Twitter and The House of Representatives:

Identifying political subgroups

Asger Vibel, Phillipp Schmalen, Simon P. von der Maase*

May 25, 2018

*Ms. Students in Economics and Political Science at the University of Copenhagen

Contents

1 Introduction(Right now just the stuff we sent Andreas - needs revision!) 3

1.1 Literature Review 5

2 The Data 6

2.1 Propublicas Congress API (Asger) 7

2.2 Twitter 7

2.3 Caucuses (Philipp) 9

3 Methods 10

3.1 Centrality measures 10

3.2 Modularity measures 10

3.3 Similarity measures 11

4 The twitter network(Simon) 12

4.1 The Full network(Simon) 12

4.1.1 Centrality 14

4.1.2 Modularity 16

4.2 The Republican subnetwork 17

4.2.1 Modularity 17

4.2.2 Similarity 17

4.2.3 Centrality 18

4.3 some bridge thingy leading to the next section 20

5 Analysis 20

6 Results 23

7 Discussion 23

8 Perspectives 23

9 Conclusion 23

10 Bibliography 24

11 Appendices **25**

11.1 The Democrat subnetwork 25

11.1.1 Modularity 25

11.1.2 Similarity 26

11.1.3 Centrality 26

11.2 Whole network 27

1 Introduction(Right now just the stuff we sent Andreas - needs revision!)

Social network relations as indicators of intra-party factions

To what extent can social media relationships between politicians be used as indicators for political subcomponents and intra-party factions in two-party systems.

The Idea In (+2) multi-party systems it is custom to measure the effective number of parties with either the Laakso or the Golosov formula (References). This formula, however, only serves to reduce the actual number of parties to a lower number of "effective parties";. As such discounting small, insignificant parties or merging parties with very similar voting pattern.

As such these methods are of very little use if we want to asses intra-party factions. In multi-party system with an electoral system of proportional representation this is of less concern since strong intra party factionalism will most likely lead to splits and splinter parties.

In electoral system of first-past-the-post it is a different matter. These systems almost exclusively enables two-party constellations. Parties thus have a stronger incentive to stick together, "not out of love but out of necessity". The risk and cost of splitting out to a third party is to great for any faction to take.

This blurs the political landscape, making it harder to identify more detailed contours. Of course domain experts such as political analysts and pundits routinely identifies intra-party factions on account of qualitative estimations. We what to see is we can augment their analytical toolbox with a more quantitative approach.

The project is simply to explore to what extent social media relationships can be used as indicators for such intra-party factions. Specifically we use the US House of Representatives and Twitter for prof-of-concept. The US is chosen since

individual politicians voting patterns are quite easily obtained. This information can then be used to validate or discredit the value of the constructed network. If the construct does prove useful, we hope that the approach can be applied to parliaments in which individual voting behavior is less accessible but where politicians are still active users of social media. In the discussion we of course comment on to what extent it is prudent to infer from the US to other countries.

More Concrete We create A network graph containing all reciprocal twitter links between House-members. We the explore this graph along two subgraphs corresponding to each party (D & R). Through various network attributes and statistics we identity central figures as well as clusters. these is compared the actual caucuses and know intra-party factions.

To further validate the usefulness of our network statistics, such as Jaccard similarity, Adamic/Adar similarity and Louvian modularity we construct a prediction scenario from given data from the Congress API and our two subgraphs (R & D).

We utilize the various network statistics to predict the shared voting behavior - within each party - of US House-members. That is; we predict to what extent two house-members (a dyad), within the same party, tend to vote the in accordance with each other.

The network constructed features are then evaluate in comparison with other features thought to influence intra-party voting behavior such as age, gender and geography.

Finally we discuss the perspectives of this approach, to what extent it can be generalized to other countries and how sensitive it is to changing allegiances over time.

something something leader roles...

1.1 Literature Review

Relevance After its launch in 2006, Twitter gained much attention by politicians who could use it mainly to communicate more directly with the public and to a broader audience. After just four years, almost 60% of congress members actively used Twitter (CITE Golbeck et al. 2010 and Straus et al. 2013). Today, around 98% have an account and use it actively, with a yearly-average of two Tweets per day (refer to descriptives, table XX) [1]. Beyond broadcasting messages, Twitter's second use is to obtain information by those you follow.

Golbeck et al. 2017 - Congressional Twitter Use Revisited on the Platform's 10-Year Anniversary
Tweet habits of congressWomen did not change within eights years from 2009 to 2017. Based on this, it is reasonable to assume that Twitter habits do not change dramatically in the near future. Hence, it is reassuring for the analysis, that its implications are not undermined by the dynamics of Twitter.

