Camilo Pedraza Jiménez – 226679

Used data: United States roads (2000) -DC

Applied Network Analysis: Data Report

General data description

The chosen data network corresponds to the District of Columbia from the Urban Areas Census of 2000. The complete data sets were assembled by Dominik Schultes and the complete description of the data can be found here.

In this network, the edges (links between nodes) represent stretches of road, while the vertices (nodes) are intersections of roads. Additionally, the network is defined as undirected and unipartite. In the one hand, regarding the undirected nature of the network, this means that the edges between nodes have no inherent direction, so the relationship between nodes is symmetric. In the other hand, the unipartite nature of the network indicates that all nodes belong to a single category or set. In this case, the single category is defined as "intersection of roads".

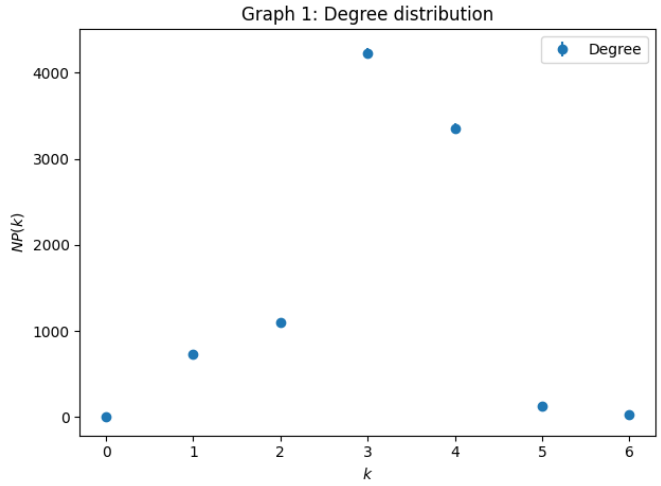
Apart from the characteristics mentioned above, the Netzschleuder website also mentions the following characteristics from the network:

- The network consists of 9,559 nodes in total. Since the network is undirected, there is no division between in and out nodes.
- There is a total of 14,909 links. Which means that the average degree of the nodes in the network (⟨k⟩) can be calculated as $\langle k \rangle = \frac{1}{N}\sum_{i=1}^{N} k_i = \frac{2L}{N} = \frac{2*14{,}909}{9{,}559} = \mathbf{3.12}$. Additionally, the standard deviation given by the website is 0.91. This means that, on average, a node in this network has 3.12±0.91 links.

Finally, from the data above, one can calculate that the maximum number of links that can be found in this network is $L_{max} = \frac{N(N-1)}{2} = \frac{9{,}959(9{,}959-1)}{2} = \mathbf{45,682,461.}$ Since this network describes the spatial relationship between roads, it is possible to understand that roads that are distant from each other cannot have a direct connection. This fact is helpful to understand why the number of links found in the network differs significantly from its L_{max}.

Degree distribution

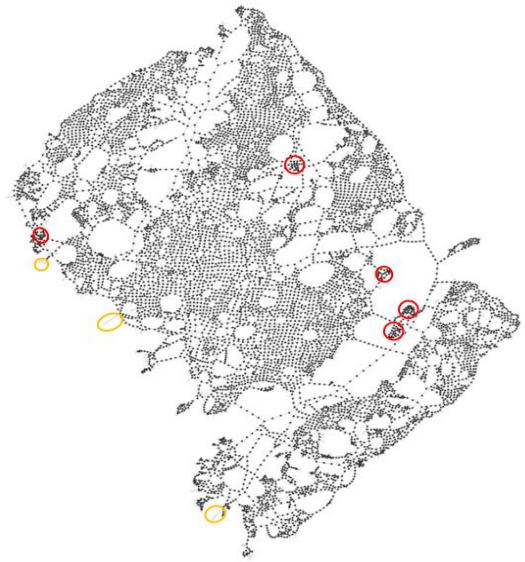
In general, the range of the degrees found in the network is 6, in which the lowest degree found is 0 and the highest degree found is 6. In this sense, Graph 1 shows the degree distribution of the degrees, providing insights into how many nodes have a specific degree. Additionally, error bars are included to help visualize the uncertainty or variability in the degree distribution. The x-axis represents the degrees, the y-axis represents the frequency or probability of nodes with those degrees, and circles with error bars indicate the data points in the plot.



