Reliability of San Francisco-Bay Area Road Network

Chung Yu Wang
Stanford University
chungyuw@stanford.edu

Hsin-Fang Wu
Stanford University
hfwu@stanford.edu

Alice Yeh
Stanford University
aliceyeh@stanford.edu

1 Introduction

Road network is an important infrastructure in society. On average, every American drives 13,476 miles a year, spending over 200 hours on traveling. Out of the 200 hours, 38 hours are wasted during congestion, wasting $121 billion on fuel cost a year.[1,2] Thus, a well-designed road network, which leads to less congestion and thus have significant economic ramifications. Unlike many other networks, road networks are generally static, yet the travel times across edges change constantly, resulting in dynamic traffic patterns. Our group is interested in how congestion occur in real road networks, where a road network is defined as a network of highways for our investigation. To gain an understanding in this field, we reviewed literature related to road networks, with topics ranging from the terminologies used to theoretical modeling of network robustness to game theoretic traffic patterns. Ultimately, we are interested in how traffic changes when the graph structure is altered by removing critical nodes.

To answer this question, we used a Bay Area Road Network dataset, which is discussed in more detail in Section 3. We modeled the road network by having roadways as edges and intersections as nodes. In this network, we first identified the critical nodes using 3 different measures: Betweenness Centrality, Mean Closeness Centrality, and Max Closeness Centrality (defined in Section 4.1).

Then we simulated how traffic would flow through the area. Finally, we investigated how route assignments of a given set of traffic are impacted after the removal of critical nodes. From our analysis, we found that travel time increased by around 50% when we removed 400 critical nodes. This increase in travel time is acceptable based on a 2009 study done by (Snelder et al., 2009).

[1] http://www.cbsnews.com/news/clogged-roads-cost-americans-121b-in-wasted-time-fuel-in-2011-report/

[2] http://www.theatlantic.com/business/archive/2013/02/the-american-commuter-spends-38-hours-a-year-stuck-in-traffic/272905/

2 Review of Prior Literature

In previous work on the topics related to road networks, we specifically reviewed those that discussed characterization of road networks, simulation of traffic, and game theory application of traffic patterns on road networks.

In order to better understand typical characteristics of road networks, we examined two papers. In (Jiang and Claramunt, 2004), the authors claimed that large urban street networks formed small-world networks but exhibited no scale-free property. This claim would be challenged later (Kalapala et al., 2006). A fundamental difference, however, lays in how road networks were represented. While Jiang & Claramunt modeled the networks with intersections as nodes and streets as edges (i.e., called primal representation), Kalapal et al. specifically highlighted that road networks would be better represented with streets as nodes and intersections as edges (i.e., called dual representation). Kalapal et al. found the road network had a heavy tail degree distribution that was well characterized by a power law distribution with an exponent $2.2 \leq \alpha \leq 2.4$. For our investigation, we modeled the road network using primal representation.

In learning more about traffic simulation, we read two papers: (Dekker and Colbert, 2004) and (Easley and Kleinberg, 2010). Whereas Dekker and Colbert examined road networks from a systems perspective, focusing on investigating the networks robustness based on node connectivity vs. link connectivity and discussing link load behavior, Easley and Kleinberg reasoned about traffic patterns in a road networks modeled as a directed graph from a game theory perspective.

Dekker and Colbert provided a framework for analysis of network reliability. They argued that node connectivity (the smallest number of nodes whose removal resulted in a disconnected or single-node graph) was more appropriate for modeling the robustness of network topologies on possible node destruction than link connectivity (the smallest number of links whose removal resulted in a disconnected graph). They concluded that the most robust network were optimally connected, which meant that the node connectivity was as high as possible.

Easley and Kleinberg, on the other hand, discussed traffic patterns in a directed graph using concepts from game theory. Two key concepts in the chapter were *Nash equilibrium*, defined as the traffic pattern in which no driver had any incentives to switch to a different route and *socially optimal*, defined as the traffic pattern that resulted in the minimum total travel time for all drivers. Easley and Kleinberg showed that for any given network, a Nash equilibrium could be proven to always exist, which could be reached by iteratively updating the best response for each driver.

