TransitTrails: Visualizing Congestion & Coverage Changes

Caitlin Bonnar **Camille Cobb** **Katie Kuksenok** **Yi Pan**

ABSTRACT

We created an interactive visualization intended to depict the effect of route cuts in Puget Sound over a year's time (specifically, March 2011- March 2012), and allow exploration of congestion data by day of week and time of day. To support this task, we designed and implemented an interactive visualization depicting historical OneBusAway data, which consist of individual bus positional check-ins every ten seconds and their deviation from schedule at that time and location. Our final visualization includes a side-by-side comparison of points on a map of Puget Sound. Points represent a location where at least one bus checked in at within a time interval specified by the user. The color of each point represents the average deviation of buses that checked in at that location. We conducted eight user studies with transit experts and regular bus riders in order to evaluate and test our design and elicit feedback about how well we supported these goals. We then performed iterative enhancement to improve interactions and layout. Although we believe that the available data were not sufficiently complete to warrant more detailed exploration, the product we created accomplishes the goals of facilitating exploration of congestion and coverage differences resulting from transit changes and serves as a useful starting point for future visualizations once data collection has been improved.

INTRODUCTION

The goal of this project is to use real-time historical bus data to support understanding of congestion patterns (measured by bus delays) and transit coverage, and how this changes at different times of day, on different days of the week, or in response to transit cuts and bus schedule or route changes. This information can help transit decision-makers understand the potential implications of proposed future changes.

The large volume and incompleteness of data (thousands of bus check-ins per day, many without geo coordinates or other relevant information) created several challenges and tradeoffs for the final visualization. For example, plotting every single bus check-in as a point on the map represents the data truthfully but occludes points on top of each other and makes it difficult to interact with each point.

This visualization is intended to depict the effect of route changes in Puget Sound over a year's time (specifically, March 2011 though March 2012) and allow exploration of congestion data by day of week and time of day.

To support this task, we designed and implemented an interactive visualization depicting historical OneBusAway data, which consist of individual bus positional check-ins every ten seconds and their deviation from schedule at that time and location. Our final visualization includes a side-by-side comparison of points on a map of Puget Sound. Points represent a location where at least one bus checked in within a time interval specified by the user. The color of each point represents the average deviation of buses that checked in at that location.

We conducted a formative suer study with eight participants, spanning both transit experts and regular bus riders in order to evaluate and test our design and elicit feedback about how well we supported these goals. We then performed iterative enhancement to improve interactions and layout. Although we believe that the available data were not sufficiently complete to warrant more detailed exploration, the product we created accomplishes the goals of facilitating exploration of congestion and coverage differences resulting from transit changes and serves as a useful starting point for future visualizations once data collection has been improved.

RELATED WORK

There are many related projects in the space of communicating transit accessibility. The urban data (urban-data.herokuapp.com) visualization displays an animated visualization of bus routes and demands (denoted by passengers-in and passengers-out) for a given route and a given day. This visualization also uses historical deviation data (in this case, the actual time a bus arrived at a particular stop) to hint at congestion during a specific day and time. We chose to convey backed-up buses in a different manner, by statically plotting bus positions for a specified time and coloring them by deviation, in order to better correlate time of day and geographic position.

In order to better evaluate the convenience of living in a specific place, Brandon Martin-Anderson et al computed transit scores (bmander.com/portofolio), besides walk scores (Carr et. al 2010) and bike scores. By taking into account of the distance to the nearby public transit stops, the frequency and type (rail, bus, etc.) of transit routes, they calculated a normalized transit score for each specific location on the maps of five U.S. cities, where public transit data is available. While transit score can help real estate agencies and normal people find places and neighbourhoods they love, it fails to consider a finer-grained traversability of multiple locations through focusing on one specific point on the map.

METHODS

Storyboards

We started by creating storyboards (Figure 1) based on the data we anticipated would be available to us (i.e., where each bus was every ten seconds and which route it was on at the time).

Our initial ideas focused on showing bus movement as animated trails overlaid on maps using trail color to encode congestion (lateness) information. We envisioned supporting comparison by showing multiple maps at once and allowing users to specify a day of week (weekday; weekend), time of day (morning commute; evening commute; non-peak hours), and date (December 2011; March 2012 -- which were before/after bus changes) for each map separately. We hoped to support filtering by bus route and highlighting individual buses.

