Zbyněk Stara (zs633@nyu.edu)
Monday, February 1, 2016

Design Study of New Techniques for Traffic Visualization
A Capstone Proposal

Abstract: The field of traffic visualization relies heavily on time series plots and geospatial maps. Little work has been done, however, to examine the merits of these two approaches, and to investigate the possible suitability of alternative visualization techniques for the field. This design study aims to identify new visualization approaches and to perform usability and complexity evaluations of those new and conventional approaches. Three additional visualization techniques are identified to provide broad variety in the design study: phylogenetic trees, cartograms, and treemaps. The visualization algorithms will be evaluated along three dimensions: influence on decision-making of users, modality, and complexity. To further validate the findings of the design study, data from three traffic data sets will be used: New York City, Massachusetts State, and Ireland. The output of the study will be a comparison of the five new and conventional visualization techniques; in particular, the identification of where each new technique excels, and how it weighs up against conventional methods.

1. Specific Aims:

When visualizing its findings, the field of traffic modeling relies heavily on time series plots and geospatial maps. Little work has been done, however, to evaluate the merits of these visualization techniques in view of user experience. At the same time, the merits of other visualization techniques have not been evaluated; therefore, an unorthodox visualization technique might prove superior to the conventional approaches, at least along some of the dimensions of evaluation.

This design study seeks to provide tangible evidence of the effectiveness of different visualization techniques as applied to traffic modeling. It is the aim of this study to ***make usability and complexity evaluations of conventional visualization as well as new techniques from the broader field of data visualization***.

Specific steps to achieve that aim are outlined in Objectives; Approach lists the aforementioned new visualization techniques to be considered by this study; expected results of evaluation of each technique are presented in the Hypotheses.

2. Background:

Visualization is, at its core, the science of representing problems and their solutions according to a logical structure that may or may not be immediately apparent. Visualization techniques seek to identify constituent elements of the problem, codify them as points in one or multiple data sets, and represent the data in a visual form that is intelligible for human readers.

The nature of this visual representation varies significantly depending on the identified relationships among data points. However, what matters is not the specific visualization technique *per se*, but the nature of the information that may be gleaned by the observer from the data. The most successful visualizations are not those that merely illustrate

the problem and/or its solution, but those that reveal information that might not have been accessible without the visual representation.

An important measure of the effectiveness of the visualization is whether it helps influence the readers' behavior. This is especially applicable in the field of traffic visualization and congestion prediction; a better visualization may allow users to avoid highly congested areas. However, the traffic modeling literature does not illustrate much variety in visualization approaches; most of the employed techniques fall into two broad categories.

One group uses time series plots. In the AITVS model, Lu, Boedihardjo and Zheng plot traffic volume, speed, and occupancy in a series of line graphs and a heatmap (167), as illustrated in Figure 1, below. Both of the graphing techniques allow one to compare speed, volume and occupancy of a stretch of road at different times of day and days of week. In the two techniques, color is semantically important, but it draws attention to different features of the two projections. In the line graphs, color distinguishes different days of the week; in the heatmap, color illustrates different levels of traffic occupancy. The traditional choice of colors green, yellow, and red conveys the meaning of low/middle/high occupancy, borrowing somewhat from the color scheme of traffic lights. An important feature of the model is the hierarchical organization of the time and space dimensions – data can be plotted in different levels of increasing granularity both in time (hours vs. days. vs. months vs. years) and in space (station vs. county vs. freeway vs. region)

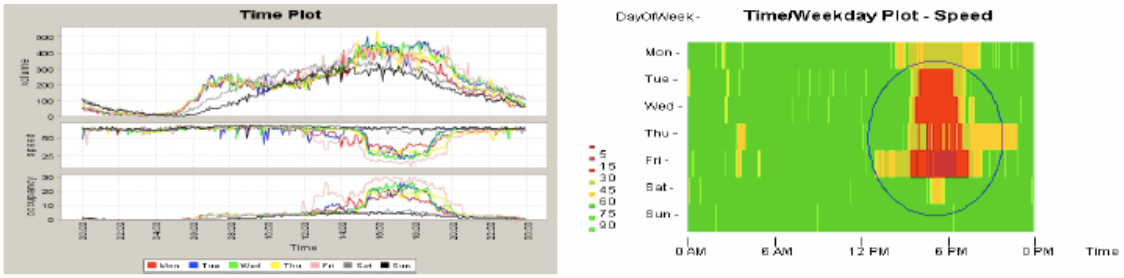

Figure 1: Line graph and heatmap visualization as presented by Lu, Boedihardjo, and Zheng (167)

The second broad group uses topographical maps to relay visual information to users. JamBayes, the traffic information and prediction system developed by Horvitz et al., uses a visual interface that assigns colors to road segments in a simplified map of Seattle road system according to current traffic congestion (4), and can be seen in Figure 2, below. The information is supplemented with clock illustrations to communicate expected time delay in more detail. Furthermore, question marks overlaying the clock illustrations are used to warn of potentially unreliable predictions, while exclamation marks are used to alert the user to potentially surprising traffic conditions, as predicted by the model.

