**Nanyang Technological University**

**MH8351 Web Analytics Project Report**

****

**Community and Key Player Detection**

**in 116th Meeting of US Congress**

**Chin Kiat Boon G1902005G**

**Ju Chuang G1901896K**

**Pua Ming Xiu G1902258C**

**Tony Liu Ran G1902349D**

Work Distribution

This project was completed by the following members:

* Chin Kiat Boon (*Section 8*)
* Ju Chuang *(Sections* 6,7)
* Pua Ming Xiu (*Sections 1,2,3,4* )
* Tony Liu Ran (*Sections 5, 9, compilation of report*)

Contents

[1 Abstract 2](#_Toc31743398)

[2 Introduction and Problem Statement 2](#_Toc31743399)

[3 Methodology and Data Description 2](#_Toc31743400)

[4 Exploratory Data Analysis 4](#_Toc31743401)

[4.1 Number of Members by Current Party 4](#_Toc31743402)

[4.2 Members by Committee Assignment 4](#_Toc31743403)

[4.3 Committee Leaders by Policy Area and Party Type 6](#_Toc31743404)

[4.4 Bills by Current Party of Sponsor 6](#_Toc31743405)

[4.5 Bills Introduced by Policy Area and Current Party of Sponsor 7](#_Toc31743406)

[4.6 Bills Introduced by Policy Area and Current Party of Cosponsors 8](#_Toc31743407)

[5 Multi-Directional Graph and Weighted Graph 9](#_Toc31743408)

[5.1 Bills Sponsored and Co-Sponsored by various Senators 9](#_Toc31743409)

[5.2 Illustration of multiple directed edges 10](#_Toc31743410)

[5.3 Illustration of weighted edges 10](#_Toc31743411)

[6 Community Detection by Louvain Clustering 12](#_Toc31743412)

[6.1 Detect Communities 12](#_Toc31743413)

[6.2 Visualization and Findings 12](#_Toc31743414)

[7 Key Players Detection 14](#_Toc31743415)

[7.1 Key Players Detection by in-degree Centrality and Insights 14](#_Toc31743416)

[7.2 Key Players Detection by In-degree 15](#_Toc31743417)

[7.3 Key Players Detection by Closeness Centrality 16](#_Toc31743418)

[8 Community and Key Player Detection for Policy Areas 18](#_Toc31743419)

[8.1 Health 18](#_Toc31743420)

[8.2 International Affairs 19](#_Toc31743421)

[9 Summary 21](#_Toc31743422)

# Abstract

Our project studied the community structure of Congressional networks of Members of Congress based on the sponsorship/cosponsorship of the legislation. The subgraphs within the networks of each communities indicated the collaborative nature of Congressmen from different parties (Republicans and Democrats) and clear demonstration of political polarization in different policy areas. Using evaluation metrics such as in-degree and closeness centrality, we detected the key players within the Congressional networks.

# Introduction and Problem Statement

The 116th meeting of United States Congress convened on 1/3/19 and will end 1/3/2021. Legislators can make public their support for a particular bill by cosponsoring it in the US Congress. Given the collaborative nature of the Members of Congress during the process of lawmaking, it is interesting to find the collaborative ties among these Congressmen. We also note the philosophical differences between the 2 parties (Republicans vs Democrats) on myriad policy areas, may lead to different Congressional networks constructed. Generally, the philosophy of a Republican is more alike to individual freedoms, rights and responsibilities. On the other hand, Democrats emphasizes more on the equality and social/community responsibility.

Using network theory, key members of the Congress who have more political influence and the communities of committees can be detected without prior political knowledge on the legislation system and membership. This project therefore aims to study the Congressional networks in the context of graph to detect the community and key player to better understand the collaborative behaviour of the Members of Congress.

# Methodology and Data Description

It is noted that the legislation contains only one sponsor with many cosponsors. To construct the Congressional network, the nodes and links between these nodes have to be defined. The nodes are represented by the Members of Congress and the edges and corresponding edge strength were based on the linkage between Congressmen who served as a sponsor or cosponsor of the same bill. This will result in a weighted network. We will then apply relevant network algorithm, i.e. Louvaine clustering, to find the community structure of the legislation sponsorship-cosponsorship network, in particular to specific policy areas such as health and international affairs. We will then detect key players using evaluation metrics such as in degree and centrality.

Our project used two of the “The 116th U.S. House of Representatives” datasets provided in Kaggle, namely the dataset pertaining to Members and Bills to construct the Congressional Networks. The Members dataset consists of 443 observations which includes information about all current and previous members of the 116th House of Representatives. The rows are indexed by name\_id and the columns include name, state, url, chamber, current\_party and committee\_assignments. The Bills dataset consists of 5806 observations which includes all bills introduced in the House of representatives indexed by bill\_id and columns include title, sponsor, cosponsors, related bills, policy\_area, subjects, committees, bill\_progress, summary, date\_introduced, number and bill\_type. Further information on the variables are indicated in the following tables.

