# Social-Democracy

## Homophily and polarisation in politics, the Italian Twitter network

MARIO LUCA

Mémoire présenté pour le Master en

Discipline: Economics and Public Policy, PhD Track

Directeur du mémoire: Ruben Durante

Année académique

2014/2015

## TABLE OF CONTENTS

## INTRODUCTION AND MOTIVATION

My interest for social networks started very far from Twitter and politics, which are only one of the many fields where these theories can be applied. It started while reading a book written in 1924 about medieval rural history: Marc Bloch's *Les Rois Thaumaturges* (1924). This great classic of the historiography of the *école des annals* describes the well-spread belief that Kings could heal with their royal touch some specific disease. Although a Bayesian mind would think that such belief is due to evaporate as the number of successes should very well be minimal, this strong credence rose and remained untouched for centuries. In particular, it seems that some substrata of the population in terms of social and geographical position were more tenacious in preserving this belief, even when other people were starting to realise it was nothing but a hoax.

Surely, the way information could spread in the XIV century has some peculiar pattern that cannot be extended to nowadays' communication, however, the bizarre co-presence of mutually exclusive beliefs as well as ungrounded ideas is a persistent phenomenon. The debate about vaccination, for instance, is a clear example of how polarised modern society can be. The credence that vaccines could be dangerous is spreading widely nowadays, albeit being incompatible with a wide scientific literature. Hoaxes about 9/11, false information about GMOs, bizarre theories on Obama's birthplace are just few examples of how lies spread in networks and become well-established truths for some niche.

The intuition that small interconnected groups are those who can sustain a false belief for a long time, being impermeable to information acquisition, is well-established in network theory. The phenomenon of *echo chambers* is, thus, the most interesting candidate to explain inconsistent beliefs within subgroups of a population. Similarly, these theories can be exploited to analyse a wider range of facts. When some communication failure occurs, instead of converging to a central pole, ideologies can have a centrifugal force towards the extremes. This will be analysed in the first section, dedicated to network models and the factors of convergence.

I wanted, then, to apply the theoretical intuitions to an empirical scenario. Therefore, I chose to collect data from Twitter and politics, to analyse the structure of the network and the evolution of political beliefs. However, Twitter does not provide as many data as one would desire. In particular, it is impossible to trace the network structure for past days, but one can only have an instantaneous screenshot of the actual situation. Thus, I have focused on the more modest task to describe the static shape of political debate, while reserving the analysis of the dynamic one for the future, when panel data will be collected. The widest part of this work will, then, be devoted to this task.

I will propose a new method, elaborated from Barberà (2015), to estimate political preferences through network data, so to have a new and reliable measure of political ideology on a continuum, that could be exploited in future to investigate polarisation and extremisms in a way that would not be possible with discrete measures.

Then, I will use the obtained estimations to identify groups, clusters and homophily between elected politicians and their followers and among the users' network. The huge dataset I have built comprehends data both of the type *user follows politician* and *user follows user*, allowing to assess the dynamics of political relationships on these two dimensions. The goal is to find whether some groups are more likely to have political echo chambers, so that one can focus on them for a dynamic analysis of polarisation in the future. I find that this is the case, and this will be the main contribution of this work. In particular, I will find that the most closed and clustered groups are those on the extreme of the political spectrum, confirming the theoretical insight aforementioned.

Then, I will start a semantical analysis of political communication. Again, this will be just a draft of what a complete work could be, as the limitations on the data do not allow a deeper investigation. Nonetheless, I will find some anecdotic hint about polarised and strategic choice of campaign topics. This short section shows how this research could be expanded to find a more rigorous result.

For the whole work, the approach I will follow will be mostly graphical. Firstly, because I believe that this kind of representation can provide more accessible insights to the reader; secondly, because a quantitative data analysis exceeds the hardware constraints due to the huge amount of data and the limits in its gathering.

This whole work has to be interpreted as a first stone for a more insightful study that can be done once some structural barriers on data collection will be overcome. More specifically, this paper is the tip of the iceberg of a wider, hidden, job. Understanding the procedures to retrieve data from Twitter was a very challenging task, as well as setting all the databases to make communicate the four main software I have used (Visual Basic, SQL, R and Matlab). This know-how investment being done, all extensions and generalisations of data can be provided at almost zero costs.

## NETWORK THEORY: CONVERGENCE AND POLARISATION

Models about the diffusion of opinions within a network are widely debated by a growing literature. Jackson (2008) provides a wide discussion on the history and the diffusion of this field as well as its frontier. The most diffused setting is inherited from epidemiology. Bailey (1957) developed the *SIS* model, where individuals can be infected of susceptible to become such. Kretschmar & Morris (1996) brought into the basic design the network structure explicitly, highlighting the dynamics of diffusion with the respect to the structure of the encounters. The further development of these class of models is the *SIR* model, where some agents is impermeable to the disease (e.g. Newman, 2002). These models can be applied to social sciences by treating a certain belief as an infectious virus that can spread through a network. In particular, it identifies the conditions in terms of network structure for the illness to spread and become epidemic. Abramson (2001) and Hethcote (2000) provide a wide introduction on the mathematical modelling of the spread of infectious diseases.

These models are often analysed by physicians and can provide some insight on the steady state amount of infected people, as well as the dynamics of the diffusion. However, ideas are not like Ebola. Not always, at least. A simple contagion theory can work for virus but it can fail in a more complex scenario. The metaphor between epidemiology and social sciences breaks quite soon if a more plausible notion of learning is not introduced.

Therefore, Bayesian models of observational learning became an interesting field of research. The first anecdotal evidences were provided by Lazarsfeld et al. (1968) that investigates the role of opinion leaders in people's electoral choices. Their focus, then, extended to the whole range of *opinions*, finding that social influence is crucial for people's ideas.

The pivotal theoretical work in this field is Bala and Goyal (1998). They study an undirected network of agents who can choose within a finite set of actions, with some uncertainty about the outcome. By observing their outcomes and their neighbours' ones, they can update rationally their beliefs on actions' profitability. The study of inferences can be problematic, especially in big networks (e.g. Gale and Kariv 2003 analysed the full structure on a three-node network only). They find that this kind of model can converge to all players adopting the same strategy or, at least, having the same payoffs. This kind of models can be applied to a wide range of phenomena, from the green revolution to malaria pills, but it still does not tell much about opinions on unverifiable events. Another interesting approach is provided by Boyd et al. (2006), who elaborate the concept of *gossip algorithms* for the diffusion of information in networks.

DeGroot model (1974) provides the theoretical skeleton for a very fruitful approach to opinion formation. In this setting, every member of a community has an a priori (e.g. the expected quality of a political candidate, of the probability of success of a treatment) and can communicate with some of his neighbours. However, the value he gives to other people's information is not homogenous but it depends from a weighting matrix. It is, indeed, plausible, to believe that not all other people's thoughts will have the same importance in one's opinion making. The reasons behind this weighting matrix can be various. Trustworthiness, relational aspects, closeness can be just few of the determinants of this matrix. Thus, DeGroot models focus on the determinants of convergence or divergence of a population.

In particular, the concept of aperiodicity becomes the most relevant factor that can allow convergence. Golub and Jackson (2010) provide an algorithm to assess the nonconvergency of a DeGroot matrix. However, the main insight is that cycles that lead to a double-counting of the same position are the best candidate for convergence failures. These echo chambers, then, can make the population diverge from a full Bayesian learning. Acemoglu et al. (2011) also prove that, when no agent is excessively influential, people will converge to the same beliefs, also with a different setting than that Jackson.

Krause (2000) provide an insightful model on *selective exposure* to information. Indeed, the paper argues that one would consider only those actors whose positions are relatively closer to theirs when updating their beliefs. This is plausible, as one can very well imagine that people pay some psychological cost for accepting too different ideologies. By generalising this to a weighting matrix that gives stronger relevance to closer opinions, it can be shown that no consensus is often reached also in well connected groups (Jackson, 2008).

The problem of double-counting of information is well analysed in DeMarzo et al. (2003), where the *persuasion bias* is presented and modelled. In the easiest scenario, one can think of three friends who have to decide what is the best restaurant in town. If both agent A and agent B have talked to agent C before meeting one another and do not take into account the fact of having been previously informed by the latter, C's opinion will drive the group in a stronger way than it should. Adjusting for repetitions of messages can be crucial: some belief might be imposed only because repeated more frequently and not because closer to the truth. This can link very well clusterisation and political extremism in a causal relationship.

Flattening the network dimension, Hegselmann and Krause (2002) provide a model based on computer simulations for polarisation and consensus. They sketch a model with bounded confidence, where opinion adjustment is proportional to one's agreement with the partner's message. They also allow for time varying confidence levels, studying the different thresholds that can allow for convergence within a well-mixed population. The condensation of people's beliefs to some focal points depends, then, on the a priori of the population. In general, they find that the smaller the confidence level, the more likely to have separating equilibria is. Also, if the bias is asymmetric, this will push the equilibria towards the dimension where confidence is lower. Increasing bias makes it even more difficult to reach a consensus.

To re-introduce networks' role, Amblard and Deffuant (2004) propose the notion of *relative agreement* that considers opinion segments, where one is receptive of information only if this falls within this segment. Running this model on small-world networks, varying the level of connectivity and randomness, they find a level of connectivity that allows single extreme convergence. The analysis of the centripetal convergence strength against the extremists' one is extended in Weisbuch et al. (2005). They take into account a multidimensional setting and find the criteria of convergence to be linked to the acceptance of further opinions. Meadows and Cliff (2012) review in a clear way the debate and the findings of the *relative agreement* and the *bounded confidence* models.

Sociophysics tries to extend these findings to a more applied context. The *founder* of this discipline, Serge Galam has elaborated a model (e.g. Galam, 2008) to describe the compresence of mutually excluding *truths* and hoaxes (Galam, 2003). He defines the conditions under which a minority can sustain a contrary-to-evidence belief, in a way that resembles the *roi thaumaturges* problem mentioned before. However, the fragility of these equilibria is shown by, among the others, Wu and Huberman (2004).

Acemoglu et al. (2010) provide the intuitions for the spread of misinformation in networks by focusing on *forceful* agents, i.e. people who influence other members of the community but who do not update their beliefs. This idea can easily be applied to concrete situations, for instance we can think of it in a Twitter network as guys who are followed by many people but do not follow any of them. The role of these asymmetries in information reception can shape and bias the convergence process.

However, this wide theoretical literature often seems too detached from concrete possible scenarios of information spreading. Nonetheless, an emerging literature is trying to apply these to large dataset that can be reached with online social networks. Facebook publishes frequently reports on information diffusion (e.g. Bakshy et al. 2012) in which they evaluate of the exposure to content influences propagation. They find a positive effect of the number of sharing friends in the probability of sharing, focusing on the impact of tie strength between users. Researchers focused on marketing also try to establish the determinants of content virality. Huberman et al. (2013) study the word-of-mouth propagation of advertisement on Twitter, finding that key players can play a substantial role in promoting a brand.