Overall, the models discussed in the papers we have read leaned more toward the theoretical side, which bore further investigation on real-life networks. The basis of our research project, thus, became an investigation of using a physical road network and then observing how simulated network traffic respond to removal of critical nodes.

3 Data Collection Process

Our original data source originated from a research group at University of Utah[3]. The data was clean but too limited for modeling the network with traffic. However, we found sufficient traffic and network information from OpenStreetMap, a crowdsourced map website[4], which provided a node ID for every intersection, edges between the nodes, and the associated geographic (longitude, latitude) coordinates, which allowed us to plot our result later for analysis. To work with this dataset, we used ArcGIS (a Geographic Information System) software for processing. We only kept the major highway routes, since most highways are consistently present. In addition to this general highway classification, there are three subclasses of highways, primary, secondary and tertiary. Upon checking the classification of each road type via Google Map, we organized the highways into the following categorizations and found the corresponding speed limit. With such classification, we calculated the base capcity of roads based on the definitions provided by U.S. Department of Transportation Federal Highway Administration[5]. Table 1 summarized the road capacities.

Number of Nodes	10,417
Number of Edges	18,940
Avg. out-degree (miles)	1.818
Avg. Distance (miles)	18.82
Diameter (miles)	74.844

Table 2: SF-Bay Area Summary of Network Statistics

We further changed the network into a directed network and only extracted out the largest strongly connected components (SCC). Table 2 summarized some important network statistics of the San Francisco-Bay Area Road Network.

In addition to the road network data, we collected the population data for each city in our network in order to generate random source and destination pairs based on population distributions. However, we later found similar results between simulations based on pairs generated randomly and pairs generated based on population distribution. Thus, for all reported simulation, only randomly generated pairs were used.

The data collection process was mostly successful, but we noticed the collected data was not entirely clean, resulting in some highways having only one direction. This limited us from doing a comprehensive analysis as will be discussed in section 8, but we were still able to analyze the network reliability in detail with realistic results.

4 Modeling & Algorithm

To find the shortest path between nodes, we used Dijkstraś algorithm. The flexibility of the algorithm enabled us to find the shortest paths based on different measures, ranging from distance to time with/without traffic.

4.1 Critical Nodes Measures

To identify critical nodes, we used 3 different measures: *Betweenness Centrality*, *Maximum Closeness Centrality*, and *Mean Closeness Centrality*.

[3]http://www.cs.fsu.edu/ lifeifei/SpatialDataset.htm

[4]https://mapzen.com/metro-extracts/

[5]http://www.fhwa.dot.gov/ohim/hpmsmanl/appn1.cfm

Type	Number of Lanes	Speed Limit(miles/hour)	Base Capacity (cars/hour)
Primary	4	65	(1700+65*10)*4 = 8200
Secondary	3	50	(1700+50*10)*3 = 6600
Tertiary	2	35	(1700+35*10)*2 = 4100

Table 1: Highway Classification

For each of the measure, we picked out the top 400 nodes with the highest values.

4.1.1 Betweenness Centrality

In calculating Betweenness Centrality, we used the following expression,

$$B(u) = \sum_{s \neq u \neq d} \frac{Path_{s,d}(u)}{Path_{s,d}} \quad (1)$$

where $Path_{s,d}$ is the number of shortest paths from node s to node d and $Path_{s,d}(u)$is the number of those paths that pass through node u in Equation 1.

To get the exact Betweenness Centrality for each node, however, was computationally expensive. Therefore, we selected 1,500 pairs of source and destination nodes randomly and identified the shortest paths based on distance to calculate each nodeś Betweenness score.

4.1.2 Mean Closeness Centrality

We defined the mean closeness centrality value for each node v to be the inverse of the mean shortest distance found from v.

$$C_{mean}(v) = \frac{1}{mean(min(distance(v,u))} \forall u \in G \quad (2)$$

However, due to computational complexity, we randomly selected 20 destination nodes from graph. To avoid repeatedly selecting the same destination nodes, we first checked that the subsequently selected destination nodes were not already visited along the paths of the previously chosen destination nodes.