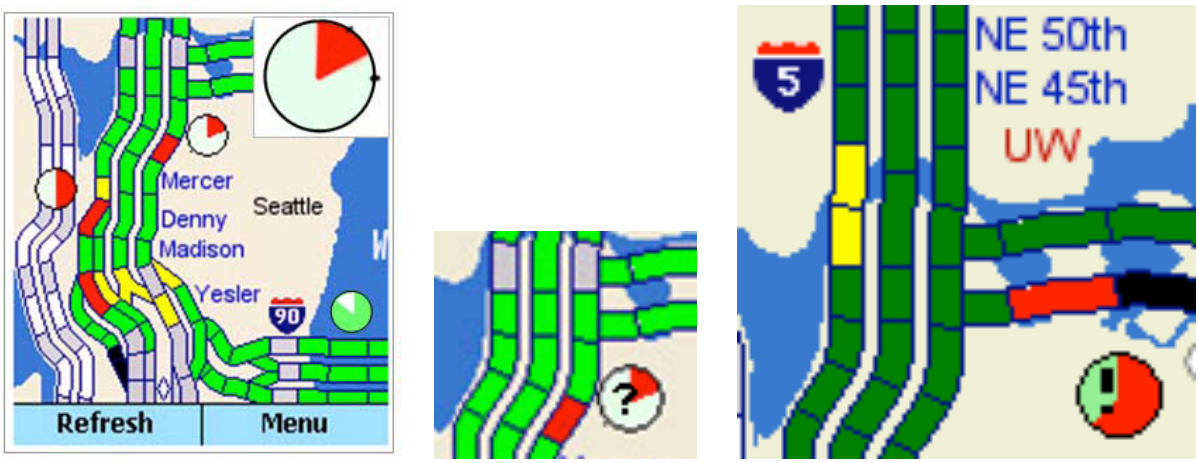

Figure 2: **Map visualization approach used by Horvitz et al. (4), including the additional features**

The most advanced traffic models combine the two visualization techniques to paint a more comprehensive, yet traditional, picture of the traffic situation. The traffic model of Wang et al. – presented in Figure 3, below, features a citywide traffic jam density map and provides traffic density heatmaps for all segments of the road network (2165).

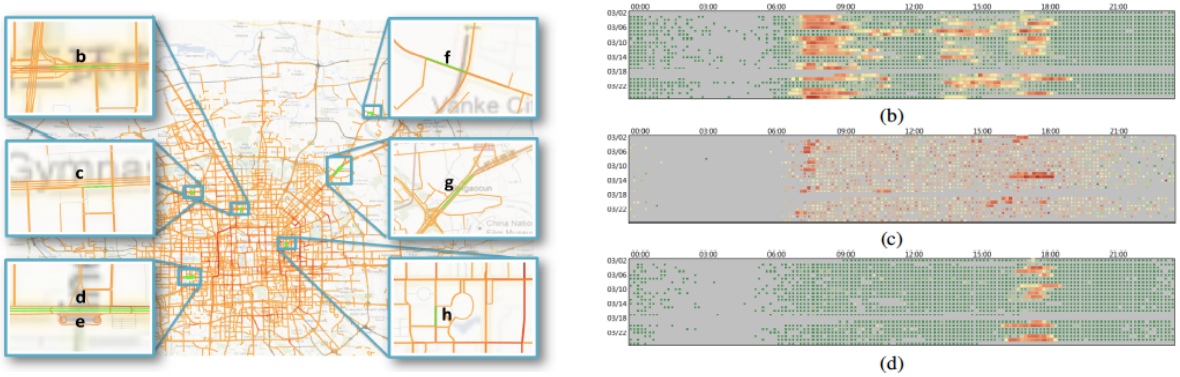

Figure 3: **A combination of a map and heatmap interface has been employed in a model by Wang et al. (2165)**

Meanwhile, TripVista by Guo et al. visualizes the "microscopic traffic data at a road intersection" (163). As can be seen in Figure 4, provided below, a topographical approach is used to map the movement of traffic across the junction, alongside a time series plot of traffic volume. Additionally, various features of the data are displayed in a parallel coordinates plot.