Halberstam et al. 2016 - Homophily, group size, and the diffusion of political information in social networks: Evidence from Twitter.
"Twitter - higher levels when measuring network segregation in terms of connections among Twitter users from different ideological groups. Social media, crowding out other forms of media–> increased polarization of the legislators"

Hemphill et al. 2013 - What's Congress doing on Twitter
"Congress appears to use Twitter as yet another broadcast mechanism rather than as a way to engage in dialogue with the public or as a "call to action" to organize constituents."

Bogle et al. 2013 - Comparison of Congressional Social Media and Cosponsorship Networks in the 112th House of Representatives
"Congressional Twitter networks tend to be structured by ideology and the

[1] also refer to http://www.tweetcongress.org/

leadership hierarchy of the parties."; "useful, but not sufficient, for ideological prediction of the members' voting patterns."

Lee 2017 - Detecting Changes in Congressional Twitter Networks over time
Even interaction via tweets is rather stable in a two year period. Short fluctuations

Mankad 2015 - Analysis of multiview legislative networks with structured matrix factorization: Does Twitter influence translate to the real world?
Motivating point: Those who are important and central within the Twitter network, also are central political figures in the real-world. Those with the highest centrality, are also mentioned more often by the media. "important politicians in Twitter networks are associated with real-world leadership positions, and that rankings from the proposed method are correlated with the number of future media headlines"

Contribution This work contributes to the literature in XX ways. Firstly, a Twitter follower network is constructed for the current, 115th, Congress. The structure of the network gives a first imporession of how the two-party system is mirrored by relations on social media. Secondly, information on caucusses is used to obtain a more nuanced view beyond the two parties. We can check if regular interaction within these groups is detectable in the social media network. [other important implications]

2 The Data

For the project at hand we have scrapped data from three different online sources; Propublicas Cnogress API, The Twitter Rest API and [insert whatever Philipp did to get the caucuses]. The following sections briefly outlines the procedure and presents argumentation regarding central choices made throughout the scrapping process.

2.1 Propublicas Congress API (Asger)

Propublicas Congress API is a channel from which you can get data about the House of Representatives and the Senate. There is a variety of data available. Hence, you can find details about the members such as gender, birthday, party and account name on different social medias together with information about the daily work of the Congress. For our project we are in the characteristics of the members and there twitter account name. By using the Congress API to get the twitter account, we ensured that the accounts is updated, and that we only scrape twitter for member how are in office.

On the Congress API every legislator has an individual ID. Using the ID's of two house-members one can obtain the voting behavior between two members. Shared voting data has in multiple papers been used to identify community structures in the House of Representatives. (Reference: Party Polarization in Congress:A Network Science Approach) Using this option we scraped information about the shared voting behavior between all members with-in each party. Using the characteristics of two members for which a shared voting was observed four additional features was created; same gender, same age same state and from neighbor state, see table X. *Agree percentage* is the level of agreement between two legislator in roll-call voting. *Same sex* is a binary variable indicating if the two members has the same sex or not. *Same age* indicates whether the age difference is less than 10 years. *Same state* and *From neighbor state* indicates whether the member is from the same state or from two states that shares a border. In line with the purpose of this project this data is later on merged with network statistics obtained form the scraping the members twitter account.

2.2 Twitter

The reason for scrapping Twitter was to get full information on the reciprocal twitter-relationship between any two House-members. As such we used Twitters REST API. Using the Twitter handles from the Congress API we looped over all House-members twitter-profiles and obtained informations regarding

	Democrats	Republicans	All
	Mean	Mean	Mean
Agree percentage	92.6	92.7	61.1
Same sex	0.565	0.829	0.68
Same age	0.437	0.497	0.47
Same state	0.066	0.036	0.04
From neighbor state	0.089	0.105	0.09
Distance btw. districts	23.69	16.17	19.83
Number data points	18,528	26,627	111,639
House-members	197	231	

Table 1: *Some text

whether House-member-A followed House-member-B and vice versa. When all information regarding the relationship between A and B was scrapped, we made sure not also to scrap the same relationship once more as B to A, thus roughly cutting the scrapping time by half[2].

The information we obtained corresponded to a directed network as House-member-A, on twitter, is free to follow House-member-B without any reciprocal gesture from House-member-B herself[3]. Alas, we where only interested in creating a simple network with no directions, as this better correspond to the notions of political sub-components. Thus we choose only created edges whenever both House-member-A and House-member-B had chosen to follow each other. This of course excluded all House-members with no followers or House-members who them self does not follow any other house-members. Just as we excluded the politicians with limited online presence when choosing to focus exclusively on politicians with twitter profiles, we now further excluded politicians with limited online activity. There are plenty reason to believe that this exclusion pattern correlates with some underlying dimensions; rural politicians could be less active online then urban politicians, older less active then younger and so on. To throughly asses such a pattern would be a project onto it self, so for now we

[2] The actual scrapping did not take that long, but Twitter only allows a relatively small number of relationships to be scrapped every 15 minutes; some 350

[3] e.g. accepting or re-following

settle for a fair warning and candid prudence.