In general, around 4,000 nodes are connected to 3 nodes, followed by over 3,000 nodes connected to 2 nodes and over 1,000 nodes connected to 2 nodes. Based on this graph, one can argue that the average degree of the network is around 3, which is consistent with the ⟨k⟩=3.12 that was found above.

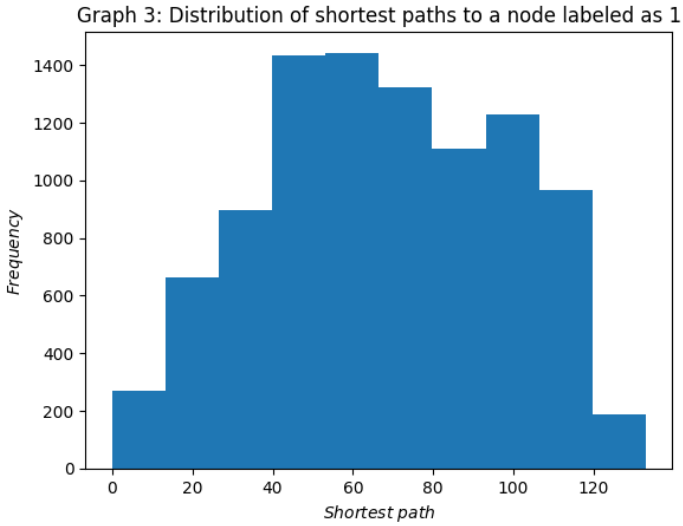
Graph 2 works as another way to represent the distribution of the degree of the network. Since the network can be analyzed spatially, it with the K-Core decomposition is possible to see the number of links that each intersection (node) has in Washington DC. Although, the number of nodes is too big to be able to see the degree of each individual node, it is possible to see that the lighter the color of the node, the lower its degree. Based on this characteristic, it is possible to identify that peripheral nodes (orange circles) tend to have fewer links, while there are some other areas of the city in which nodes with a higher degree are concentrated (red circles).

Graph 2: K-Core decomposition

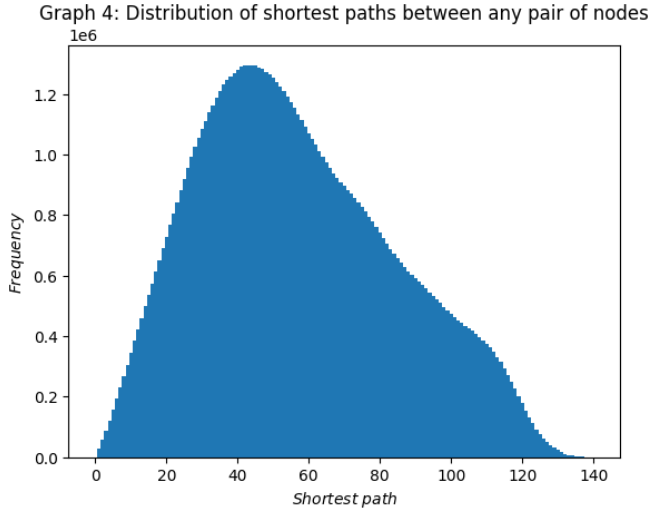


Shortest paths

Regarding the shortest paths in the network, graph 3 can be useful as a first approach. In this graph, it is possible to see the distribution of the shortest paths from any other node to a random node labeled as 1. Based on the results, one can argue that the distribution of the shortest paths tends to be normally distributed, and the highest frequency of shortest path is around 60.



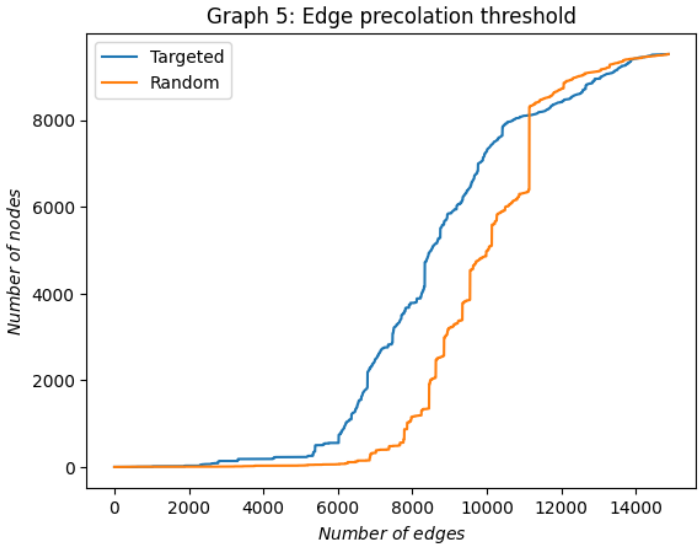
Additionally, Graph 4 shows the general distribution of the shortest paths between any pair of nodes in the network. Based on this graph, one can see that the most frequent shortest paths are around 50 connections.



