Spanish Members of Parliament Show Homophily in Twitter

Candidate Number: 1033256

April 29, 2019

1 Introduction

In 1954, Lazarsfeld and Merton coined the term homophily to capture a phenomenon that may seem intuitive for all of us: people tend to be attracted and develop connections with similar others. In fact, as McPherson, Smith-Lovin and Cook (2001) detail, this same idea can be found on Aristotle's Nichomachean Ethics ("love those who are like themselves") and Plato's Phaedrus ("similarity begets friendship"), but also on multiple fiction works. This phenomenon, often referred with the popular expression "birds of a feather fock together", has become a central research topic for various social sciences.

It is no surprise, however, that homophily is especially important in politics. Politicians tend to relate with others of similar ideologies. Even if this is an obvious tendency, extreme homophily could also indicative of a highly polarized political arena. This is, probably, the case of Spain. A series of economic and social problems, coupled with the arrival of new political parties, has resulted in a highly fragmented parliament. Last general elections, held in 2015, had to be repeated because all negotiations to produce a stable government coalition failed. The current government had to call elections after failing to obtain enough support to pass the General State Budget.

Motivated by this case, we will study the homophilic tendency that different parliamentary groups in the Spanish Crongress are showing on Twitter. Our hypothesis are:

1. That there is an homophilic effect, meaning that MPs from a particular parlamentary group are more likely to follow other MPs of the same group than MPs of other groups on Twitter.
2. That this effect is strong enough that there are clearly identifiable clusters for the major parliamentary groups.

1.1 Data Collection

The process starts with a simple algorithm that allows to map a Twitter social network. This method makes use of the Twitter API through the smappR (Barbera, 2013) package, which offers the critical advantage of allowing the use of different API credentials for Twitter data extraction. This characteristic, not available in other popular packages like twitteR or SocialMediaLab, makes feasible the mapping of big networks in relatively low time.

The official account of the Spanish Congress, @congreso_es, has created some lists that include the official parliamentary groups in the current legislature: PP (right), PSOE (centre left), Podemos (left), Ciudadanos (centre right), PNV (Basque nationalists), ERC (Catalan nationalists), and Mixto (various smaller parties). The total number of accounts included in these lists is 315, when the total number of MPs in Spain is 350. Furthermore, it is possible that these lists may not be completely updated, containing accounts of politicians that are no longer MPs and excluding others who are. For instance, the ex-president Mariano Rajoy is included in the list, when in fact he left Congress more than ten months ago. In any case, the coverage of this network appears to be fairly reasonable, and all the major political figures seem to be included.

The first step was collecting these lists and merging them into one that contained all the 315 MPs with a Twitter account. Apart from the Twitter name, handle, ID, description, and location, we got the language of the account and the parliamentary group corresponding to the account. Gender was manually coded for each MP.

The second step was collecting all "friends" or "followees" of each MP (i.e. people that are followed by the MP). This required gathering a total of 641,924 that are followed by MPs, something that took approximately 5 hours. However, only the friendships between the 315 MPs were kept, while all the links with other 51,0921 accounts were removed. Note that all the data processing and analysis was done using the `igraph` R package.

1.2 Data Pre-processing and Descriptive Statistics

This initial network is directed, has no self-loops (i.e. it is impossible to follow oneself n Twitter), and had 315 nodes and 15,792 edges. However, from an initial visualization it is quite clear that there are 10 components. The only big component had 306 MPs, while the smaller ones corresponded to MPs that did not follow and were not followed by the rest of the users in the network. For simplicity, we decided to remove these users from the study, resulting in a network with 306 nodes and 15,792 edges.

This initial network is directed, has no self-loops (i.e. it is impossible to follow oneself n Twitter), and had 315 nodes and 15,792 edges. However, from an initial visualization it is quite clear that there are 10 components. The only big component had 306 MPs, while the smaller ones corresponded to MPs that did not follow and were not followed by the rest of the users in the network. For simplicity, we decided

to remove these users from the study, resulting in a network with 306 nodes and 15,792 edges.

We plotted it using a Fruchterman-Reingold algorithm (Fruchterman, and Reingold, 1991), which considers each edge as an attractor between two nodes also applying a repulsive force between all nodes. Therefore, connected nodes tend to appear together while disconnected nodes are separated. The result is quite intuitive: it appears that the MPs of the four big parties (PP, in blue; PSOE, in red, Podemos, in purple; and Ciudadanos, in orange) tend to follow each other more than the other's parties. Similarly, ERC (in yellow) appears close to Podemos, something not entirely surprising considering that both are close in the political spectrum.