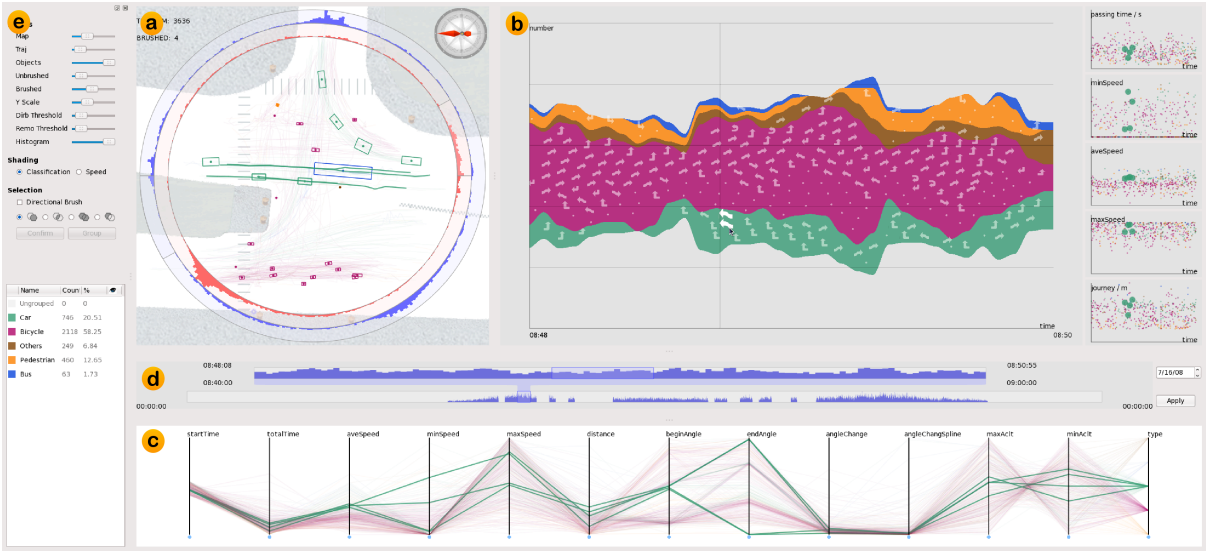

Figure 4: TripVista by Guo et al. Combines topographical approach with a time series plot and a parallel coordinates plot (163)

As the two last examples illustrate, it cannot be said that the field of traffic visualization is not evolving. It does, however, seem limited by the two *modi operandi*, broadly defined above. It is the aim of the proposed Capstone research to explore the usefulness of other visualization techniques – outlined below – as applied to traffic. Precisely because this field has not previously been investigated in depth, it has the potential of yielding new approaches that may provide useful fresh insight into the problem.

3.1. Objectives:

1. ***Identify and develop new methods of traffic visualization.*** A preliminary list may be found in the Approach section of this proposal.

2. ***Modify existing traffic data to work within the framework of the algorithms of the visualization approaches.*** This includes the newly identified visualizations as well as the conventional visualization approaches outlined in the Background section of this proposal.

3. ***Present the visual output of the investigated visualizations to subjects in a user study.*** Based on the results of the study, each of the visualizations will be evaluated along the following three dimensions:
    - Does the visualization succeed in ***influencing the decision-making process*** of the users of the traffic model?
    - Does the visualization succeed only in ***specific modes of use***? For example, is a particular approach more suited for mobile devices as opposed to desktops or tablet devices?
    - What is the ***complexity*** of generating the visualization? How many resources does it require?

Several challenges have been identified as potential obstacles on the way to attain the stated objectives:

- The quality of the data sets may not hold sufficient information to be useful – no matter the visualization used.
- The quality of the data sets may be inconsistent. Different data sets may provide different amount of information, and at different levels of time granularity. To assure fair comparison in the user study, this may require recourse to the lowest common denominator of the data sets, potentially prompting the aforementioned problem of ata quality.
- The benefit of each of the proposed new visualizations may bring to the investigation has to be precisely identified, to avoid "application bingo" (Munzner 140). All design decisions have to be carefully evaluated and, above all, validated, throughout the design study.
- Precise implementation of all of the diverse visualization techniques will be required to ensure the successful execution of the project. Indeed, one of the dimensions of evaluation of the visualization algorithms is algorithmic complexity and performance.

3.2. Preliminary Data:

The traffic model underlying this project will be supplied with three primary sources of open data; each of these includes two distinct but equally necessary varieties of data. The first is a network of road *segments* – akin to the one illustrated in Figure 5 – for which traffic variables are monitored, and the second is a set of discrete *readings* of the values of those variables at different timestamps.

The first source of data is the one provided by the New York City Department of Transportation for a network of major roads in the five boroughs of the city. This is the primary set for the investigation since it has been monitored for the longest time; the most traffic readings have been collected for this set. The data consists of 151 traffic *segments*, where a segment is a collection of latitude and longitude coordinates forming a polyline. As can be seen in Figure 5, data for different directions is stored as different segments; however, no directionality information is provided, and thus will have to be inferred circumstantially from the relative position of the endpoints of each individual traffic segment. The *readings* offer data on time needed to traverse each segment, as well as average speed of traffic flow. Each of the readings in the data set provides information for a particular segment only; there is no way to interpolate the data for subdivisions of that segment, however useful that information could be for analysis.

The second data source – from Massachusetts Department of Transportation – provides traffic data for a network of highways in the State of Massachusetts. It differs from the New York data set in the following three aspects:

- It is more comprehensive than the one of New York, as it spans 229 segments.
- The segments in the data set are, in general, shorter than those of New York City, relative to the area covered.
- The readings contain information about the travel speed corresponding to free flow, while this has to be inferred indirectly in the New York data set.