The script regarding both scrapping and the construction of edges can be found in **appendix X**

2.3 Caucuses (Philipp)

A caucus is a group of congress members that forms a subgroup of the main political party. Its members meet regularly to discuss legislative objectives and align individual views with broader group interests. On this basis, its members interact with each other and are likely to have similar goals.
Possibly, the Twitter network, as well as the voting behavior, mirrors those subgroups. It would add a more nuanced view beyond the bipartisan structure.
Ideally, data for each member should contain affiliations with caucuses. Unfortunately, this is neither provided by the Congress API, nor is there a comparable centralized database available. Still, Wikipedia gives updated information on the major caucusses and their members. Each Wikipedia page of a caucus is structured differently, but commonly includes a list of members, where each member is displayed as a link to her own Wikipedia page. Knowing this and making a list of the caucusses by hand, information is collected in two steps. First, all link texts are retrieved from a caucus' page. Second, those texts are compared to the list of currently serving congress members. Then, the resulting intersection of the two sets gives the caucus members.[4]
With the caucus data, it can be tested if closer interaction within a subgroup, implies a sub-cluster in the Twitter or voting behavior network. For the latter, the information gain by adding the caucus data could be limited, due to low variation of the voting data. However, for the twitter network there could be more pronounced clustering. The repeated interaction of members with similar interests and goals, hypothetically increases the likelihood of strong ties on social media. A simple indication of this would be to color code the caucuses in the network graphs and look for distinct patterns.

[4]The script can be found in XY. (add reference)

3 Methods

3.1 Centrality measures

Betweenness Centrality is given by:

$$Degree\ Centrality = \frac{\text{number of neighbors a node has}}{\text{number of neighbors a nodes could possibly have}}$$

As such Degree Centrality is a measure of how well connected a node is taking into account to the size of the network. Nodes with a high Degree Centrality correspond to highly connected nodes[5].

$$Betweenness\ Centrality = \frac{\text{num. shortets paths through node}}{\text{all possible shortest paths}}$$

Betweenness Centrality captures bottleneck-nodes rather then highly connected nodes. Together Degree Centrality and Betweenness Centrality can help us identify central nodes in our network. In figure 2 we plotted the two measurements against each other and mark the top 3 politicians from each party.

3.2 Modularity measures

Louvain Modularity To identify clusters in networks, the Louvain method is used (CITE Blondel et al. 2008). It is based on the modularity measure, which indicates how a network is split into subgroups. High modularity of a network means that there are subgroups, which are densely connected within, but sparsely connected to points outside of their subgroup. The optimzation procedure of the Louvain method has two steps. First, it optimizes the modularity of a given network partition. Second, it aggregates nodes of a subgroup and generates a new network with each subgroup as a node.

[5] If there - as in this case - is no self-loops then the max number of neighbors a node can have is $all\_nodes - self$

Best Patition: By "Best Partition" we mean computing the partition of the network graph which maximizes the Louvain modularity**(ref. community documentation...)**. Thus given the measurement of Louvain modularity a "Best Partition" splits the network up in most pronounced clusters.

3.3 Similarity measures

first Rand and adjusted Rand, then Jaccard for partition then for the network.. **You also you the Jaccard score to compare the partitions...**

The Jaccard coeficient - in a network setting - mearsurement for the numbers of common neighbors between to nodes. That is, to what extent does $node_1$'s set of nieghors overlab with $node_2$'s set of nieghbors:

$$\text{score}(node_1, node_2) := \frac{|\Gamma(node_1) \cap \Gamma(node_2)|}{|\Gamma(node_1) \cup \Gamma(node_2)|}$$

Adamic/Adar...

The Ademic/Adar coeficients is a somewhat related measure. It is defined as shuch:

$$\sum_{\text{z : neighbors shared by } node_1, node_2} \frac{1}{log(frequency(z))}$$

The point is that this measurement gives a higer wieght to rarely shared neighbors, such that:

$$\text{score}(node_1, node_2) := \sum_{z \in \Gamma(node_1) \cap \Gamma(node_2)} \frac{1}{log|\Gamma(z)|}$$

In studie XXX YYY found that it out performed Jaccard coeficeint.

Rand index/SMC....

4 The twitter network(Simon)

In the following sections we will first present the full twitter network of the House of Representatives. Here after we proceed to the central task of identifying clusters within the sub-networks of the two separate parties. Exploring the full network first serves as a proof-of-concept or a sanity-check of the methods applied. More concretely, the full network is easily labeled according to party lines and the respective leaders are easily identified[6]. Thus whether our techniques captures political salient structures is easily evaluated on the full network. This contrasts the central problem of identifying political sub components in the subnetworks, as we here only have tentative indicators such as the caucuses and shared voting percentage. Furthermore, the split between to two parties are bound to be greater then any intra party fractionazation, point being; if we are not able to cluster the House Members according to party lines, there is little hope of identifying more subtle cluster in the subnetworks. As such the first section below is acts as both as a presentation of the full network and as benchmarking, anticipating the central analyses following it.