The interaction between internet and politics has been studied under many perspectives. Campante et al. (2013) focus on the impact of voter turnout, exploiting the differences in ADSL technology in Italy, highlighting the multifaceted effect of online activism on the political sphere. Falck et al. (2014) confirm instead a positive role of internet on turnout, finding also some effect on the share of votes, especially for right wing parties. The increasing role of Twitter in political campaigns (Smith, 2013) can classify the latter as a source of news that has the potential to influence the democratic process (Kohut et al. 2012). Therefore, it seems promising to use the political blogosphere as a battlefield for competing theories about networks and polarisation. Indeed, the data that can be extracted by Twitter are so detailed that one can try to map the theory to empirics in a fruitful and concrete-problems-oriented way. Especially, it seems right to treat Twitter as a *news media* (Kwak et al., 2010), applying to it the theories of selective exposure used for media (e.g. Bryant & Miron, 2004).

Studies about homophily on Twitter are gaining momentum in the research. Conover et al. (2011) distinguishes the segregation between information creation and its diffusion, finding a stronger effect on the latter. Himelboim et al. (2013) finds that polarisation in Twitter increases with the portion of partisan users. De Choudhury (2011) spots out that homophily in interests is stronger than in any other attribute (e.g. race, gender, localisation). Halbestrom and Knight (2013) find that social media are highly segregated along the ideological lines, finding a stronger probability for users who follow candidates from a party to follow other users who lean towards that political force. Stroud (2010) supports the claim that people look for and accept only positions that are not too far from their original one, corroborating the idea of political segregation.

Particularly interesting is the work of Himelboim et al. (2013) on emotional clusterisation that can identify how people associate one another with similar attitudes towards politically charged news. The patterns of communication hereby considered go further than a simple political identification and also can take into account the level of agreement or disagreement to a topic. A similar approach is done by a case study of a politically charged event regarding an abortionist doctor's shooting (Yardi & Boyd, 2010), highlighting how emotional reactions tend to group with like-minded people.

Focusing on echo chambers (Pariser, 2011; Sunstein, 2009), a growing literature is studying the radicalisation effects of homophile networks. Messing and Westwood (2012) show how political endorsement shape the diffusion of news, possibly having macro-level social implications. Indeed, the combination of homophile groups and news double-counting could have a great role on politics. Colleoni et al. (2014) find evidences of echo-chambers in the US political debate, without any clear insight on the polarisation effects. However, their way to identify voters seems quite too simplistic and subject to false positives. Conover et al. (2012) make, to the best of my knowledge, the most extended study of Twitter political clustering, confirming the idea of an exacerbating political debate.

However, a consensus on the polarisation effect of Twitter has not been reached. In a more recent work, Halbestrom and Knight (2014) conclude that social media could expose people to a more heterogeneous source of information, albeit the exposure on like-minded remains stronger and disproportional. Flaxman et al. (2013) point out that the magnitude of segregation is limited and not as relevant as many other authors claim. Brundinge (2010) finds that, through inadvertent exposure, internet increases the heterogeneity of political discussion, enlarging one's political sphere. This is a particularly interesting insight. Indeed, there is no need to think that *online* social networks should be less homophile than *offline* ones. While in normal life one is bound to be surrounded by similar people, internet offers the possibility to be reached by ideas of all kind. Moreover, in a not mediated network as Twitter, the possibility to have some unfiltered *invasion* of one's usual part of the ideological spectrum becomes likely.

The role of *passive* exposure to various news could outbalance the tendency to create homophile networks. The exposure to dissonant information could have a centripetal effect, as Barberà (2014) argues. Using a panel design, this paper tries to overcome the limits of not having an exogenous source of political diversity exposure. The results seem to encourage the idea of a moderator effect of social media.

Therefore, two possible, mutually exclusive, theories try to describe the multifaceted relationship between social media and politics. What one should do in order to find what is the most accurate one, the centripetal or the centrifugal, is still unclear and up to debate. Firstly, one should tackle the issue of the static description of the network. Indeed, if no ideological clustering was to be found, one could confidently advocate against the echo-chamber paradigm. Thus, a first step should be to take a screenshot of the actual situation and check for hints of segregation. This is what I will attempt to do in the successive pages of this work.

Then, one should see the dynamic effect of the existing network. In particular, one should find a source of heterogeneity in like-minded exposition, to estimate its role on polarisation. Given that it is difficult to find an instrument of a shock that could provide sufficient leverage, one should try to exploit panel data so to find an overtime effect of isolated groups. The path to this analysis is full of difficulties. First, Twitter does not encompass the totality of news gathering of agents. Indeed, some unobserved channel of opinion shaping could move users' positions. I think that for this purpose it would be interesting to distinguish online only and real life connections, so to measure exactly all the forces that shape the process. The effects of homophily in politics should be followed also offline.

Hence, an analysis of Twitter has to be interpreted as the tip of the iceberg of a wider, deeper, question: do people only communicate with their alter-egos? If so, how can they get away from their status quo? Online social media are just one of the many form of networks that researchers should focus on. However, given the possibilities of gathering data that they provide, they can be a privileged battlefield where one can test and perfect more general theories.

### THE DATA

I have chosen Italy as the target for this analysis as it presents a very peculiar political scenario. Indeed, as in many other countries in Europe, Italy saw a great uprising of new political actors during the last few years. Among those, the most successful, *Movimento Cinque Stelle*, made a wide use of internet, as its founder, former comedian Beppe Grillo, runs the most followed blog in the country. The great mixture of e-democracy and anti-system rhetoric contributed to its development and attracted to new media a wide sector of the population who did not use to be politically active. Moreover, it obliged other parties to rapidly develop new techniques of communication, making Twitter widely diffused in the political debate. Just to make a simple comparison, while François Hollande does not reach one million followers, David Cameron barely reaches 1,02mln and Mariano Rajoy 784.000, both Matteo Renzi and Beppe Grillo are followed by more than 1,8 million users (more than the US Republican runner-up Mitt Romney).

Apart from the pivotal players, one can find a wide diffusion of Twitter among Italian politicians, even though not all parties have been able to exploit it proficiently. While collecting the data, it appeared clear that politicians from the right wing and from Southern Italy are less represented in social media than in reality.

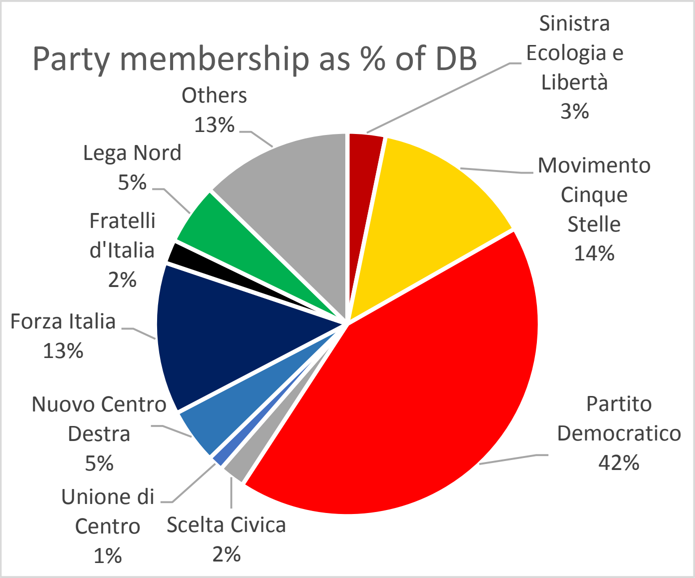


*Figure 1 Party membership as a percentage of the Database*

I have personally built a wide dataset trying to encompass the widest class of politicians possible. Although not necessary for the purposes of this work, I believe that this list will become useful in future stage of the research, allowing a micro-founded analysis of political communication. I have, then, looked for the Twitter account of every politician elected in the 2013 national elections; of those who took part to one of the last four government (Renzi, Letta, Monti and Berlusconi IV), all the ministries secretaries and undersecretaries; the member of the national secretary of the main parties, even if not elected; the European MPs elected in May 2014; all politicians elected at regional level as of January 2015, classifying them for their role and keeping track of their position in the region; The *assessors* of each region's capital city; the majors of all cities above 15.000 inhabitants. For the three latter categories, I kept track of the region of provenience, so to be able in future to estimate the geo-localisation of voters.

Given the great instability of Italian system, labelling the parties has been particularly challenging. To make an example, the right wing main parties as of 2009 were "*Alleanza Nazionale*" and "*Forza Italia*", who merged together in "*Popolo della Libertà*". However, during the last five years part of this PDL became independent, with the names "*Futuro e Libertà*" and "*Nuovo Centro Destra*", while the whole

PDL changed its name again in “*Forza Italia*”. Similarly, given the different electoral rules and alliances at National and local level, some parties merged temporarily only for an electoral campaign. Thus, the name of the party in the official lists of the elected is often inaccurate or in the need of clarification. To try to unify under a single label such a unstable system, I have often used the biography of the politician to assess his actual position. Moreover, a wide quantity of minor parties are present also at National level. Therefore, I have reduced the analysis to those with more than 10 candidates on Twitter, reducing from 26 to 9 the amount of parties (nonetheless, I still have the full information in my dataset).

This wide dataset, including 1402 accounts is highly self-selected, but generally speaking it reflects the relative share of elected people in Italy. The most underrepresented party is *Forza Italia*, the biggest right wing party, who exhibited the highest share of elected people without an account. On the other hand, *Movimento Cinque Stelle* seems a bit overrepresented, especially as it is almost not present at a local level, because it relies highly on internet communication.