These three differences might prove advantageous for the construction of the different visualizations methods; they can be used to present data that is both more detailed and accurate than that of New York. However, these effects may be counterbalanced by two

shortcomings of the Massachusetts data. First, its readings have been collected for a shorter time than those for the New York data set; second, it describes a set of highways with relatively few segment intersections compared to New York. This means fewer interrelations in the visualizations, and possibly fewer insights these might offer.

The last data source is the National Roads Travel Times data set provided by Ireland's Open Data Portal. Compared to the other two data sources, it is the smallest – it only has 61 segments –, but it provides the most metadata about the segments, and the most traffic information in its readings. This includes not only the travel time and the free flow travel time, but also the normally expected travel time. This data set provides the most interrelations between segments, however the small overall number of traffic segments and the short period of data collection restricts the amount of useful information available for the model's analysis.

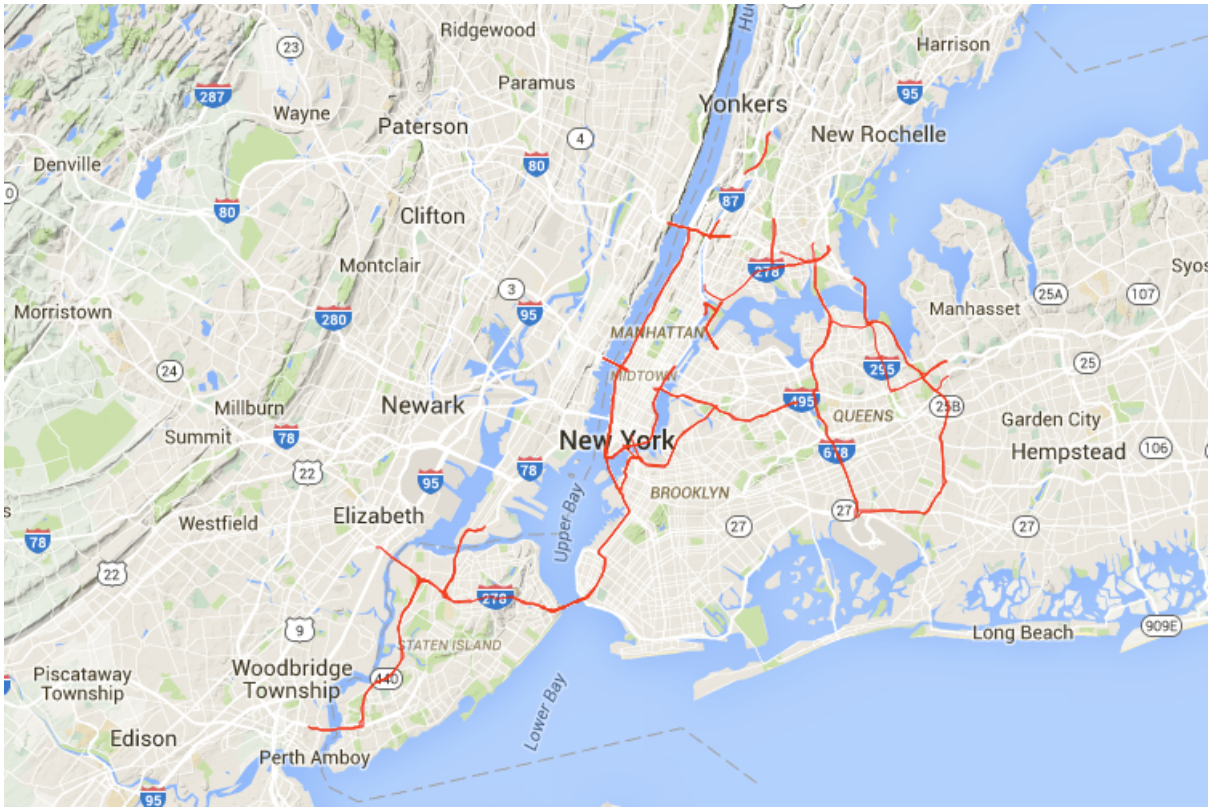

Figure 5: **A topographical plot of the data set provided by New York City Department of Transportation. In the figure, all 151 monitored road segments are presented. A live version of this visualization may be accessed at http://www.zbynekstara.info/capstone.**

3.3. Approach:

Three visualization techniques were identified as candidates for investigation. Each of these visualizations focuses on a different aspect of the data; the techniques were chosen to offer broad variety of visualization approaches to present in the user study. These will be provided with the outputs of the traffic model based on the preliminary data outlined in the previous section.

The first visualization technique is the phylogenetic tree, presented in Figure 6. As Fitch and Margoliash explain, this is a binary tree based on arranging data points according to the amount of changes necessary to bring it close to other data points (279). The most ready example is the example of git commits. These records only store the minimal changes necessary to arrive at the final data point (text, in this case) starting from the original. Phylogenetic trees were used in research, too, however, in topics ranging from molecular biochemistry to origami.

