The Efficacy (or Lack Thereof) in Using Betweenness Centrality as a Clustering Heuristic on Road Networks

R Oren Pincock

A common heuristic used in vertex-attack clustering[1] is betweenness centrality. The nodes with the highest centrality are good candidates for an attack set.

Betweenness centrality is computationally expensive. If weight isn't an issue, then Brandes[2] runs relatively quickly in O(VE) time (though it is not parallizable). But it is in the nature of road networks that weighting is important. Distance is one of the natural ways that we classify two cities as separate. For weighted graphs, we must use Johnson's[3], which runs in a slower O(V^2logV) time, but it is parallelizable.

However, any road networks large enough to cluster tend to be very large, and a higher-order time complexity can be devastating. The state of California contains 1.6 million intersections[5][7]. The Bay Area alone has 430,694[4][7]. Running Johnson's on the latter takes almost a week of CPU time.

Johnson's is based on Djikstra's which is already used on road networks to find shortest path of travel. These shortest paths tend to have a bias towards Interstates and Highways, and so nodes lying on highways tend to have a higher centrality.

Yet this is not an effective way to cluster cities. Highways can be find inside and outside cities. And there are still back-country roads that have a low centrality and don't belong in any defined metropolitan area. In fact, betweenness centrality may be a completely useless measure due to the peculiar nature of road networks.

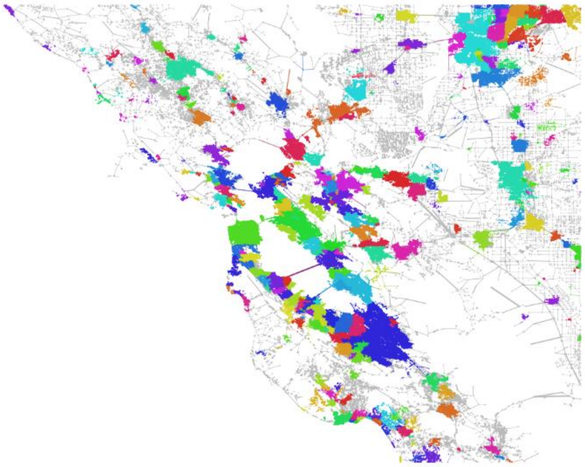

Figure 1 - Bay Area

This study utilizes road networks from the Bay Area (Fig 1) and the Los Angeles Metropolitan Area (LA) (Fig 2)[4]. A base truth for sorting the tables was obtained from the Census Bureau[3], which supplies shapefiles outlining all municipalities in California (including unincorporated census designated places). These are used to color Fig 1 & 2. Each municipality has a unique color and any node outside a municipality is colored gray.

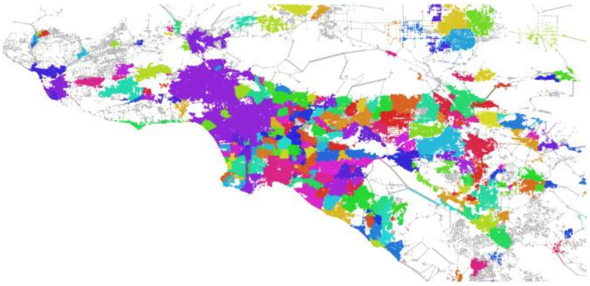

Figure 2 - Los Angeles Metropolitan Area

Examining the centralities of nodes belonging inside and outside these areas provides no distinction between the two. Most nodes have a centrality of almost 0 (Table 1, Fig 3). To help better illustrate true shape, histograms showing just the top 1% and 2% of nodes with highest centrality are also given for the Bay Area (Fig 5, 7) and LA (Fig 4, 6).

Notice the shape of the distribution is highly similar regardless of whether the node is in a city or in a rural area. Also, the top 59 most central nodes in the Bay Area are all in cities.

In LA (Fig 2, 6, 7), the lack of disparity between the centralities of nodes in and outside of cities is still present, but the entire chart has a lower average centrality. Comparing the maps for the Bay Area (Fig 1) and LA (Fig 2), it is apparent that LA has one supercluster of adjacent municipalities, whereas the Bay Area is more distributed. This suggests that when clusters are viewed more holistically, the average centrality will tend to be lower. This is likely due to the higher concentration of residential areas in cities, which tend to have a lower centrality because of their grid structure. However, this trend is impossible to observe on the scale of a local node community, and is therefore only useful as an unreliable verification measure after a cluster has already been determined.

In conclusion, although the unique characteristics of road networks promote certain quirks and even faint patterns in centrality, the large amount of computation required coupled with the lack of clear results to interpret obviates betweenness centrality as a viable metric to apply to these types of networks.