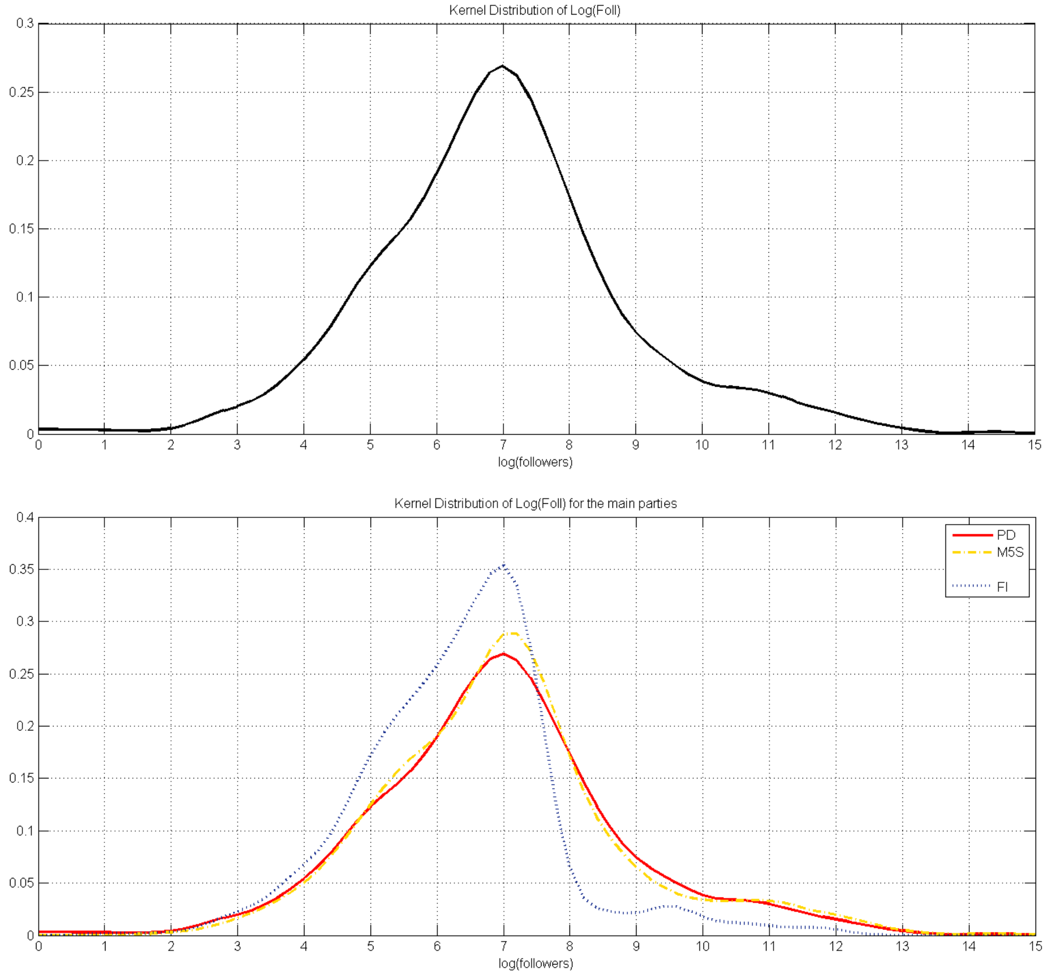


*Figure 2.1 and 2.2 Kernel distribution of Log(followers) for the whole population (2.1) and for the three main parties (2.2)*

These two figures help in understanding the distribution of the amount of followers of the politicians in the database. The distribution presents fat tails, in particular on the right side, meaning that there is a high amount of very popular accounts (which is reasonable, as the dataset includes more politician on the national level than on the local one), while the modal amount of followers seems to lie around

1000. To corroborate the thesis of left-wing biased database, one can disentangle the distribution for the three main parties. While the left parties (PD and M5S) have a distribution similar to the whole population, the right wing exhibits a sharp drop after the modal value. Indeed, among the right-wing we do not find many of the major players, who do not have a Twitter account; on the top of that, those who are on the platform and that should be expected to have the same popularity of their left-counterpart (e.g. former ministers), have sensibly fewer followers. Again, this suggests that right-wing politicians are less present on social networks and that also their voters are less active.

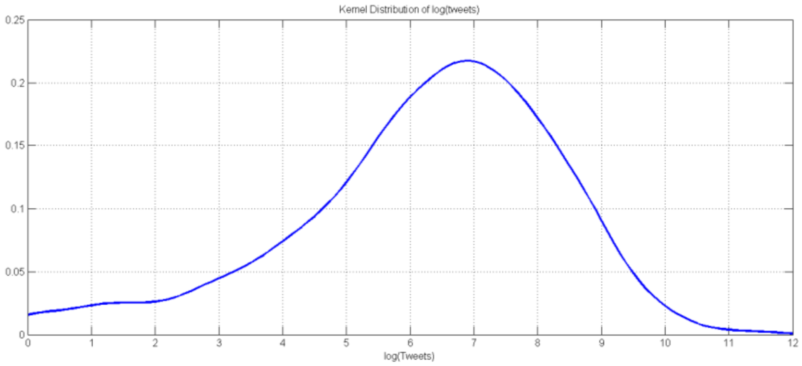


*Figure 4 Kernel distribution of the log number of tweets*

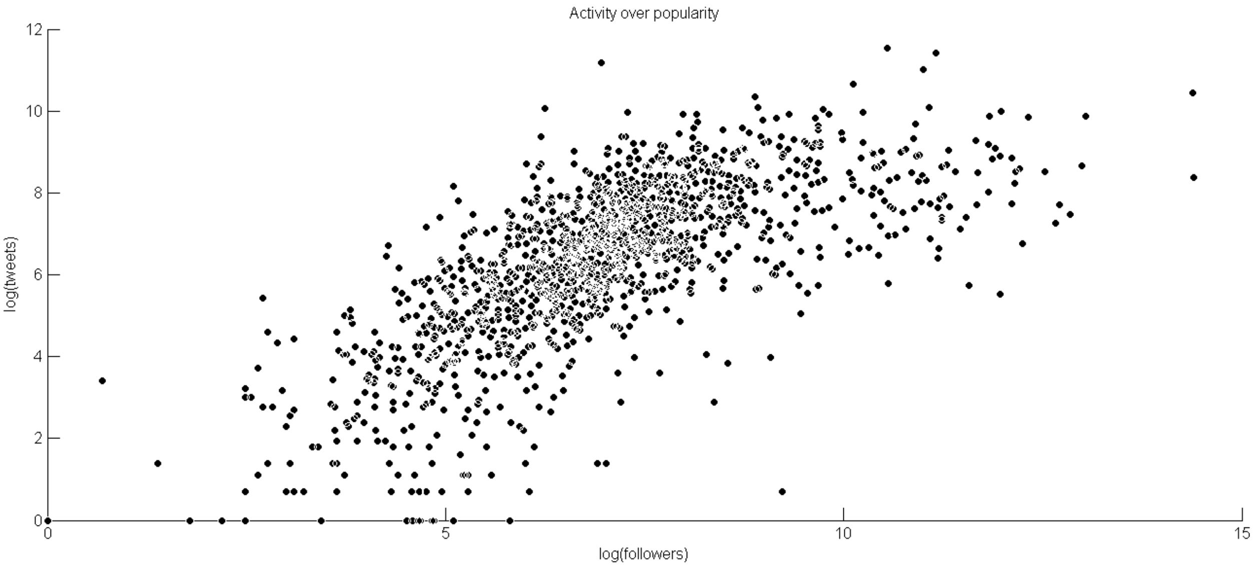


*Figure 3 number of tweets over number of followers, values in log*

It is also noticeable that a great share of the people in our database are not active users. Indeed, above 19% of the politicians have not sent more than 100 tweets (this is particularly true at local level). However, their presence in the model will not impact negatively the estimates as they do not focus on whether one follows a candidate because of the quality of his tweets, but only on the relative political closeness. The correlation between the number of followers and tweets in log is 0.704, confirming the intuition that the two dimensions should be linked: more popular actors are those who actually use Twitter more often, hence following does not just reflect the popularity of a candidate outside the social network. Surely, one cannot find a causal relationship, but it is reasonable to state that popularity and activity go in the same direction.

Once the list of politician was constructed, a long work of programming was required to get the data from Twitter. I have chosen to use the API wrapper Tweetinvi, so to handle the data collection through a C# procedure. I have used VisualStudio with the help of ReSharper to build a complex program that could repeat the queries to Twitter API, store the data and interact with a SQL database to write them. Unfortunately, Twitter caps the number of information that it is allowed to share. Indeed, although almost every information is in theory public, a great limit is posed by the query cap. For instance, a query like https://api.Twitter.com/1.1/followers/ids.json which is used to retrieve the list of followers of a certain user, will return only the last 5000 followers and can be called at most 15 times in a quarter of hour. Therefore, I had to build a cursored query that would ask for the *next* 5000 followers each time and that could take into account the limit of demands per timespan. This was a particularly challenging task, that one can see in the annexes. Anyway, even once the program was set, the data collection required a couple of weeks of unstopped cycles. However, the procedure I have elaborated is implementable with new tokens so to speed up the running time.

Unfortunately, even if the data about the following relationship are given in a sequential way, there is no way to get any piece of information on when the link was created. Therefore, one cannot build a panel data of the relationship status, as it was at first my intention. Nonetheless, by running the same code everyday for a long period, one can take notice of the changes in the list and derive an actual panel data for new followers. I could not follow this route for this work, as I could not collect data for a long period, but it is a promising data source for a future analysis.

I end up with a number of different users that follow at least a politician of 6041763. However, it is extremely likely that a great mass of these will be inactive users. In total, I have more than 14 millions relationships of the kind "*user follows politician*". Therefore, so to have an easier to manage dataset, I reduce my attention to those who follow at least 20 politicians in the list, have published at least 100 tweets and are followed by $150 \leq n \leq 5000$ other users. I keep a maximum of 5000 followers for two reasons: the first one is so to remove from the list public figures or institutional accounts, as I want to focus on *normal* users; the second is technical, as 5000 is the query limit for the followers' query. This reduces the number of *informative* users to barely more than 53 thousands. However, to get the information about published tweets and followers of the whole mass of users, one has to preliminarily run a new set of queries, which takes approximately a dozen days of uninterrupted work.

### THE POLITICAL SPACE MODEL

Let us now focus on the procedure I am going to implement to retrieve information about the political ideas of users and politicians. While some indicator of this dimension can be given for candidates, whose party affiliation is *in se* a declaration of political ideas, nothing can be said about users. Indeed, Twitter does not provide any self-identification field (which is present on facebook, for instance). One might try to look in the description provided by the user for some party-specific word. However, this strategy is extremely reductive, as many people, albeit politicized, might now write it explicitly in their profile. Moreover, the possibility of content injection is high. For instance, querying for descriptions that contain the word "*PD*" gives only 1119 results out of the 6 million + database. Among those, roughly an half are elected representatives or official account of local sections of the party. On the top of that we have about one hundred account that use the word "*PD*" without any political meaning (e.g. PD is also the short name for the city of Padua) or that write it ironically (e.g. saying "*I don't vote for PD as I'm a real socialist*"). Similarly, attempting an analysis of the tweet content can be disappointing.

Conover et al. (2010) show that the use of hashtags is highly difficult to predict one's actual opinion on a subject.

Hence, I focus on an original approach from Barberà (2015) who is capable to predict one's political position exploiting the network structure. Indeed, what is fascinating about this approach is the fact of having a homogenous measurement scale for both normal users and political actors, which does not rely on any prior information of any of those, but instead it is just derived by some assumptions of the determinants of the following relationship. The author exploits the fact that networks tend to be homophile (McPherson et al., 2001) and that users will be more likely to follow people of closer political ideology (Bryant and Miron, 2004).

This makes sense: as Twitter does not filter tweets and we can assume that people have a finite amount of time to spend on their feed and prefer to have relevant tweets, it makes sense that one should be more likely to follow people he is actually interested into. Surely, something as simple as the relative share of followed politicians that belong to a certain party is not a good measure of the political ideas. If, for instance, some party has more popular candidates (e.g. because it is the party at the government), many users of other parties would follow them because of their popularity without sharing their political thought.

Therefore, we need a measure that takes into account the popularity of candidates and considers the structure of the relationships as a whole. Barberà proposes this specification for the probability $P(y_{ij} = 1)$ of user $i$ following political actor $j$

$$P\left(y_{ij} = 1 \middle| \alpha_j, \beta_i, \gamma, \theta_i, \varphi_j\right) = logit^{-1}(\alpha_j + \beta_i - \gamma \left\|\theta_i - \varphi_j\right\|^2)$$

Where $\alpha_j$ is the *outdegree* of the user, capturing his political interest, while $\beta_i$ is the *indegree* popularity of a politician. These parameters take into account the fact that some users might be more keen on following a high amount of politicians, whereas some political actor might have a lot of followers because of his activity on the network or its role in society more than for his political ideas.