When a dissimilarity metric is applied to the outputs of the traffic model – for example, "Starting with a given time step, how closely does the development of traffic volumes and congestion copy the development in other traffic segments?" –, a phylogenetic tree may thus be constructed, one that records the hypothetical succession of changes in the data set. Thus, possibly hidden interrelations might be revealed in the resulting arrangement of tree nodes. Then, if one knew that all branches at one side of a tree are congested, one would opt to use a different set of nodes of the phylogenetic tree, one that indicates the cluster is less congested individually and overall.

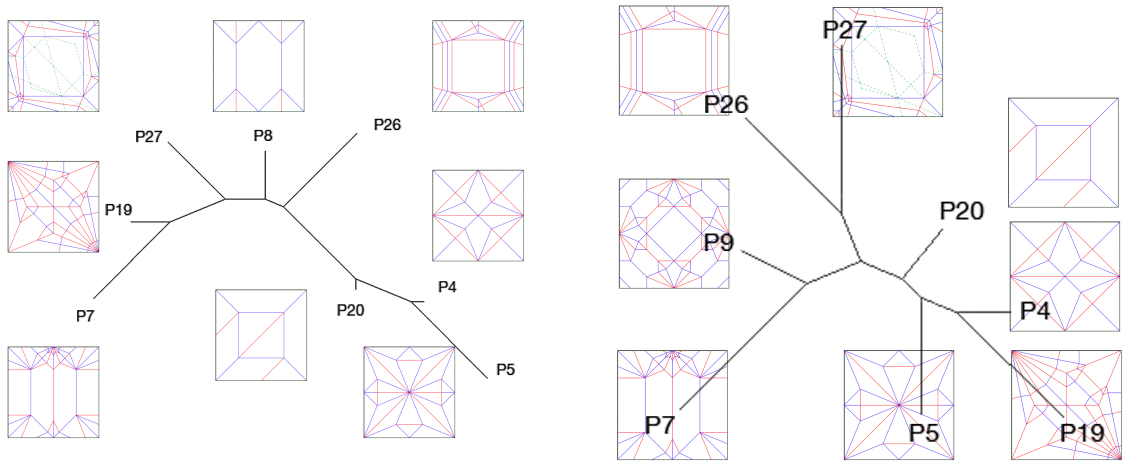

Figure 6: **Two phylogenetic-tree reconstructions of origami crease patterns, found in Oh et al. (390, 393). Patterns close to each other in this visualization are more similar than those farther apart.**

The second technique is the technique of the cartogram, elaborated upon significantly by Gastner and Newman (7502), and presented in Figure 7. Cartograms begin with an ordinary map, however, the visualization algorithm then substitutes topographical accuracy for an illustration of values of a variable of interest – population density, for example. The map is distorted based on the visualized variable; the higher the value, the more prominent its position becomes in the visualization. Conversely, if the value of the variable is relatively small in a given location, its area in the cartogram is shrunk proportionately. The technique represents the information in an easily decipherable, pseudo-graphical, manner, as long as features of the original map are discernible in the cartogram and able to supply visual clues to the reader.

This visualization would use the traffic data to determine the degree to which a particular traffic segment should be emphasized. The more congested a segment is, the more

it would distort geography around itself; drivers would be able to quickly identify the worst routes, and opt instead for a detour through the least emphasized areas. An additional assumption is built into this visualization – when there is heavy traffic at a particular route, traffic is expected to be heavy at neighboring roads as well. A congested road would exercise influence over the whole of surrounding geography, succinctly suggesting that neighboring roads are likely to be congested even though they are not actually part of the traffic data set.