4.1.3 Maximum Closeness Centrality

We defined max closeness centrality value for each node v to be the inverse of the longest shortest distance found from v. Unlike the Mean Closeness Centrality measure, Max Closeness Centrality was an experimental metric for centrality that we conceived for our investigation.

$$C_{max}(v) = \frac{1}{max(min(distance(v,u))} \forall u \in G \quad (3)$$

Similar to the computation for the Mean Closeness values, the computation for the Maximum Closeness values also used 20 unvisited destination nodes.

4.1.4 Comparison of Different Critical Node Measures

We plotted our findings in Figures 1, 3, and 2. The red dots denote the *top most* important set of critical nodes, the orange dots denote the *second most* important set of critical nodes, the green dots denote the *third most* important set of critical nodes, and the purple dots denote the *fourth most* important set of critical nodes. Across all three different measures, the Bay Bridge was consistently chosen as the most critical section of our road network. In Figure 1 in particular, we originally postulated that the Bay Bridge would be the most important followed by the other bridges (San Mateo Bridge, Dumbarton Bridge, Richmond-San Rafael Bridge). However, with some surprise, we found that the 101 route along the left side of the Bay (in the Redwood City, San Mateo region) actually had higher Betweenness scores. This suggested that in our 1,500 randomly picked routes, there were more pairs that sought to travel between the North Bay and the South Bay than between the West Bay to the East Bay. On the other hand, for the two Closeness measures in Figures 3 and 2, respectively, we saw that the critical nodes were mainly all concentrated on and connecting to the Bay Bridge with Figure 3 having a more scattered distribution.

Analysis of how different measures of identifying the critical nodes affected route re-assignments is discussed later in Section 6.1.

4.2 Traffic Simulation Modeling

4.2.1 Modeling

To model the traffic pattern of 4 million cars[6] in the Bay Area, we ideally would like to assign a route for each car by iteratively choosing the best response for the car given the current traffic pattern