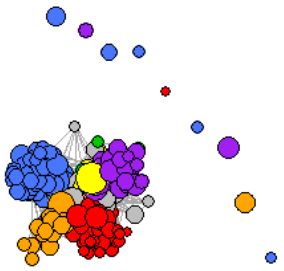

Figure 1:

A first look into de network allows us to clearly see important differences between MPs in the vertex properties that have been included: parliamentary group and number of Twitter followers. First, it is clear that not all parliamentary groups are of the same size. There are 100 PP MPs, 81 from PSOE, 62 from Podemos, 32 from Ciudadanos, 8 from ERC, 4 from PNV, and 14 from the Mixed group. Second, we see that some accounts have much more followers than others. The number of followers ranges from 395 to 2,272,550. However, note that the number of Twitter followers is a very long-tailed distribution, and only 18 MPs have more than 100,000 followers.

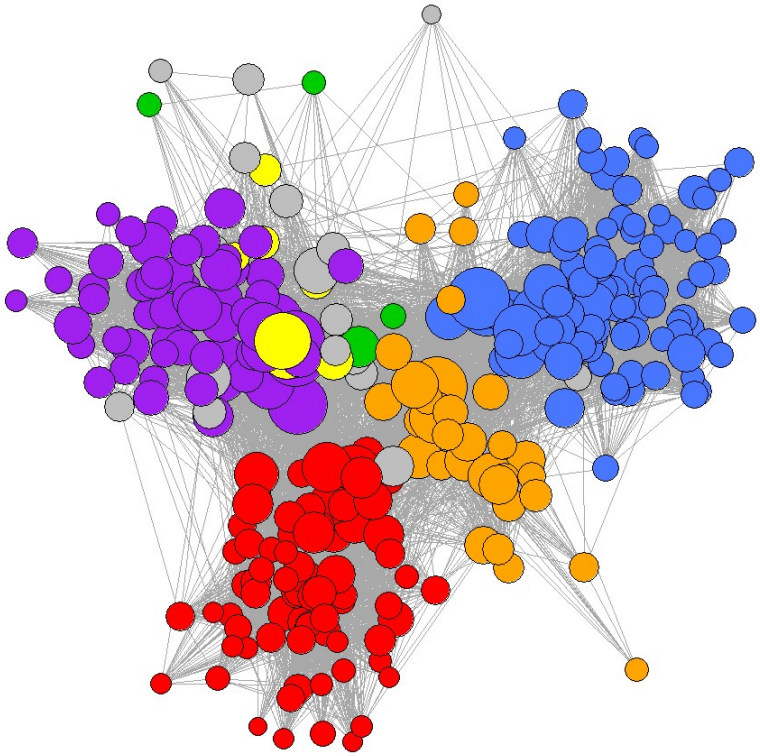

Figure 2:

Before starting any analysis, we will take a closer look at the properties of the network. We have already seen that the network has 306 vertices and 15,792 edges. It presents a density of 0.17, meaning there is a 17% probability of an edge (i.e. as we are talking about a directed graph, an edge means the directed tie between i and j, or vice versa) between two randomly sampled edges. The transitivity, also called global clustering coefficient, is 0.54, meaning that if two vertices are connected to a third, the probability of an edge between the two is 54%. The reciprocity is of 0.64, meaning that the probability of i being connected to j given that j is connected to i is 64%. The graph is, as seen in Figure 2, relatively well connected, with an average shortest path of 1.91.

Additionally, we will study the centrality of the vertices using four different centrality measures:

- In-degree centrality (larger is better), simply reporting which vertices have more connections directed towards them.
- Betweenness centrality (larger is better), which refers to the number of times each vertex is part of the shortest paths between the other nodes, measuring what we may call gatekeeping potential.
- Closeness centrality (smaller is better), referring to the average length of the shortest past between each vertex and all the other vertices.
- Page rank (larger is better), famously developed by Larry Page serving as a basis for Google Search, which measures how likely a random surfer is to reach a certain vertex.