The quadratic term, instead, reflects the ides of *selective exposure* I have aforementioned. When one's political idea matches perfectly the candidate's one, this term will be zero and minimized. The further one gets from the candidate, the less like he will follow it. $\gamma$ is just a scaling parameter that multiplies the importance of political closeness.

The estimation of this model is highly complicated. Indeed, using the whole database of 1402 politicians and 6 million users, it would require to estimate more than 12 millions parameters. Even subsampling the users database, this would still require an intractable problem with most methods. I have tried many alternative ways and specifications for this position, trying to exploit some patterns of the data, but at the end I realised that a Monte Carlo Markov Chain method would have been the only solution.

In particular, I have used a Hamiltonian MCMC with a NUTS procedure, following the original paper's hint. This model draws i.i.d. set of parameters for the variables to estimate. In a nutshell, the model will investigate the neighbourhoods of the proposed parameters, with a trial and error chain of random proposals for the unknown parameters that will be accepted if they make the simulated and the actual data match better. For a more detailed introduction to MCMC and Metropolis-Hastings algorithms Andrieu et al. (2003) or Gilks et al. (1996) can be very helpful.

| party | phi |
|---|---:|
| sel | -1 |
| 5s | -0,75 |
| ex5s | -0,75 |
| idv | -0,75 |
| pd | -0,5 |
| rad | -0,5 |
| ind | -0,25 |
| psi | -0,25 |
| gal | 0,25 |
| gpa | 0,25 |
| local | 0,25 |
| misto | 0,25 |
| null | 0,25 |
| pensionati | 0,25 |
| pi | 0,25 |
| pop | 0,25 |
| fi | 0,5 |
| GrandeSud | 0,5 |
| ncd | 0,5 |
| sc | 0,5 |
| udc | 0,5 |
| frat | 1 |
| lades | 1 |
| lega | 1 |

The original work treats the mean and the variance of such distributions as parameters to be estimated, but to simplify the computation I assumed that the hyperparameters all come from $\sim N(0,1)$. I have further simplified the original model as the computation time was too high for a personal computer. To give a flavour of the running times: to estimate the model with an optimized and simplified version of the model, at an average RAM usage of 8GB with an i7 quadcore processor, it took me more than 8 days to get the results. This is surely the greatest limit of this approach, as it takes a huge amount of time and does not give any feedback before the procedure is completed. Thus, if for instance some chain does not converge well, one can find it out only when it is too late. At each simulation, indeed, more than 20 thousands parameters have to be estimated, so to allow the model to converge to the fittest values.

The choice of the a priori is particularly important to reduce the convergence time. Hence, I started from some slightly informative a priori for the political position of candidates. I have used a left/right relative positioning from Openpolis to assign starting values $\in [-1,1]$. The most important factor is the sign, indicating that a certain politician is on the left (negative sign) or on the right. The table on the left reports these initial values for all the parties we had in the database. Some values can be arguable, as Italian system is very volatile and is becoming more and more tri-polar (one can hardly insert the second biggest party, *Movimento 5 Stelle*, in a left/right logic). However, as it will be clear later, these a priori will not determine the final values when the MCMC converges well.

However, one cannot provide any a priori on the *users*. I have randomly selected a subset of 10 thousands *informative users*, but the best a priori in this case is given by a random normal parameter. However, this part of the estimation is more accurate to retrieve the values for politicians, from whom one will be able to compute the *users'* in a second time. For the popularity of the politician and the politicization of users, instead, I have used the logarithm of the number of links they have in the adjacency matrix.

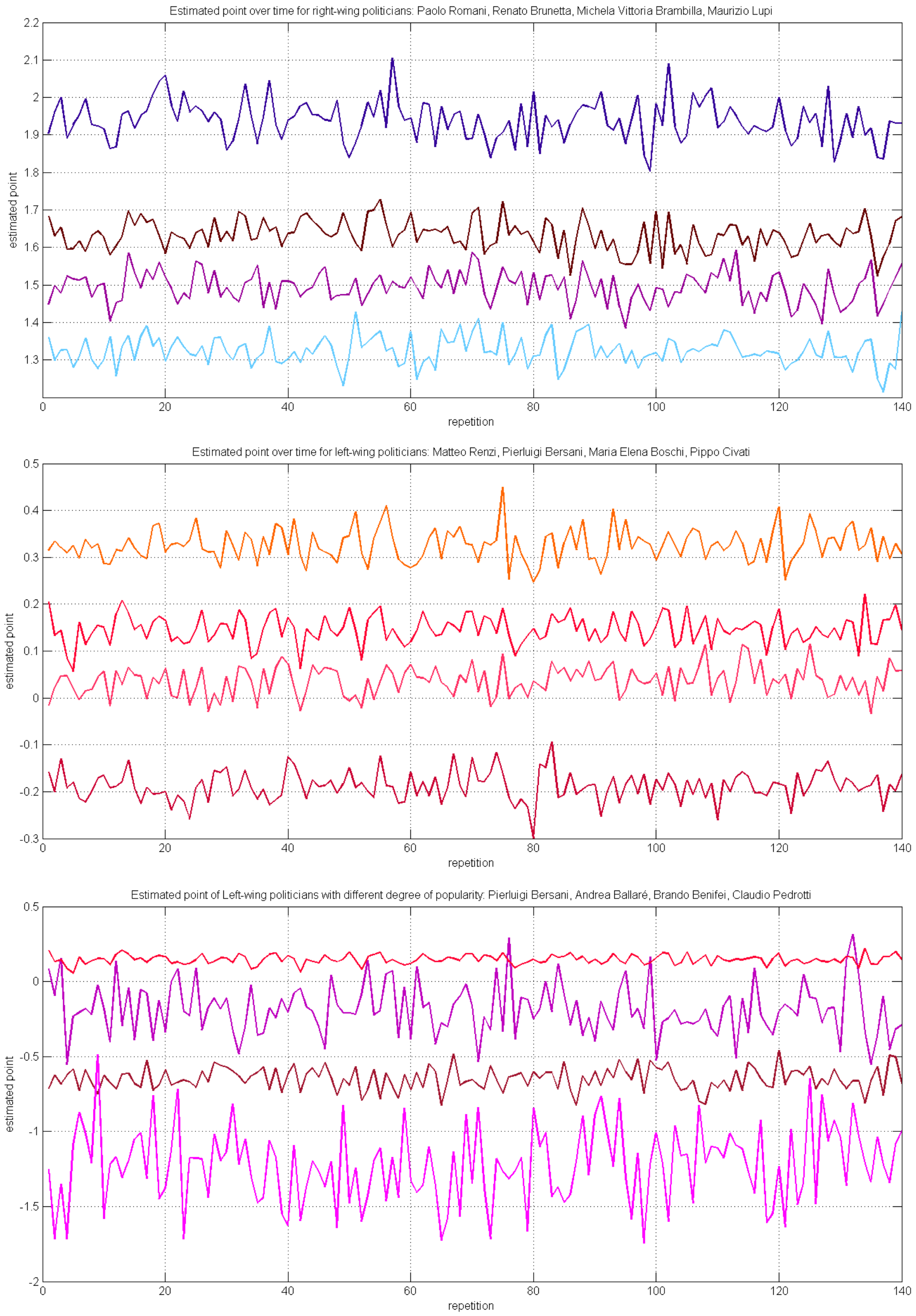


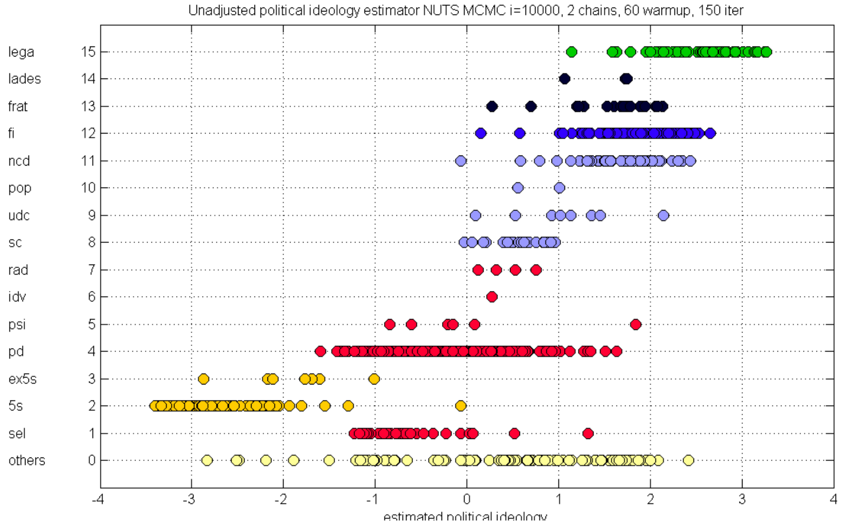
*Figure 5.1-5.3: estimation of the parameter $\varphi_j$ for some representative politician for the right-wing (5.1), of the left-wing (5.2) and of left-wing with different levels of popularity: national level (Bersani), EU parliament (Benifei) and city major (Ballaré and Pedrotti)*

To run this model, I have used the rstan package on RStudio. I tried to write a NUTS MCMC algorithm on Matlab, but its efficiency was too low, so I opted for an already existing package. I have set for two chains (results represent the average of the two), warmup of 60 and 140 iterations. One would very much desire to iterate the procedure more times, however it would have taken more than a month of computation with the means I had a disposition to estimate the model with more than 500 iterations. Surely, more powerful computers would be able to handle the simulation is lesser time.The first and the second figure show the evolution of the estimated political point over time. All the politicians in the same figure had the same a priori, however, the starting points are different as the first 60 repetitions of warmup already calibrated the model. Indeed, it is noticeable that the values fluctuate around the starting point, although they do not seem to become less volatile over time as one should expect. Thus, for the estimation we will use the average of the last three values of this series. One can remark that the relative position on the spectrum is particularly accurate in an ordinal point of view. From the four representative politician picked for the centre-left, for instance, the model correctly puts Matteo Renzi in a more central position, where Bersani is more on the left and Civati is the most extreme of the three. This exactly represent their political positions and affiliation in the current scenario. One can also notice that the lines of politicians in the same graph tend to move in a similar way, this reflects the fact that they share a huge amount of followers, which corroborates the initial assumption on homophily in the following relationship. For instance, while the correlation between Michela Vittoria Brambilla and Renato Brunetta, who were in the same right-wing government and the same party, is 0.4265, that of Brambilla and Pippo Civati, who is in the left side of the democratic party, is only 0.0753.