After calculating the longest shortest path, the result is 138, which is consistent with the distribution observed in Graph 4. Furthermore, the calculated Average Path Length is 28.65, which is lower than the most frequent value seen in Graph 4. This result implies that the distribution of the path lengths is skewed towards shorter distances in the network, which suggests that there are more short than long paths in the network.

Percolation threshold

In general terms, percolation is a concept that involves the gradual formation or removal of connections in a network, in which nodes or edges are added or removed based on a probability distribution. In the case of this specific network, Graph 5 suggests that a targeted edge percolation strategy (e.g., removing edges that connect the most nodes) is more effective in breaking down the connectivity of the network compared to the random edge percolation strategy, before around 11,000 edges are reached. The fact that the targeted line is above the random line indicates that the network is more vulnerable to targeted attacks on specific edges than to random ones, although it is important to mention that both lines tend to be close to each other. This vulnerability insight is valuable for enhancing the robustness of the network by addressing structural weaknesses and fortifying critical connections. Additionally, it is important to mention that both strategies have an exponential increase in when specific numbers of edges are affected. For the targeted strategy, such increase happens at around 5,000 edges, while for the random strategy it happens at around 8,000 edges.



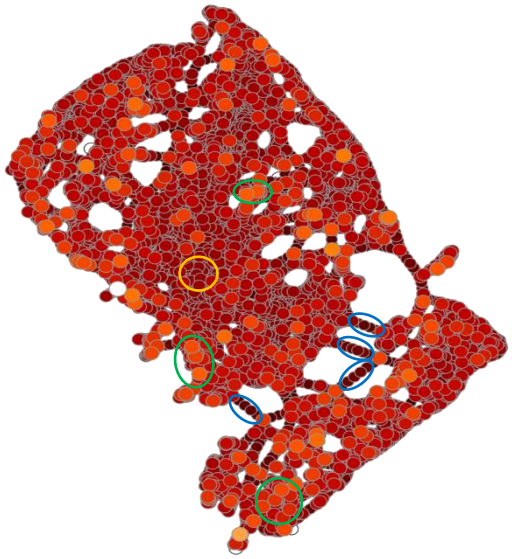
Centrality measures

The previous percolation analysis serves as an introduction to understanding the relationship between nodes and the relevance that specific nodes can have within the network. In this sense, the centrality measures help quantify the importance or influence of nodes within such network. These measures help identify nodes that play significant roles in connecting and mediating information flow within a network.

As a first centrality measure, the **PageRank** measures the importance of a node based on the importance of its neighbors. In this sense, nodes connected to other high-ranking nodes receive higher PageRank scores. Graph 6 indicates that the PageRank of the different nodes tend to be homogeneously distributed across Washington DC. However, it is important to mention that intersections located in bridges (blue circles) tend to have a lower PageRank score. In the same

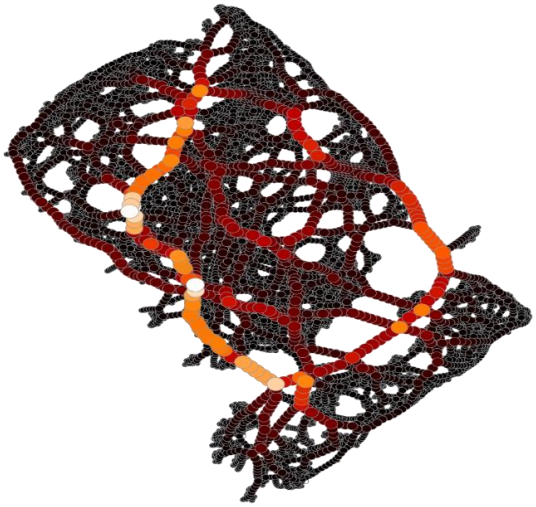
way, intersections located in the city center (orange circle) do not appear to have higher PageRank scores compared to some intersections located in more peripheral areas (green circles).

