphor, its entry in the ranking lists (if listed) and more information on the panel of Fig. 1(f). Also, a circular region can be defined on the map to select a group of stations.

In the 10-month timeline, each cell represents an aggregation over entire days. Since a simple average is not informative, we aggregated the data for each day by six different reduce operators: average, minimum, maximum, range, and time of minimum and maximum values. Such operators show the time of the day during which extreme values were first registered.

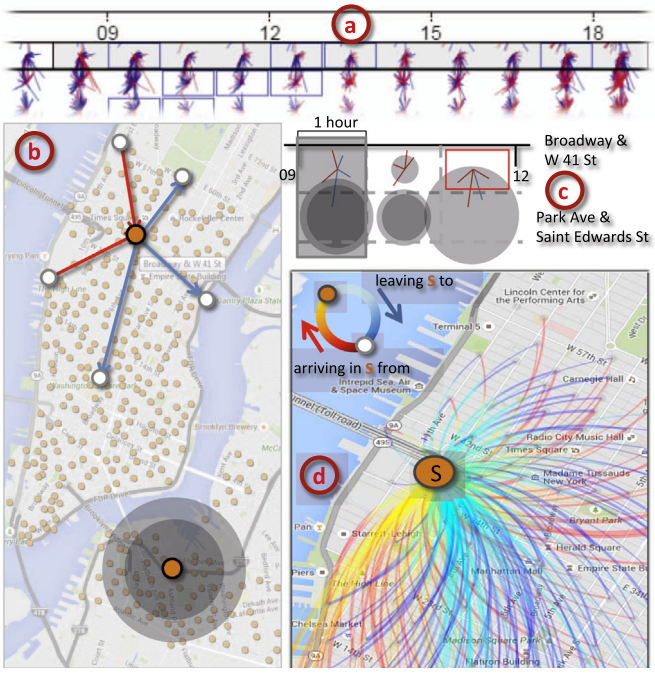

Fig. 6. Trips timeline matrix. (a) Trips are represented by star-plot glyphs in their respective cell of the matrix. (b) Incoming and outgoing trips are given by red and blue lines respectively. Cyclic trips are drawn using a transparent circle with a radius proportional to the trip distance. (c) Two stations selected in the matrix in a 1-h interval. The red rectangle is a full outage at the Broadway and W 41 St station. (d) For periods longer than one hour, trips leaving from the selected station are drawn in curved lines. The map shows incoming and outgoing trips of roadway and W 41 St station between 8am and 6pm. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

4.2. Trips timeline and origin–destination matrices

Trips contain origin and destination stations and times of pickup and drop-off. We use two matrix visual designs to explore the outgoing directions of trips and the relationship between origin and destination stations. This design allows inspecting the station commuting behavior during the day and identify the trip directions that might lead to station outages.

4.2.1. Trips timeline matrix

We use a timeline matrix similar to the station state matrix to encode the outgoing direction of trips, with rows and columns representing stations and time intervals, respectively. Partial reordering of this matrix is similar to the state station matrix but now using derived variables from the trip data. In each cell of this matrix, we draw a glyph to encode the directions of trips from that given station in the corresponding time interval. In Fig. 6 we illustrate a portion of the timeline for two stations in the 9–12am period. We use a 1-h aggregation interval for the trips, instead of the 15-min interval used for the station states. The center of the glyph corresponds to the station associated with the current row. Every incoming trip at this station generates a red segment, coming from the relative direction of the origin station. Similarly, outgoing trips are colored in blue. Cyclic trips are represented with a particular design. We use semi-transparent station-centered gray circles that cover the area that a bike could reach if moving along a straight line at an average speed of 2.7 m/s. Also, station outages are encoded in the cell border color (red and blue are full and empty outages respectively). This matrix timeline provides a summary of how many trips started or ended at each station at every hour, which are used to identify interesting time intervals that can be selected for further inspection. Selected stations are highlighted in the map view, with trip directions drawn using curved lines, to avoid overlap of incoming and outgoing trips between the same two stations, as used in the representation of Wood et al. [17]. However, in our scheme, colors are used to differ between outgoing and incoming trips: outgoing ones are drawn in blue lines in clockwise order, while incoming trips are drawn in red lines in anti-clockwise order (Fig. 6). Also, lightness is a cue of direction: outgoing trips begins in the station at the cyan extreme ending at the station in the blue extreme, while incoming trips go from red (origin) to yellow (destination). Note that the selected station is always at the lighter extreme of the curve.

4.2.2. Trips origin–destination matrix

The second matrix visual design is an origin–destination (OD) matrix, with rows and columns representing outgoing and incoming stations respectively. Cell colors are mapped to a trip variable aggregated at the selected time. Possible variables include: number of trips, trip duration, balance difference and station capacity difference. The number and duration of trips can be used to identify preferred stations. The difference of balance from the incoming and outgoing stations encodes the states of the stations involved in the trip. The balance difference varies from -1 to 1. A value of 1 indicates that the incoming station is full and the outgoing station is empty (critical case). On the other hand, a value of -1 shows the opposite; a full station at the origin and an empty station at the destination (ideal case). Closer to zero values means that the two stations have similar balances. The capacity difference helps distinguish trips that occurred from bigger to smaller stations (or vice versa), or between stations with equivalent capacities. Aggregate values can be shown for days of the week, weekdays, weekends, months, seasons, or simply a particular day of the year. Partial reordering is similar to the station state

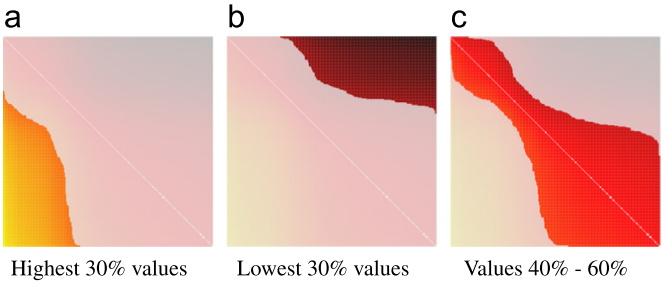

Fig. 7. Trips OD matrix automatic selection by value percentage: balance difference aggregated over weekdays for September 2013.