This estimation system, however, becomes more inaccurate where the number of followers decreases. The last figure depicts the convergence pattern of politicians of different degrees of popularity. The almost flat line is associated to Pierluigi Bersani, winner of 2013 elections and former secretary of PD; the third one is associated to a young member of European Parliament with an average degree of popularity (6847 followers), while the second and the fourth lines are those of majors of medium size cities (Novara and Pordenone). Their volatility is remarkably higher. Bersani has a standard deviation of 0.0305, while Ballaré 0.1728. Moreover, the ordinal political positioning seems inaccurate. They are considerably too much on the left with respect to the national politicians.

The model seems to work pretty well also for politicians of the *Movimento 5 Stelle*, which is located on average around $-3$. On the other hand, it fails significantly in estimating the position of the movement's leader, Beppe Grillo, who is situated according to the simulation in the PD area. It is true that, being particularly famous, he has a number and variety of followers that might drive him towards the centre, but this estimation is particularly erroneous. I think that this might be due to the particular shape of the Italian system, where it is pretty difficult to put in a left/right spectrum the three main parties. Moreover, the presence of a 3-party system unavoidably pushes one of the party around zero in the estimations, making it quite difficult to interpret them. Indeed, this model has reflection invariance: the scale can be reverted left to right and might give results that are of the right absolute value, but with the wrong sign. This can be easily solved in a bipolar system, by changing the signs whenever the result goes against the priors, however if a party is around zero, it becomes more difficult to interpret it.

Nonetheless, the results are surprisingly good, as they identify the parties in a logical ordinal way. It is particularly interesting that the model is able to correct for wrong ordinal priors. For instance, I put *Sinistra Ecologia e Libertà* more on the left that *Movimento 5 Stelle*. However, it is common knowledge that the electorate of the last has much more in common with *Partito Democratico*, with whom it was allied at 2013 elections, than how *M5S* does. Therefore, the estimation correctly situated the latter in a more extreme position.



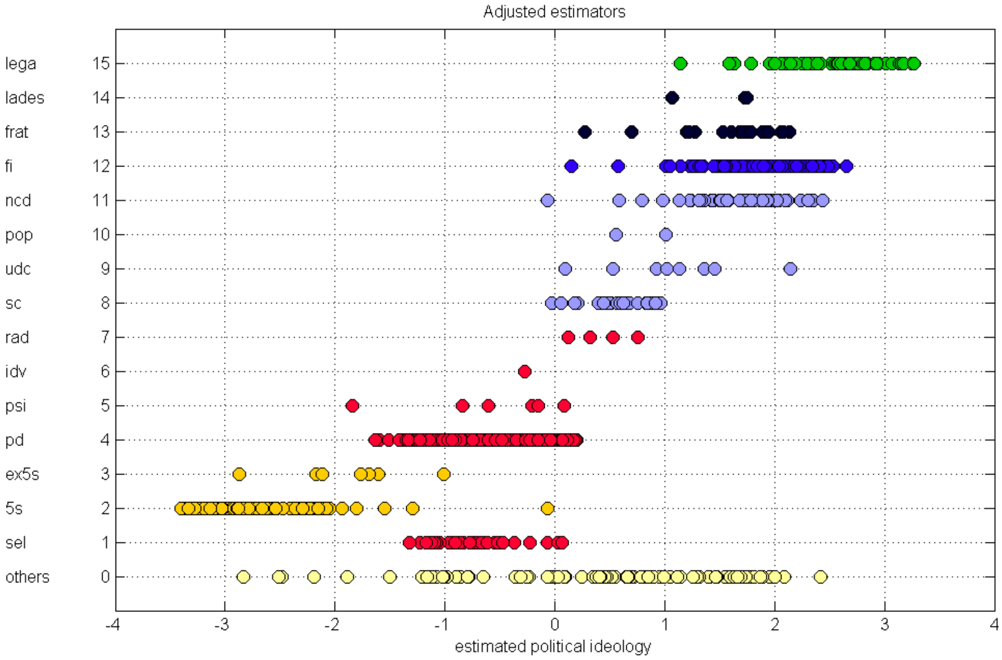
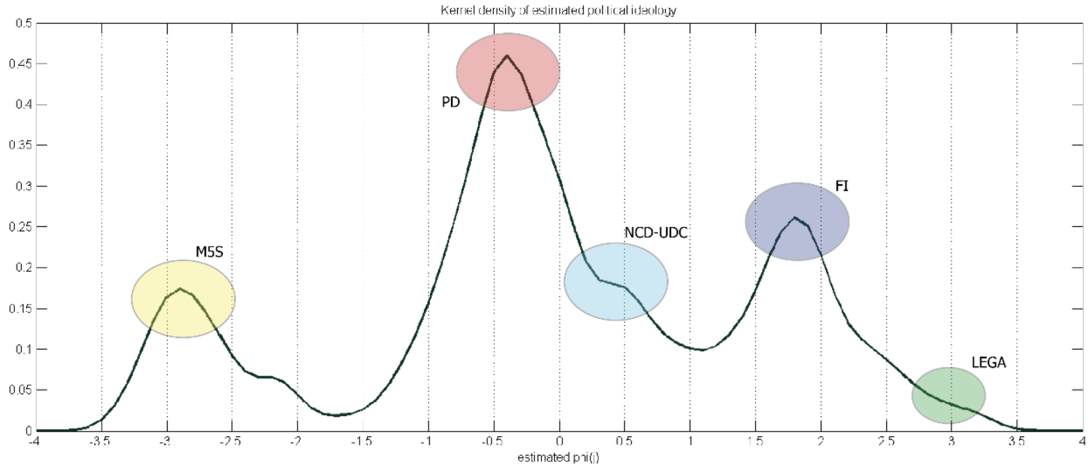


*Figure 6. and 6.2: estimated political ideology from the MCMC procedure; unadjusted for sign mismatch (6.1) and adjusted (6.2) colours represent appurtenance to the same part of the political spectrum*

Arbitrarily correcting for mismatching signs for parties (reversing the signs when the final result is too much on the right for someone belonging to a left party and viceversa), we have a neater and very sound representation. It is nevertheless noticeable that more popular candidates tend to be more on the centre with respect to the other members of the same party. This could be probably solved by a more sophisticated model that gives more weight to the popularity, but its implementation is far too demanding in terms of time of computation to be interesting to develop. In particular, I have limited the variation of the parameter weighting the relative importance of political ideology with respect to popularity and the popularity indicators in a $N{\sim}(0{,}1)$ distribution. This is arbitrary, as one cannot know nor guess the mean and, more importantly, the variance of such parameters for the MCMC procedure. Allowing for an estimation of the meta-parameters would surely provide a more balanced indicator, but it would slow down a procedure that was already difficult to be handles by a single computer. This estimation required already a huge amount of time, increasing the number of estimated parameters would make it unfeasible.

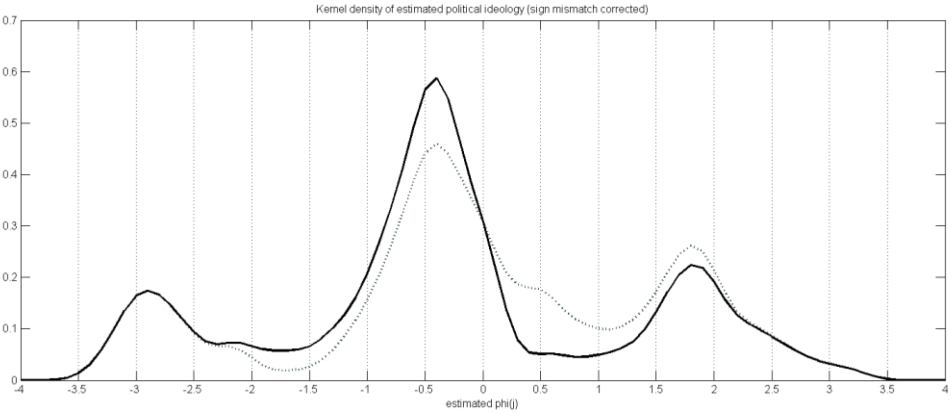
The graph below depicts the density function of the estimated political subspaces of all politicians, while the median area associated to each party is highlighted with a circle. We have a particularly good positioning of the mass, with three peaks corresponding to the main parties. The order and the relative closeness make sense and confirms that the estimation strategy fits the reality. If we arbitrarily correct for sign mismatch for left wing politician estimated with positive coefficient, we end up with a slightly changed figure that accentuates the three-party system.



Adjusting for the possible sign mismatch will slightly change the density in the middle of the figure, making even sharper the identification of the three main political areas. Controlling with the party affiliation I have on the database, one has a very neat identification of the areas, confirming the fact that the procedure succeeds in identifying the political area. The mean and standard deviation for the main parties are as follow:

| | **Mean** | **StdDev** |
|---|---|---|
| *M5S* | -2.7236 | 0.4359 |
| *PD* | -0.4447 | 0.3309 |
| *FI* | 1.8276 | 0.3871 |

notice that PD is almost equidistant from the other two parties.



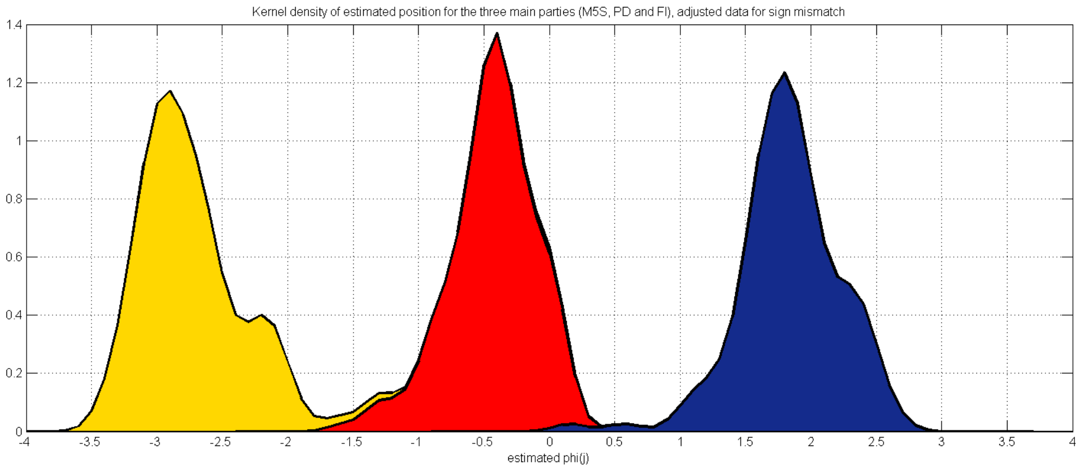


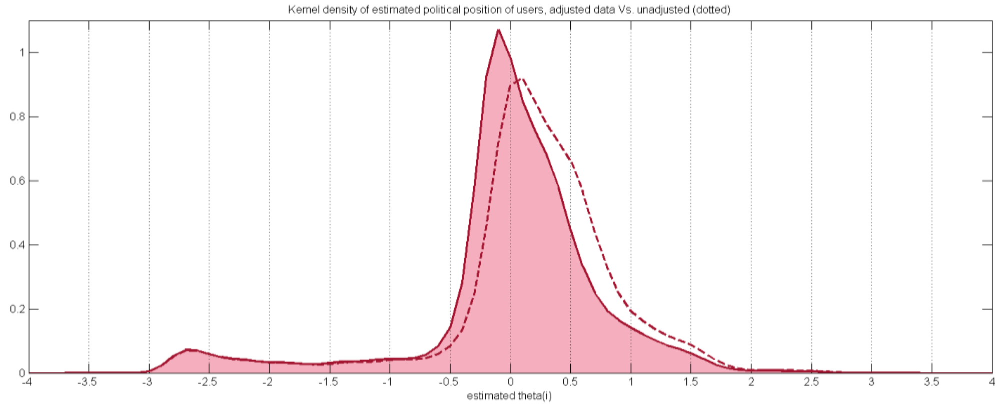
*Figure 7.1-7.3; Kernel density estimation of the estimated political position. 7.1 shows the rough data and highlights the area where we should expect the main parties to lie. 7.2 corrects for sign mismatch, making sharper the identification of centre-left candidates, unadjusted data in dotted line. 7.3 shows the Kernel density for the subgroup of politicians belonging to the three main parties: those in the area where we should expect party x are actually affiliated to that party.*