[6]http://www.mtc.ca.gov/maps_and_data/

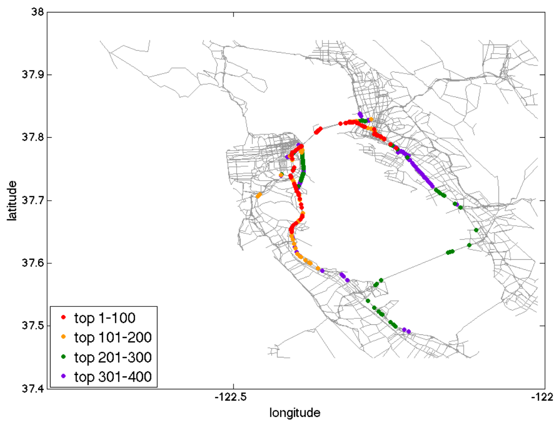

Figure 1: Top 400 Critical Nodes-Betweenness

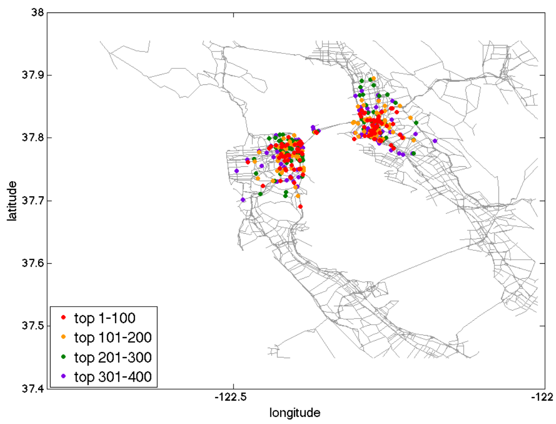

Figure 2: Top 400 Critical Nodes-Mean Closeness

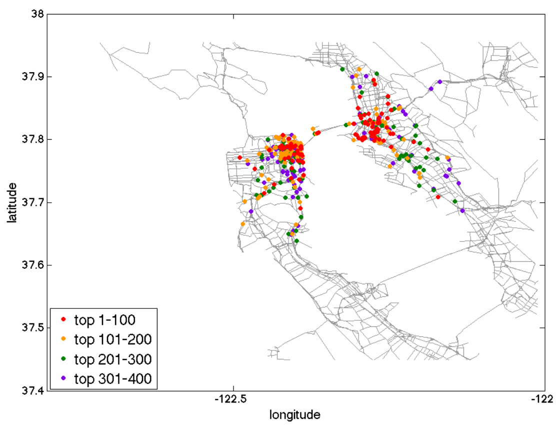

Figure 3: Top 400 Critical Nodes-Max Closeness

until reaching Nash Equilibrium. However, this would be computationally prohibitive, and therefore instead, we iteratively chose the best response for a batch of cars in every iteration. The number of cars per batch decreased and the number of iterations for re-assignment increased as we refined toward our final route assignment. Our final batch size is 1000 cars, with 4 iterations of re-assignment for the final batch size.

To model the travel time as a function of traffic volume, we used link congestion function from the Bureau of Public Roads (BPR) (Davis and Xiong, 2007). This function is commonly known as the PBR function and is of the following form:

$$T(q) = T_o(1 + \alpha\frac{q}{c})^\beta \tag{4}$$

where T_o is the free-flowing travel time (edge distance divided by speed limit), q is the volume, c is the capacity, and α and β are congestion parameters (where $\alpha = 0.15$ and $\beta = 4$ is fitted using traffic data referenced from (Davis and Xiong, 2007)).

In simulating traffic and understanding how the overall system cost of the Bay Area Network, we defined 3 different scenarios, Time-based, Distance-based, and Traffic-based. As their namesake suggest, each scenario seeks to model the cost based on different measures.

Time-Based System selects optimal path from source node to destination node based on shortest time travelled. Note that each edge has a different speed limit and length and thus the total travel time is the aggregate of all the edges distance divided by its speed limit (free flow travel time). Selection of path is independent of others batches whereabouts and based solely on path with the smallest aggregate time.

Distance-Based System selects optimal path from source node to destination node based on shortest distance traveled. Selection of path is independent of others cars' whereabouts and based solely on shortest path found.

Traffic-Based System selects optimal path from source node to destination node based on smallest travel time, calculated from Equation 4. Note that for this scenario, every batch's route assignment depends on other batches' route assignments.

4.2.2 Convergence

Finding route assignments was an iterative process, and therefore to find if the traffic system was reaching Nash equilibrium, we measured the convergence by looking at the difference in two metrics, route change and volume change, between the previous and the current iterations.

Volume Change When choosing the best response action for each car, volume on each edge is a critical factor since it strongly influences the actual travel time across the edge. We defined volume change as the RMS value of the difference in volume on all edges in the previous and the current iterations.

$$\Delta V = \sqrt{\frac{\sum_{edge \in G} (V_{prev}(edge) - V_{cur}(edge))^2}{\#edges}} \quad (5)$$

Route Change A slightly different, potentially more stringent, measure was to measure the route change for each batch of cars as the sum of the number of new edges visited in the current iteration and the number of visited edges in the previous iteration (but not visited in the current iteration) divided by the number of edges in the previous route assignments. More succinctly, this measured the number of elements in set symmetric difference between previous and current edge sets divided by the number of elements in the previous edge set.

$$\Delta Route = \frac{\sum_{c \in cars} Route_{prev}(c) \triangle Route_{cur}(c)}{\#cars} \quad (6)$$

Figure 4 and Figure 5 show that the between-iteration metrics for three different number of total iterations, each with different number of iterations for different number of cars per batch. The blue curve has the longest total iteration with 5 iterations for 4000 cars/batch, 10 iterations for 2000 cars/batch and 10 iterations for 1000 cars/batch. The red curve has 2 iterations for 4000 cars/batch, 4 iterations for 2000 cars/batch and 8 iterations for 1000 cars/batch. Finally, the green curve has 1 iteration for 4000 cars/batch, 2 iterations for 2000 cars/batch, and 4 iterations for 1000 cars/batch.