4.1 The Full network(Simon)

As descried in section 2.2 the foundation of our twitter graph is the collocations of edges between House-members indicating that both politicians follow each others. of course the nodes can be inferred from the edges and additional meta-data - such as "party" and "voting with party percent" - are append from the Congress API. It should be noted - once more - that all politicians with no edges - $degress = 0$ - have been effectively sorted out.

A Raw and relatively un-analyzed the network is presented in figure 1. The network is color-coded along party-lines. The structure of the nodes are given by the Fruchterman-Reingold force-directed algorithm **ref to paper and elaboration**.

[6] using the Congress API

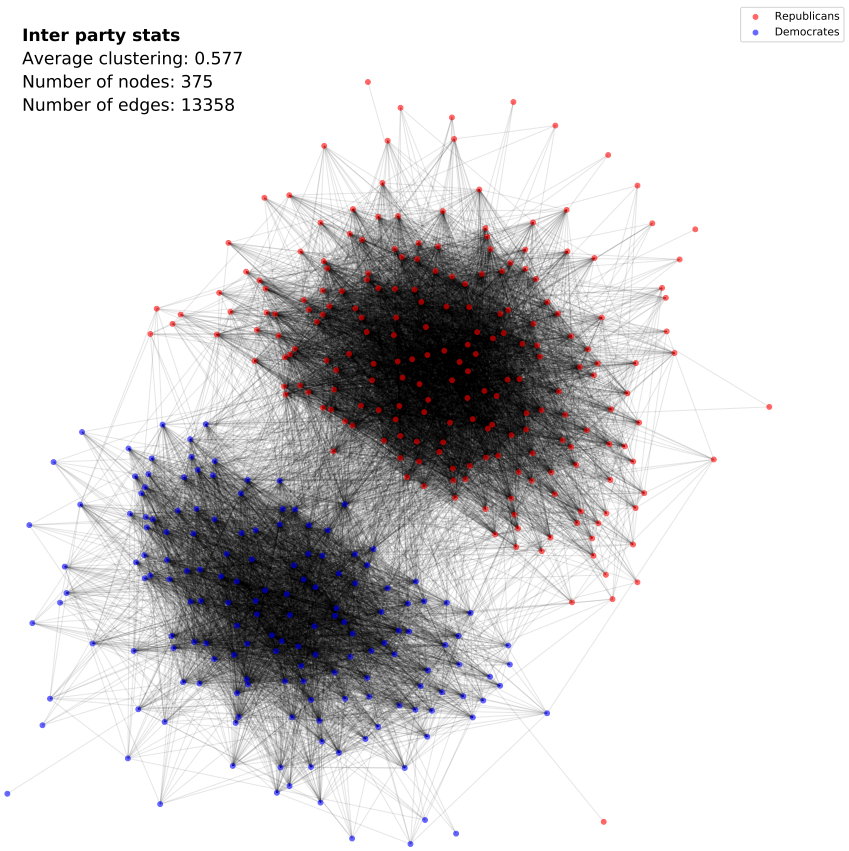

Figure 1: Something cleaver....

As seen, the graph is split almost perfectly according to party-lines - with one republican exception to whom we shall return in later. Of cause the fit of the party clustering can be examined more systematically with Louvian Modularity, which we shall elaborate to in section 4.1.2. In table 2 we have summaries central statistics regarding the network and the two parties.

As noted we do not know why some politicians do not have a twitter; have few fellow politicians following them or do not themselves follow any fellow politicians. Thus It is also hard to say why more Democrats then Republicans have been sorted out. But looking at the statistics it does seem that the Republicans are general more connected. This is a pattern that will continue to appear

	Democrats	Republicans	All
House-members(Congress API)	200	249	449*
pct left out	19%	15%	17%
Nodes	162	213	375
Degrees (sum)	11037	15679	26716
Degrees (mean)	68.13	73.61	71.24
Degree Centrality (mean)	0.18	0.20	0.19
Betweenness Centrality (mean)	0.0027	0.0024	0.0025

Table 2: *The keen reader will notice that the US House of Representatives only have 435 seats at any one time. The reason we have a 14 more observations is as laid forth in section 2.1

throughout the project and will be discussed latter in section. For now though, the difference we see are minor and might just be a artifact of the republican majority and noise. Interpretation with due prudence is thus advised.

None-the-less it does carry a warning for the analyses to come; some intra-party subgroups might be more virtually connected then others, without this translating to a more tight-knit factions in real life.

As noted in section 3.1 the two last measures displayed in table 2 can be seen as denoting the importance of a given node in a network. As mentioned, one step in identifying subgroups can be to identify potential political or idealogical "centroids" around which political factions can form. Thus we now test to what extent it is possible to identify the official leaders of the two parties using the centrality measures.