1. Members Dataset

|  |  |
| --- | --- |
| Variables | Description |
| name\_id | unique name\_id of member |
| name | name of member |
| state | name of state |
| url | member url on congress.gov |
| chamber | currently limited to the House |
| current\_party | Republican, Democratic or Independent |
| committee\_assignments | list of committees to which the member is assigned |

1. Bills Dataset

|  |  |
| --- | --- |
| Variables | Description |
| bill\_id | unique bill or resolution id in the form H.R.#, H.RES.#, H.CON.RES.# or H.J.RES.# |
| title | Official title of bill or resolution |
| sponsor | name\_id of bill sponsor |
| cosponsors | list of name\_id of cosponsors |
| related bills | list of related bills |
| policy\_area | A single policy area assigned to every bill or resolution |
| subjects | list of subjects assigned to every bill or resolution |
| committees | list of committees to which the bill was referred |
| bill progress as of the last update | bill progress as of the last update |
| summary | bill summary |
| date\_introduced | date bill was introduced |
| number | number of bill or resolution |
| bill\_type | one of H.R., H.RES., H.J.RES. and H.CON.RES. |

# Exploratory Data Analysis

This section analyses some of the variables namely current\_party, committee assignments, sponsor, cosponsors and policy areas in the dataset to (i) better understand the legislative membership and (ii) explore the relationship between the policy area of the bills and the current party of members.

## Number of Members by Current Party

A screenshot of a cell phone

Description automatically generated

Out of 443 members of the legislation, 203 (45.8%) and 239 (54.0%) are from the party of Republican and Democratic respectively. There is only 1 independent legislator. This is in line with the majority win by the Democratic Party in the House during November 2018 midterm elections.

## Members by Committee Assignment

A picture containing object, comb

Description automatically generated

The top 5 committee assignments are in the areas of transportation and infrastructure, financial services, armed services, energy & commerce and appropriations. It was observed that the Republican and Democratic parties are fairly distributed in terms of their committee assignments. Although Democrats are more represented in majority of the committee assignments, it is notable that “Selected Committee on the Modernization of Congress” and “Ethics” committees are equally represented by both parties. Interestingly, the ‘minority leader’ and ‘minority whip’ are represented by the Republicans while ‘majority leader’ and ‘majority whip’ are represented by the Democrats. This seems to be in line with the respective political ideologies of the parties in terms of social and human ideas. For instance, Democrats believes more in community and social responsibility while Republican advocates on individual rights and justice.

## Committee Leaders by Policy Area and Party Type

A screenshot of a cell phone

Description automatically generated

We can see that the committee chairs are all represented by the Democrats while all the ranking members are represented by the Republicans. Being the majority party, the Democrats are to select the Committee Chairs and Republican as the minority party will select the Ranking Members to lead them. We can already see some collaborative ties within the same party in the election of their leaders.

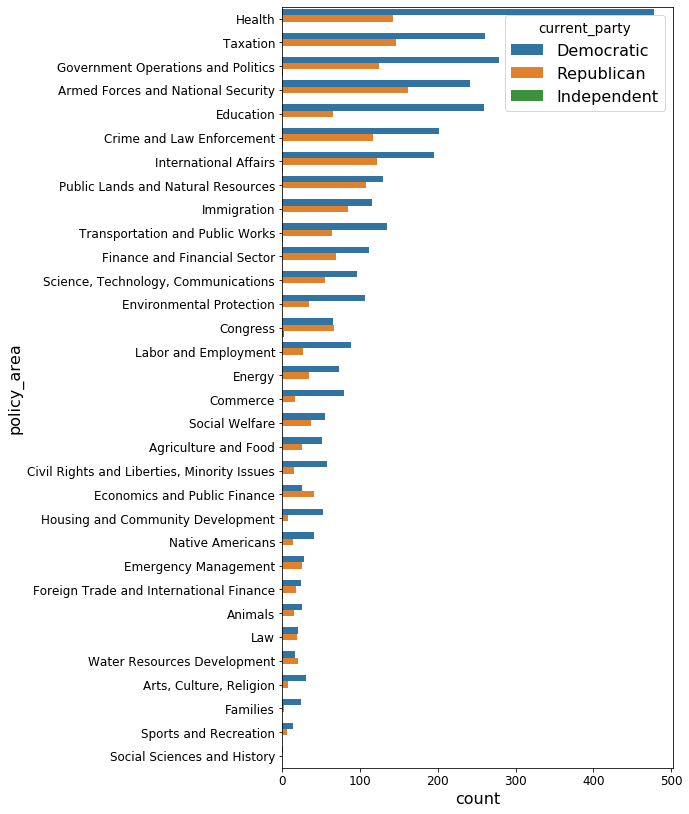
## Bills by Current Party of Sponsor

A screenshot of a cell phone

Description automatically generated

Bills under the Sponsorship of Democratic Party (67.4%) were more than twice than that of the Republicans Party (32.4%). This indicated the legislative control of Democrats in the House of Representatives.