	Bay Area	Los Angeles
All Nodes	95.00%	94.64%
Inside Municipalities	95.15%	95.29%
Outside Municipalities	93.42%	94.61%

Table 1. Percent of Nodes with Centrality less than 0.001

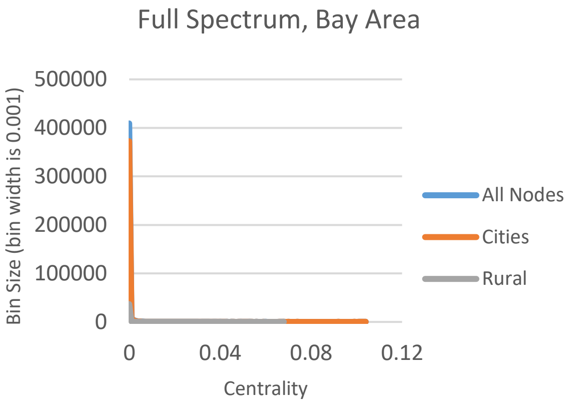

Figure 3

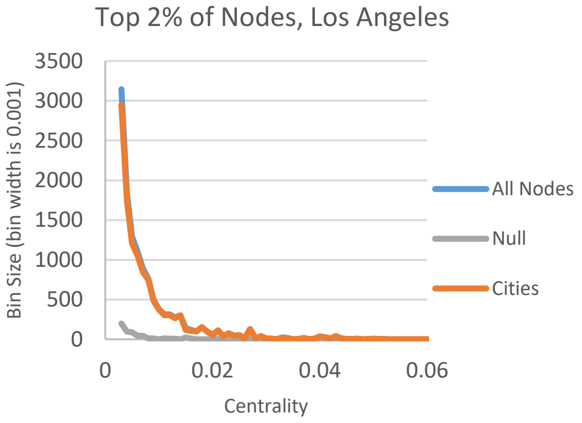

Figure 4

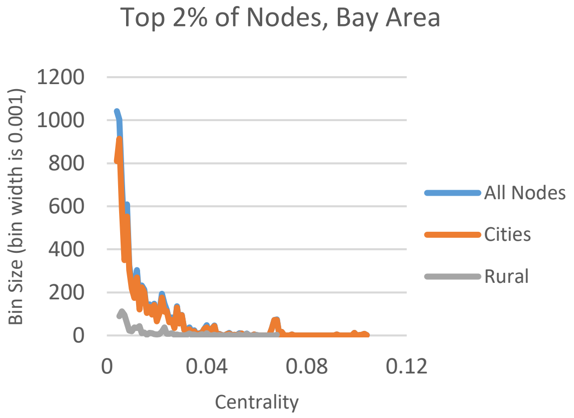

Figure 5

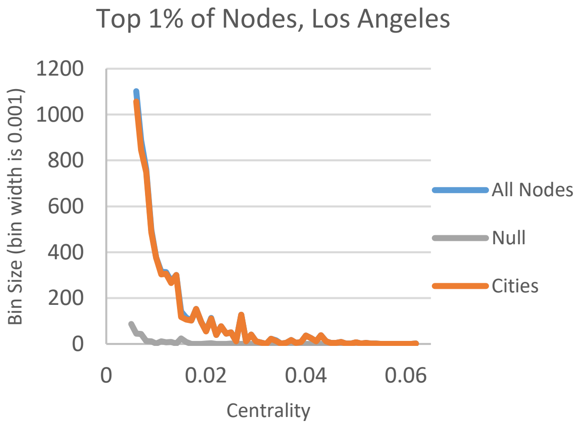

Figure 6

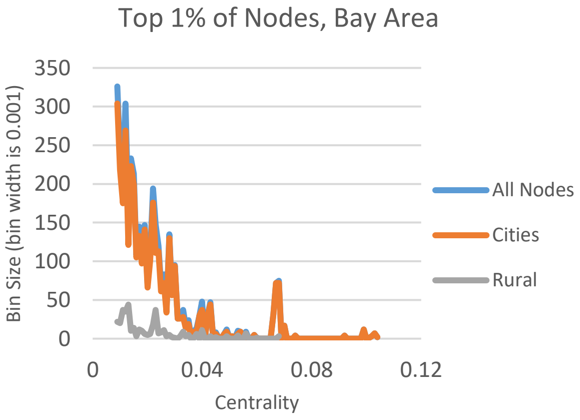

Figure 7

[1] John Matta, Jeffrey Borwey, and Gunes Ercal. (2016), *The Vertex Attack Tolerance of Complex Networks*. RAIRO Operations Research

[2] Brandes, Ulrik. (2001) *A Faster Algorithm for Betweenness Centrality*. University of Konstanz

[3] Johnson, Donald B. (1977), *Efficient algorithms for shortest paths in sparse networks*, Journal of the ACM

[3] https://www.census.gov/geo/maps-data/data/cbf/cbf_place.html

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

[5] http://download.geofabrik.de/north-america/us/california-latest.osm.pbf

[6] https://github.com/gka/pyshpgeocode/blob/master/shapegeocode.py

[7] https://github.com/nightduck/GephiConvScripts/blob/master/osm_gdf_conv.py

[8] https://github.com/nightduck/GDFBetweenness