4.1.1 Centrality

Given that we know who the official leaders of the two parties are, the following section severs to illuminate to what extent it is possible to identify central political figures using attributes from the full network; namely Betweenness Centrality and Degree Centrality.

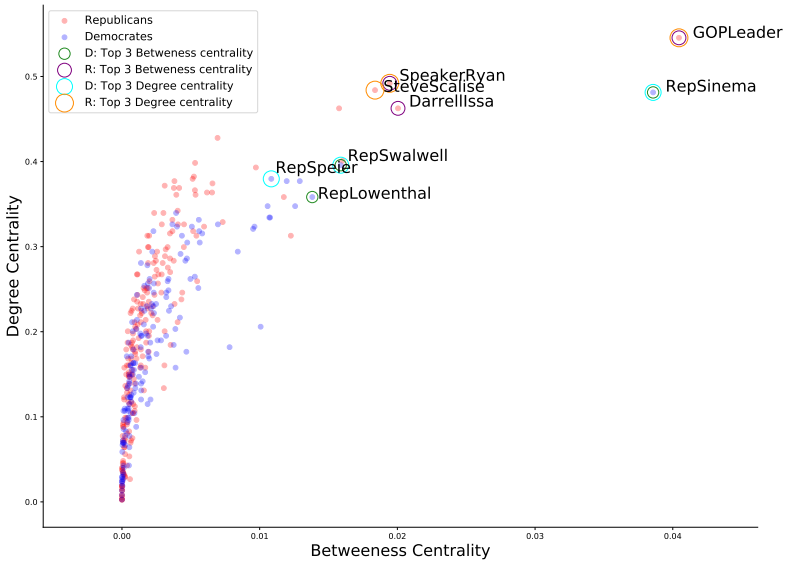

Figure 2: Something cleaver....

Inspecting 2 it is clear that there are some relationship between the two measures[7].

Looking at the annotated Republican we recognize three of the leaders: Kevin McCarthy (GOPLeader), Paul D. Ryan (SpeakerRyan) and Steve Scalise (SteveScalise). Contrasting, no official leader of the democrats appears to be high-scores of the two measurements. in table 3 we present all official party leaders from both party along there ranking in respectively Betweenness Centrality and Degree Centrality.

Here we thus find the fourth republican leader Luke Messer way down in the top 200's. Looking at the democrats, it is interesting that the leaders do not op obtain a better score - especial considering who well the measurements are able to capturer the republican leadership.

Thoughts? Some kind of conclusion.

[7] Clearly not a linear relationship, but something akin to a very heteroskedastic logarithmic relationship

	Name	Rank(BC)	Rank(DC)	Role
R	Kevin McCarthy	1	1	Majority Leader
R	Luke Messer	137	124	Policy Committee Chair
R	Pual D. Ryan	4	2	Speaker of the House
R	Steve Scalise	5	3	Majority Whip
D	James E. Cluburn	120	53	Assistant Leader
D	Steny H. Hoyer	10	17	Minority Whip
D	Nancy Pelosi	34	50	Minority Leader

Table 3: Only official leaders included. Rank(BC) denotes Betweenness Centrality ranking out of 375. Rank(DC) denotes the Degree Centrality ranking out of 375

4.1.2 Modularity

We shall now turn to more direct clustering. For this we utilize the the Louvian Modularity measure on both the party split and a "Best Partion" split. In appendix **XXX** we have color-coded the network according to "best partition", but the curious reader will note that it is almost identical to the party-color-coded network presented in figure 1. With three exceptions; we now have three clusters since two republican gets there own group; and the republican who only had one edge - one to a democrats - is of course grouped with the rest of the democrats. Thus the clustering of the republicans is sumbject to a minimal amout of noise. All democrates, on the other hand, are grouped neatly together. To further illustrate the similarity table 4 presents the Louvian Modularity of both the party split and the "best partition" split. These a exactly th same.

	Louvian Modularity
"Best Partition"	0.415
Parties	0.415

Table 4:

Given the the Louvian Modularity score goes from -1 to 1 a score of 0.415 is not bad for our party split. Further more it is practically indistinguishable from the most optimal spilt possible given the network at hand. The similarity between the party split and the "Best Partition" indicates that at least in this

round the network attributes succeed in "classifying" the nodes according to salient political clusters. Thus we proceed to the two sub networks.

4.2 The Republican subnetwork

In two following sections we will turn the approach somewhat around; first identifying clusters using "best partition" and then compare this to the a Caucus partition using Louvian Modularity. Afterwards the "Best Partitions" will be compared to the Caucuses partition using Jaccard similarity and the Rand Index. Lastly we utilizing the two centrality measures to find central figures within each of these clusters, and comment on examine weather we these central persons can be said to constitute different political factions.

4.2.1 Modularity

Figure 3 shows the republican sub-network color-coded according to the best partition.