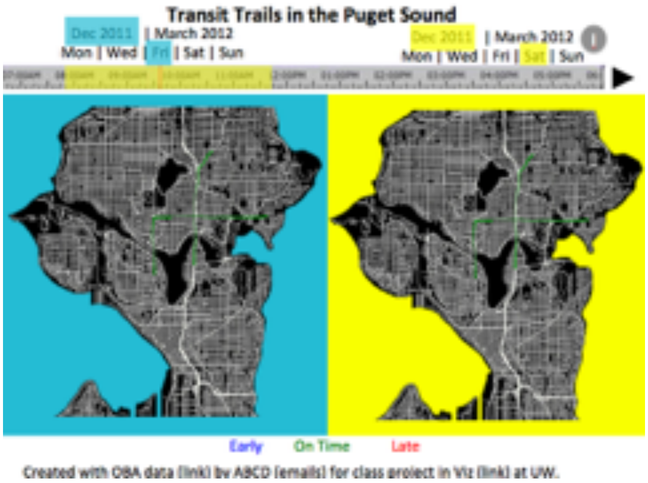

Figure 1. Initial high-fidelity sketch.

Data Wranglin

The most challenging part of the technical process was curating the data, and figuring out how much of it was suitable to use. The data files were stored as tab-delimited text files on a server, many of which had null values or miscategorized agency or trip IDs (attributes which were crucial to joining with static Google Transit Feed Specification (GTFS) data, which includes information about bus stops, bus routes, and schedules. After several attempts, we found that including route numbers would cause us to lose up to 90% of the data.

As an alternative, we also examined data collected more recently; however, the current versions of the data collection scripts also do not collect good data. It was not within the scope of this project to fix the data collection; this has many complexities and will take significant engineering effort, time, and, possibly, coordination with the agencies responsible for making this data available in the first place. We modified our design to account for this incomplete data. For example, we decided that we could not provide functionality for filtering buses by route.

We selected several days of data to visualize -- a weekday, Saturday, and Sunday during each of three different periods surrounding transit changes (March 2011, December 2011, and March 2012). This allowed us to choose days where the data were less sparse, which we hope provide a reasonably realistic picture of general transit patterns. To add validity to comparisons between congestion and coverage on the days selected, we also chose days that had similar weather and no large city-wide events or holidays.

IMPLEMENTATION

In our progress presentation, we asked for suggestions on good mapping APIs that would support vector overlays and allow us to dynamically plot points based on latitude and longitude. Leaflet, a Javascript library that could be embedded easily into HTML, was highly recommended. We created an HTML page that instantiated two Leaflet maps side by side and simple Javascript functions to plot and remove points based on user-specified options (HTML radio buttons and a JQueryUI slider).

Interaction with filters and the slider triggered data fetching into the maps from an API that we implemented for a database hosted on a remote server. This API draws from a database of pre-curated data (to improve efficiency). As we developed the system, we realized that we needed to do something to cut down on the amount of data being transferred upon instantiation, because load time for the data was very long. We did this by (1) randomly sampling 2000 data points from each day and (2) combining bus check-in points that were within a small distance of each other. Although we recognize that this is potentially misleading (for example, if day 1 only had 2000 total points vs. day 2 has 10,000 data points, we show a much larger portion of data from day 1). In spite of the potential drawbacks of this data sampling methodology, this change was necessary in order to achieve reasonable loading times for the final visualization.

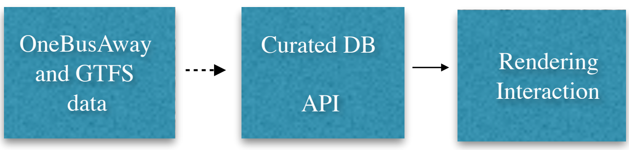

Figure 2. System architecture. In order to reduce the overhead of loading and processing raw data client-side, we implemented a server-side database and custom API for delivering the congestion data we successfully wrangled out. The database on the server contained a curated subset of the total data, and helped improve visualization startup. The added overhead of the network API call, however, necessitated random sampling.