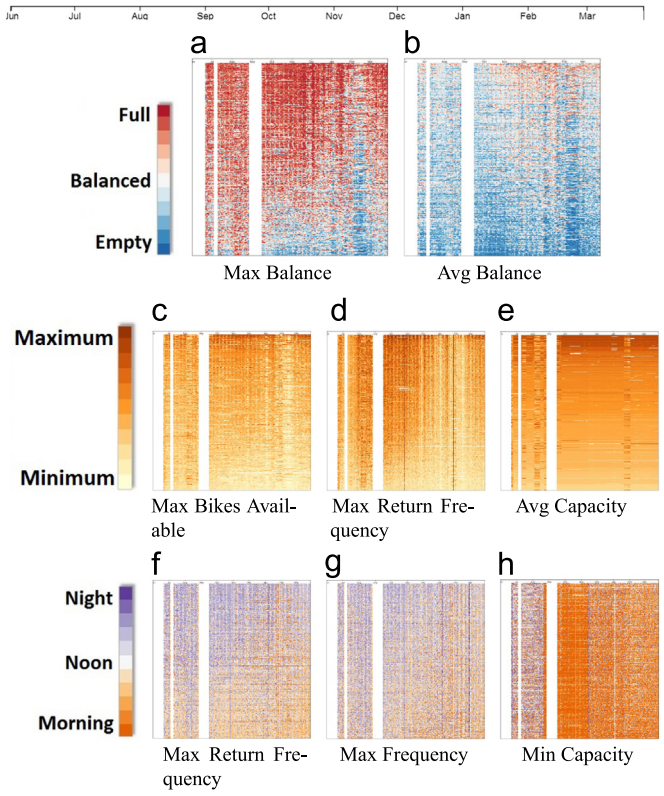

Fig. 8. Calendar view displaying system usage over 10 months with different variables. Top row shows the 10-month time scale shown at the top of each profile. High maximum balance values at early months (top row), decrease in available bikes during February (middle row), earlier occurrences of max return frequency and min capacity increase at the colder months (lower row). (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

matrix, using the following variables: station trips, capacity, latitude, longitude, number of trips, trip duration, and balance.

The OD trip matrix can be inspected using a manual or automatic selection. In the manual selection, the user navigates over the trips OD matrix to find relevant patterns. The vertical and horizontal brushes select the stations that serve as start or end points, respectively. To conserve geospatial context along with the trips matrix view, we use a map that shows each station as a circle whose color is associated with their role in the selected trip. Outgoing stations are drawn in blue, incoming stations in red, and stations that serve as both source and sink in purple. The second mode uses a variable to automatically select stations. This allows making comparisons based on the percentage of a variable range. For example, automatic selection with a range of 70–100% for the balance difference finds the highest 30%, in this case with values between 0.4 and 1. This and other two selections are illustrated in Fig. 7.

5. Results

Next, we describe several analyses performed over the NYC Citi Bike data. We show a high-level calendar overview analysis, from an in-depth exploration of specific days of the week in different periods (months or seasons), to answering queries of stations and circulation patterns.

5.1. Calendar view analysis

The calendar view supports an overview analysis as the initial stage of the exploration pipeline. By combining any of the derived variables extracted from the station state feeds with a choice of how to reduce it for each day, it gives different perspectives on how the Citi Bike program developed along the 10 months after inception. Fig. 8 shows different visual profiles, with color scales chosen according to the nature of the variable displayed. Column patterns are visible in almost all profiles and serve as good points to start the analysis of the data.

The first profile row displays maximum and average balance. The max balance reveals that in the early months (summer), it is common for the stations to get close to a full outage at least once a day. This pattern changes as fall and winter approach and system usage decreases. Since balance is aggregated over a whole day in this view, we cannot argue about the severity of the outage (its duration), such analysis needs to be done in the 24 h timeline view presented ahead. The average balance has a predominance of neutral values (lighter colors) at early months. This pattern begins changing slightly when Fall arrives, with stations showing darker red and blue patterns.

In the first profile of the middle row, we show the daily maximum bikes available at each station. There is a column pattern of lighter colors in February, which indicates a smaller number of bikes at stations. This could mean that either bikes were removed from the system, or more bikes were circulating. By looking at other profiles, we observe that the first explanation is more likely. Profile (d) shows a decrease in the maximum return frequency during the same period. This profile also reveals a period of outliers at the end of January. The regularity of the anomaly makes the operational activity be the likely cause for the pattern observed.

In the third row, we also display return frequency like in profile (d), but instead we encode the time of the day when the maximum value was registered. The purpose of this perspective is to see which stations are destinations by morning (orange) and night (purple). Profile (g) displays the time of maximum using total frequency (returns plus rentals). We observe which stations are more used during early and late hours, regardless if the station is origin or destination. Both perspectives show stations becoming popular at early hours during winter (the increase of orange color during this period), but only the second one brings up the system-wide anomaly that happened at the ending of February (a dark purple column). The profiles (e) and (h) are based on the station capacity, which is supposed to be constant, but the profiles show otherwise. In profile (e) we observe changes in the average capacity (e.g., end of February). The time of minimum capacity of the stations in profile (h) reveals a pattern of the predominance of late hours until the end of August. There is a sudden change at early hours, changing again in November, with a single-day-wide purple column followed by a random pattern until March.