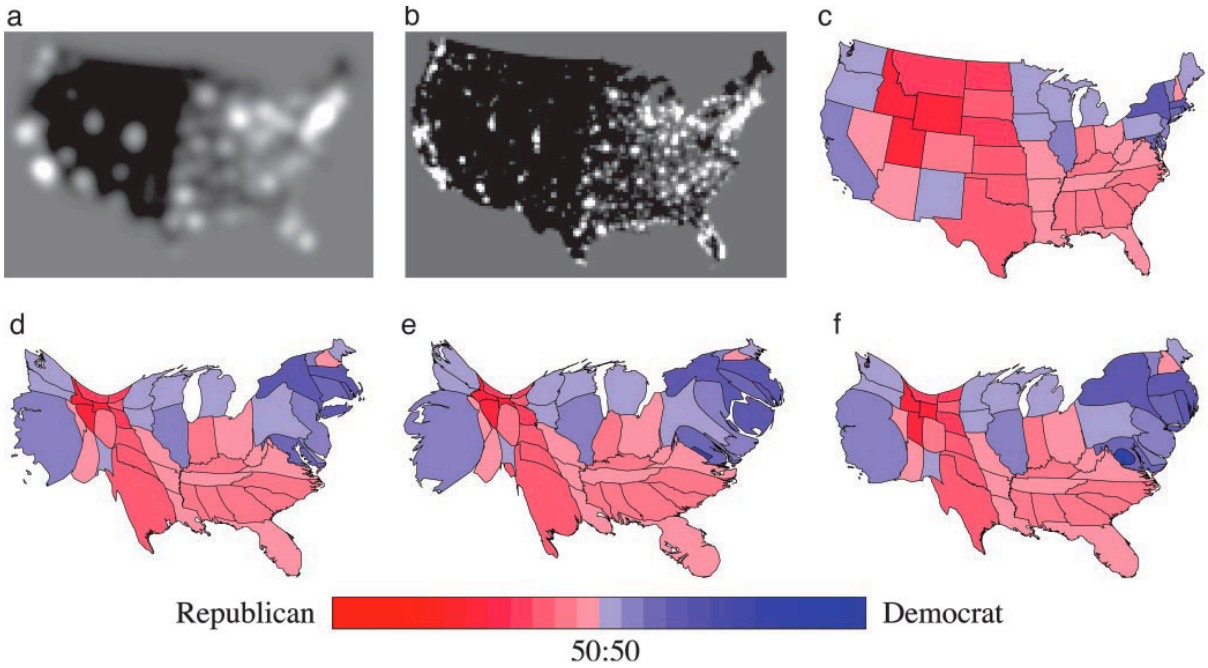

Figure 7: Cartograms illustrating the results of the 2000 US presidential election, from Gastner and Newman (7502). A map of the United States (c) is distorted according to population distribution data at two different levels of granularity (d, e), and according to the distribution of electors in the US Electoral College (f).

The final promising technique is treemapping, introduced by Johnson and Shneiderman (284) and improved by Bruls, Huizing and van Wijk (33). The technique, presented in Figure 8, represents nodes of trees as rectangles, while child nodes are represented as rectangular subdivisions inside their parent rectangles. The algorithm progresses recursively through the tree nodes, producing increasingly smaller subdivisions as the tree depth increases.

The treemapping technique is ordinarily used to represent hierarchical data, but there are two ways in which it can be applied to the traffic data set. The first is to utilize the hierarchical nature of the variables used – hours vs. days vs. weeks vs. months vs. years and station vs. county vs. freeway vs. region, to illustrate with the decisions of Lu, Boedihardjo and Zheng (167) – to show congestion data. For example, each day of the week might be represented by a rectangle, with the days that experience the most congestion the biggest. Each one of the *day* rectangles might then have a subdivision for each traffic segment, with the most congested segments having the biggest rectangles. The user might then determine the most congested areas by seeking the biggest rectangles, and avoiding the corresponding real-life areas.

The second option is to construct the tree map as an alternate visualization of a phylogenetic tree. Starting with such a phylogenetic tree, each split would represent a new split in the parent rectangles. A tree map would be produced, such that adjacent rectangles would represent traffic segments that are the most similar in terms of congestion developments. Thus, the user would be able to know that if a particular traffic segment is congested, other segments – the ones that are close to the congested segment in the tree map – are likely to be congested too, and adjust their path accordingly.

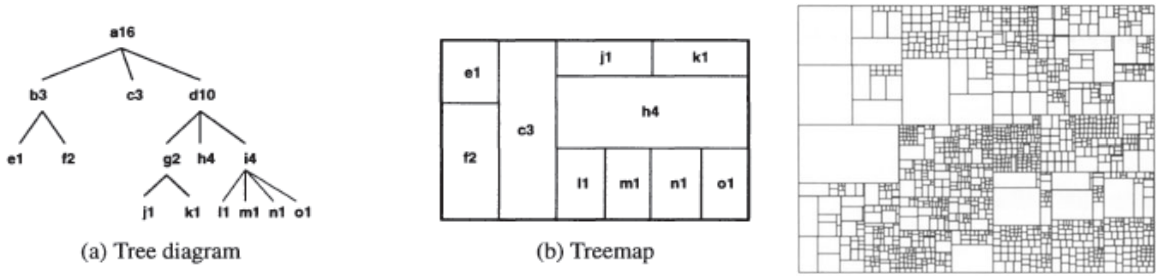

Figure 8: Illustrations of treemaps from Bruls, Huizing and van Wijk: A treemap with corresponding tree diagram (34), and a complex treemap (39). The authors acknowledge that graphical aids – not shown here – are often necessary for viewers to understand the underlying hierarchy (39).

To evaluate the different visualization approaches – including the traditional ones described in the Background section of this paper, and the alternative ones described in this section – a user study will be performed.

The user study participants will be interviewed individually. The group of study subjects will be split into ***six*** groups; subjects in different groups will work with data from a different pair of data sets. The first study to be presented to each participant will alternate.

Participants will be presented with each of the visualizations separately. The order in which the visualizations will be presented will be different for each participant. With the visualization at hand, each participant will be asked to perform the following tasks. Unless specified otherwise, the participant will be able to view the visualization in each of the three modes of presentation – at a mobile phone, at a tablet, and at a desktop computer.

