Predicting song popularity using track features

Miguel Cozar Tramblin A20522001
Carlos Muñoz Losa A20521562

13th February 2023

Abstract

Social Networks are widely used today. Millions and millions of users use different types of social networks on a daily basis to communicate, get information, entertain or collaborate. There are an endless amount of applications that involve social connections between people, companies, governments, etc. In this project we have decided to study U.S. senators' connections between them in Twitter. Code publicly available in GitHub [1]).

1 Data Collection

The first stage of our project consist in crawling US senators' data. We took the list of senators scrapping Wikipedia [2] using Pandas [3]. With that data we were able to build the network using the library networkx [4] for python and the library Tweepy [5] for checking the relations with the Twitter API in python. We decided to use 100 nodes because we have to check the relations for each pair of governors and that widely exceeds the API's request's limit. Using the US senate we finally built a graph with 100 nodes and 2378 edges 1.

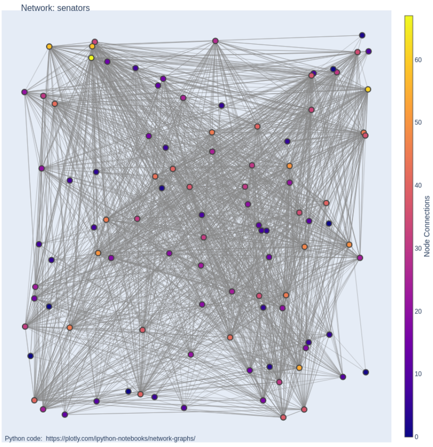

Figure 1: Network Draft

2 Data Visualization

We believed that the visualization was the key part of the project. Therefore, we have invested a vast amount of time to get several plots that we believe explain very well how the US senators interact between themselves on Twitter.

It's well-known that the US geography has influenced everything historically, and politics and their relationship is another example. For the purpose of providing evidence of this, we believed that our graph had to be plotted over a map. To do this, we leveraged Geoapify [6] and a public dataset of US zipcodes [1]. We organized the senators by state and picked a random zipcode of their state. With Geoapify, we can get a pair of coordinates given a zipcode. Due to the lack of data in the zipcode dataset, we had to take some of the zipcodes manually. This whole proceeding is on a jupyter notebook called Senators Dataset.

Once we had a pair of coordinates and all the data from each senator, we plotted the data over a map. This was done by leveraging python plotly library [7] and the perfect sinergy that this library has with Mapbox [8]. We created interactive maps stored in .html files publicly available [2].

The first plot is shown in Fig 2, where it is observed that the biggest points are in both coasts, specially in the north east. The most followed node is senator Bernie Sanders, an independent candidate. The Democratic party biggest nodes are in New York and New Jersey, whereas the republicans have its biggest influence in the south, with senators Ted Cruz and Marco Rubio as most followed senators.