Graph 6: PageRank score



Another centrality measure that is useful fur the analysis is the betweenness. This measure calculates the extent to which a node lies on the shortest paths between other nodes. Nodes with higher betweenness centrality act as bridges or intermediaries in the network. In this sense, Graph 7 shows the main intersections that serve as the best roads to connect the different areas of Washington DC.

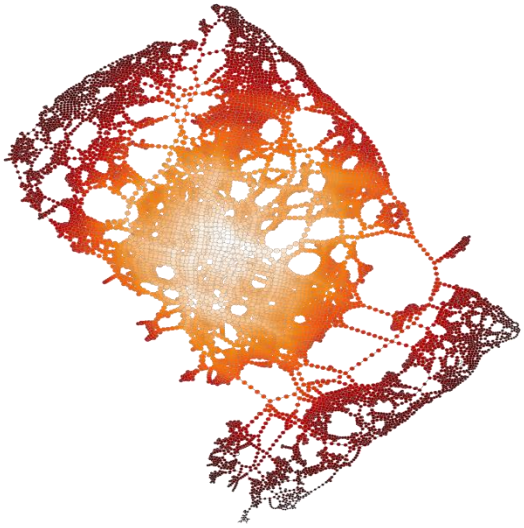
Graph 7: Betweenness Centrality



Finally, closeness is a centrality measure that evaluates how quickly a node can reach all other nodes in the network. Nodes with higher closeness centrality are closer (in terms of shortest path length) to other nodes. Since the analyzed network is based on a geospatial interaction, it is

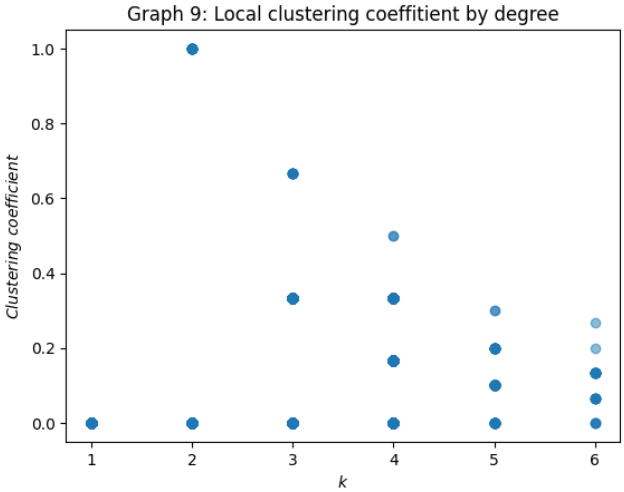
expected that the nodes located in the city center have a higher closeness centrality compared to the ones located in the periphery of the city. This characteristic can be seen in Graph 8.