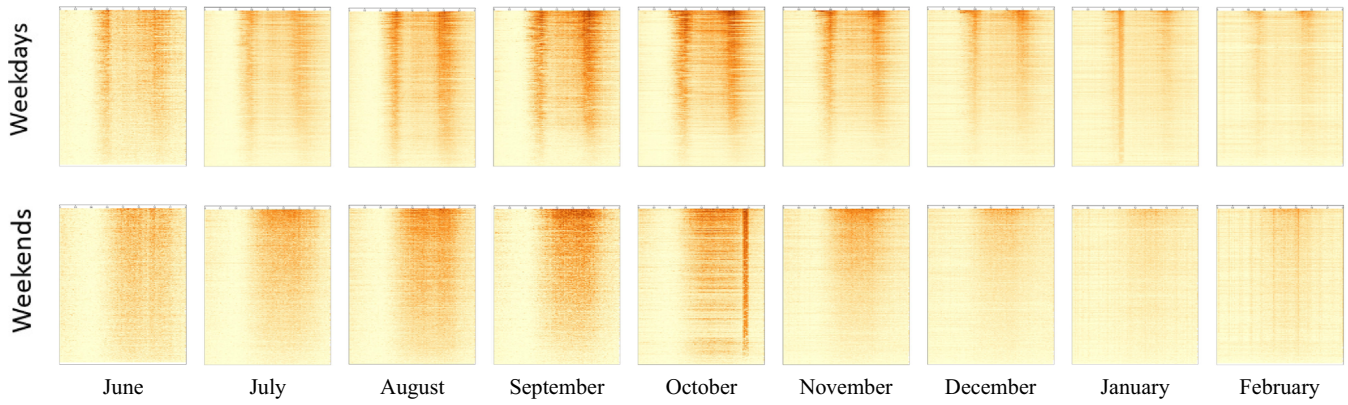

Fig. 9. Monthly frequency during weekdays and weekends. System usage increase is observed when profiles get darker from June to October, and lighter during the winter. Some profiles show interesting results: October weekends have a peak of use around 8pm, the 9am peak on weekdays of January is more evident than at 8pm, and weekends of February have two short peaks around 11am and 3pm. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

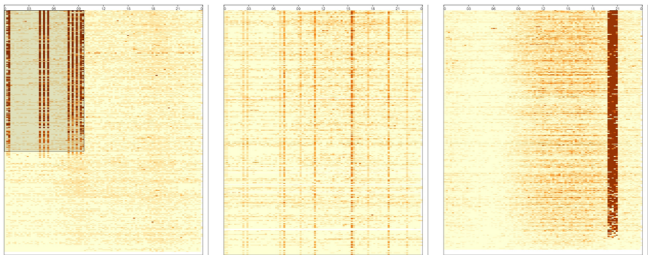

Fig. 10. Three different days with outlier behavior detected. First a curious pattern of strong stripes shows a peak of bike arrivals of more than half of the program's stations in the morning of 27 January. Then a more subtle set of stripes showing synchronized peaks of frequency for most stations at different times of the day in 16 February. And a strong one-hour-long frequency peak around 8pm on 26 October.

5.2. Daily view analysis

In the calendar view analysis, we identified smooth changes along the 10-month timeline and anomalies in the different profiles of the station state data. Daily timeline views allow narrowing the analysis to specific days in which unexpected behaviors happened. Fig. 9 shows weekday and weekend profiles in monthly frequency. There is a clear difference between the system usage during workdays and weekends. Frequency on weekdays has two peaks, one around 9am when commuters go to work, and another at 6pm when they return home. The usage during the day is higher than early morning or late night, but lower at rush hours. During weekends, there is a single wider, lower, and smoother peak that begins later than weekdays, at 10am, and ends later, at 9pm. Since the same color scale and extreme values were shared between the profiles, we observe that Fall had the most intense activity (overall darker colors) with a strong decrease during Winter. We notice the contrast increasing from June to October and decreasing again to February. Also, the duration of the weekend peak during Winter was shorter than the other months (11am–6pm), probably due to shorter days and longer nights.

There is a peak between 9 and 10pm only seen during weekends in the Fall. To further inspect these outliers we drill down specific days of the week. Fig. 10 compares the frequency of Saturdays and Sundays in October, revealing that the anomaly comes from the former. By looking into the profiles of each Saturday of October, we found out that it only happened on the 26 October. Fig. 10 also shows other two outlier days with different patterns. Due to the synchronization and intensity of those patterns, it is unlikely that they were caused by the commuters activity.

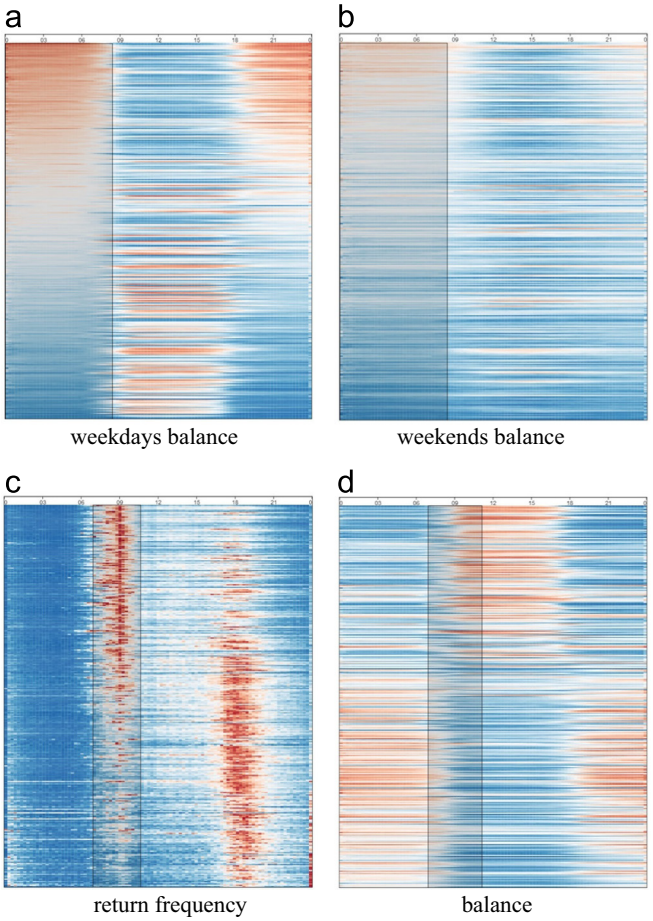

Fig. 11. Partial ordering of four profiles in a 24-h period: (a) ordering stations (rows) considering the balance from the first third of the day. A pattern reveals that stations with high balance (red) change into low balance (blue) after 8am, and return back at 6pm. The same pattern is not visible for weekends (b), with no clear definition of roles for the stations, and an even distribution of balance. Another partial reordering of weekdays around 9am is given in (c). We observe that stations with low return frequency around 9am have a high return frequency at 6pm, which is a common commuter pattern. The balance in (d) using the same ranking of (c) allows to find provider and receiver stations. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