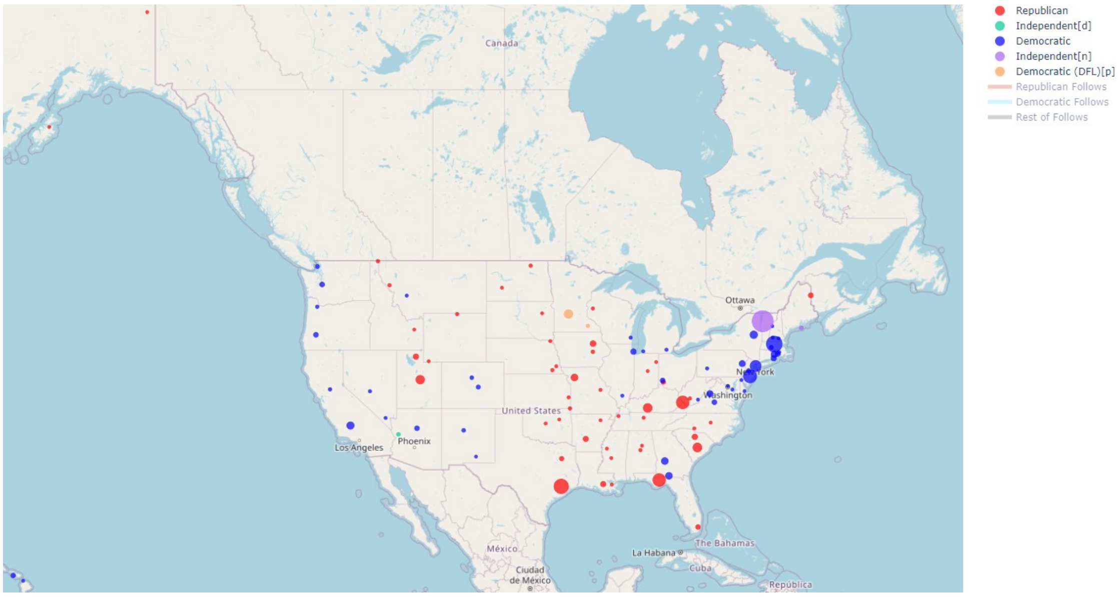

Figure 2: Senators by Number of followers

Continuing with the analysis, we plotted all the edges of the graph, and divided them later into democratic and republican follows. Fig 3. As you may observe, a big insight observed is the tendency of the red lines (republican links) to stay in the centre of the

[1] https://raw.githubusercontent.com/scpike/us-state-county-zip/master/geo-data.csv

[2] https://drive.google.com/drive/folders/1cvbFbPy6yMWUkAQpj3FoBuggHL5Wm9im?usp=share$_link$

US. However, blue links are crossing from west to east coast and also going to Hawaii, with democratic influence.

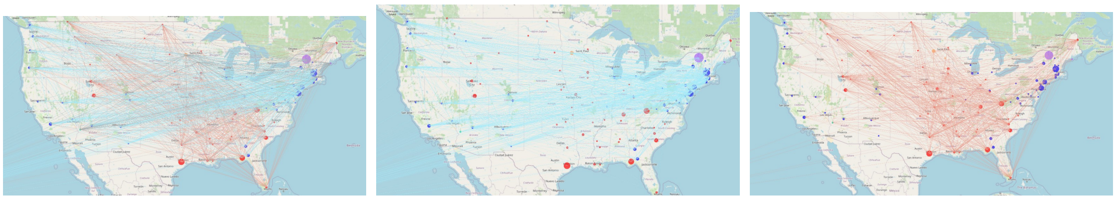

Figure 3: All the follows in the network, all democratic party follows, and all republican follows [4]

At this point, we thought we needed to represent the same map but scaling the points on how many followers had each node inside the graph, abandoning the general popularity on Twitter. This might give us a view of how popular are senators among other senators. For instance, Bernie Sanders (12M followers) might eventually be less appreciated by their senator colleagues. In addition, another interesting insight to check was the inter-party connections. This is: republicans following democrats and vice-versa.

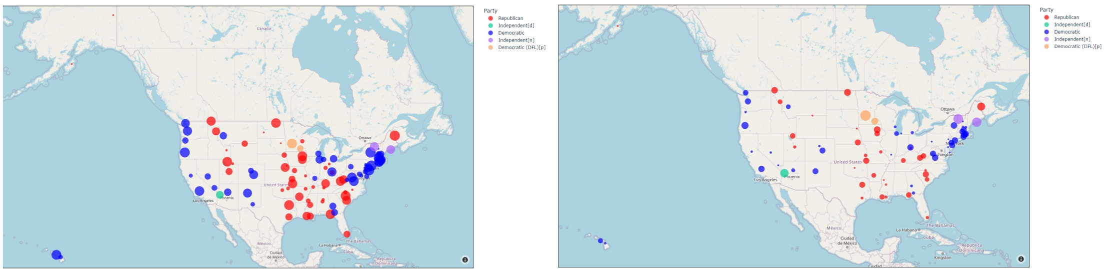

Figure 4: On the left, the representation of the senators scaled by number of followers inside the US Senate, on the right, the inter-party representation

Figure 4 confirms our suspects. We can observe that the fact of being popular among the people does not assure the same in the senate. In this case, the democrats from New York and New Jersey are leading the followers race with Cory Brook having 50 followers. Nevertheless, if we look to the right we will say that again the tendency is different. In this case, it is curious that senators from the West coast seem to have a better relation with republicans, and the same happens for republicans from the North of the US. It is specially curious how senator Ted Cruz, top 1 republican senator in Twitter follows, has only 3 democrats following him.