While the Leaflet API worked well for importing data and creating overlays, it did not support our initial ideas of creating animated bus “trails” with the opacity of the trails fading over time. For the sake of this short-term project, it was suitable to plot points and use colors, locations, and number of points to communicate the congestion and coverage information.

Because there were problems with occlusion when plotting multiple busses that checked in at the same location during the same time interval, we decided to cluster points in the

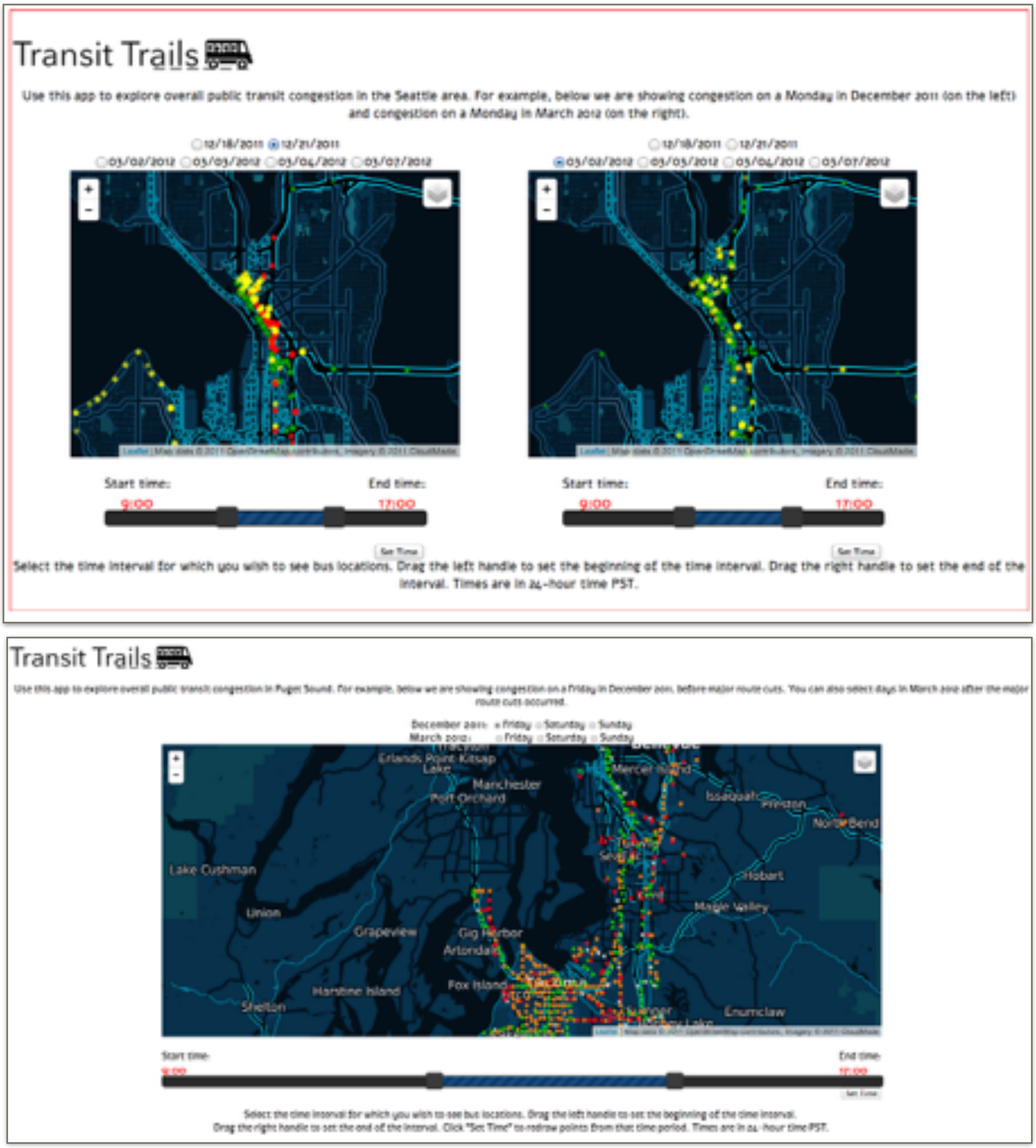

Figure 3A shows implemented prototype A (top) and an alternative implemented prototype B (bottom).

data processing stage rather than handle clustering client-side. We computed the average deviation from schedule of the clustered buses and assigned the plotted point a color according to the average deviation. This decision has the tradeoffs of avoiding occlusion at the cost of losing information when check-in locations are dense.