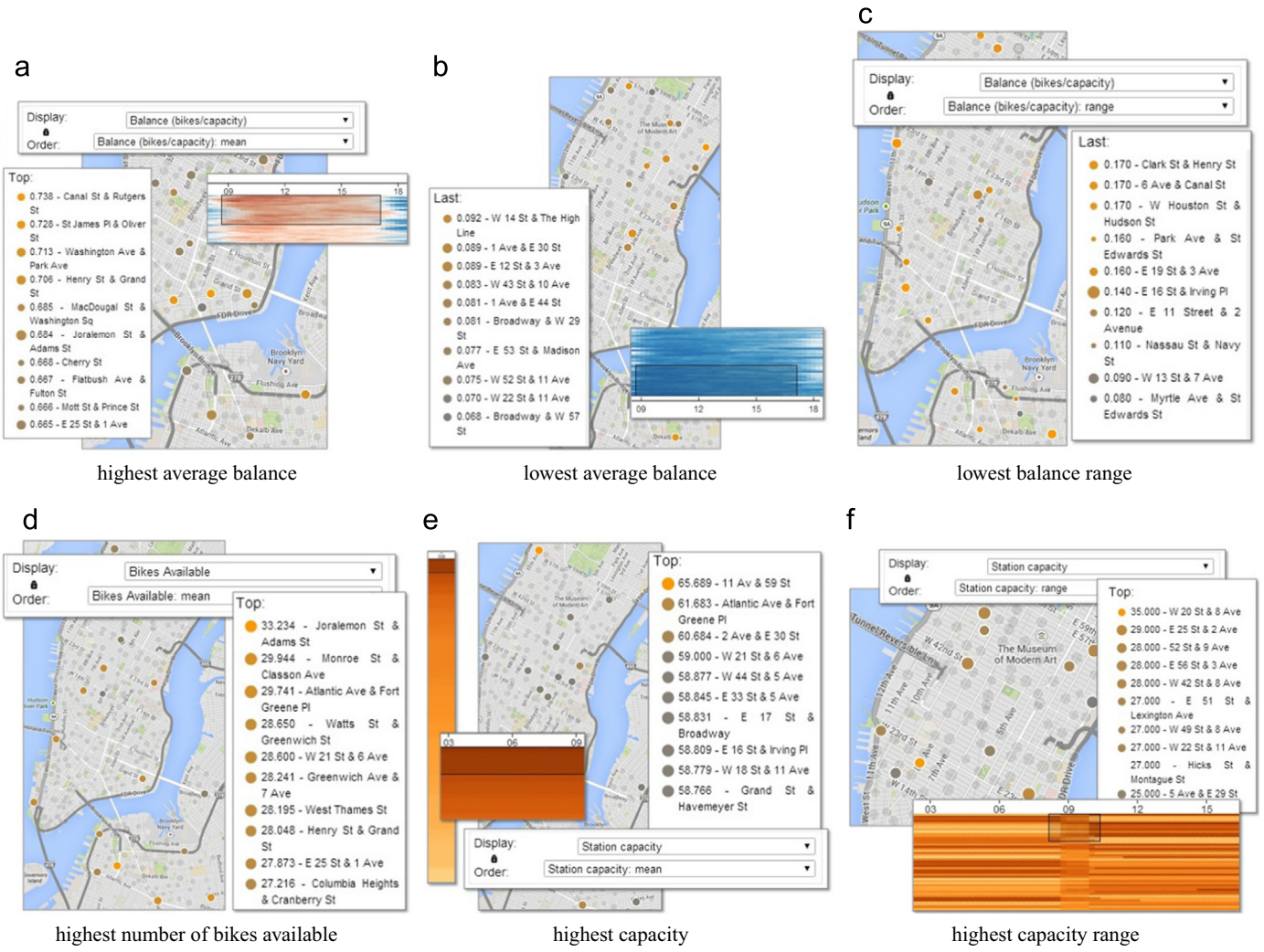

Fig. 12. Station analysis. All station data can be queried using different variables and aggregations. Timeline and station brushes allow narrowing the period and highlight stations on the map. (a), (b) Highest and lowest balance average of full and empty stations during working hours of Wednesdays during the Summer. A separation on the map can be seen between nearly full stations below midtown and empty ones at north. (c) Stations with the lowest range of balance during the day. (d) Stations with the highest number of bikes during working hours. Highest station capacity for Wednesdays during Fall (e) and on 01/28/2014 (f). This particular day has an unusual change for some stations around 9am. Using the time brush to select this interval allows finding the most affected stations.

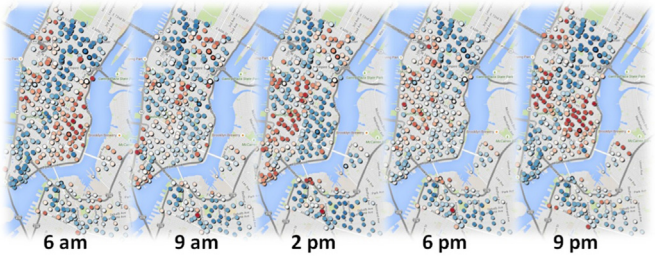

Fig. 13. Station roles. Stations are ordered by average balance at different times in a typical weekday in the Fall of 2013. We observe a change of station roles (providers or receivers) during the day. Early in the day, most riverside stations are almost full (red circles). This situation changes as time goes by and it is almost the opposite at 2pm, returning to the initial setting after 9pm. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

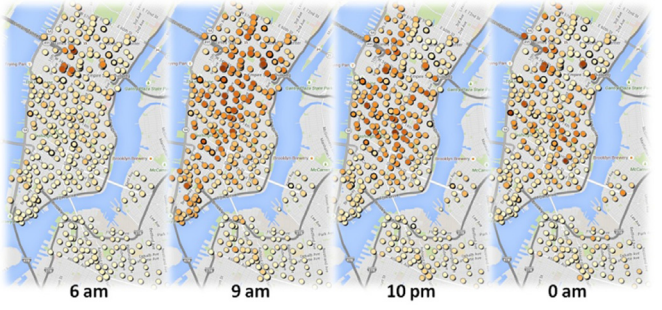

Fig. 14. Usage frequency on weekdays at different times. Stations around Penn station show early movement. As the day progresses, the frequency increases in other parts of Manhattan. At night, it is not completely back to normal in areas with entertainment options.

