Democratic National Committee E-mail Network

Walter Roberto Vives Castro *Data Engineering Data 8° "A" Universidad Politécnica de Yucatán(*Licenciatura, Km. 4.5. Mérida-Tetiz Tablaje. CP 97357 Ucú,Yucatán, México*)*

***Abstract*— In this paper we will analyze a network of emails from the Democratic National Committee. The characteristics of the network will be analyzed, centrality measures will be applied, the analysis of its distribution, the simulation of the network will be done using the Erdos-Renyi model and the analysis of communities will be done.**

Keywords—Network, Community detection, Degree distribution, Models of network, Centrality measures, Graph Theory.

Introduction

The Democratic National Committee (DNC) is the governing body of the United States Democratic Party. The committee coordinates strategy to support Democratic Party candidates throughout the country for local, state, and national office, as well as works to establish a "party brand". It organizes the Democratic National Convention held every four years to nominate a candidate for President of the United States and to formulate the party platform. While it provides support for party candidates, it does not have direct authority over elected officials. [1]

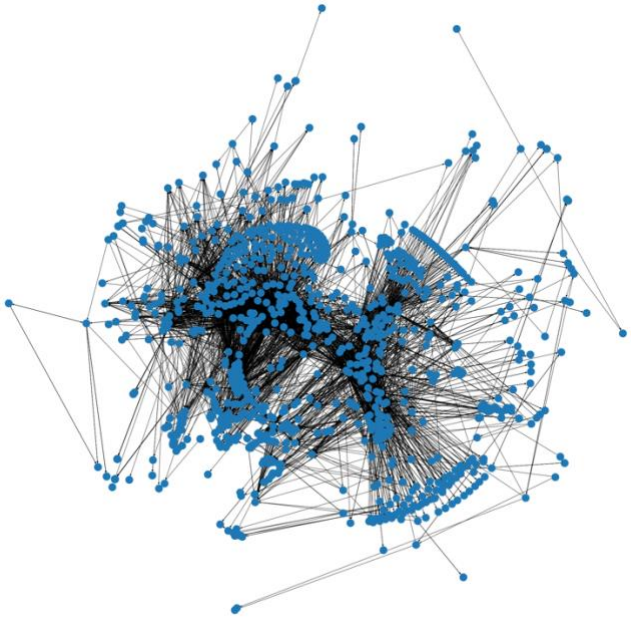
The DNC is responsible for articulating and promoting the Democratic platform and coordinating party organizational activity. When the president is a Democrat, the party generally works closely with the president. In presidential elections, it supervises the national convention and, both independently and in coordination with the presidential candidate, raises funds, commissions polls, and coordinates campaign strategy. Following the selection of a party nominee, the public funding laws permit the national party to coordinate certain expenditures with the nominee, but additional funds are spent on general, party-building activities. There are state committees in every state, as well as local committees in most cities, wards, and towns (and, in most states, counties). [1]

The network used was "Democratic National Committee email", it was retrieved from networkrepository.com.

Description of the network

Each node represents a person, the connection between nodes (edges) represents the email communication. Therefore, the network type is directed since the email is send from a node A to a node B.

This is the directed network of emails in the 2016 Democratic National Committee email leak. The Democratic National Committee (DNC) is the formal governing body for the United States Democratic Party. A dump of emails of the DNC was leaked in 2016. Nodes in the network correspond to persons in the dataset. A directed edge in the dataset denotes that a person has sent an email to another person. Since an email can have any number of recipients, a single email is mapped to multiple edges in this dataset, resulting in the number of edges in this network being about twice the number of emails in the dump.



I. Network Characersitics

- Number of nodes: 1,891
- Number of edges: 5,598
- Max degree: 580
- Min degree: 1
- Average degree: 5.920676890534109
- Average Clustering: 0.176958616167228

The maximum connection of a node is 580. In this study case means that an email was sent to 580 people. The minimum degree was 1. An email was sent to only one person.

It can be noticed that the average clustering coefficient has a low value. Clustering coefficient of a vertex (node) in a graph quantifies how close its neighbours are to being a clique (complete graph). Therefore, it can be deduced that the network has a few cliques.