1. ***Identify the three most congested segments of the traffic network*** based on insight gleaned from the visualization at hand. The accuracy of the three guesses relative to actual traffic situation would be recorded as a proxy metric and used to measure success of the visualization in influencing the decisions of the user.
2. ***Identify the three times of day in which the traffic network, as a whole, seems the most congested.*** The accuracy of the three guesses relative to actual traffic situation would be evaluated in the same fashion as discussed in Task 1.
3. ***Plan a route through the system*** based on what the user perceives to be normal traffic. After performing the task, the user will be asked to rank the visualizations based on perceived influence on decision-making process. For each of those, the user will also be asked to determine the medium, found to be the most helpful.
4. ***Plan a detour in the system based on actual traffic*** at that moment in a simulated driving environment – which means without access to a desktop system. The user would then be asked to evaluate the visualizations similarly to the first task.

After completing each task, participants will be asked to identify the mode of presentation they found to be the most helpful in informing their decision.

Finally, after having been presented with all visualizations, the participants will be asked to subjectively evaluate which visualization had the most influence on their decisions.

3.4. Hypotheses:

The hypothesized results of this design study may be plotted in a table, presented below as Table 1. The columns hold the dimensions of evaluation, while the rows hold the different visualization techniques that are the subject of the design study.

	Influence on Decision-Making Process	**Modality**	**Complexity of Generation**
Map	*High*	*Mobile*	*Simple*
Time Series Plot	*High*	*Desktop*	*Simple*
Phylogenetic Tree	Moderate	Desktop	Moderate
Cartogram	High	Mobile	Hard
Treemap	Moderate	Tablet	Moderate

Table 1: The hypothesized outputs of the study. Modality refers to the mode of use in which the users are expected to find particular visualization technique the most useful.

It is presumed that the two visualization techniques that are currently the most used in traffic model visualization research – maps and time series plots, emphasized by italics – will prove the most optimal overall. – After all, that is why they are already so widely used for traffic visualization; at the same time, sharp differences are expected to arise among the other visualization techniques. The aim of the design study is to quantify these differences, and to identify the situations in which each might be more appropriate than either of the two conventional approaches.

Maps and time series plots are expected to exercise relatively the most influence on decision-making process of users, simply by virtue of their proliferation in various areas of human activity; they do not need to be explained. Maps are expected to be perceived as more useful at mobile platforms, by virtue of providing real-time traffic situation. On the other hand, time series plots – which include line graphs, parallel coordinates plots, and heatmaps – are expected to be more useful in a desktop setting, as they hold more general insights about the traffic situation, informing the general route selection, but do not provide ready insight into the current congestion along that route. Both of the conventional visualizations are simple to generate – data can be presented to the user with minimal modifications.

Most of the alternative visualization approaches are expected to exert less influence on decision-making processes, by virtue of the reduced familiarity of the general population with these methods. The only exception is the cartogram, which communicates its findings through the well-known language of topography.

Similarly, the alternative visualization methods are not expected to prove the most useful in the mobile medium because they provide summaries of that same long-term data, rather than real-time information. The cartogram is expected to be the mobile exception due to its inherent similarity to the mapping visualization technique. At the same time, the treemap is expected to be the most useful on tablet devices, as the visualization is expected to be quite expansive and/or require zooming – and in those situations, the extra screen space of tablets relative to the mobile screens is expected to be important.

Complexity of generation is expected to be the primary bane of the alternative techniques, as the data produced by the traffic model would need to be extensively adapted before proving useful to users. The two related techniques of the phylogenetic tree and treemap require the data to be clustered according to a similarity metric, at the same time, the cartogram requires distortion of underlying map image – the results cannot be simply plotted on a Google Map, for example.

4. Potential Impact and Summary:

According to Arnott and Small, "time spent ensnarled in traffic is not simply time wasted; for most of us, it is time miserably wasted." They offer a simple calculation to estimate the magnitude of the problem. About one third of all driving in metropolitan areas takes place in congested conditions, during which average speed decreases to half of the traffic segments' free-flow value. Even without considering the cost of additional fuel, accidents, air pollution, and other losses due to congestion, the economic finding that drivers are willing to spend "about 1.33 USD to save 10 minutes [of] travel time" puts the annual cost of driving delays in the United States at 48 billion USD (446).

If it were possible to influence the traffic decisions of individual drivers, they could choose routes that are more efficient, based on current and long-term traffic situation in the area. It is important to identify the traffic visualization methods that allow drivers to make optimal decisions quickly and confidently in a variety of scenarios and in a variety of settings.

The proposed design study strives to quantify the potential benefits of five techniques of traffic visualization. In doing so, it seeks to compare the traditional approaches of using maps and time series plots to three promising alternatives, and thus to illuminate the most efficient modes of communication of traffic data for specific purposes and in specific situations. In doing so, this study hopes to help effect change of decisions of individual traffic agents – reducing the immense losses of time and money that could be spent much efficiently in most any other way.

5. Budget and Justification:

User studies are the core of this design study. Two iterations are expected, the second one to build upon the experience of the first, to provide space to fix any procedural shortcomings of the first iteration as they might be discovered during data analysis. The cost per hour of volunteers' time is based on the general rates for NYU Abu Dhabi research studies.