Fig. 11 shows how to use partial reordering to find temporal patterns. The brush defines the extent of the partial reordering in the matrix. Different patterns arise for weekdays and weekends. On weekdays, top-ranked balance stations are full in the morning, become empty during working hours, and back to full again at night. Lower-ranked stations have the opposite behavior. However, during weekends this pattern is not visible, as bikers have an unpredictable behavior, riding more for leisure than for working. In the bottom row, we display the return frequency on weekdays and order the rows by the same 8–10am interval (morning rush hour). The result clearly shows working hours destinations at the top and origins at the bottom.

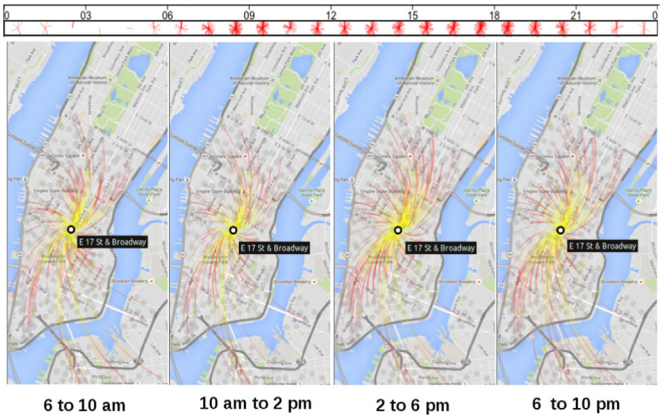

Fig. 15. Incoming trips on weekdays at different times for station E17 & Broadway. Using the trips matrix we identified E17 & Broadway as a commuting hub. Being a central station, it sustains intense activity from 6am to 10pm, as a destination of trips from almost every part of the city.

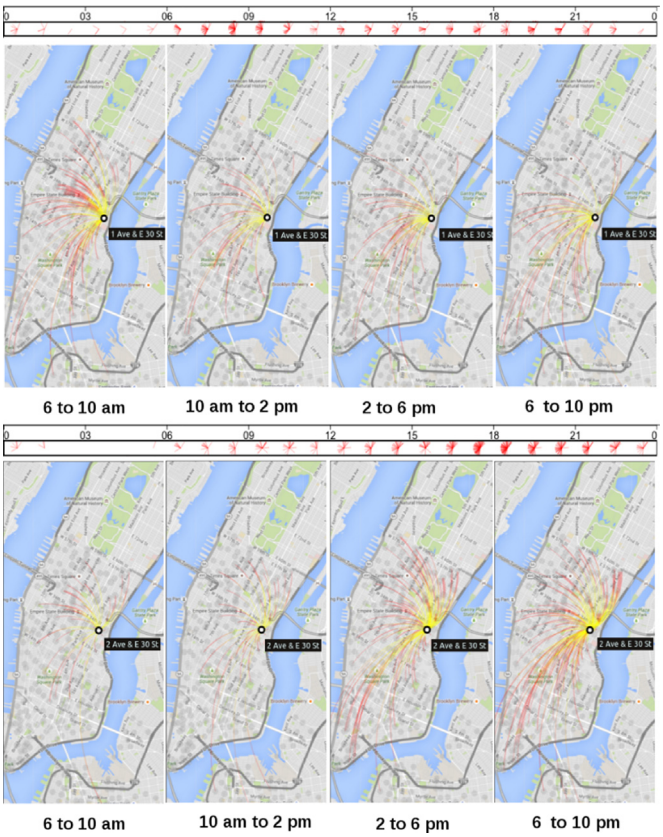

Fig. 16. Incoming trips on weekdays at different times. Top: Station 1 Ave & E30 St. Bottom: Station 2 Ave & E30 St. As the day goes by, the frequency of incoming trips decreases for the top station and increases for the bottom station, despite being nearby stations. At night, we observe a high number of incoming trips coming from the southern part of the city for the 1 Ave & E30 station.

5.3. Station analysis

An important design requirement is to be able to identify the stations that become bottlenecks of the system by having outages. Our prototype provides a myriad of ways to look at this data. It is possible to select which variable to inspect, different aggregation periods, narrow the data scope using brushes and evaluate the station rankings generated using partial reordering. In Fig. 12, we illustrate six queries using different criteria. In (a) and (b) the brush is defined between 9am and 5pm. Stations are ordered by

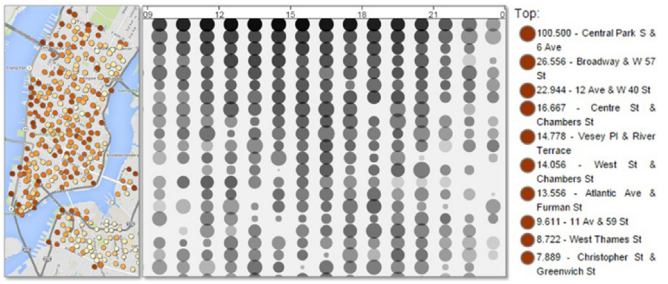

Fig. 17. Cyclic trips on Sundays in the Summer of 2014. Stations are ranked by the average number of cyclic trips. Cyclic trips are evenly distributed around 4pm. The Central Park station has the highest number of cyclic trips per hour.