We created two working prototypes for use in a small user study, described in the following section. Both prototypes allowed users to select one of several dates (using radio buttons) and a time range (using a timeline with two movable sliders). The data feeding these prototypes was incomplete because we chose to perform a join to obtain route information, as described above. We were able to plot all available data this way (since there were a max of 2000 data points per day); however, check-in points were much more sparse in Seattle than in Tacoma, which seems to be due to the mislabeling of most of King County Metro's trip data. Both showed the same type of data on Leaflet maps, which have pan and zoom functionality and allow users to choose a light or dark color scheme for the underlying map.

Prototype A (Figure 3A): Two maps side by side, each with separate date and time selection and separate panning and zooming. Dots were colored green if buses checked in on time or early, yellow if slightly late, and red if more than 5 minutes late. Dates are listed in a "DD/MM/YYYY" format, without specifying the day of week. Dots each represent just one bus check-in, so occlusion of nearby dots is a problem; however, clicking a dot shows the bus's route number, if it is available.

Prototype B (Figure 3B): One map, which requires users to change the data selection in order to make comparisons. Dots are green if on time, blue if early, orange if slightly delayed, red if more than 5 minutes delayed, and grey if the delay is unknown. Bus check-ins are clustered with other nearby check-ins; thus, average deviation is shown when dots are clicked, but route numbers cannot be shown. Dates are listed by day of week.

User Study and Design Iteration

We designed and carried out a small user study to get feedback on the usefulness of these prototypes. We conducted short (~30 minute) semi-structured interviews with 8 participants (4 male, 4 female; 4 transit experts, 4 regular transit riders). We asked (1) about familiarity with OneBusAway, (2) about familiarity with the February 2012 transit changes and whether these cuts impacted them, (3) what questions they would want to answer with access to the type of data that OneBusAway uses.

All participants were familiar with OneBusAway. Participants had a wide range of familiarity with the transit changes: some were able to name specific ways in which routes and schedules changed, some were aware of but unaffected by the changes, and some had no previous knowledge of the changes.

Participants who were transit experts were much more interested in changes to the overall picture of buses in the Puget Sound. Regular riders were mostly interested in using this data to answer questions directly relevant to their daily lives.

We asked participants to use and explore the two prototypes. We took notes of (1) which features and interactions seemed unintuitive, difficult, or unapparent to participants; (2) participants' specific suggestions for improvements; (3) participants' preferences for certain features that differed between the two prototypes; and (4) other interesting or unexpected insights gained during these interviews. For example, we noted when participants found an intriguing pattern, when our system spurred participants to ask new and interesting questions, or when participants considered how they might use this system differently if data collection were better.

This study exposed improvements we could make to our system. These insights allowed us to iterate on our design and improve the final product.

FINAL DESIGN

The final code for our system can be accessed on GitHub at https://github.com/CSE512-14W/fp-cobbc12-cbonnar-pany5-kuksenok or used as an interactive visualization at http://homes.cs.washington.edu/~cbonnar/viz/busviz.html.

The final version of our system is similar to the Prototype A, described in Section 3c -- it uses colored dots on two maps to show bus check-in locations colored based on delay and allows users to specify a date and time range for each map. We found that comparison was an important attribute to our study participants, and more readily supported our goal of visualizing route cuts. However, we incorporated several features from Prototype B and made several design improvements based on feedback from the user study and class poster session.

Features that we incorporated from Prototype B include: (1) clustering geographically close check-in points; (2) a selector for which day of data uses day of week rather than date (to differentiate better between the three types of typical bus schedules and facilitate cross-comparison); (3) a different color for early and very early, instead of merely on time and late.