It also displays the modularity score of "Best Partition" (4 clusters; 0.137) and the Caucus partition (6 cluster; 0.008). Two things are clear from this. The network seems a lot harder to split into partitions, and the Caucus partition is performing virtually no better then a random partition. The question however whether this two partitions still share some common ground.

4.2.2 Similarity

As mentioned we utilize the Jaccard score in two somewhat different settings; one of predicting edges, and one of assessing two sets of partitions. table 5 presents the similarity between the two partitions using respectively the Jaccard score and the Rand index both raw and adjusted.

The result in table 5 is not inspiring. As noted the Jaccard can take values

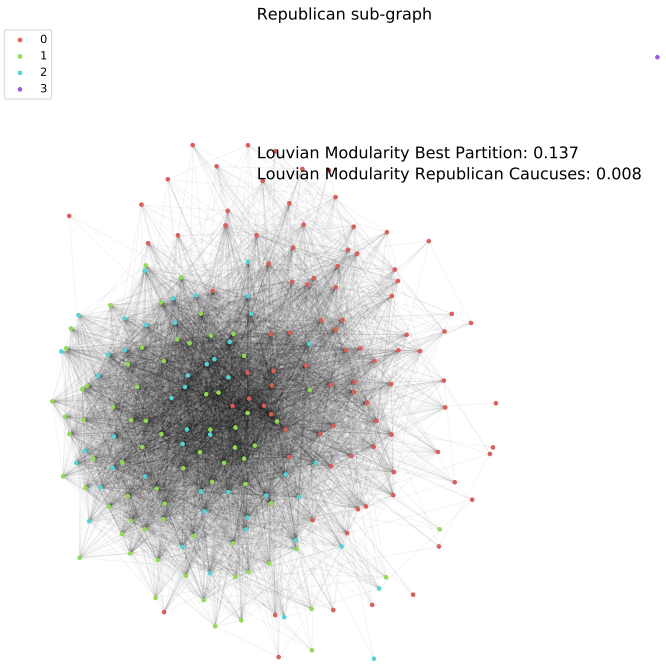

Figure 3: Color-coded according to "Best Partition". 5 major Caucuses are included plus 1 for None

between 0 and 1, thus a score of 0.21 must be regarded as a poor fit. Looking at the two implementations of the Rand index it is clear that the fit between the two partitions does roughly as well as random.

As mentioned, however, there are we can not discard the communities identified by "Best Partition" on account of not matching up to caucuses. Given the promissing results regarding th full network we might indeed have identified some more latent political boundaries - or we might just have to fitted noise and randomness. This is where we turn to centrality.

4.2.3 Centrality

Instead of comparing our "Best Partition" with the caucuses, we here identify the most central[8] nodes from each community in the "Best Partition". Turning

[8] given our the measurements and hand

Jaccard	0.21
Rand Index (SMC)	0.56
Adj. Rand index	0.013

Table 5:

to more qualitative domain knowledge we then asses whether these central nodes in any way correspond to different political or ideological division within the given party.

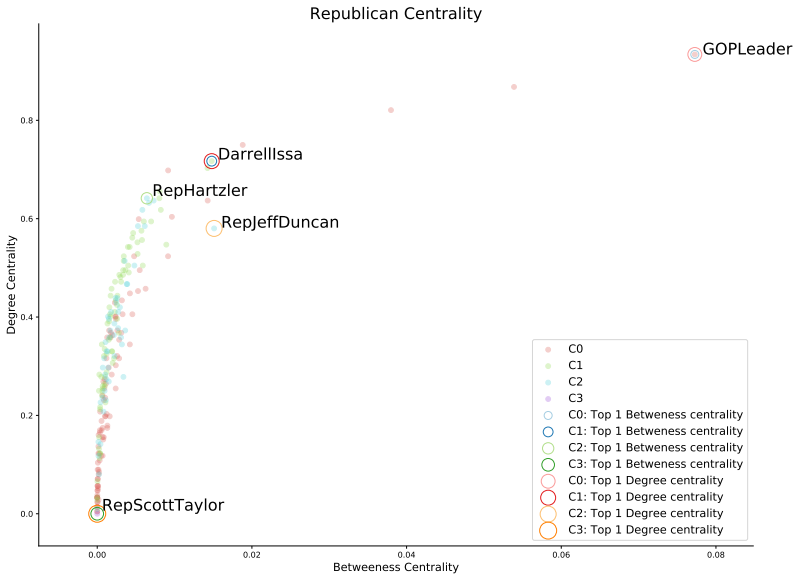

Figure 4: something....

Lets first address Representative Scott Taylor. The Keen reader will have guessed that this is of course the republican with no edges to the republican part, and thus he have been appointed his own community where he is the top score everything.