In both Figure 4 and Figure 5, blue curves have two major peaks at iteration 6 and 16 and the red curves have a major peak at iteration 7. These are the iteration numbers when we transitioned from a larger number of cars per batch to a smaller number of cars per batch. Given that we would like 1000 cars/batch in the final iteration, these peaks signify that some iterations with the previous number of cars per batch were wasted, since when we transitioned to a new number of cars per batch, we would need to start from a higher value of between-iteration change again. To save computation time, we altered the number of iterations in each number of cars/batch to reduce the wasted

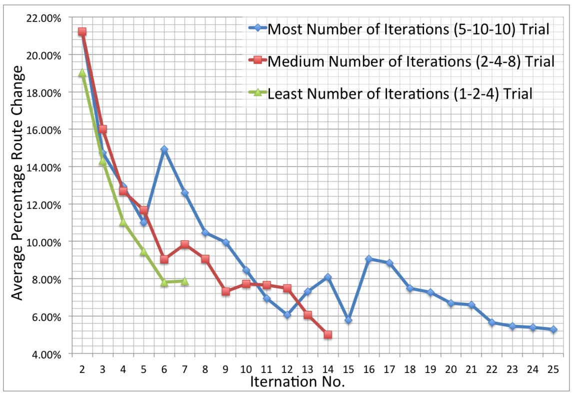

Figure 4: Percentage Route Change

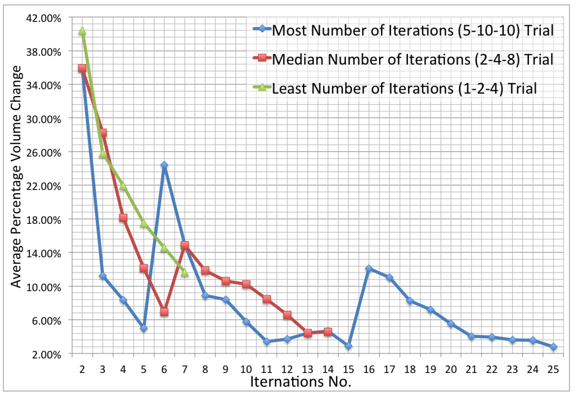

Figure 5: Percentage Volume Change

computation time from the blue curve, to red curve and finally to the green curve with just 7 total iterations.

In both Figure 4 and Figure 5, we observed that our traffic simulation is converging to equilibrium. Certainly with more iterations, we could obtain more converged traffic assignments, but with 7 total iterations shown in the green curve, we are confident that the final traffic assignment is within about 10% from the true traffic equilibrium. Hence, for all of the traffic simulations, we used 7 total iterations.

5 Results

5.1 System Cost Comparisons

We calculated the overall costs for placing all 4 million cars (broken down in batches) into our Bay Area Road Network.

In Table 3, we removed different number of critical nodes, out of the top 400 we identified. Then we identified 1000 randomly chosen (source, destination) pairs to model their changes based on the 3 different scenarios. We saw that there were some pairs that became unfeasible given the removal of different number of critical nodes with respect to the different ways of identifying critical nodes. In particular, there are 35 unreachable pairs using Betweenness Centrality, 72 unreachable pairs using Max Closeness centrality, 79 unreachable pairs using Mean Closeness centrality. Thus, we took away those routes and performed the simulations based only on routes that were still feasible given the removal of all 400 nodes.

5.2 Route Re-Assignment Visualization

To better visualize how route assignments are affected by the removal of critical nodes for the 3 different scenarios, we plotted the route assignment changes using Betweeness value as the critical node identification measure. In particular, we picked two different (source, destination) pairs.

North Pair To see how our simulation impacted travelers who primary travelled on the northern half of the Bay Area, we picked a (source, destination) pair, from 21 Coggins Dr, Walnut Creek to Valencia St to 17th St, San Francisco, to plot for the three different scenarios. This result is shown in Figures 6, 7, and 8. Before the Top 100 critical nodes were removed, the three scenarios all choose to pass through the Bay Bridge. After the removal of critical nodes and Bay Bridge was not available anymore, the Distance-based and Traffic-based scenarios chose to pass through the San Rafael Bridge (northest), while the Time-based scenario selects the San Mateo Bridge (south) to pass through. Note that for the Traffic-based scenario, different car batches would be assigned to different routes given the same source and destination pair.

South Pair To see how our simulation impacted travelers who primary travelled in the southern part of the Bay Area, we picked a (source, destination) pair, from Beach Park Blvd & Edgewater Blvd, Edgewater Place Shopping Center, Foster City, to Whipple Road Exit from Interstate 880 in Hayward, to plot for the three different scenarios. Here we remove top 400 critical points. The result is shown in Figures 9, 10, and 11.