In-degree centrality, betweenness, and PageRank reported relatively similar measures, with the major political figures showing the largest scores in all of them. Closeness measures were also related to the other three, but to a lesser degree, something that can intuitively be seen in the table below. A potential reason for this is that closeness centrality is, to some degree, penalizing for homophily: the vertices that are very connected to their own parliamentary group but lack connections with the other groups will see their average shortest path increase.

In Degree	**Betweeness**	**PageRank**	**Closeness**
sanchezcastejon	marianorajoy	sanchezcastejon	antonipostius
marianorajoy	pablocasado_	sorayasds	aliciapiquer
sorayasds	AAlvarezAlvarez	marianorajoy	marianbeitia
Pablo_Iglesias_	sanchezcastejon	Pablo_Iglesias_	jagirretxea
pablocasado_	Pablo_Iglesias_	pablocasado_	Arevalo80MT

Table 1:

2 Inferential statistics: ERGM

To study the network characteristics, including its homophily, we will use a family of statistical models called Exponential Random Graph Models, fantastically im-

plemented in the R package `ergm`. ERGMs offer the possibility to select complex models with different statistics that allow us to understand relevant properties of the network. In this case, we will start including in our model two basic statistics: edges, which captures the density of the network, and then mutual, which estimates the tendency of ties to be reciprocated in a directed network. Note, however, that our research questions does not directly concern the density or the reciprocity. To study the homophily in this network, we will later add the nodematch parameter, which in this case will explore if an MP from a certain parlamentary group is more likely to follow another MP given that he or she belongs to the same parlamentary group. In this first approximation we will focus in what is called uniform homophily, referring to the fact that we are assuming each political group has the same propensity for homophily. Additionally, we will also calculate the nodemarch parameter relative to the gender of the MP.

Although obtaining the maximum-likehood explicitly is theoretically possible, in almost all cases of certain complexity we use numerical methods that obtain an MLE approximation based on Markov Chain Monte Carlo. Except for the first case, the models will be estimated using MCMLE.

Term	**Estimate**	**Std Err**	**MC Err**	**p**	**Lo (95%)**	**Hi (95%)**
edges	-1.591	0.009	0	<0.001	-1.608	-1.574

The first model only considers the edges statistic, implying that porbability of a particular tie existing is exp(-1.59)/(1+exp(-1.59)) = 17%, which logically corresponds with the density of the network that was calculated in the previous section.

Term	**Estimate**	**Std Err**	**MC Err**	**p**	**Lo (95%)**	**Hi (95%)**
edges	-2.530	0.0147	1	<0.001	2.558	-2.500
mutual	3.102	0.0318	1	<0.001	3.0396	3.164

In a second model, we see that the reciprocity statistic is also significant. Given that one account follows another, the probability of the latter following the former is incremented exp(3.08)=21.76 times, from exp(-2.53)/(1+exp(-2.53)) = 7% to exp(-2.53+ 3.08)/(1+exp(-2.53+ 3.08)) = 63%.

Term	**Estimate**	**Std Err**	**MC Err**	**p**	**Lo (95%)**	**Hi (95%)**
edges	-3.218	0.0220	1	<0.001	-3.261	-3.175
mutual	1.866	0.0346	1	<0.001	1.798	1.934
nodematch.party	2.452	0.022	1	<0.001	2.408	2.496
nodematch.gender	0.081	0.020	1	<0.001	0.042	0.121

The third model considered the effects of homophily. First, we see that the edges and mutual statistics are still significant. The probability of one MP following another if the latter does not follow the former, and if they are not from the same party and gender is only exp(-3.22)/(1+exp(-3.22)) = 4%. This probability increases to exp(-3.22+1.85)/(1+exp(-3.22+1.85)) = 20% if the latter is already following the

former. However, the biggest effect corresponds to the parliamentary group homophily. The probability of one MP being connected through Twitter to another MP is incremented exp(2.45) = 11.59 times if they are both of the same parliamentary group. That is, the probability of one vertex to have an edge with another vertex being both part of the same parliamentary group ranges between exp(-3.19 + 1.86 + 2.45)/(1+exp(-3.19 + 1.86 + 2.45)) = 75% if the latter is following the former and exp(-3.19 + 2.45)/(1+exp(-3.19 + 2.45)) = 32% otherwise, given that both MPs are of different gender. If they are of the same gender, these probabilities increase to exp(-3.19 + 1.86 + 2.45+0.085)/(1+exp(-3.19 + 1.86 + 2.45+0.085)) = 77% and exp(-3.19 + 2.45+0.085)/(1+exp(-3.19 + 2.45+0.085)) = 34%, respectively. From this, we can conclude that gender homophily is quite small, although the effects are still significant. In particular, a MP is exp(0.85) = 1.09 times more likely to follow another MP if they are of the same gender.