### Estimation of users' political preferences

Let us now proceed to the estimation of the political position of users. Despite one would like to use a method similar to the one adopted to estimate politician's ideal points, the great amount of time needed to compute these data forces us to find a simpler way to get some parameter about users. Therefore, I use the sum of the coefficients of followed politician weighted for their relative popularity.

$$\theta_i = \frac{\sum_{j=1}^{J} y_{ij} \log\left(1+\sum_{i=1}^{I} y_{ij}\right)\varphi_i}{\sum_{j=1}^{J} y_{ij} \log\left(1+\sum_{i=1}^{I} y_{ij}\right)}$$

This specification takes into account the fact that people tend to follow more those who are closer to them, while correcting for the fact that more popular candidates should tell less about one's ideology, as they might be followed because of their role more than because of their position. This indicator is pretty rougher than what I wanted to build at first place, but the technical limits due to the computation time feasibility do not allow me to do more than this. In particular, it is likely that the overall distribution of users will be too much on the centre, where the Democrats are, because of a centripetal force due to the fact that candidates on the extremes will outbalance one another. However, as I will discuss shortly, the final shape of the density distribution is satisfying, once the limits of the method are taken into account.



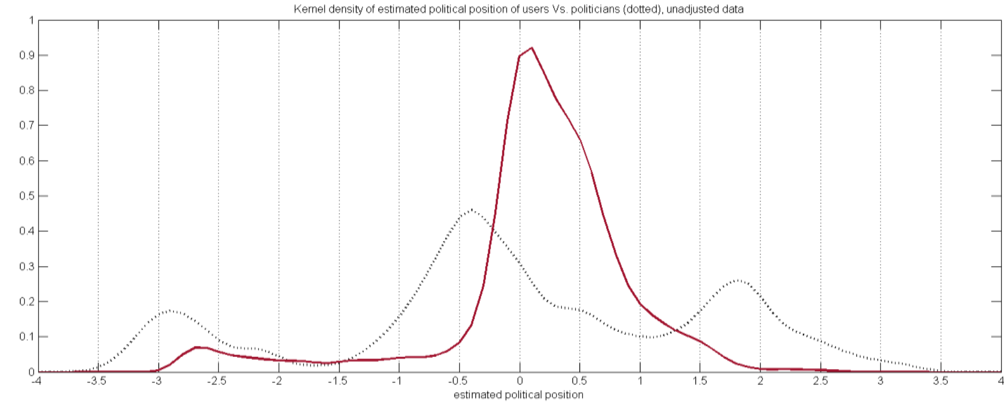
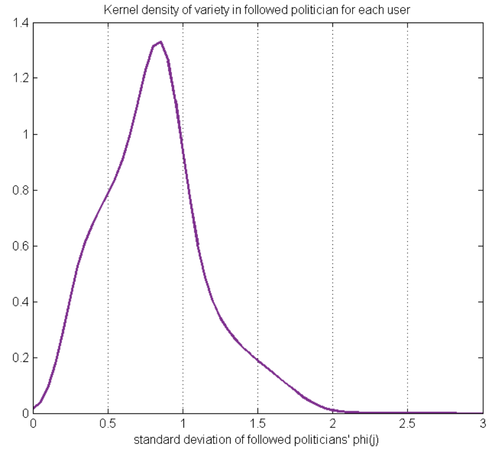


*Figure 8.1-8.3: 8.1, kernel density estimation of the political position of users, given their following choices confronted with that computed for politicians in 7.1 (dotted), computed taking into account of mismatching signs (8.2) and density of the standard deviation of the followed politicians weighted for their popularity (8.3).*



The final result shows, as depicted in the graph, a great mass of users around zero, between the two big parties in the left/right spectrum. This is on the one hand plausible, as it is common knowledge that the electorate tends to be less partisan than political actors. However, it is not clear if the relative closeness of the majority of people to the *Partito Democratico* is due to the fact that the median voter is represented by the latter (which might be the case, seen the remarkably better results this party had in the last elections) or only to the mathematical effect of being in the middle and capturing also all voters following a great variety of politician. The interpretation of the standard deviation, weighted for popularity, of the political values $\varphi_j$ of the politicians followed by user $i$ tell that a wide amount of people have a relatively low variety of politicians in their following list. Therefore, one can be more confident in using the estimators derived with this method.

Trying other refinements, for instance changing the weighting function or eliminating the users with a too wide standard deviation, does not change significantly the shape of the distribution, nor the single values. The most remarkable thing that one can notice while reducing the sample to users with homogenous following scheme is that the small bulgy area around -2.7 (the mean point of M5S candidates) remains untouched, while all other areas far from zero tend to disappear. This gives a hint on the fact that the electors of that party might show a greater level of homophily in tastes. This might just be a mechanical effect of users with extreme values to have high coefficients only because less exposed to diversity. Nonetheless, it is difficult to explain why this phenomenon would be present on extremes corresponding to M5S only and not on the right area. The insight is that there might be something specific about voters of that party that is not just due to mechanical effects of the estimation procedure.

This measure of users' relative position is surely less accurate than that of politicians, in particular because it is obtained through a direct transformation of politicians' estimators, while a more complex second stage MCMC with a model similar to that used for followed actors would provide a more reliable indicator. Therefore, for the time being, I propose to restrict the population to three main categories, flattening the relative position within a party's influence zone. I thereby select the population of politically active users by labelling as party *x*'s voters only those whose estimator lies within a standard deviation from the party's mean. i.e.

$$l_i = x \leftrightarrow \overline{\varphi^x} - \sigma(\varphi^x) \leq \theta_i \leq \overline{\varphi^x} + \sigma(\varphi^x)$$

With this criterion, I obtain a list of 1126 who I label as PD voter, 250 for FI and 346 for M5S. The fact that PD's voters are many more than those of the other parties is surely due to the central position of this party in the results of the parameters estimation. As argued before, due to the imperfect specification of the users' parameter and the peculiar tri-polar shape of Italian system, people emerging as PD's partisan might instead be users who follow a very various amount of politicians, as extreme candidates at different poles would cancel out one another.

In particular, it is not possible to unequivocally say that a voter of M5S party would consider himself closer to PD than to FI (and vice versa), as a linear model would suggest. Indeed, the analysis of political trends show a very volatile and unpredictable scenario. However, I believe that these trends of swing voters from the right to M5S are captured by a population that is very underrepresented in terms of age and status by Twitter users. Nonetheless, I would like to explore a multi-dimensional model that uses instead of a left/right scale a PD/M5S; M5S/FI; FI/PD one. This could be done by removing from the MCMC simulation at the first stage all affiliates to the parties in the spectrum to be excluded, and then trying to harmonise the obtained data.

Looking at the Kernel density estimator of users' position and the mean of party affiliates, it is clear that M5S and FI would have fewer voters, as we are on the tail of the distribution. However, by randomly checking the profile of some of the *voters* and trying to guess their political ideas by reading their feed, I have had a good feedback in terms of accuracy of the estimator. Although this is more a rule of thumb than a formal way to determine the correctness of the procedure, it confirms that it can be safe for the time being using these data for a descriptive analysis of the political network.

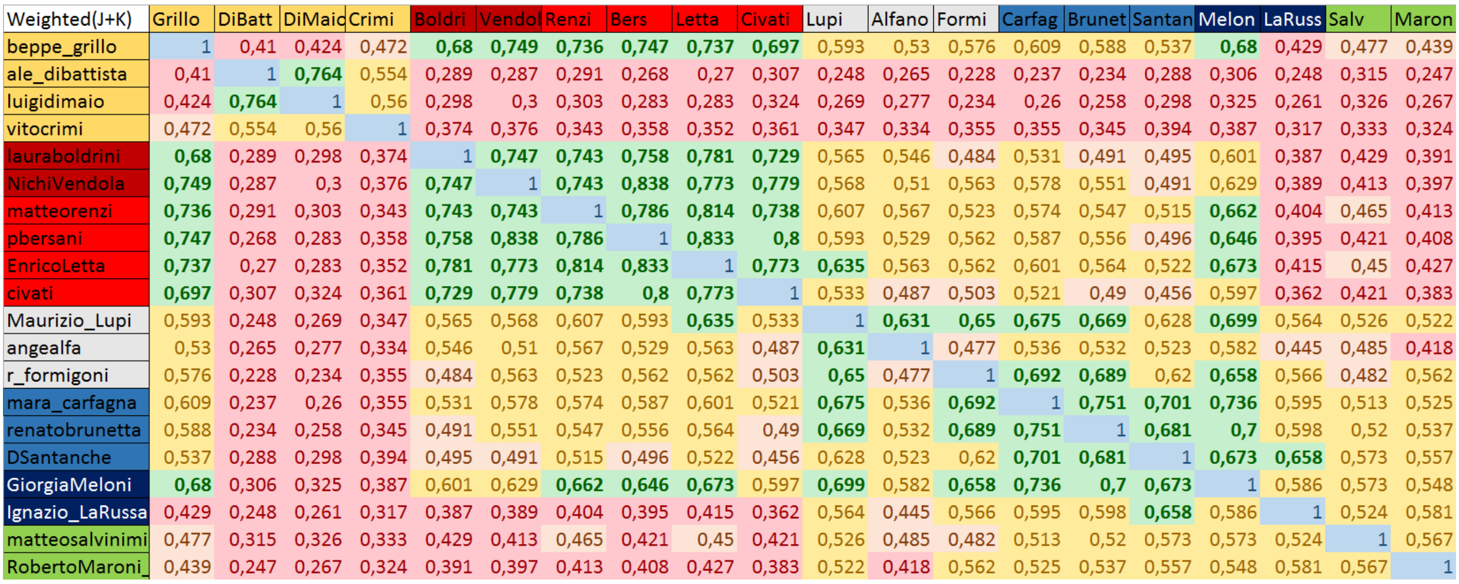
For this stage of the analysis I have used as list of users the same one I adopted for the MCMC model that estimated politicians' positions, with the addition of 4000 new randomly selected accounts. I did so because of convenience, as the extrapolation of these datasets is extremely costly in terms of time. One should be aware that by using the same list of users that has been exploited by the first stage model, the estimators of political activity of users already tell us something about the politicians they follow. There might be some echoing problem between these two dimensions, as more clustered in political tastes users will tend to increase the absolute value of $\theta_j$. This could be very well resolved by selecting a completely new set of accounts for the political ideology estimation of users. Indeed, by doing so, we would rule out any possible contamination due to this reflection problem. By checking the Kernel density for the newly inserted users only, however, I had a distribution very similar to that of users exploited in the first place, so I am confident in using this list, albeit imperfect.