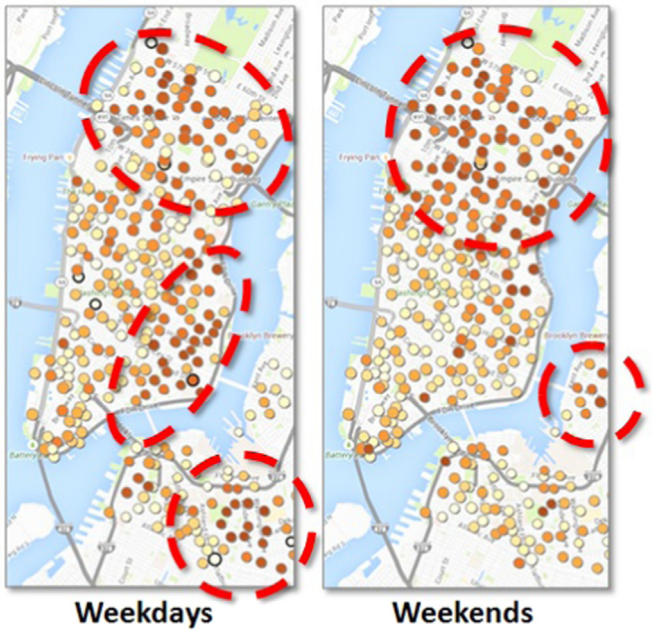

Fig. 18. Outages. Stations are ranked by the number of outages. Areas with a higher concentration of stations suffer from outages on weekdays: Midtown and East Village in Manhattan, and between Fort Greene Park and the Pratt Institute in Brooklyn. Midtown and in Williamsburg have outages on weekends. Also, notice the difference between the overall distribution of outages in weekdays and the one in the weekends. It has a larger spread in the weekdays.

the average balance in that period of the day. Using the station brush (selecting rows of the matrix) we order stations according to their average balance. In (a) the brush is limited to the highest rows of the matrix, which reveals on the map the stations that are usually full at the selected period (Wednesdays during the Summer). In (b), the brush selects the lowest rows to reveal empty stations. We observe a division between full and empty stations in the lower and upper parts of the area covered by the program respectively. In (c) rows are ordered by their range of balance values during the day. By selecting the lowest rows, we found that these stations remained at a nearly constant balance level. An interesting task is to find out where the majority of bikes of the program are at a given interval. In (d) the number of bikes available in the stations is displayed. Ordering the rows by the number of bikes available and selecting the highest rows with the brush allows drawing in the map the stations with the largest number of bikes. Stations of higher capacity are shown in (e) for Wednesdays during Fall. Since the capacity is constant for each station, the expected pattern is a smooth vertical gradient. However, the capacity results for a single day revealed changes in this data, In (f) there is an anomaly in the capacity profile on 01/28/2014

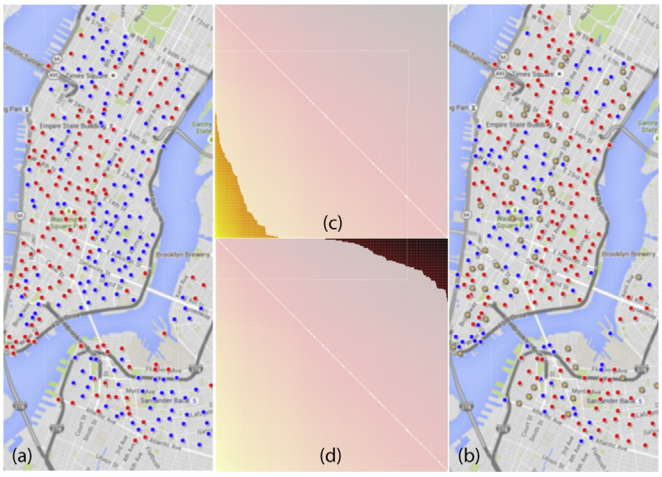

Fig. 19. Balance difference in the OD trip matrix for weekdays at noon of March 2014. Blue and red stations are respectively the origin and destination of trips. Rows in the OD trip matrix are ordered by station balance, and colored from high (yellow) to low (red) values. (a) Stations with highest 20% at noon showing trips that leads to outages and (c) corresponding OD matrix. (b) Stations with lowest 20% and (d) corresponding OD matrix. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

around 9am. Selecting this interval in the timeline brush and ordering the rows by capacity allows finding the affected stations.

5.4. Circulation and outage analysis

Bike movement can be observed during the day by looking at changes in the average balance. An alternating pattern of station roles was observed in Fig. 11. To complement this analysis, in Fig. 13 we display station data on the map. Red and blue circles on the map correspond to full and empty stations respectively. At 6am most bikes are located at riverside stations. At 9am few bikes are parked, probably because bikers are commuting to work. At early afternoon, the opposite happens. There is a concentration of full stations in the middle of Manhattan, while riverside and Brooklyn stations are empty. Later in the day (after 9pm) the system nearly returns to the original morning situation, with slightly higher concentration at Williamsburg and East Village.

Fig. 14 displays the overall usage frequency during weekdays in the Fall of 2013. The day begins with little activity, except for stations nearby Penn station. Usage frequency increases in Manhattan towards the financial district and midtown around 9am, decreasing again only late in the night. Also, stations near Broadway are shown to be popular destinations, as expected. There is some increase in the Brooklyn downtown area, but not as intense as in Manhattan.

While balance derived data can tell a lot about commuters circulation, trip data add distance, duration and origin–destination information. We used the trips timeline and OD matrix to explore and relate both balance and trip data. With the timeline showing trips as glyphs, we could easily find patterns in the direction of trips of the stations. In Fig. 15 the timeline of trips shows how the station at E17 & Broadway has a constantly high flow of incoming trips coming from every direction, while the map shows the trip directions in different time intervals. Fig. 16 now relates the arrival patterns of two neighbor stations. As the incoming flow at 1 Ave & E30 St. station decreases, the flow at Station 2 Ave & E30 St. increases proportionally.