In the Time-based and Distance-based scenarios, before we removed the Top 400 critical nodes, route assignments were along the San Mateo Bridge. For Traffic-based scenario, some car batches were assigned passing through San Mateo Bridge, while others were assigned to go further south and pass through the Dumbarton Bridge instead. After the removal of Top 400 critical nodes, with the Bay Bridge and San Mateo Bridge both unavailable, all three scenarios chose to pass through the Dumbarton Bridge.

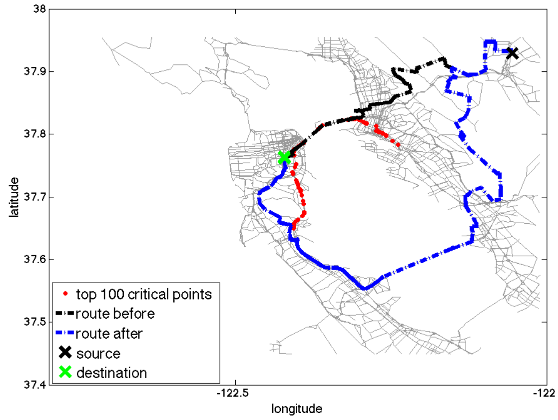

Figure 6: Route Re-Assignments(N)-Time

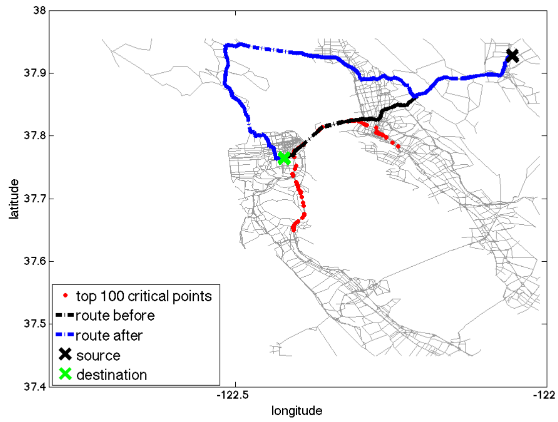

Figure 7: Route Re-Assignments(N)-Distance

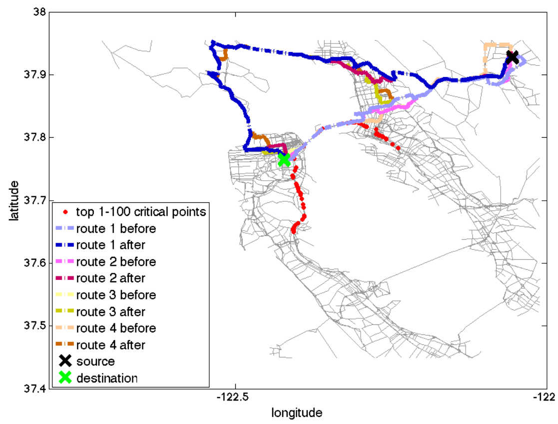

Figure 8: Route Re-Assignments(N)-Traffic

# Nodes	%Time			%Distance			%Traffic		
Removed	Btw.	C-Mean	C-Max	Btw.	C-Mean	C-Max	Btw.	C-Mean	C-Max
100	28.55	26.89	13.83	27.34	25.56	13.07	20.77	18.97	7.97
200	29.90	27.13	15.17	28.97	25.79	13.94	21.46	18.06	9.97
300	51.82	29.67	18.78	49.53	26.86	16.07	38.74	17.23	9.66
400	53.57	31.44	23.67	50.19	28.32	19.42	42.14	18.77	11.21

Table 3: Summary of System Cost % Changes

6 Analysis & Evaluation

Based on the critical nodes we found, we removed them in batches of 100. We found that the SF-Bay Area's road network is reliable, incurring only around 50% increase of the original system cost upon removing 400 critical nodes. We have also found that the *Betweenness* measure for identifying the critical nodes incurred the most significant changes in system cost, indicating that it was the optimal measure for identifying critical nodes.