Graph 8: Closeness Centrality



Clustering analysis

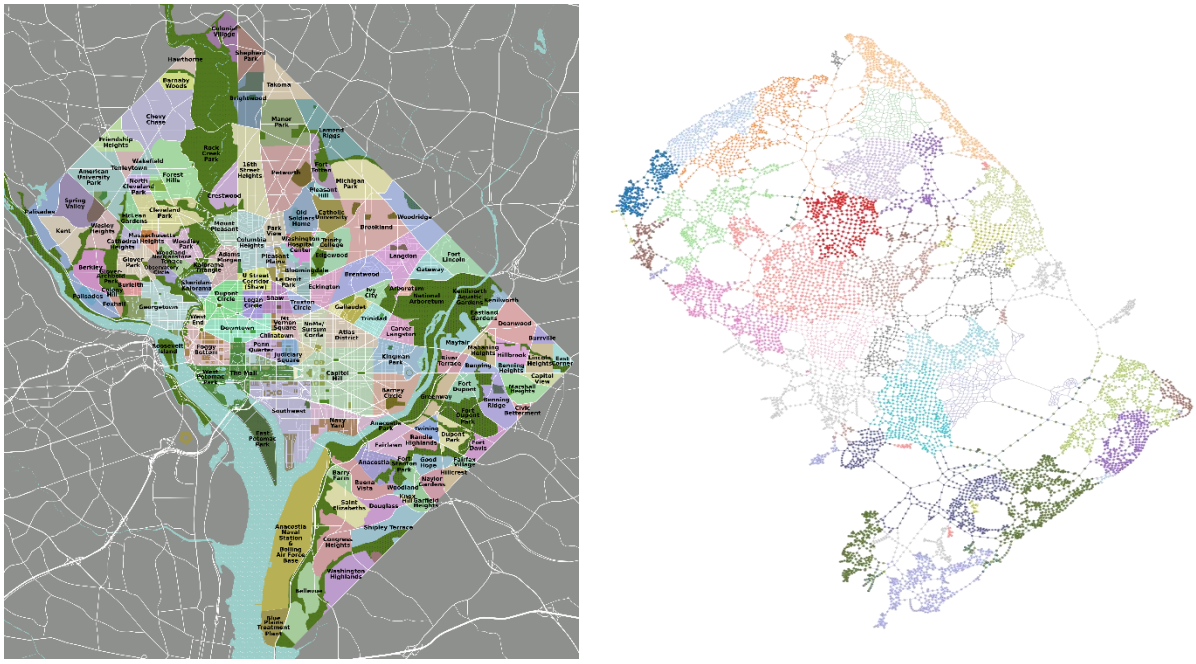
The different centrality measures are helpful to analyze the existence of communities within this network. In this sense, it is important to explain first that a community refers to a subgroup or subset of nodes within a network that are more densely connected to each other than to nodes outside the subgroup. Thus, the concept of communities helps identify and understand the modular or clustered structure within a network. In this regard, Graph 9 illustrates the distribution of the local clustering coefficients for the different nodes of the network, regarding their degree. Two main conclusions can be drawn from this graph. First, the variance of the local clustering decreases as the degree of the nodes increases. Second, the same trend can be seen for the average clustering coefficient by degree: the higher the degree, the lower the average clustering coefficient.



The average clustering coefficient calculated for the whole network is **0.039** with a variance of 0.0011. Since the clustering coefficient is a value between 0 and 1, it could be argued that the coefficient for this specific network is low. This could be related to the nature of this network being geospatial and the specific characteristics that this type of network can cause to the nature of the node and edges. Thus, it is important to compare the calculated coefficient to the another that could be found in a network with similar characteristics to understand if the connectivity of this network is actually low.

To better understand how clusters are formed in this network, blockmodeling can be used as a way of simplifying the structure of the network by grouping nodes into blocks or clusters based on their connectivity patterns. This process tires to identify patterns of connections within the network and represent them in a more compact form. In general terms, this method can reveal underlying structures or communities within the network. In this respect, Graph 10 compares the community composition of the intersections (blockmodeling) and the real neighborhood composition of Washington DC. It can be seen that the communities created through blockmodeling, based on the connectivity patterns of the intersections, closely resemble the actual neighborhood division of the city. Additionally, it could be argued that the communities from the blockmodel process might group different real neighborhoods of Washington DC, and not only one neighborhood.

Graph 10: Real Neighborhood Map vs. Blockmodel



Additional to the normal blockmodel, a nested blockmodel goes further by allowing blocks to contain sub-blocks, creating a hierarchical or nested structure. This hierarchical organization helps to capture more nuanced relationships within a network and reveals multi-level community structures. In this sense, Graph 11 shows that there are many communities of similar size in the hierarchical structure, which suggests a balanced or equitable distribution of nodes across the

different levels of the hierarchy within this network. This characteristic may indicate that the network does not have dominant or disproportionately large communities.

Graph 11: Nested Blockmodel

Furthermore, the different connections that can be seen between nodes from different communities could reflect small-world properties. Based on this, it is important to clarify that a small-world network is characterized by short average path lengths between nodes, high clustering, and the presence of a few long-range connections. Thus, it could be argued that the existence of connections between the different communities could lead to relatively short paths between them.

Conclusions

Although a deeper mathematical analysis could be useful to understand better how well connected this network is, one can qualitatively argue that the presence of direct links among the communities generated in the network may lead to short path lengths that can connect the network well. In this regard, comparing the average clustering coefficient of this network with the one that can be calculated for other geospatial networks may be beneficial to conclude whether the value of such coefficient is actually small. Furthermore, the centrality measures used in the analysis help to understand the existing differences among peripheral intersections, avenues, and intersections located in the center of Washington DC. Finally, knowing that the network is more vulnerable to targeted attacks on specific edges than to random ones is valuable for policy measures that address the structural weaknesses of the network to improve its robustness.