As observed before, there is a significant difference in station usage during weekdays and weekends. During weekends, there is no clear partitioning of the city into regions of contrasting behavior. However, an evident pattern is the increase of cyclic trips. In Fig. 17, we inspect Sunday trips in the Summer of 2014. Stations are ranked by the average amount of cyclic trips between 9am and midnight. We observe that cyclic trips happen more frequently around leisure spots such as the Central, Battery, and Brooklyn Bridge Parks. Another difference between weekdays and weekends is the places where outages occur. As can be seen in Fig. 18, outages happen during weekdays in the upper north stations in Midtown, East Village and a region of Brooklyn. Over the weekends, outages happen in Midtown and Williamsburg.

A different, trip-centered, outage analysis can be done using the OD Trips Matrix. By aggregating day intervals in three defined periods of times we can show different behaviors during the day. We considered four-hour long intervals: 6am–10am (morning), 10am–2pm (lunch), and 5pm–9pm (evening). To identify the highest values, we used the automatic selection with a percentage range of 80–100%, which means that only the highest 20% will be selected. Similarly, for the lowest values, we used the 0–20% range. Balance difference (BD) was the selected variable to inspect because it can be used to identify problematic trips, i.e. trips that contribute to outages in the system, taking a bike from an almost empty station to another one almost full. In Fig. 19 we used the OD trip matrix to find such trips, the ones with high BD. The opposite kind is also shown, with lowest BD values: full stations at the origin and empty stations as the destination. March 2014 was used for this test, showing that evening trips from the middle region to the north or south are more guaranteed to avoid outages.

6. Discussion

Our prototype led to several insights in the usage of the Citi Bike program for the first 10 months of use, demonstrating the potential to analyze Bike-Sharing Systems in general. The solution can be applied to other similar BSSs, where datasets represent the state of stations and trips between them, as a problem of flow of commodities in a graph. The 10-month timeline (Fig. 8) showed how to follow commuters behavior as the program matured and the city underwent weather changes (design requirement **R4**). The same trends can be found by looking for patterns in different periods using the 24-h view (Fig. 9). The straightforward analysis that followed was the comparison among different days of the week, same days of the week in different periods, and weekdays against weekends. Anomalies were easily identified in specific days and hours as sudden changes of color in the different matrix views as shown in Fig. 10. Due to the regularity of the anomalies, lack of spatial correlation among affected stations, we concluded that they were related to operational activities. A possible cause could be special events in the city at the given periods of time, however, we could not find such case in the news that would explain the outliers.

Another problem is finding stations that fit a given criteria, such as empty or full outages (design requirement **R1**). We showed that brush-wise reordering of the rows in the matrix by different variables was useful to identify locations on the map that exhibit the behavior we are looking for. We found full and empty stations during working hours on the map in Fig. 12(a) and (b); identified stations with almost no balance change in usual days in Fig. 12(c); showed stations that kept the highest amount of bikes in Fig. 12(d) and found unexpected changes in the capacity of stations comparing Fig. 12(d) and (e). More than helping in the search of patterns, the reordering brushes was essential to create time lapses that show the evolution of rankings through time. Figs. 13 and 14 showed how bikes moved among different city regions (associated with the requirement of finding station roles **R3**), and the system usage in a typical working day. We observed changes in popularity among the regions of Manhattan, Brooklyn, and Williamsburg in Fig. 18 by looking at the station rankings for the frequency of outages in both weekdays and weekends.

Seasonal trends were clearly observed, with a considerable decrease in the rate of commuting during cold months (even

though it was still surprisingly high given the harsh weather). We found no explanation for the anomalies identified by relating them to unusual events in the city (design requirement **R2**), resting on the assumption that those were caused by operational issues. Finally, the results validated the hypothesis that data from bike-sharing can be used to provide many cues of the city lifestyle, and that our proposal is adequate for this analysis.

7. User scenarios

We believe that the main contribution in this work is more than the individual visual designs or the user interactions described in Section 4, but instead their combination into an integrated tool suitable for the analysis of Citi Bike data. We list below possible scenarios of users that can benefit from our system:

- *Scenario* 1 – *BSS administrator*: The administrator can use the station state timeline matrix with partial re-ordering to forecast a group of top ranked stations' outages, and use this information to prioritize staff re-balancing assignments. This analysis heavily depends on the day of the week and the time of the day considered. The flexibility of the interface combined with the user interface offers the user the ability to process the different groups of top ranked stations generated in different time searches. As a result, the administrator can assign rebalancing staff and resources on a weekly, monthly, or seasonal basis.
- *Scenario* 2 – *Rebalancing researcher*: In Section 2.1 we list papers that propose re-balancing algorithms for BSS's. Most of these algorithms work with real data from BSSs, but generate simulated data that forecast alternative re-balancing scenarios. Our system could be used to further inspect such simulated scenarios. For example, deletion or insertion of stations can be simulated in these algorithms, and our tools can be used to inspect the impact of the results at specific stations or group of stations.
- *Scenario* 3 – *Big data analyst*: Our system offers a flexible way to inspect trips (e.g. the timeline matrix or trips OD matrix). The patterns of trip movements can be used by a big data analyst to find correlations and customer preferences. This analysis can be used to improve marketing along important routes, suggest revenue opportunities, among other business benefits.

8. Conclusion

Operating a bike-sharing system deployed in a big city is a challenging task due to the intense commuting dynamics and its complexity. In such scenario, when the number of outages increase, rebalancing requires more effort as the system is usually larger (more stations and bikes) and the rebalancing fleet is subject to traffic jams. This is even more important when the system provides 24-h service and rebalancing must be done on-the-fly. A deeper understanding of the system dynamics may help the operational efforts to provide a better service. For this purpose, we introduced visualization designs to support the exploration of a dataset with station usage footprint and user trips from NYC Citi Bike. Data can be visualized for specific days, or aggregated over different time periods. The designs we proposed introduced different matrix visualizations of the data, combined with a brushing interface that created partial re-orderings of the data. We presented results to validate the applicability of the proposed solution over a 10-month long period, which can be used to justify the adoption of the bike-sharing as a new transportation mode. Even though we did not introduce a solution to rebalancing, we believe that our analysis tool may improve rebalancing schemes by adding instance-specific knowledge to current solutions.