## Bills Introduced by Policy Area and Current Party of Sponsor



We merged the two datasets “Members” and “Bills” to explore the relationship between the bills that are introduced by their policy areas and the current party of the sponsor of bill.

We found that the 5 most common policy areas of bills introduced in the House in 2019 are Health, Taxation, Government Operations and Politics, Armed Forces and National Security and Education. Majority of the bills introduced were under the sponsorship of Democrats in most of the policy areas. The Democratic legislation is not surprising due to the dominance of Democrats in the House. It is notable that in areas such as Congress, Economic & Public Finance and Water Resources Development, there are more bills introduced under the sponsorship of Republicans than Democrats.

## Bills Passed by Policy Area and Current Party of Sponsor

A screenshot of a cell phone

Description automatically generated

The bills that were introduced have to be passed by the House of Representatives which comprised vastly of Democrats before being voted in by the Senate Republicans and passed up to the President to be enrolled. We will look at the number of bills passed in the house to explore the political behaviours of these Congressman.

The top 5 bills passed by the House are in the policy areas of Government Operations and Politics, International Affairs, Finance and Financial Sector, Armed Forces and National Security and Emergency Management. We can see that despite the large number of bills introduced (Fig 4.5), only a small number of bills were passed through by the House of representatives. This is so as the passing of the bills require bipartisan approval which is often hard to achieve due to the philosophical differences in the two parties and degree of polarisation in political behaviours among members of the same party.

From the chart, majority of the bills that were passed were sponsored by Democrats. In policy areas Native Americans and Law, there were more bills passed that were under the sponsorship of Republicans than Democrats. Interestingly, the Republicans had introduced lesser bills than the Democrats in these areas but had a higher number of bills passed, indicating a higher pass through rates. This may indicate some degree of political polarisation in the group of Democrats where some of them exhibit more conservative behaviours in these policy areas despite their political orientations and were more likely to vote with the opposite party, resulting in a higher pass through rates for the bills introduced by the Republicans.

Even though there were more bills introduced under the sponsorship of Republicans than Democrats in policy areas of Congress, Economic & Public Finance and Water Resources Development, none of the bills introduced by the Republicans were passed by the House. This may indicate some degree of political polarisation in the group of Republicans where some of them exhibit more liberal behaviours in these policy areas, resulting in the 0 pass through rates in bills introduced by the Republicans in these areas.

The next few sections aim to modify the raw structured data for graph analysis, look into the collaborative nature and identify key players of the Congressional networks.

# Multi-Directional Graph and Weighted Graph

This section aims to highlight novel techniques used to transform relevant features of the raw structured data into multi-directional graph followed by weighted edges graph. Thereafter, community detection algorithm is applied to the transformed data.

### Bills Sponsored and Co-Sponsored by various Senators

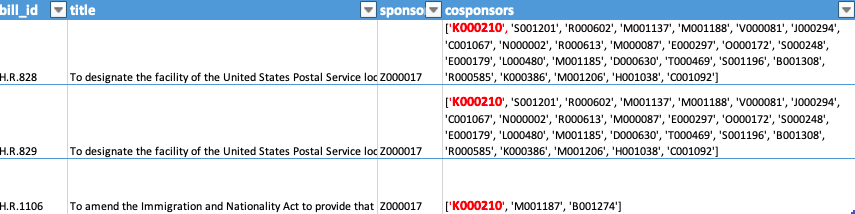


Table 5.1 contains an extraction of raw data “house\_legislation\_116” which documents every single bill sponsored and cosponsored by various party members. As seen in the table , a single member (Z000017) who sponsored one single bill (H.R.828) were co-sponsored various members in the first row. This fits our understanding of having a hub receiving many indegrees in typical community detection techniques.

However , the key difference is that a single co-sponsor (K000210) can support the same member (co-sponsor) three times for three different bills (H.R.828, H.R.829, H.R.1106) tabled as illustrated 5.2 below.