### CLUSTERS AND HOMOPHILY BETWEEN POLITICIANS

Focusing now on clusters, I will attempt to build an indicator of clustering between politicians. I cannot use the standard definitions, as this network sees two classes of actors *users* and *politicians* and we want to map how many *users* are in common between two *politicians*. The modified version of the canonical index that I propose is:

$$m_{jk} = \frac{2\sum_{n=1}^{N} e_j^n e_k^n}{\sum_{n=1}^{N} e_j^n + \sum_{n=1}^{N} e_k^n}$$

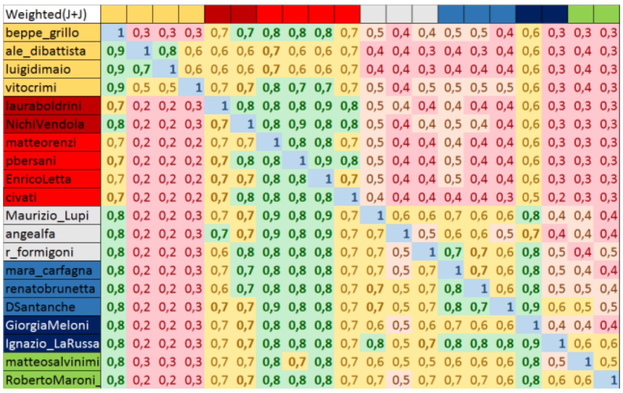
Where $e_i^n$ is equal to one iff user $n$ follows politician $i$.



| Weighted(J+K) | Grillo | DiBatt | DiMaic | Crimi | Boldri | Vendol | Renzi | Bers | Letta | Civati | Lupi | Alfano | Formi | Carfag | Brunet | Santan | Melon | LaRuss | Salv | Maron |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| beppe_grillo | 1 | 0,41 | 0,424 | 0,472 | 0,68 | 0,749 | 0,736 | 0,747 | 0,737 | 0,697 | 0,593 | 0,53 | 0,576 | 0,609 | 0,588 | 0,537 | 0,68 | 0,429 | 0,477 | 0,439 |
| ale_dibattista | 0,41 | 1 | 0,764 | 0,554 | 0,289 | 0,287 | 0,291 | 0,268 | 0,27 | 0,307 | 0,248 | 0,265 | 0,228 | 0,237 | 0,234 | 0,288 | 0,306 | 0,248 | 0,315 | 0,247 |
| luigidimaio | 0,424 | 0,764 | 1 | 0,56 | 0,298 | 0,3 | 0,303 | 0,283 | 0,283 | 0,324 | 0,269 | 0,277 | 0,234 | 0,26 | 0,258 | 0,298 | 0,325 | 0,261 | 0,326 | 0,267 |
| vitocrimi | 0,472 | 0,554 | 0,56 | 1 | 0,374 | 0,376 | 0,343 | 0,358 | 0,352 | 0,361 | 0,347 | 0,334 | 0,355 | 0,355 | 0,345 | 0,394 | 0,387 | 0,317 | 0,333 | 0,324 |
| lauraboldrini | 0,68 | 0,289 | 0,298 | 0,374 | 1 | 0,747 | 0,743 | 0,758 | 0,781 | 0,729 | 0,565 | 0,546 | 0,484 | 0,531 | 0,491 | 0,495 | 0,601 | 0,387 | 0,429 | 0,391 |
| NichiVendola | 0,749 | 0,287 | 0,3 | 0,376 | 0,747 | 1 | 0,743 | 0,838 | 0,773 | 0,779 | 0,568 | 0,51 | 0,563 | 0,578 | 0,551 | 0,491 | 0,629 | 0,389 | 0,413 | 0,397 |
| matteorenzi | 0,736 | 0,291 | 0,303 | 0,343 | 0,743 | 0,743 | 1 | 0,786 | 0,814 | 0,738 | 0,607 | 0,567 | 0,523 | 0,574 | 0,547 | 0,515 | 0,662 | 0,404 | 0,465 | 0,413 |
| pbersani | 0,747 | 0,268 | 0,283 | 0,358 | 0,758 | 0,838 | 0,786 | 1 | 0,833 | 0,8 | 0,593 | 0,529 | 0,562 | 0,587 | 0,556 | 0,496 | 0,646 | 0,395 | 0,421 | 0,408 |
| EnricoLetta | 0,737 | 0,27 | 0,283 | 0,352 | 0,781 | 0,773 | 0,814 | 0,833 | 1 | 0,773 | 0,635 | 0,563 | 0,562 | 0,601 | 0,564 | 0,522 | 0,673 | 0,415 | 0,45 | 0,427 |
| civati | 0,697 | 0,307 | 0,324 | 0,361 | 0,729 | 0,779 | 0,738 | 0,8 | 0,773 | 1 | 0,533 | 0,487 | 0,503 | 0,521 | 0,49 | 0,456 | 0,597 | 0,362 | 0,421 | 0,383 |
| Maurizio_Lupi | 0,593 | 0,248 | 0,269 | 0,347 | 0,565 | 0,568 | 0,607 | 0,593 | 0,635 | 0,533 | 1 | 0,631 | 0,65 | 0,675 | 0,669 | 0,628 | 0,699 | 0,564 | 0,526 | 0,522 |
| angealfa | 0,53 | 0,265 | 0,277 | 0,334 | 0,546 | 0,51 | 0,567 | 0,529 | 0,563 | 0,487 | 0,631 | 1 | 0,477 | 0,536 | 0,532 | 0,523 | 0,582 | 0,445 | 0,485 | 0,418 |
| r_formigoni | 0,576 | 0,228 | 0,234 | 0,355 | 0,484 | 0,563 | 0,523 | 0,562 | 0,562 | 0,503 | 0,65 | 0,477 | 1 | 0,692 | 0,689 | 0,62 | 0,658 | 0,566 | 0,482 | 0,562 |
| mara_carfagna | 0,609 | 0,237 | 0,26 | 0,355 | 0,531 | 0,578 | 0,574 | 0,587 | 0,601 | 0,521 | 0,675 | 0,536 | 0,692 | 1 | 0,751 | 0,701 | 0,736 | 0,595 | 0,513 | 0,525 |
| renatobrunetta | 0,588 | 0,234 | 0,258 | 0,345 | 0,491 | 0,551 | 0,547 | 0,556 | 0,564 | 0,49 | 0,669 | 0,532 | 0,689 | 0,751 | 1 | 0,681 | 0,7 | 0,598 | 0,52 | 0,537 |
| DSantanche | 0,537 | 0,288 | 0,298 | 0,394 | 0,495 | 0,491 | 0,515 | 0,496 | 0,522 | 0,456 | 0,628 | 0,523 | 0,62 | 0,701 | 0,681 | 1 | 0,673 | 0,658 | 0,573 | 0,557 |
| GiorgiaMeloni | 0,68 | 0,306 | 0,325 | 0,387 | 0,601 | 0,629 | 0,662 | 0,646 | 0,673 | 0,597 | 0,699 | 0,582 | 0,658 | 0,736 | 0,7 | 0,673 | 1 | 0,586 | 0,573 | 0,548 |
| Ignazio_LaRussa | 0,429 | 0,248 | 0,261 | 0,317 | 0,387 | 0,389 | 0,404 | 0,395 | 0,415 | 0,362 | 0,564 | 0,445 | 0,566 | 0,595 | 0,598 | 0,658 | 0,586 | 1 | 0,524 | 0,581 |
| matteosalvinimi | 0,477 | 0,315 | 0,326 | 0,333 | 0,429 | 0,413 | 0,465 | 0,421 | 0,45 | 0,421 | 0,526 | 0,485 | 0,482 | 0,513 | 0,52 | 0,573 | 0,573 | 0,524 | 1 | 0,567 |
| RobertoMaroni_ | 0,439 | 0,247 | 0,267 | 0,324 | 0,391 | 0,397 | 0,413 | 0,408 | 0,427 | 0,383 | 0,522 | 0,418 | 0,562 | 0,525 | 0,537 | 0,557 | 0,548 | 0,581 | 0,567 | 1 |

*Table 1 number of shared followers between 20 representative politicians (those with most followers within their party). Colours represent the index belonging to a certain percentile range, described below. Politicians grouped by party and ordered left/right*

This table displays the matrix of shared followers, computed with the procedure described above. I have used different colours to highlight values over the 75th percentile (bold and green), between 50th and 75th (yellow), between 25th and 50th (pink) and below the 25th (red). Also, I have disposed the users grouping for their party and in the order left/right provided by the model. This indicator of political clustering is surely complicated to interpret, as the number of followers varies a lot between the agents. Indeed, if an actor has considerably less followers than another, the indicator will be small, even though all the small actor's followers might be in the big one's list. Conversely, if such correction for the size of both users was not to