3 Network Measures Calculation

In order to analyze the network we have performed and plotted some interesting metrics such as Degree Distribution, Clustering Coefficient, Pagerank, Diameter, Closeness and Betweeness. We have performed this metrics using the networkx library functions. We also have collected the top 5 nodes for each measure and collected all of them in a table in Annex4.

3.1 Degree Distribution

Studying the number of edges connected to each node. In the representation shown in figure 5 we can see that the majority of nodes has between 0 and 80 edges, and there are a few between 80 and 100 edges. There are a large number of nodes that have around 40 incoming connections meanwhile the out coming connections are more spread among the nodes in the graph. This means that there are some nodes that have a lot of nodes pointing to them and so we can infer that this senators have more influence with the data that correspond in our graph. Specifically we can see that Cory Booker has great influence because he has a lot of in and out connections.

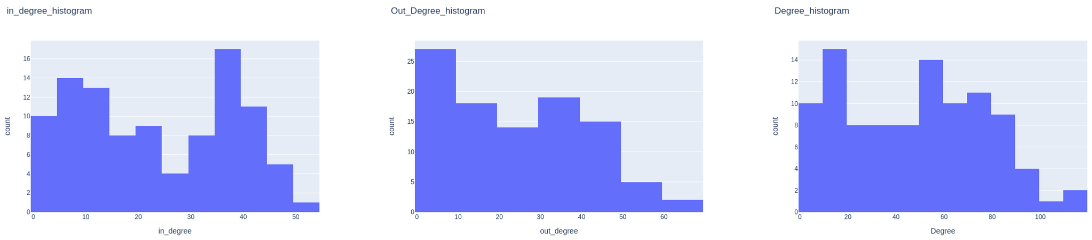

Figure 5: Degree distribution, including incoming connections and outgoing connections

3.2 Clustering Coefficient

The clustering coefficient represents the grouping tendency that the nodes in a network have. So the nodes that tend to cluster together have higher clustering coefficient that the ones that keeps away from others. In figure 6 we can see that most nodes have a clustering coefficient over 0.4. There is also a notable amount of nodes with a clustering coefficient of 0.62 which could be caused by a large number of nodes forming a cluster.

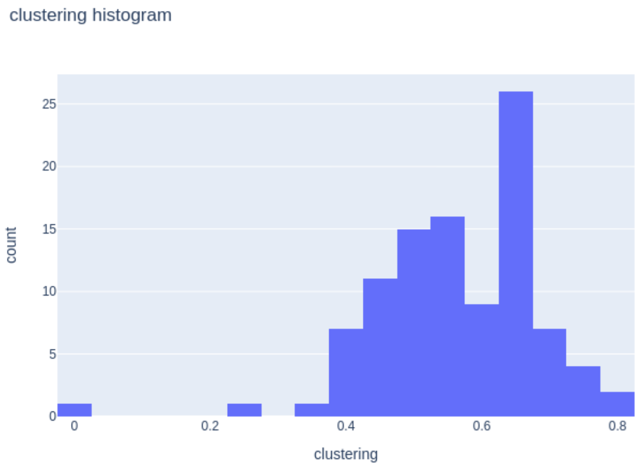

Figure 6: Clustering coefficient

3.3 Pagerank

PageRank computes a ranking of the nodes in a graph based on the structure of the incoming links. It was originally designed as an algorithm to rank web pages. In our case we have designed a directed graph so we have in-out edges depending on which senator is following or being followed by. The values of the pagerank are very close to 0 but it is true that there seems to be two groups, one with pagerank below 0.1 and other over that value. That means that there are a group of nodes with more and better connections in the network.7