## Illustration of multiple directed edges

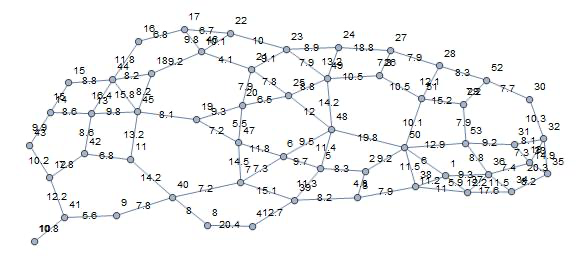


For the purpose of graph analysis , this insight obtained has been converted to [multi-directed graph](https://networkx.github.io/documentation/networkx-1.10/reference/classes.multidigraph.html). The algorithm captures the number of parallel edges (relationships) between the sponsors and co-sponsors. Each edge can hold optional data or attributes. Finally , in order to simplify the graph and reduce the number of edges, the multi-directed graph are converted to weighted edges for analysis as seen in illustration 5.3 below.

## Illustration of weighted edges



As seen above , the three directed edges have been reduced to one single edge with a weightage of three. This translates to an in-degree of three and the sum of all the directed weighted edges from the neighbours to a given node is the main determinant of centrality measurements. Illustration 5.4 below gives an example of nodes with weighted edges when constructed.



In essence, sponsors and cosponsors who tend to support each other's bills are more densely connected. This fits the situation well as same party members show a higher tendency to support each other bills for the party’s vested interests in Congress. Upon applying community detection technique, more densely connected nodes tend to form the same cluster.

Intuitively , we believe that same clusters should consist of members of the same party. In the next section of the report , we will apply Louvain community detection techniques to verify our hypothesis.

# Community Detection by Louvain Clustering

## Detect Communities

*#use louvain community detection algorithm to detect communities in G*

communities =[]

louvain = community.best\_partition(G, weight = 'weight', random\_state=42)

for i in set(louvain.values()):

nodelist = [n for n in G.nodes if (louvain[n]==i)]

communities.append(nodelist)

set(louvain.values())

From the above codes, we can get the communities groups by the ‘community.best\_partition()’. It will give each nodes a value of integers such as 0, 1, 2…. For the nodes with same value, they belongs to the same community. ‘set(louvain.values())’ will print out the unique values assigned to each nodes and the output is shown as below:

{0, 1, 2}

So we can conclude that there are three communities in total and then we separated and saved the nodes of each community in variable ‘communities’, namely the first list in communities is community 0, second list in communities is community 1 and third list is community 2.

From the data, we know that there are three parties. Let us take a look at the composition of the three communities:

A screenshot of a cell phone

Description automatically generated

From the table above, we can find that the members in community 0 are mainly from Republican party while those of both communities 1 and 2 are mainly from Democratic party.

## Visualization and Findings

Let us visualize the communities to help to have a good understanding:

*# function for setting colors of nodes and edges*

def get\_paired\_color\_palette(size):

palette = []

for i in range(size\*2):

palette.append(plt.cm.Paired(i))

return palette

The above codes are defining a function to generate required number of different colors.

*#make plot using matplotlib, networkx spring\_layout, set\_colors using cluster\_count and get\_paired\_color\_pallette*

clusters\_count = len(set(louvain.values()))

plt.figure(figsize=(10, 10))

light\_colors = get\_paired\_color\_palette(clusters\_count)[0::2]

dark\_colors = get\_paired\_color\_palette(clusters\_count)[1::2]

g = nx.drawing.layout.spring\_layout(G, weight = 'weight', seed = 42, threshold = .0000000001)

for i in set(louvain.values()):

nodelist = [n for n in G.nodes if (louvain[n]==i)]

print(len(nodelist))

edgelist = [e for e in G.edges if ((louvain[e[0]]==i) or (louvain[e[1]]==i))]

node\_color = [light\_colors[i] for \_ in range(len(nodelist))]

edge\_color = [dark\_colors[i] for \_ in range(len(edgelist))]

nx.draw\_networkx\_nodes(G, g, nodelist=nodelist, node\_color=node\_color, edgecolors='k', label = i)

nx.draw\_networkx\_edges(G, g, edgelist=edgelist, alpha=.5, edge\_color=edge\_color)

plt.title('Louvain clustering: Unfiltered', fontdict={'fontsize': 25})

plt.legend()

plt.axis('off')

plt.show()

The above codes to select three colors for the nodes and three colors for the edges and then draw the communities by the network drawing function.

Following is the visualization output:

A picture containing text

Description automatically generated

From the graph we can see that the red nodes (community 2) are mixed with green nodes (community 1). Based on the previous composition table, we know that the majority of both communities 1 and 2 are from Democratic party. So the graph match what the data shows. And it tells us that there are two factions inside Democratic party.