II. Centrality Measures

Here are some different ways to **measure centrality**: **Degree centrality**: This is simply the number of edges of the edge. The more edges, relatively speaking within the graph, the more important the node. **Closeness centrality**: This measure is the inverse of the sum of all shortest paths to other vertices. [2]

Closeness centrality measures each individual's position in the network via a different perspective from the other network metrics, capturing the average distance between each vertex and every other vertex in the network. Assuming that vertices can only pass messages to or influence their existing connections, a low closeness centrality means that a person is directly connected or "just a hop away" from most others in the network. In contrast, vertices in very peripheral locations may have high closeness centrality scores, indicating the high number of hops or connections they need to take to connect to distant others in the network. [2]

In graph theory, betweenness centrality (or "betweeness centrality") is a measure of centrality in a graph based on shortest paths. For every pair of vertices in a connected graph, there exists at least one shortest path between the vertices such that either the number of edges that the path passes through (for unweighted graphs) or the sum of the weights of the edges (for weighted graphs) is minimized. The betweenness centrality for each vertex is the number of these shortest paths that pass through the vertex. [3]

The betweenness centrality captures how much a given node (hereby denoted u) is in-between others. [4]

Eigenvector centrality is used to measure the level of influence of a node within a network. Each node within the network will be given a score or value: the higher the score the greater the level of influence within the network. This score is relative to the number of connections a node will have to other nodes. Connections to high-scoring eigenvector centrality nodes contribute more to the score of the node than equal connections to low-scoring nodes. [5]

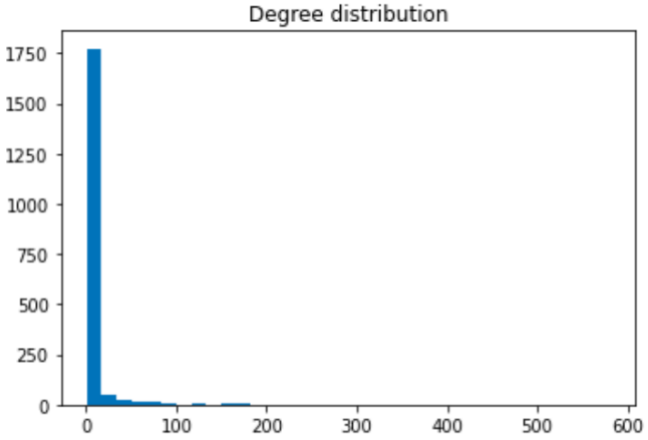
- Max closeness centrality: ('1669', 0.2580557072637903)
- Max Betweenness Centrality: ('1669', 0.12887116023101908)
- Max Degree Centrality: ('1874', 0.3068783068783069)
- Max Eigenvector Centrality: ('1874', 0.3205937512612431)

It can be noticed that the node "1669" has the maximum Closeness Centrality and Betweenness centrality, this is since the node "1669" is the node that is more central compared with all the other nodes and that same node is an important node given that it has considerable influence within a network by virtue of their control over information passing between others. It is the one whose removal from the network will most disrupt communications between other vertices because it lies on the largest number of paths taken by messages.

It can be noticed that the node "1874" has the maximum Degree Centrality and Eigenvector Centrality.

Eigenvector Centrality is a generalization of the degree centrality: a weighted degree vector that depends on the centrality of its neighbors (rather than every neighbor having a fixed centrality of 1). In other hand, Degree Centrality recall that a node's degree is simply a count of how many social connections (i.e., edges) it has. The degree centrality for a node is simply its degree. A node with 10 social connections would have a degree centrality of 10. A node with 1 edge would have a degree centrality of 1. [6]

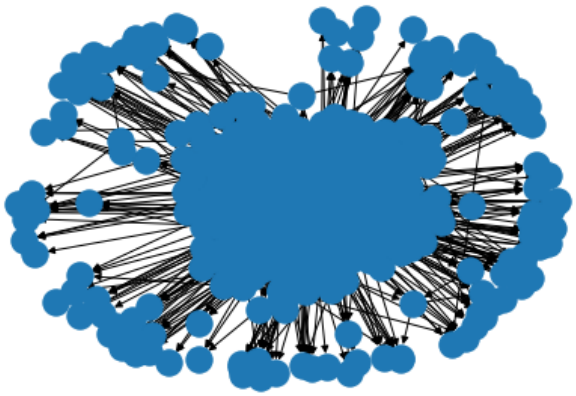
III. Models of Networks



The degree distribution plot shows us a positive skewness, since the maximum number of edges in this network is 580 and the minimum 1. Most of the nodes has less than 50 edges. Therefore, most of the people sent less than 50 emails. The average was 5 emails per person.

If we try to simulate the network as a random network using Erdos-Renyi model, we take as parameters n which are the numbers of nodes, p which is the probability that a connection between nodes is created.