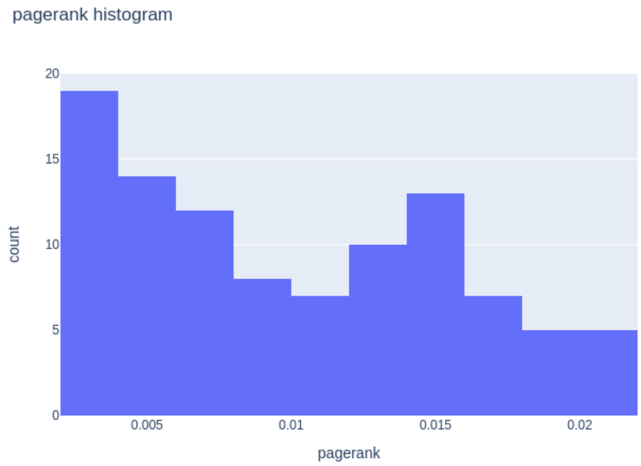

Figure 7: Pagerank

3.4 Diameter

We couldn't calculate the diameter of our network because the network is the library Networkx launches an error with the message *"infinite path length because the digraph is not strongly connected"*.

3.5 Closeness

The closeness parameter is the inverse of the mean distance of the optimal path between each node to all its reachable nodes. This means that the nodes with highest closeness coefficient require less distance to follow to reach other nodes. In figure 8 we can see that the closeness values are very close to 0.5 that means that the network is well connected and all the nodes require shorts path to be reached. In fact it is interesting to notice that Cory Booker has a high closeness coefficient, this fact can be due to his large number of connections, as we saw in the Degree distribution of the network. We can say that the node representing Cory Booker has remarkable importance in the US Senate.

3.6 Betweeness

The parameter betweenness centrality of a node is the sum of the fraction of all-pairs shortest paths that pass through each node. That reflects whether the node is essential for the network and makes the network very well connected or not. In our network all nodes have a lower value of betweeness except one. This means that that node has

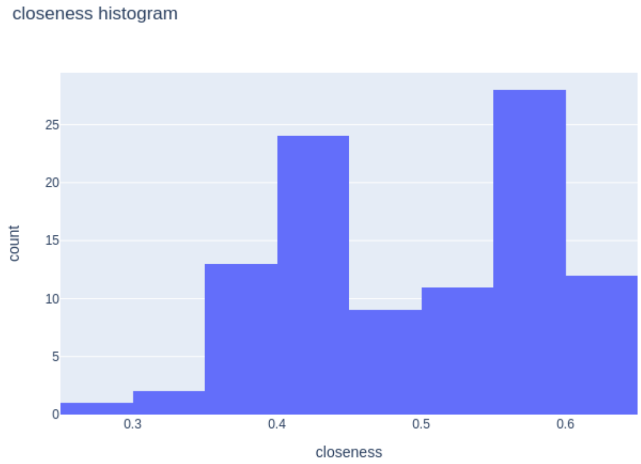

Figure 8: closeness

significantly more paths passing through than the rest of the nodes in the network. In annex 4 table we can see that Cory Booker appears again. This reinforce the importance of Cory Booker.

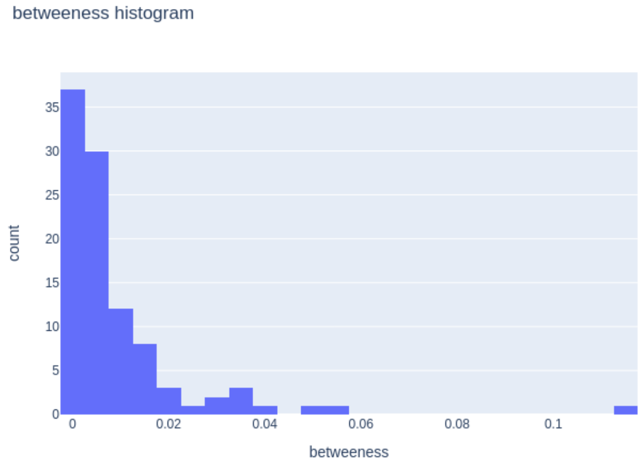

Figure 9: Betweeness

4 Conclusions

We have performed a first overview of how the US Senators make relationships between them using Twitter and online social network analysis. We believe we have showed an overview of how the US geography influences the senators popularity, and also how being popular in among people in Twitter won't assure the same in the Senate. We have studied how popular senators are likely to not have a great impact on the rival party.