6.1 Comparison of Different Critical Nodes Measures

When we removed the critical nodes based on the different measures of critical node identification, we saw that the Betweenness measure provided the largest increase in system cost in all 3 different scenarios. This suggested that the Betweenness measure (Equation 1) was the best measure for our road network reliability investigation. On the other hand, Max Closeness measure gives the most gradual increase in overall system costs. Note that Max Closeness measure not only gave the smallest *inter*-measure increase, it also gave the smallest *intra*-measure increase. Lastly, Mean Closeness measure was less sensitive than Betweenness measure but more sensitive to Max Closeness measure. Note that all 3 measures produced acceptable system costs increases if we used a similar definition for acceptable delay defined by (Snelder et al., 2009)[7].

6.2 Comparison of Different Scenarios

Given the numbers presented in Table 3, Traffic-based scenario was the least sensitive to the removal of critical nodes. This was because for both Time and Distance-based scenarios, there was only one unique route (shortest time or shortest distance) for the same source and destination. However, for the Traffic-based scenario, there was not a single optimal path for the same source and destination pairs (we allowed up to 4 different routes per source and destination pair). Therefore, when a critical node was removed, it was less likely to impact a large number of drivers since drivers were already dispersed to avoid traffic. Furthermore, we noticed that the difference in travel time between using any of the four routes for the same source and destination pair all have coefficients of variation that is less than 10%, suggesting that all routes are similarly optimal at every iteration. Therefore, if one of the routes was congested, it could redistribute its load to other original routes, which were similarly optimal to its original route, with some second order redistribution. Therefore, the overall change was less significant. Hence, we saw a smaller impact on the Traffic-based scenario.

6.3 Number of Nodes Removal

As expected, with increasing number of critical nodes removed, across all different measures of identifying critical nodes and in both Time and Distance scenarios, there was an increase in the overall system cost. We note that for the Betweenness measure, the largest increase across all 3 scenarios occurred between removing 200 and removing 300 nodes. This was due to the fact that the Top 201-300 nodes are located along the San Mateo Bridge (shown in Figure 1). With the Bay Bridge already removed and now with the San Mateo Bridge also removed, any car that sought to cross the water must then take round-about alternatives.

7 Conclusion

Our simulation suggests that the Bay Area road network is fairly reliable, incurring only about 50% increase in system cost when 400 critical

[7] *acceptable delay* due to a road closure is 21 minutes during peak hours and 16 minutes during off-peak hours for a trip that would normally take 30 minutes, which translates into 70% and 53% travel time increase for peak and off-peak hours.