The total node of our network is: 1891 and the density of our network is: 0.0015663166377074362. We know that $\langle d \rangle = p$. Therefore, if we simulated our network using n $= 1891$ and p $= 0.0015663166377074362$ we get the next network:



Comparation between real network and the network created using Erdos-Renyi model:

	Real	Expected
Nodes	1891.000000	1891.000000
Edges	5598.000000	5614.000000
Max_degree	580.000000	16.000000
Min_degree	1.000000	0.000000
avg_degree	5.920677	5.937599
avg_clustering	0.176959	0.000910
Transitivity	0.077784	0.000940
Max_closenness	0.258056	0.174709
Max_betweennes	0.128871	0.025586
Max_degree_centrality	0.306878	0.008466
Max_eigenvector_centrality	0.320594	0.082805

How it can be noticed the number of nodes is the same, and the number of links between the nodes is almost the same, they only differ from 16 links.

In the case of max degree, in the real network there is an outlier, there is a node with 580 connections, in the Erdos-Renyi model this type of atypical data cannot exist because a general probability is passed to it for the generation of the Node connections, if we put a high probability p, the nodes will have a higher probability but that does not mean that there may be the possibility that some node has a higher probability. Everyone is equally likely to have connections.

The average degree was practically the same, because the parameter that was passed to the Erdos-Renyi model was the density of the real network. $\langle d \rangle = p$.

The other metrics vary due to the position of the nodes.

If we realize, the structure of the network is similar, it has a greater number of nodes in the center and they are connected to nodes around it, this simulates the email network because there are central nodes (people) that send emails to other people and these people do not forward it, so those nodes remain around the network.

IV. Community Detection

The total communities detected by Girvan–Newman algorithm was 57. It can be seen that there are many communities, so it was decided to make a list of the modularity of each partition to see the number of clusters with an acceptable modularity to be able to make the graph of the network with a number less than 57.
E.g.,

	Cluster	Modularity
0	2	0.359961
1	3	0.348784
2	4	0.363248
3	5	0.401298
4	6	0.392505
5	7	0.376233
6	8	0.358317
7	9	0.341716
8	10	0.324786

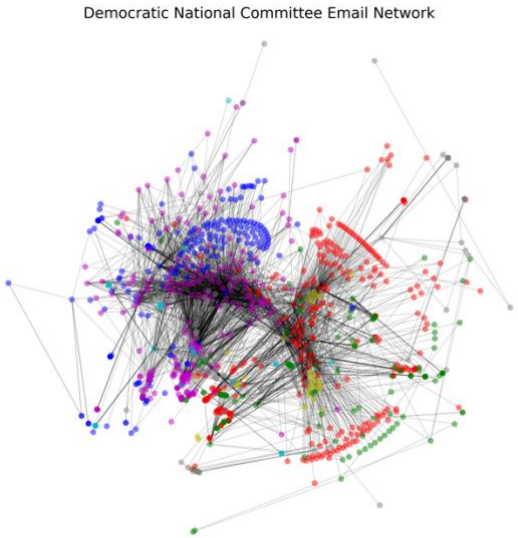
Number of Clusters and Modularity of a Network

This could not be possible due to computational cost. Iterate over all possible partitions and extract its modularity is complex, which is why another method was chosen to select the most representative communities.

The length of each of the 57 clusters was measured to see which community has the greatest number of elements, the result was the following.

[442, 367, 323, 276, 226, 107, 38, 20, 5, 5, 4, 4, 4, 4, 4, 3, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]

It can be noted that there are 6 communities with an acceptable number of elements, on the other hand, in most of the 51 remaining communities there is only 1 element in that community. Therefore, the 51 communities were grouped into one. With this, the total number of



communities was 7.

V. Conclusion

Thanks to the various metrics in the network analysis, we can find important information, we can detect communities, which nodes are the most influential, which node connects the different communities, how many connections each node has. This can be applied in various fields of the industry, for example if we want to release a new product, we can analyze a network of our products, the connections between the nodes would be if the products were purchased in the same purchase. With this we can know which product is more related to the other nodes and with that we can create a marketing strategy to enhance products. Network analysis is extremely interesting and complex, if used wisely, interesting insights can be obtained

References

[1] Wikipedia contributors. (2021, June 10). *Democratic National Committee*. Wikipedia. https://en.wikipedia.org/wiki/Democratic_National_Committee

[2] Derek L. Hansen, ... Itai Himelboim, in Analyzing Social Media Networks with NodeXL (Second Edition), 2020

[3] Wikipedia contributors. (2021b, July 18). Betweenness centrality. Wikipedia. https://en.wikipedia.org/wiki/Betweenness_centrality

[4] Charles Perez, Rony Germon, in Automating Open Source Intelligence, 2016

[5] Shaw, A. S. (2019, July 13). Understanding The Concepts of Eigenvector Centrality And Pagerank. Strategic Planet. https://www.strategic-planet.com/2019/07/understanding-the-concepts-of-eigenvector-centrality-and-pagerank/

[6] Jennifer Golbeck, in Introduction to Social Media Investigation, 2015

[7] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.