Performing the measures has brought several conclusions. The network that we have created has some relevant nodes that have more connections than others, this results in an increase of the importance of those nodes for the network because many paths converge through those nodes. In translation to the senators data we can say that the influence of some senators is remarkable and they can get easily to almost every other senator.

References

[1] *Github repository.* URL: `https://github.com/carlosmlosa/Twiter_socialMediaProject`.

[2] *List of current United States senators.* URL: `https://en.wikipedia.org/wiki/List_of_current_United_States_senators`.

[3] *pandas - Python Data Analysis Library.* URL: `https://pandas.pydata.org/`.

[4] *Networkx library.* URL: `https://networkx.org/`.

[5] *Tweepy library.* URL: `https://www.tweepy.org/`.

[6] *Maps, APIs and components — Geoapify Location Platform.* URL: `https://www.geoapify.com/`.

[7] *Plotly: Low-Code Data App Development.* URL: `https://plotly.com/`.

[8] *Maps, geocoding, and navigation APIs SDKs — Mapbox.* URL: `https://www.mapbox.com/`.

Annex 3 Top 10 senators by category

	By Twitter Followers			By Senator Follows			By Interparty Follows (excluding independent senators)		
1	Bernie Sanders	Independent[n]	12.5M	Cory Booker	Democratic	50	Susan Collins	Republican	25
2	Elizabeth Warren	Democratic	7M	Kirsten Gillibrand	Democratic	49	Ron Wyden	Democratic	20
3	Ted Cruz	Republican	5.9M	Chuck Schumer	Democratic	48	John Hoeven	Republican	19
4	Cory Booker	Democratic	4.8M	Chris Murphy	Democratic	48	Chuck Grassley	Republican	18
5	Rand Paul	Republican	4.6M	Susan Collins	Republican	47	Jim Risch	Republican	17
6	Marco Rubio	Republican	4.5M	Ron Wyden	Democratic	45	Kirsten Gillibrand	Democratic	17
7	Chuck Schumer	Democratic	3.4M	Tim Scott	Republican	44	Sherrod Brown	Democratic	16
8	Lindsey Graham	Republican	2.2M	John Cornyn	Republican	44	Mark Warner	Democratic	16
9	Mitch McConnell	Republican	2.1M	Amy Klobuchar	Democratic (DFL)[p]	44	Dianne Feinstein	Democratic	16
10	Mitt Romney	Republican	2M	Brian Schatz	Democratic	44	Chuck Schumer	Democratic	16

Table 1: Top 10 senators

Annex 4 Top 5 senators by measure

Degree	**In Degree**	**Out Degree**	**Clustering**	**Pagerank**	**Closeness**	**Betweeness**
Senator John Cornyn	Cory Booker	Senator John Cornyn	Maggie Hassan	Mike Lee	Kirsten Gillibrand	Senator John Cornyn
Cory Booker	Kristen Gillibrand	Cory Booker	Senator Jack Reed	Sen. Susan Collins	Cory Booker	Cory Booker
Amy Klobuchar	Chris Murphy	Dr. Roger Marshall	Chris Van Hollen	Marco Rubio	Sen. Susan Collins	Amy Klobuchar
Tim Scott	Chuck Schume	Amy Klobuchar	Senator Bob Menendez	Senator John Cornyn	Chuck Schumer	Mike Lee
Sen. Susan Collins	Sen. Susan Collins	Captain Mark Kelly	U.S. Senator Cindy Hyde-Smith	Kirsten Gillibrand	Tim Scot	Dr. Roger Marshal

Table 2: Top 5 senators by measure