Turning to the other - more interesting - central nodes, we find McCarhty (GOPLeader), Darrell Issa (DarrellIssa), Vicky Hartzler (RepHartzler) and Jeff Duncan (RepJeffDuncan). Three of these republicans are part of the caucus "Republican Study Committee" while McCarhty is member of the "Freedom

Caucus". Both these caucuses are highly conservative, though the Fredom Caucus might be the most conservative. Two of the politicians are from California (East coast); McCarhty and Issa, while the Hartzler represents Missouri's 4th district (the Midwest) and Duncan represents South Carolina's 3rd district (East Coast). Thus we might very well have caught geographical splits more then we did political - a proposition we shall touch upon again in section **ML thing**.

Another possibility is that we caught some party hierarchy. Looking at figure 4 it is clear that one community (C0) exhibits some notable properties. The four absolute top scores - McCarthy, Ryan, Scalise and Black are all from this community. Though this proposition is intriguing, it is outside the scope and subject of the project at hand.

For the sake of brevity, the corresponding section regarding the Democrats can be found in **appendix XX** in section 11.1. But the Essen of it is largely the same as for the republicans - if the communities we found have any substantial meaning, it does not correspond to anything captured be the caucuses.

4.3 some bridge thingy leading to the next section

Regarding the Caucuses, it should be noted that there is no certainty that these are the most politically salient subcomponents in the House of representatives. The Caucuses are just one scale of evaluation. That is to say, just be course the Best Partition and the Caucuses does not match up does not mean that the best partition does not capture something.

Thats is why we also utilize shared voting percentage ect.....

5 Analysis

START: Bare så jeg ikke hele tiden skal op og kigge: To further validate the usefulness of our network statistics, such as Jaccard similarity, Adamic/Adar

similarity and Louvian modularity we construct a prediction scenario from given data from the Congress API and our two subgraphs (R & D).

We utilize the various network statistics to predict the shared voting behavior - within each party - of US House-members. That is; we predict to what extent two house-members (a dyad), within the same party, tend to vote the in accordance with each other.

The network constructed features are then evaluate in comparison with other features thought to influence intra-party voting behavior such as age, gender and geography.

SLUT:Bare så jeg ikke hele tiden skal op og kigge.

To construct a prediction scenario from the Congress API and the two subgraphs we started by merging the network statistics and the datasets decried in table X we end up having two inter-party datasets. The dataset contains 12,720 and 13,777 observations for the democrats and republicans respectably. However for the democrats we observed that 4,855 connections had a jaccard on zero. For the republicans this was 6,778 pairs. Even though we initially wanted to include these we decided not to as these might blur a possible connection between the jaccard and voting behavior. We conducted a robustness analyses where we included pairs with a zero jaccard which will be decried below. For an Hence that final data that will be used investigate whether the link between the network statistics and voting behavior contains 7,865 and 13,777 for the democrats and republicans respectably.

Some initial investigation of the two dataset was done by looking at the correlation between the variables..

To investigate whether the network statistics can say anything about the political landscape we tried out different ways to measure the importance of the variables we included in our data. Table X shows the coefficients of the variables found using a OLS estimation and the feature importance found using a random forest algorithm. These two measures both say something about which features

Figure 5: Democrats

Figure 6: Republicans

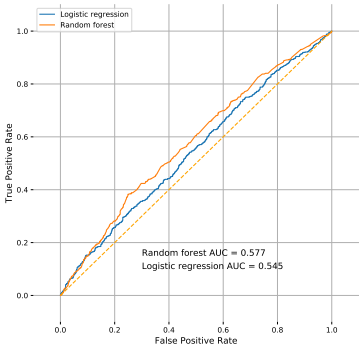

Figure 7: Democrats

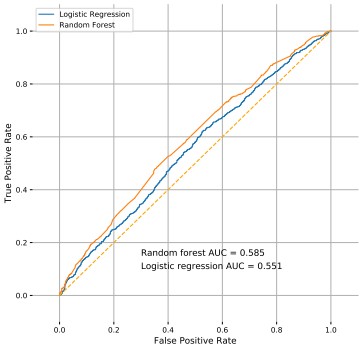

Figure 8: Republicans

are informative but from two different perspective.

The random forest is basically just multiple decision trees. A decision trees is build by a number of optimal splits on the features. To identify the optimal split a measure call information entropy is used. For each tree the contribution from each feature can be computed and by avering over all trees in the forest the features can be ranked in terms of informational gain. (taget fra: http://blog.datadive.net/selecting-good-features-part-iii-random-forests/, Skal uddybes noget mere)

As expected, the plot suggests that 3 features are informative, while the remaining are not.

(Opdateres når philpp har distancerne)

	OLS	Random Forrest		
	Democrats	Republicans	Democrats	Republicans
Agree percentage	92.6	92.7	92.7	92.7
Same sex	0.565	0.829	92.7	92.7
Same age	0.437	0.497	92.7	92.7
Same state	0.066	0.036	92.7	92.7
From neighbor state	0.089	0.105	92.7	92.7
Number data points	18,528	26,627		
House-members	197	231		

Table 6: *Some text

6 Results

7 Discussion

8 Perspectives

9 Conclusion

10 Bibliography

11 Appendices

11.1 The Democrat subnetwork

Here we briefly handle the democratic sub-network in similar fashion to the Republican in section 4.2. We proceed in same order; identifying and analyzing clusters, then central nodes.