Changes based on user feedback include: (1) a diverging, more visually appealing color scheme for dots; (2) maps can be synchronized for panning and zooming; (3) defaulting to the "light" map color scheme (we elected to keep the "dark" scheme for accessibility reasons); (4) map color schemes are called "light" and "dark" rather than "day" and "night"; (5) data selection updates when time slider handle is released rather than requiring users to click "set time" button; (6) added a legend for dot color meaning; (7) improved layout of explanatory text; (8) icon to indicate when data is loading; (9) larger maps that take up more screen space.

Our participants had differing opinions on what colors should be used (whether they should be the same as OneBusAway colors or not, and whether early and late should be colored similarly or differentiated). In the end, we chose a balanced diverging color palette with colors that related more to traffic congestion than OneBusAway arrival times (recent user studies on OneBusAway's colors hint that they are largely ineffective; also, we believe they serve a different purpose, where an early bus is red because it is "worse" than a late one in terms of having to make it to the stop in time). Early buses are represented as blue in order to differentiate them from on-time and late buses, which also brought out an interesting problem in the data: some bus ids seem to be mislabeled, perhaps switched with the bus that is supposed to follow after them on the same route.

DISCUSSION

The small user study we conducted as well as interactions with users during our class poster session revealed useful insights and feedback for the different iterations of our design. Beyond specific feedback that we incorporated into design improvements for the final system (discussed in Section 4), many participants expressed interesting ideas and examples of how this system could provide new insights. Some participants envisioned other features that