User Study – First Iteration		
Item:	**Quantity:**	**Total:**
Compensation to volunteers (50 AED/hour) (30 volunteers expected)	30	400 USD
	Total	**400 USD**

User Study – Second Iteration		
Item:	**Quantity:**	**Total:**
Compensation to volunteers (50 AED/hour) (30 volunteers expected)	30	400 USD
	Total	**400 USD**

An Apple Developer account would be necessary to test the apps to present the visualizations in the mobile and tablet environments, two of the three modes of presentation required in the user study.

Apple Developer Account		
Item:	**Quantity:**	**Total:**
One-year subscription	1	100 USD
	Total	**100 USD**

The remaining budget would be spent on attending 1 to 2 conferences on Information Visualization to acquire the most recent knowledge related to the field as well as get to know researchers with whom collaborations can be built.

Conference 1 – IEEE PacificVis, April 19-22, 2016		
Item:	**Quantity:**	**Total:**
Student week pass	1	500 USD
Roundtrip flight tickets to Taipei, Taiwan	1	800 USD
Accommodation (40 USD/night)	5	200 USD
	Total	**1500 USD**

Conference 2 – IEEE InfoVis, October 23-28, 2016		
Item:	**Quantity:**	**Total:**
Student week pass	1	500 USD
Roundtrip flight tickets to Baltimore, USA	1	500 USD
Accommodation (50 USD/night)	5	250 USD
	Total:	**1250 USD**

6. Ancillary Travel and Collaboration:

Conference travel:

- 2016 IEEE Pacific Visualization Conference, Apr 19-22, 2016, Taipei, Taiwan
- 2016 IEEE Information Visualization Conference, Oct 23-28, 2016, Baltimore, USA

Works Cited

Arnott, Richard, and Kenneth Small. “The Economics of Traffic Congestion.” *American Scientist* 82.5 (1994): 446-55. Print.

Bruls, Mark, Kees Huizing, and Jarke J. van Wijk. "Squarified Treemaps." *Data Visualization 2000: Proceedings of the Joint EUROGRAPHICS and IEEE TCVG Symposium on Visualization.* By Willem Cornelis de Leeuw, and Robert van Liere. Vienna, Austria: Springer, 2000. 33-42. Print. *Eurographics.*

Fitch, Walter M., and Emanuel Margoliash. "Construction of Phylogenetic Trees." *Science* 155.3760 (1967): 279-84. Print.

Gastner, Michael T., and M. E. J. Newman. "Diffusion-based Method for Producing Density-equalizing Maps." *Proceedings of the National Academy of Sciences* 101.20 (2004): 7499-504. Print.

Guo, Hanqi, et al. "TripVista: Triple Perspective Visual Trajectory Analytics and its application on microscopic traffic data at a road intersection." *Visualization Symposium (PacificVis).* Hong Kong, China: IEEE Pacific, 2011. 163-70. Print.

Horvitz, Eric J. et al. "Prediction, Expectation, and Surprise: Methods, Designs, and Study of a Deployed Traffic Forecasting Service." *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence (UAI2005).* Arlington, VA, USA: AUAI, 2005. 1-8. Print.

Johnson, B., and B. Shneiderman. " Tree-maps: a space-filling approach to the visualization of hierarchical information structures." *Visualization, 1991. Visualization '91, Proceedings., IEEE Conference on.* San Diego, CA, USA: IEEE, 1991. 284-91. Print.

Lu, Chang-Tien, Arnold P. Boedihardjo, and Jinping Zheng. "AITVS: Advanced Interactive Traffic Visualization System." *Proceedings of the 22nd International Conference on Data Engineering (ICDE’06).* Atlanta, GA, USA: IEEE, 2006. 167. Print.

Munzner, Tamara. "Process and Pitfalls in Writing Information Visualization Research Papers." *Information Visualization: Human-centered Issues and Perspectives.* By Andreas Kerren, John T. Stasko, Jean-Daniel Fekete, and Chris North. Vol. 4950. Berlin, Germany: Springer, 2008. 134-53. Print. *Lecture Notes in Computer Science.*

Oh, Seung Man, et al. “A Dissimilarity Measure for Comparing Origami Crease Patterns.” *Proceedings of the 4th International Conference on Pattern Recognition Applications and Methods (ICPRAM).* Lisbon, Portugal: SCITEPRESS, 2015. 386-93. Print.

Skog, Tobias, Sara Ljungblad and Lars E. Holmquist. "Between aesthetics and utility: designing ambient information visualizations." *Proceedings of the IEEE Symposium on Information Visualization 2003 (INFOVIS’03).* Seattle, WA, USA: IEEE, 2003. 233-40. Print.

Wang, Zuchao, et al. "Visual Traffic Jam Analysis Based on Trajectory Data." *IEEE Transactions on Visualization and Computer Graphics* 19.12 (2013): 2159-68. Print.