From the visualization we can find that different parties are separated quite clearly. It shows that the community detection results are not bad as it can differentiate different parties properly.

# Key Players Detection

## Key Players Detection by in-degree Centrality and Insights

*#set community\_of\_interest, change to 1 or 2 for the rest two communities*

community\_of\_interest =0

*#create subgraph with nodes limited to the community\_of\_interest*

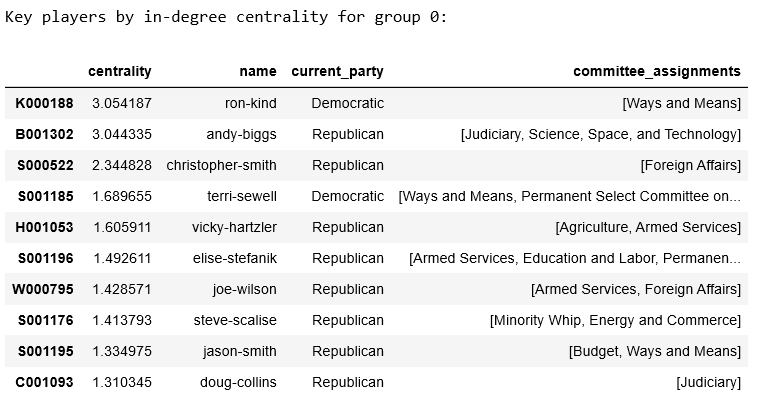
subgraph0 = MD.subgraph(communities[community\_of\_interest])

*#sort members of community\_of\_interest by in\_degreee centrality*

pd.DataFrame.from\_dict(nx.algorithms.centrality.in\_degree\_centrality(subgraph0), orient = 'index', columns = ['centrality']). merge(members[['name','current\_party’, 'committee\_assignments']], how = 'outer', left\_index = True, right\_index = True).sort\_values(by= 'centrality',ascending = False).head(10)

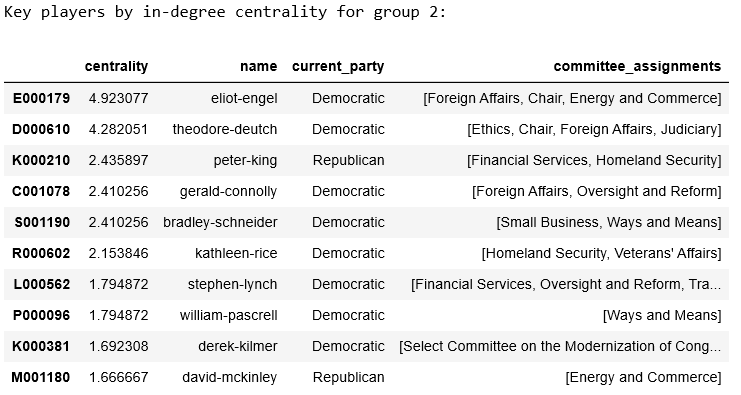
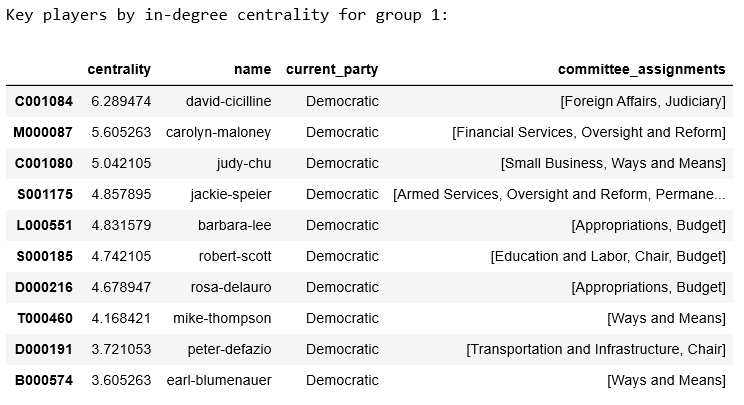
By the above codes, it calculated the in-degree centrality and sort the data frame based on it. It is for community 0. If you want to get the results for community 1 and 2, then change the community\_of\_interest to the corresponding value.

Following is the output for community 0:



The top 5 key players are circled by the dotted red rectangle. What is interesting, in the top 5 key players in group 0, there are two people are from Democratic party even though the majority of the members in community 0 are from Republican party. This could be an indication of [Bipartisanship, sometimes referred to as nonpartisanship](https://www.thebalance.com/bipartisan-definition-benefits-examples-4589699), in which opposing political parties find common ground through compromise. It also could be due to hidden political motive in light of the power struggle in the US Congress or they may be just undercovers.