This last model was checked for degeneracy, and the MCMC diagnostics looked satisfactory. Its goodness-of-fit was also assessed, with good results. The corresponding plots can be consulted in the appendix.

These results indicate that there is a strong homophilic effect. However, it could be the case that not all parliamentary groups showed the same level homophily. To test this, we fitted the same model, but modifying the nodemarch statistic to include differential homophily.

Term	**Estimate**	**Std Err**	**MC Err**	**p**	**Lo (95%)**	**Hi (95%)**
edges	-3.206	0.030	1	<0.001	-3.265	-3.147
mutual	1.843	0.0520	1	<0.001	1.741	1.945
nodematch.party .Ciudadanos	2.600	0.133	2	<0.001	2.339	2.860
nodematch.party .ERC	4.215	2.063	2	0.041	0.171	8.259
nodematch.party .Mixto	1.7459	0.146	3	<0.001	1.459	2.033
nodematch.party .PNV	2.386	0.766	2	0.002	0.884	3.887
nodematch.party .Podemos	2.563	0.0716	2	<0.001	2.423	2.703
nodematch.party .PP	2.365	0.0419	2	<0.001	2.283	2.447
nodematch.party .PSOE	2.510	0.0494	3	<0.001	2.413	2.607
nodematch.gender	0.075	0.032	2	0.017	0.0134	0.137

The results appear to indicate that there is no homophily differences between the major parliamentary groups. We see that the effect of homophily is not significant for the ERC party members, but this is probably explained by the fact that the party has little parliamentary representation. The homophily in the Mixto group appears to be lower, which is reasonable considering this group contains MPs from very different parties. In any case, it is important to note that the MCMLE estimation did not converge, meaning that the estimated coefficients may not be accurate. Additionally,

the MCMC diagnostics still look acceptable, but the goodness-of-fit diagnostics look worse. Therefore, this evidence is not conclusive and further research would be needed to really clarify the potential differences in homophily of different parties, probably requiring more data that records past MPs.

3 Community Detection

We've seen that there is a clear homophilic tendency in the Spanish MPs. In this section, we will explore how far that tendency is: even if we lost the political group that corresponds to each MP, we would be able to reestablish this data accurately just based on the structure of the network. This is done with a technique called community detection, a process that allows to group set of nodes in communities, such that each community is densely connected internally.

Kolaczyk and Csárdi (2014) explained that different community detection methods provide different answers, and that it is the responsibility of the user to decide which one represents the reality most accurately. In this case, we will show the results of three community detections algorithms:

- Fast Greedy Modularity Optimization: Modularity is a measure of how well a network is divided into clusters. That is, a network with high modularity will have dense connections between the nodes within the same cluster and little connections to the nodes in the other clusters. Therefore, modularity-based approaches intend to maximize the modularity of the obtained partition. Exact optimization of modularity is possible, but too computationally expensive in large graphs. Therefore, there are different approaches that can approximate the optimum. Clauset, Newman, and Moore (2004) proposed the fast greedy modularity optimization, which starts by every vertex being its own cluster and in concurrent steps these clusters are merged in such a way that the value of modularity undergoes the largest increase. It stops when it is not possible to increase the modularity any more. Note, however, that this approach does not consider the direction of the edges in the network. Other approaches include the so-called leading eigenvector method (Newman, 2006), which starts with every vertex being part of a single giant community that is iteratively divided into smaller fragments, and the spinglass approach (Reichard Bornholdt, 2006), based on models from statistical physics.

- Louvain: Developed by Blondel, Guillaume, Lambiotte, and Lefebvre (2008), it also intends to optimize modularity initially assigning each vertex to a separate community and iteratively moving vertices between communities in such a way that modularity increases. This method runs extremely fast, which makes it especially popular with large networks, and tends to produce accurate results. Yang, Algesheimer, and Tessone (2016) performed an extensive comparison of the accuracy and computing time of different community detection algorithms, concluding that the Louvain approach outperformed the others. Additionally, in some cases this method is especially interesting because it is able to find a list of clusterings at different resolutions scales. Note,

however, that it only considers undirected graphs, and therefore we had to also transform our network to undirected.