11.1.1 Modularity

Just as the case with the republican sub-network we see a relatively low score for the "Best Partition" and a score pretty close to zero for the caucus partition.

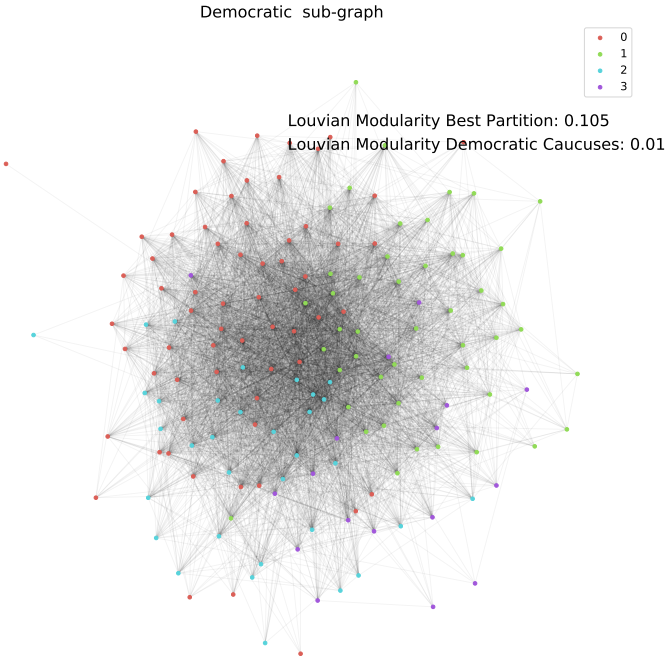

Figure 9: Color-coded according to "Best Partition", 3 major Caucuses are included plus 1 for None

The conclusion are thus the same; the network is harder to partition and the network structure does not correspondent well to the caucus structure.

11.1.2 Similarity

The similarity measures between the "Best Partition" and the caucus partition too, are very comparable to the results from the republican sub-network.

Jaccard	0.19
Rand Index (SMC)	0.60
Adj. Rand index	0.037

Table 7:

The results here closely mirrors those we obtained analyzing the republican sub-network. It is a poor fit between the two partitions, not mush better then random.

11.1.3 Centrality

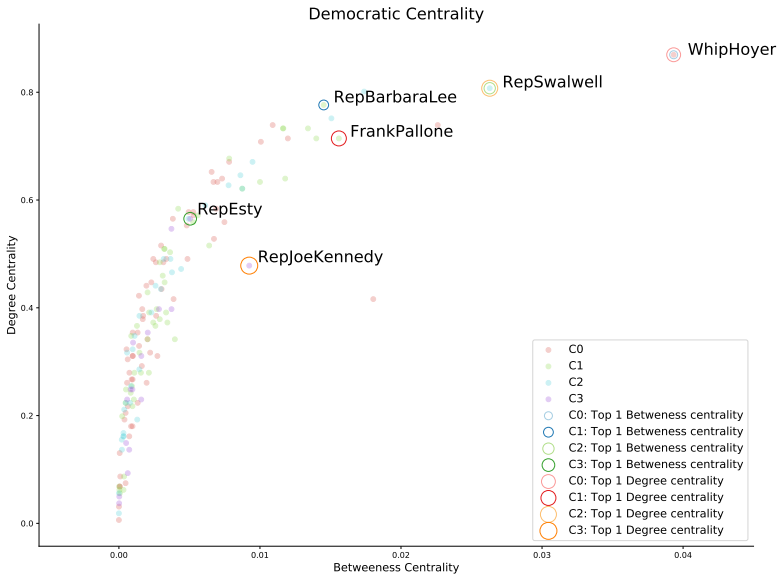

Figure 10: something...

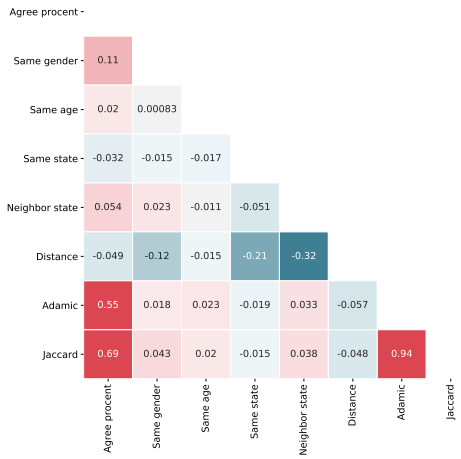

Figure 11: something...

11.2 Whole network

Se the pages below.