Let us take a look at the results for communities 1 and 2:



From community 2, we can also find one Republican party member among the Democrats, This should be the similar reason talked previously.

## Key Players Detection by In-degree

*# key players using degree for community 0*

sortdeg = sorted(subgraph0.in\_degree, key=lambda x: x[1], reverse=True)

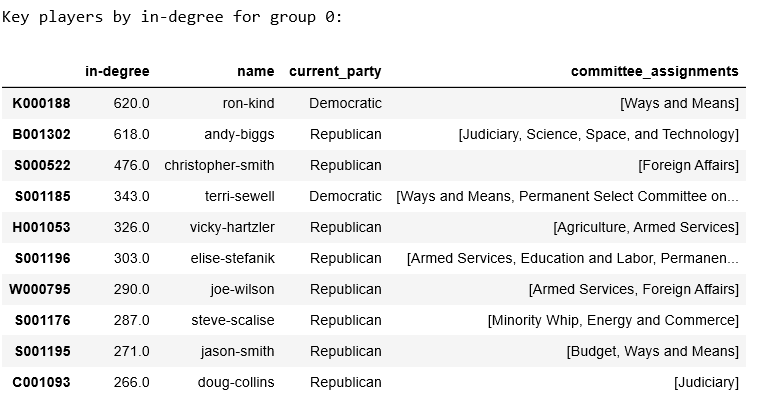
deg =pd.DataFrame(sortdeg).rename(columns={1: “in-degree"})

deg.set\_index(0, inplace=True)

deg.merge(members[['name','current\_party', 'committee\_assignments']], how = 'outer', left\_index = True, right\_index = True).sort\_values(by= ‘in-degree',ascending = False).head(10)

Let us take community 0 as an example. If you want to investigate on communities 1 or 2, you can do that by change the subgraph0 to subgraph1 or subgraph2 instead.

Based on the following output, we can find that the top 5 key players are the same as previous based on in-degree centrality due to similar technique:



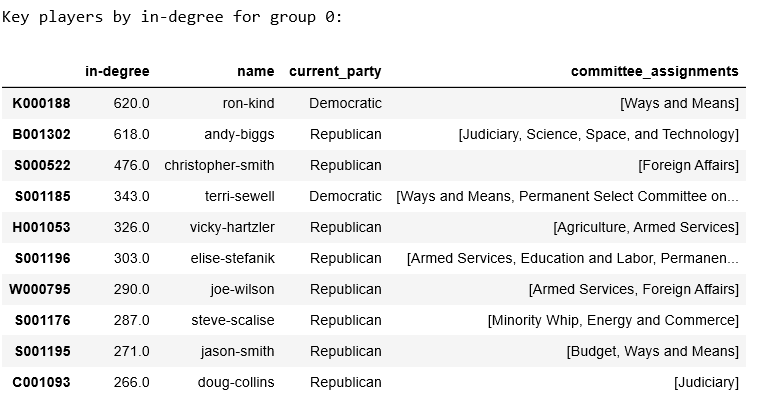
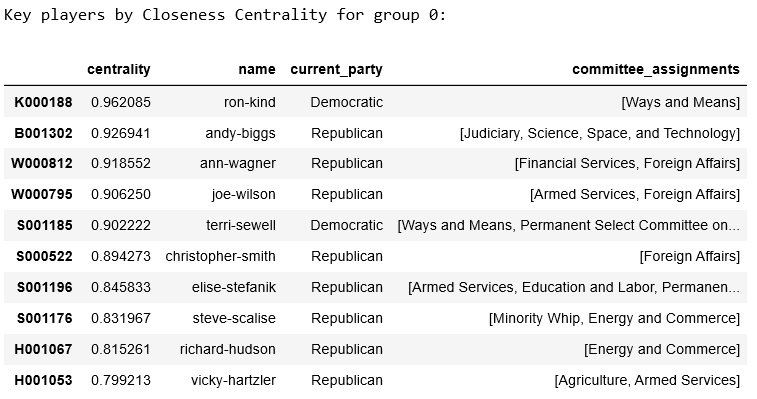
## Key Players Detection by Closeness Centrality

pd.DataFrame.from\_dict(nx.closeness\_centrality(subgraph0),orient='index',columns=['centrality']).merge(members[['name', 'current\_party','committee\_assignments']], how='outer', left\_index=True, right\_index=True). sort\_values(by= 'centrality',ascending = False).head(10)

Similarly, we will use community 0 to demo this technique. If you want to take a look at communities 1 or 2, then just modify the subgraph0 to subgraph1 or subgraph2.

Closeness Centrality is the reciprocal of the sum of the length of the shortest paths between the node and all other nodes in the graph. The more central a node is, the closer it is to all other nodes. And normally we will do normalization for the results.