- Walktrap: This relatively simple approach by Pons and Latapy (2005) performs short random walks under the assumption that these walks are more likely to remain within the same cluster given that there are many edges between the vertices that are members of the cluster and little that go outside the cluster.

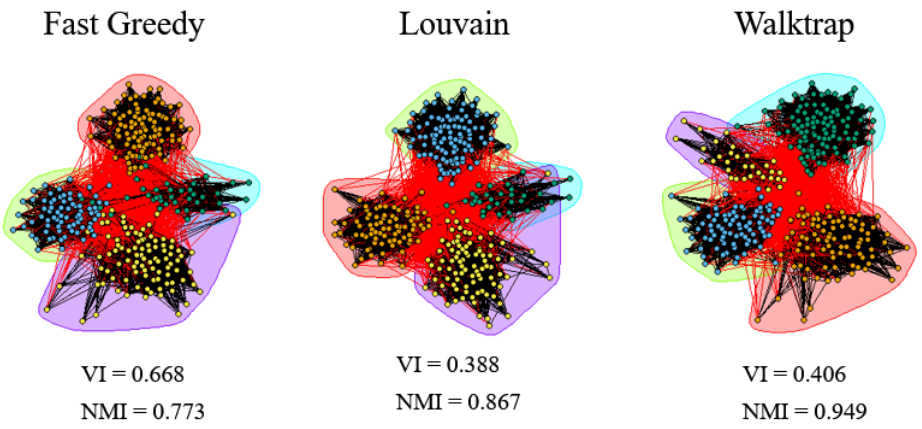

Figure 3:

In general, we see that the three algorithms identified 4 clusters: one largely corresponding to PP, one for PSOE, one for Ciudadanos, and one for Podemos, ERC, PNV, and Mixto. These results are not surprising considering that Podemos and the other MPs corresponding to smaller parliamentary groups share many political views.

Additionally, we reported two external quality measures: the Variation of Information (VI) measure and the Normalized Mutual Information (NMI). The two measures compare how well the identified clusters correspond to the ground truth – that is, the official parliamentary groups in the Spanish congress. The first one ranges form 0 to log n (i.e. $\log(305) = 2.48$), and measures the distance of two clusterings from one another, meaning that lower is better. The second one ranges from 0 to 1, a higher score meaning that the identified clusters correspond to the ground truth. Unsurprisingly, VI are relatively low and NMI is more or less close to one in all the cases, meaning that the clusters identified by the three community detection algorithms largely resemble the political groups in the Spanish Congress. However, note that Louvain and Walktrap performed better than Fast Greedy.

It is important to note that this is not an extensive exploration of community detection algorithms. There are many which have not been included in the analysis, as the classical Girvan-Newman algorithm, the Label propagation algorithms or the InfoMap algorithm.

4 Conclusion

We have demonstrated that there is a clear homophilic tendency between the Twitter accounts of MPs in the Spanish congress. Furthermore, we have shown that this effect is so pervasive that popular community detection algorithms are able to identify the parliamentary groups with a high accuracy just based on the network structure.

It is hypothesized that this homophilic effect is, in part, a product of a polarized political arena. However, it would be necessary to study the Twitter structure of past Congress MPs to understand if homophily has increased. Another future direction would be to study compare with the corresponding networks in other countries instead of across time.

5 References

Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, P10008.

Clauset, A., Newman, M. E., & Moore, C. (2004). Finding community structure in very large networks. Physical review E, 70(6), 066111.

Fruchterman, T. M., & Reingold, E. M. (1991). Graph drawing by force-directed placement. Software: Practice and experience, 21(11), 1129-1164.

Kolaczyk, E. D., & Csárdi, G. (2014). Statistical analysis of network data with R. New York: Springer.

Newman, M. E. (2006). Finding community structure in networks using the eigenvectors of matrices. Physical review E, 74(3), 036104.

Pons, P., & Latapy, M. (2005). Computing communities in large networks using random walks. In International symposium on computer and information sciences (pp. 284-293). Springer, Berlin, Heidelberg.

Reichardt, J., & Bornholdt, S. (2006). Statistical mechanics of community detection. Physical Review E, 74(1), 016110.

Yang, Z., Algesheimer, R., & Tessone, C. J. (2016). A comparative analysis of community detection algorithms on artificial networks. Scientific reports, 6, 30750.