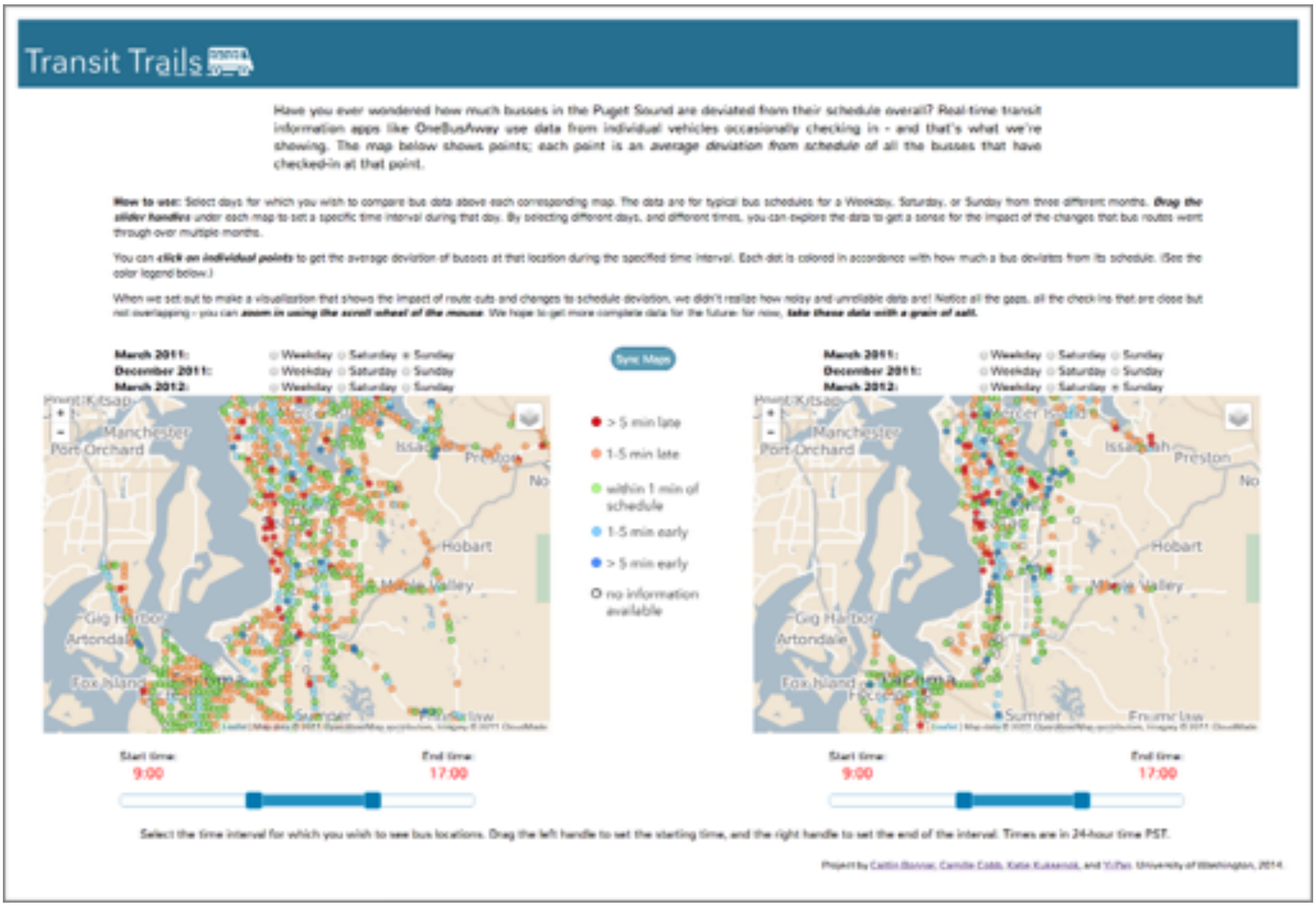

Figure 4. A screenshot of the final implementation of the congestion-comparison visualization, TransitTrails.

could improve this system if it could be built on a more complete dataset.

Both transit experts and regular transit riders were interested in using this visualization to explore aspects of their personal transit experiences (for example, zooming to their neighborhood or usual bus routes). Many users were interested in exploring their commute times or peak vs off-peak hours; our system supported this task reasonably well, though incomplete data was problematic, especially since most participants live in Seattle where the data used for prototypes was more sparse. Several participants noted that they would like to view the frequency of buses on the same route, and the times of day when this frequency changes: "If I don't have OneBusAway, like my phone battery is dead, how long would I have to wait?" Although this information could have been generated from GTFS schedule data, the incompleteness and inability to include route information in our visualization made this task impossible to support at this time. Some participants found interest beyond their current transit situation, for example knowing whether an apartment they are looking at would be convenient to buses that are relatively on time: "oh I want to live here, but there's no coverage." Another participant noted that it would be interesting to see transit patterns on big football game days or during other city-wide events.

FUTURE WORK

The next step for this system is to show it to King County Metro officials. In addition to providing them with an interesting new visualization of their data, we anticipate feedback leading to useful system improvements.

Another future direction is adding additional data as layers. For example, information about average economic status throughout the city could facilitate understanding of how transit cuts impact the citizens who rely most heavily on buses. Although buses that serve these areas may not have the highest ridership and thus be likely candidates for cuts, economic factors and the value of supporting equal access may be reason to reconsider. A simpler additional layer that many participants suggested they would appreciate is bus stop locations.

The clustering of nearby bus check-ins into a single point is something for which we would like to further explore. For example, we could plot one larger point to represent several smaller points and separate the points when users zoom in.

We could, alternatively, consider showing dots colored based on the median or largest deviation at that location.

The most important improvement, however, relates to improvement of the underlying data. This will require revisions to data collection scripts and additional data collection. With more complete data collection, it will be worthwhile to reconsider and explore many of our initial design ideas, like filtering by routes, showing how frequently buses run along each route, or revisiting the idea of showing animated "trails" for each bus (this would require a change from the Leaflet API or creating our own plugin for Leaflet). We also hope to explore ways of incorporating live bus data to enable comparison of current vs. historical congestion and coverage information.

REFERENCES

[1] B. Ferris, K. Watkins, and A. Borning, "OneBusAway: results from providing real-time arrival information for public transit," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2010, pp. 1807–1816.

[2] B. Ferris, K. Watkins, and A. Borning, "OneBusAway: A Transit Traveler Information System," in Mobile Computing, Applications, and Services, Springer, 2010, pp. 92–106.

[3] Seattle Transit Blog. "A Closer Look At Metro's Cuts: Seattle and North King County." Nov, 2013. Accessed Feb 2014. http://seattletransitblog.com/2013/11/08/a-closer-look-at-metros-cuts-seattle-and-north-king-county/

[4] L. J. Carr, S. I. Dunsiger, and B. H. Marcus, "Walk scoreTM as a global estimate of neighborhood walkability," American journal of preventive medicine, vol. 39, no. 5, pp. 460–463, 2010.

The columns on the last page should be of approximately equal length.