The results are shown as below (let us put it together with the output measured based on in-degree to do a comparison):



We can find that 3 people in the top 5 key players measured by in-degree also in the top 5 key players measured by closeness centrality. It make sense because if a node has a high degree, it will be close to many other nodes which makes its closeness centrality larger. Similarly liking the air routes illustrated below:

A close up of a map

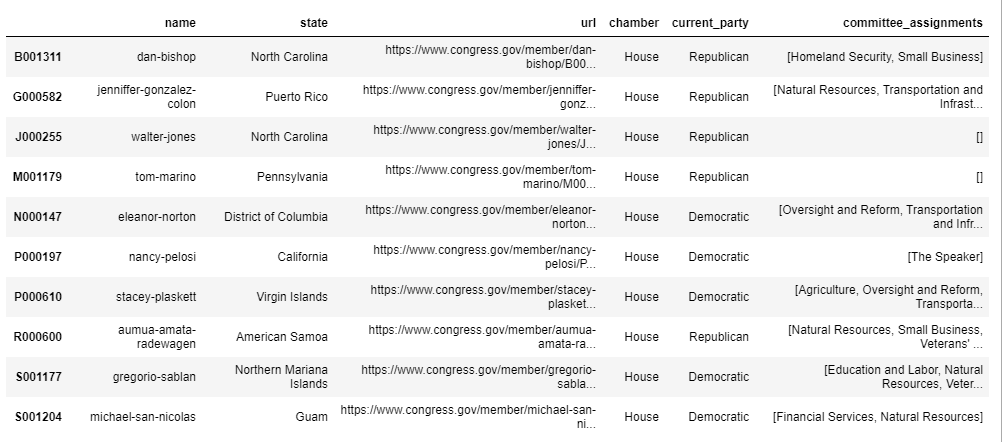
Description automatically generated

Liking the hubs, they have a much higher degree which means they are directly connected to many other nodes and make the distance between them and other nodes shorter. That will result in a higher closeness centrality for them.

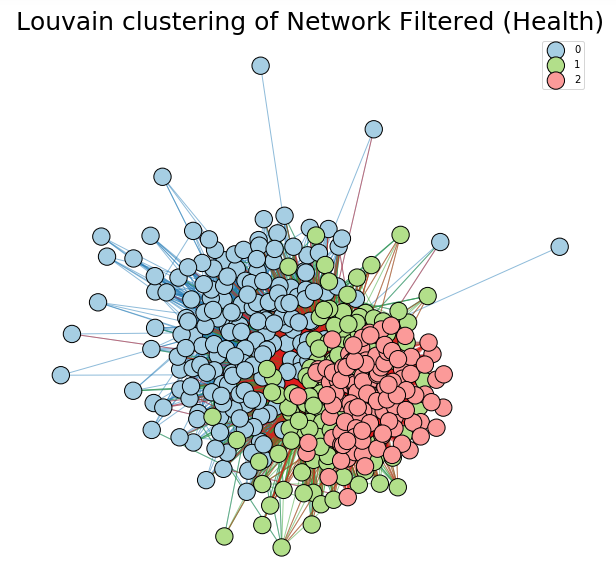
# Community and Key Player Detection for Policy Areas

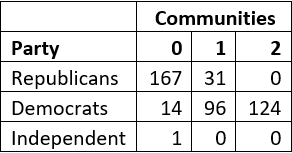
## Health

The following Members of Congress (MOC) were observed to not have been involved in any bills related to health policies:



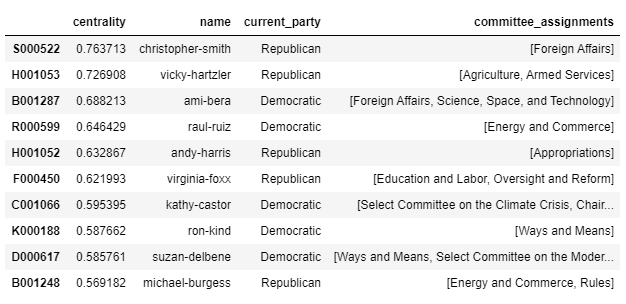
With the exception of the above 10 members, we looked into the interaction of the remaining MOCs in this policy area. In order to do that, we used the Louvain method for community detection, resulting in the 3 communities as shown below.





According to the figures, across the communities, we can see that the Republicans largely reside in Community 0 (167 members or 84.3%) while the rest are in Community 1. Meanwhile, the Democrats are more divided, with the largest group being only 124 members (53.0%) in Community 2. 96 of them (41.0%) formed Community 1 along with only 31 Republicans. As members have a tighter sponsor-cosponsor relationship for each other’s legislation bills within each community, this suggests the possibility of different groups forming within the Democratic Party, possibly divided due to their views on how the health sector should be run in the country.