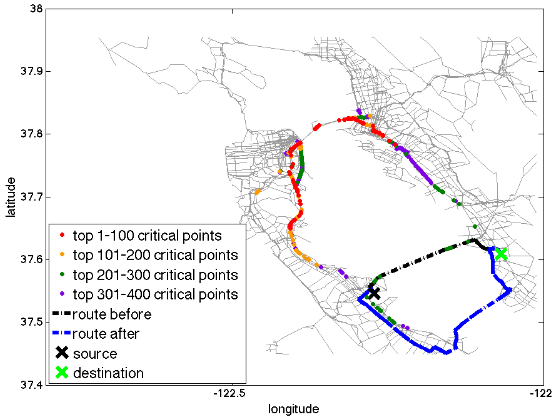

Figure 9: Route Re-Assignments(S)-Time

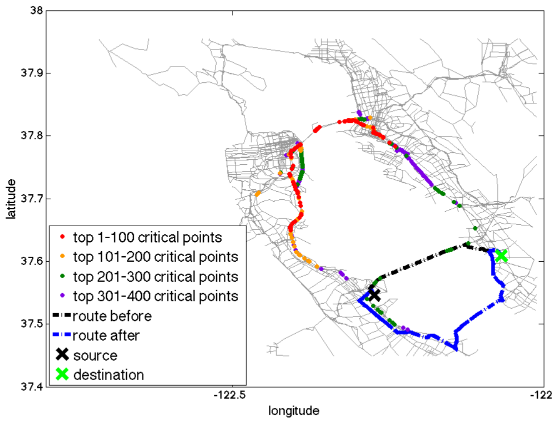

Figure 10: Route Re-Assignments(S)-Distance

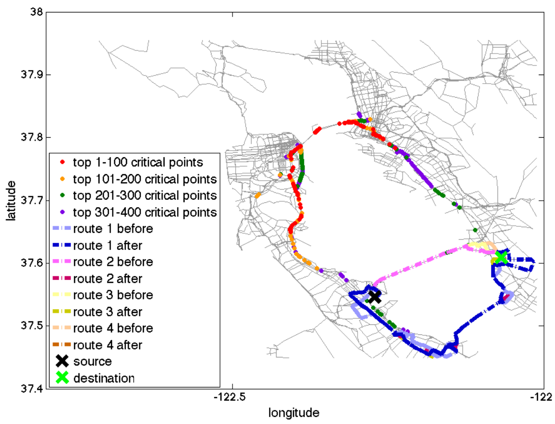

Figure 11: Route Re-Assignments(S)-Traffic

nodes were removed. Among the three different scenarios, Time-based, Distance-based and Traffic-based, we found that the Traffic-based scenario was the least affected by the removal of nodes, with a maximum of 42% increase in system cost. Since the Traffic-based scenario is a more realistic scenario, this further suggests the Bay Area road network is reliable in reality. From our analysis, we also found that Betweeeness is the most efficient measure of criticalness of a node, consistently incurring larger change in system costs. Thus, when analyzing a road network, with respect to finding the ciritical nodes, Betweenness Centrality should definitely be included as one of the measures.

8 Future Work

Currently, our analysis used the critical nodes that were found in one single batch instead of iteratively removing the earlier critical nodes before finding the subsequent ones. The current batch approach yielded critical nodes under regular daily routine. On the other hand, the iterative approach would yield critical nodes under a hypothetically more well-planned attack situation. Both have their values for analysis. In our latter half of project, we attempted the iterative approach, but we found our network to be disconnected after only removing a node on the Bay Bridge and the Richmond San Rafael Bridge and no nodes from the San Mateo Bridge or the Dumbarton Bridge. After investigation, we realized that both the bridges in the south only included the west-bound direction; therefore, the network was no longer strongly connected. Although we could conceivably still predict the next critical nodes to be on those bridges, the locations of the subsequent critical nodes were not as apparent and the actual predicted increase in system cost would be less accurate. Therefore, further clean up of the data would be expected along for further investigation of iteratively removing critical nodes.

In addition, OpenStreetMap has regions of map, and in this case, we picked the SF-Bay Area Map. However, such region may be too small to effectively measure the actual increase in system cost. For example, our current network did not include the southern end of the peninsula that is connected to the mainland; therefore once all bridge connections are broken, then the network becomes disconnected. Therefore, it would be interesting to incorporate a larger network (of the size that drivers would be willing to take detour for) to see how much additional cost is incurred.

Contributions

All of us contributed work into this project equally. Alice Yeh specifically looked into data collection process and analysis of data. Chung Yu Wang and Hsin-Fang Wu focused more on finding critical nodes and the subsequent simulations.

Acknowledgement

We gratefully acknowledge the help from Jason Wu, Claire Lin, David Medeiros and all the GIS staff from Branner library for their extraordinary help with helping us prepare our data.

References

Gary Davis and Hui Xiong. 2007. Access to destinations: travel time estimation on arterials.

Anthony H Dekker and Bernard D Colbert. 2004. Network robustness and graph topology. In *Proceedings of the 27th Australasian conference on Computer science-Volume 26*, pages 359–368. Australian Computer Society, Inc.

David Easley and J Kleinberg. 2010. Modeling network traffic using game theory. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*, pages 229–247.

Bin Jiang and Christophe Claramunt. 2004. Topological analysis of urban street networks. *Environment and Planning B*, 31(1):151–162.

Vamsi Kalapala, Vishal Sanwalani, Aaron Clauset, and Cristopher Moore. 2006. Scale invariance in road networks. *Physical Review E*, 73(2):026130.

Maaike Snelder, JM Schrijver, LH Immers, and B Egeter. 2009. Designing robust road networks. In *Transportation Research Board 88th Annual Meeting*, number 09-2718.