Acknowledgments

The authors wish to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper. This work was supported in part by an IBM Faculty Award, Moore-Sloan Data Science Environment at NYU, NYU Tandon School of Engineering, NYU Center for Urban Science and Progress, AT&T, NSF awards CNS-1229185, CCF-1533564 and CNS-1544753, CNPq processes 140983/2011-2, 246197/2012-9, 449555/2014-3 and 308851/2015-3.

References

[1] Citi bike 2013 summary. URL ⟨http://www.nycbikemaps.com/spokes/citi-bike-2013-summary/⟩; 2013.
[2] Rainer-Harbach M, Papazek P, Hu B, Raidl GR. Balancing bicycle sharing systems: a variable neighborhood search approach. In: Middendorf M, Blum C, editors. Evolutionary computation in combinatorial optimisation – 13th European conference, EvoCOP 2013. Lecture notes in computer science, vol. 7832. Springer; 2013. p. 121–32. ⟨http://link.springer.com/chapter/10.1007%2F978-3-642-37198-1_11⟩.
[3] Papazek P, Raidl GR, Rainer-Harbach M, Hu B. A PILOT/VND/GRASP hybrid for the static balancing of public bicycle sharing systems. In: Computer aided systems theory – EUROCAST 2013. Springer; 2013. p. 372–9. ⟨http://link.springer.com/chapter/10.1007%2F978-3-642-53856-8_47⟩.
[4] Raidl GR, Hu B, Rainer-Harbach M, Papazek P. Balancing bicycle sharing systems: improving a VNS by efficiently determining optimal loading operations. In: Blesa MJ, et al. editors. 8th international workshop on hybrid metaheuristics, HM 2013. Lecture notes in computer science, vol. 7919 . Springer; 2013. p.130–43. ⟨http://link.springer.com/chapter/10.1007%2F978-3-642-38516-2_11⟩.
[5] Schuijbroek J, Hampshire R. Inventory rebalancing and vehicle routing in bike sharing systems; 2013.
[6] Urli T. Balancing bike sharing systems (bbss): instance generation from the citibike nyc data. CoRR 2013; abs/1312.3971.
[7] Guenther MC, Bradley JT. 20th international conference on analytical and stochastic modeling techniques and applications, ASMTA 2013. Proceedings, Ghent, Belgium, July 8–10, 2013. Springer, Berlin, Heidelberg; 2013. p. 214–31.
[8] Froehlich J, Neumann J, Oliver N. Measuring the pulse of the city through shared bicycle programs. In: International Workshop on Urban, Community, and Social Applications of Networked Sensing Systems (UrbanSense08), 2008. ⟨http://www.citeulike.org/user/tnhh/article/3385555⟩.
[9] Froehlich J, Neumann J, Oliver N. Sensing and predicting the pulse of the city through shared bicycling. In: Proceedings of the 21st international joint conference on artificial intelligence. IJCAI'09, 2009. p. 1420–6.
[10] Borgnat P, Abry P, Flandrin P, Robardet C, Rouquier JB, Fleury E. Shared bicycles in a city: a signal processing and data analysis perspective. Adv Complex Syst 2011;14(03):415–38.
[11] O'Brien O, DeMaio P. The bike-sharing world map. ⟨http://www.bikesharingworld.com⟩; 2009.
[12] Zaltz Austwick M, O'Brien O, Strano E, Viana M. The structure of spatial networks and communities in bicycle sharing systems. PLoS ONE 2013:8.
[13] Wellington B. Mapping citi bike's riders, not just rides.⟨http://iquantny.tumblr.com/post/81465368612/mapping-citi-bikes-riders-not-just-rides⟩; 2014.
[14] Kaufman S. Citi bike and gender. ⟨http://wagner.nyu.edu/rudincenter/2014/05/citi-bike-and-gender/⟩; 2014.
[15] Beecham R, Wood J, Bowerman A. Identifying and explaining interpeak cycling behaviours within the London cycle hire scheme. In: Workshop on progress in movement analysis: experiences with real data, 2012. p. 15–6.
[16] Wood J, Beecham R, Dykes J. Moving beyond sequential design: reflections on a rich multi-channel approach to data visualization. IEEE Trans Vis Comput Graph 2014;20:12.
[17] Wood J, Slingsby A, Dykes J. Visualizing the dynamics of London's bicycle hire scheme. Cartographica 2011;46(4):239–51.
[18] 5.5 million journeys at nyc bike share. ⟨http://oobrien.com/2014/04/5-5-million-journeys-at-nyc-bike-share/ ⟩; 2011.
[19] Ferzoco J. How new yorkers and tourists use citi bike on two nice days. ⟨http://ny.curbed.com/tags/jeff-ferzoco⟩; 2014.
[20] Hurter C, Ersoy O, Fabrikant S, Klein T, Telea A. Bundled visualization of dynamic graph and trail data. IEEE Trans Vis Comput Graph 2014;20(8):1141–57.
[21] Ferreira N, Poco J, Vo HT, Freire J, Silva CT. Visual exploration of big spatio-temporal urban data: a study of New York city taxi trips. IEEE Trans Vis Comput Graph 2013;19(12):2149–58.
[22] Wang Z, Lu M, Yuan X, Zhang J, Wetering Hvd. Visual traffic jam analysis based on trajectory data. IEEE Trans Vis Comput Graph 2013;19(12):2159–68.
[23] Guo D, Chen J, MacEachren AM, Liao K. A visualization system for space–time and multivariate patterns (vis-stamp). IEEE Trans Vis Comput Graph 2006;12(6):1461–74.
[24] Citibike system data. ⟨http://www.citibikenyc.com/system-data⟩; 2014.