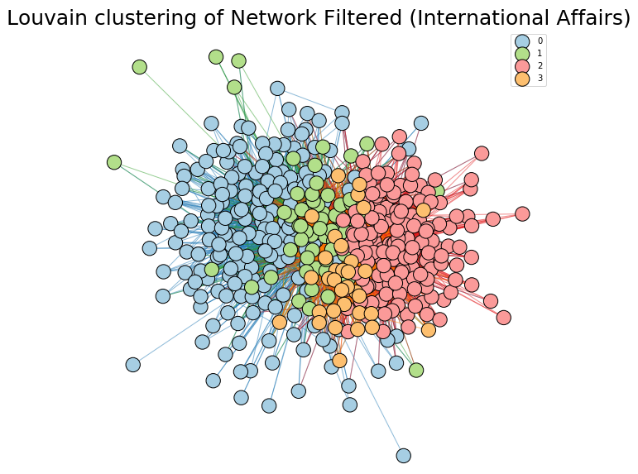
This trend can be further seen from the key players detected, taking Community 0 as an example.

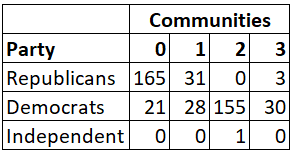


Similar to 7.1, we tabulate the top 10 members in Community 0 based on the centrality metric, sorted in descending order. We note that although the Democrats make up only 15.7% of the community, they form half of the top 10 members, who are the key players in this community. This suggests that these Democrats have actively been receiving and giving support to the bills proposed by the Republican-denominated community, and are a central part of it. It implies that these Democrats are in fact more aligned with the Republicans in this policy area.

## International Affairs

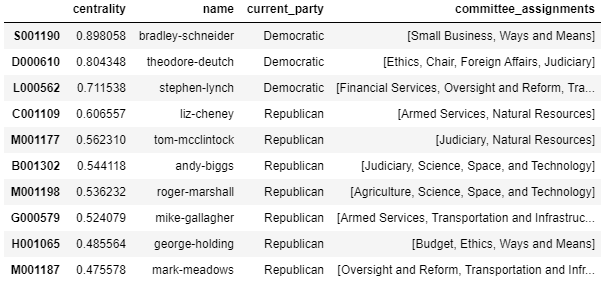
Amongst the 434 members that have participated in proposing and supporting bills with regard to International Affairs, we used the Louvain method again, which results in 4 communities this time round.





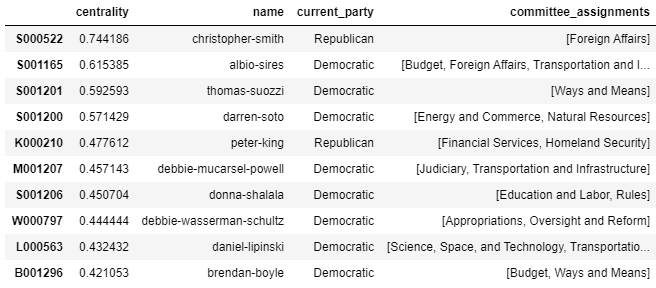
As with the communities seen for the Health policy area, we can see a larger spread of the Democrats across all communities, compared to the Republicans. 66.2% of the Democrats (155 members) are in Community 2, with the rest split relatively evenly across the other communities. Meanwhile, Community 0 holds 82.9% of the Republicans, implying that these members are more aligned with the party’s position on issues related to International Affairs.

As with 8.1, we take the data of Community 0 to generate the key players.



It is perhaps even more apparent this time round, that although Community 0 is Republican-dominant, the most supportive key players are actually Democrats.

However, at the same time, looking at Community 3 which is Democrat-dominant, we obtain the following table.



The same situation as above has occurred as well, where although the Democrats make up the bulk of the community, the most central key player happens to be a Republican. This suggests that some of the Republicans actually do share different views with the rest of their party members, although the proportion is still much lesser as compared to the Democratic Party.

# Summary

Based on the report , one may deduce that community detections through sponsorship and co-sponsorship has provided some preliminary assessment into the true political orientation of party members in the Congress. On the other hand, the analysis also shed some light on the extent of Bipartisanship in each policy area for which Republicans and Democrats could potentially work together to pass the bills.

The report documents an in-depth understanding of the dataset and choice of selected features, application of multi-directional graph with parallel and weighted edges, basic analysis through the Louvain Clustering and identification of key influencers via various metrics. Nevertheless, additional data is required to provide new features and derive a more justified analysis.