| Weighted(J+J) | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| beppe_grillo | 1 | 0,3 | 0,3 | 0,3 | 0,7 | 0,7 | 0,8 | 0,8 | 0,8 | 0,7 | 0,5 | 0,4 | 0,4 | 0,5 | 0,5 | 0,4 | 0,6 | 0,3 | 0,3 | 0,3 |
| ale_dibattista | 0,9 | 1 | 0,8 | 0,6 | 0,6 | 0,6 | 0,7 | 0,6 | 0,6 | 0,7 | 0,4 | 0,4 | 0,3 | 0,4 | 0,3 | 0,4 | 0,6 | 0,3 | 0,4 | 0,3 |
| luigidimaio | 0,9 | 0,7 | 1 | 0,6 | 0,6 | 0,6 | 0,7 | 0,6 | 0,6 | 0,7 | 0,4 | 0,4 | 0,3 | 0,4 | 0,4 | 0,4 | 0,6 | 0,3 | 0,4 | 0,3 |
| vitocrimi | 0,9 | 0,5 | 0,5 | 1 | 0,7 | 0,7 | 0,8 | 0,7 | 0,7 | 0,7 | 0,5 | 0,4 | 0,5 | 0,5 | 0,5 | 0,5 | 0,6 | 0,3 | 0,4 | 0,3 |
| lauraboldrini | 0,7 | 0,2 | 0,2 | 0,3 | 1 | 0,8 | 0,8 | 0,8 | 0,9 | 0,8 | 0,5 | 0,4 | 0,4 | 0,4 | 0,4 | 0,4 | 0,6 | 0,3 | 0,3 | 0,3 |
| NichiVendola | 0,8 | 0,2 | 0,2 | 0,3 | 0,7 | 1 | 0,8 | 0,9 | 0,8 | 0,8 | 0,5 | 0,4 | 0,4 | 0,5 | 0,4 | 0,4 | 0,6 | 0,3 | 0,3 | 0,3 |
| matteorenzi | 0,7 | 0,2 | 0,2 | 0,2 | 0,7 | 0,7 | 1 | 0,8 | 0,8 | 0,7 | 0,5 | 0,4 | 0,4 | 0,4 | 0,4 | 0,4 | 0,6 | 0,3 | 0,3 | 0,3 |
| pbersani | 0,7 | 0,2 | 0,2 | 0,2 | 0,7 | 0,8 | 0,8 | 1 | 0,9 | 0,8 | 0,5 | 0,4 | 0,4 | 0,5 | 0,4 | 0,4 | 0,6 | 0,3 | 0,3 | 0,3 |
| EnricoLetta | 0,7 | 0,2 | 0,2 | 0,2 | 0,7 | 0,7 | 0,8 | 0,8 | 1 | 0,7 | 0,5 | 0,4 | 0,4 | 0,5 | 0,4 | 0,4 | 0,6 | 0,3 | 0,3 | 0,3 |
| civati | 0,7 | 0,2 | 0,2 | 0,2 | 0,7 | 0,8 | 0,8 | 0,8 | 0,8 | 1 | 0,4 | 0,4 | 0,4 | 0,4 | 0,4 | 0,3 | 0,5 | 0,2 | 0,3 | 0,3 |
| Maurizio_Lupi | 0,8 | 0,2 | 0,2 | 0,3 | 0,7 | 0,7 | 0,9 | 0,8 | 0,9 | 0,7 | 1 | 0,6 | 0,6 | 0,7 | 0,6 | 0,6 | 0,8 | 0,4 | 0,4 | 0,4 |
| angealfa | 0,8 | 0,2 | 0,2 | 0,3 | 0,7 | 0,7 | 0,9 | 0,8 | 0,9 | 0,7 | 0,7 | 1 | 0,5 | 0,6 | 0,6 | 0,5 | 0,7 | 0,4 | 0,4 | 0,4 |
| r_formigoni | 0,8 | 0,2 | 0,2 | 0,3 | 0,6 | 0,8 | 0,8 | 0,8 | 0,8 | 0,7 | 0,7 | 0,5 | 1 | 0,7 | 0,7 | 0,6 | 0,8 | 0,5 | 0,4 | 0,5 |
| mara_carfagna | 0,8 | 0,2 | 0,2 | 0,3 | 0,7 | 0,7 | 0,8 | 0,8 | 0,8 | 0,7 | 0,7 | 0,5 | 0,7 | 1 | 0,7 | 0,6 | 0,8 | 0,5 | 0,4 | 0,4 |
| renatobrunetta | 0,8 | 0,2 | 0,2 | 0,3 | 0,6 | 0,7 | 0,8 | 0,8 | 0,8 | 0,7 | 0,7 | 0,5 | 0,7 | 0,8 | 1 | 0,6 | 0,8 | 0,5 | 0,5 | 0,4 |
| DSantanche | 0,8 | 0,2 | 0,2 | 0,3 | 0,7 | 0,7 | 0,9 | 0,8 | 0,8 | 0,7 | 0,7 | 0,5 | 0,7 | 0,8 | 0,7 | 1 | 0,9 | 0,6 | 0,5 | 0,5 |
| GiorgiaMeloni | 0,8 | 0,2 | 0,2 | 0,3 | 0,7 | 0,7 | 0,8 | 0,8 | 0,8 | 0,7 | 0,6 | 0,5 | 0,6 | 0,7 | 0,6 | 0,6 | 1 | 0,4 | 0,4 | 0,4 |
| Ignazio_LaRussa | 0,8 | 0,2 | 0,2 | 0,3 | 0,7 | 0,7 | 0,8 | 0,8 | 0,8 | 0,7 | 0,8 | 0,5 | 0,7 | 0,8 | 0,8 | 0,8 | 0,9 | 1 | 0,6 | 0,6 |
| matteosalvinimi | 0,8 | 0,3 | 0,3 | 0,3 | 0,7 | 0,7 | 0,8 | 0,7 | 0,8 | 0,7 | 0,6 | 0,5 | 0,5 | 0,6 | 0,6 | 0,6 | 0,8 | 0,5 | 1 | 0,5 |
| RobertoMaroni_ | 0,8 | 0,2 | 0,2 | 0,3 | 0,7 | 0,7 | 0,8 | 0,8 | 0,8 | 0,7 | 0,7 | 0,5 | 0,7 | 0,7 | 0,7 | 0,6 | 0,8 | 0,6 | 0,6 | 1 |

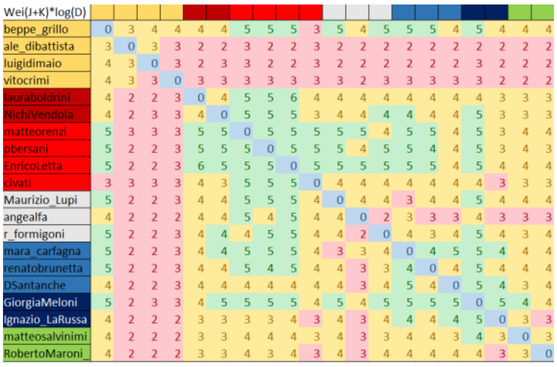
*Table 2 same table as Table 1, reweighted using the number of followers of columns only.*

take place, we would have very high indexes for people with few followers (if a has only 2 followers and b has 1.4 millions, it is extremely likely that $m_{ab}$ would be 1 if no correction was provided).

However, this table points out some interesting fact about the following network of some politicians. One can easily notice that there are two big groups of highly correlated politicians in Table 1, corresponding to the centre-left and the centre-right blocks roughly. Quite interestingly, Giorgia Meloni a far right politician, shares a wide part of her followers with politicians from all the spectrum. Surely, this might just be due to her (relative) popularity, as she is the 15th most followed politician, and the 9th biggest within this list; nonetheless, this is a politically interesting phenomenon.

Similarly surprising is the fact that *Movimento 5 Stelle* leader Beppe Grillo, who is known for his extremist positions, shares many followers with all representatives of the centre-left. This means that, being pivotal on the social media, he will attract many followers who do not share his political ideas, but who follow him because of his role in the political debate. Nonetheless, it is striking that he has a lower index with members of his own party than with very further actors as right-wing exponent Daniela Santanché. This is not due only to a problem of mismatch in sizes between the followers list, as she has an amount of followers comparable to that of the others M5S members. This figure also suggests that there might be a wide mass of people who follow all the most popular actors disregarding their political ideology. This also suggests that one should not be surprised if Grillo's estimated position is very close to that of centre-left politicians, as the two share a vast majority of followers.

To interpret better the dynamics, Table 2 represents an easier indicator, that is weighted on one dimension only. It represents the share of people who follow the actor in a row that also follow that in the column. Here we see, as expected, a different distribution of the coefficients. It emerges, for instance, that the pool of electors of M5S and centre-left is well connected. Not surprisingly, almost everyone who follows a medium-size politician, also follows the most popular ones (those corresponding the green lines, barely speaking). Hence, at least for top players, it appears that popularity plays a greater role than ideology. This is taken account of in the model, as it does not consider the relative closeness only. Therefore, correcting for *indegree* is crucial.



| Wei(J+K)*log(D) | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| beppe_grillo | 0 | 3 | 4 | 4 | 4 | 4 | 5 | 5 | 5 | 3 | 5 | 4 | 5 | 5 | 5 | 4 | 5 | 4 | 4 | 4 |
| ale_dibattista | 3 | 0 | 3 | 3 | 2 | 2 | 3 | 2 | 2 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| luigidimaio | 4 | 3 | 0 | 3 | 2 | 3 | 3 | 2 | 2 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 3 | 2 | 2 | 2 |
| vitocrimi | 4 | 3 | 3 | 0 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 2 | 3 | 3 | 3 | 3 | 3 | 2 | 2 | 2 |
| lauraboldrini | 4 | 2 | 2 | 3 | 0 | 4 | 5 | 5 | 6 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 3 | 3 | 3 |
| NichiVendola | 4 | 2 | 3 | 3 | 4 | 0 | 5 | 5 | 5 | 3 | 4 | 4 | 4 | 4 | 4 | 4 | 5 | 3 | 3 | 3 |
| matteorenzi | 5 | 3 | 3 | 3 | 5 | 5 | 0 | 5 | 5 | 5 | 5 | 5 | 4 | 5 | 5 | 4 | 5 | 3 | 4 | 4 |
| pbersani | 5 | 2 | 2 | 3 | 5 | 5 | 5 | 0 | 5 | 5 | 5 | 4 | 5 | 5 | 4 | 4 | 5 | 3 | 4 | 3 |
| EnricoLetta | 5 | 2 | 2 | 3 | 6 | 5 | 5 | 5 | 0 | 5 | 5 | 5 | 5 | 5 | 5 | 4 | 5 | 4 | 4 | 4 |
| civati | 3 | 3 | 3 | 3 | 4 | 3 | 5 | 5 | 5 | 0 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 3 | 3 | 3 |
| Maurizio_Lupi | 5 | 2 | 2 | 3 | 4 | 4 | 5 | 5 | 5 | 4 | 0 | 4 | 4 | 3 | 4 | 4 | 5 | 4 | 4 | 4 |
| angealfa | 4 | 2 | 2 | 2 | 4 | 4 | 5 | 4 | 5 | 4 | 4 | 0 | 2 | 3 | 3 | 3 | 4 | 3 | 3 | 3 |
| r_formigoni | 5 | 2 | 2 | 3 | 4 | 4 | 4 | 5 | 5 | 4 | 4 | 2 | 0 | 4 | 3 | 4 | 5 | 4 | 3 | 4 |
| mara_carfagna | 5 | 2 | 2 | 3 | 4 | 4 | 5 | 5 | 5 | 4 | 3 | 3 | 4 | 0 | 4 | 5 | 5 | 4 | 4 | 4 |
| renatobrunetta | 5 | 2 | 2 | 3 | 4 | 4 | 5 | 4 | 5 | 4 | 4 | 3 | 3 | 4 | 0 | 4 | 5 | 4 | 4 | 4 |
| DSantanche | 4 | 2 | 2 | 3 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 3 | 4 | 5 | 4 | 0 | 5 | 4 | 3 | 4 |
| GiorgiaMeloni | 5 | 2 | 3 | 3 | 4 | 5 | 5 | 5 | 5 | 4 | 5 | 4 | 5 | 5 | 5 | 5 | 0 | 5 | 4 | 4 |
| Ignazio_LaRussa | 4 | 2 | 2 | 2 | 3 | 3 | 3 | 3 | 4 | 3 | 4 | 3 | 4 | 4 | 4 | 4 | 5 | 0 | 3 | 3 |
| matteosalvinimi | 4 | 2 | 2 | 2 | 3 | 3 | 4 | 4 | 4 | 3 | 4 | 3 | 3 | 4 | 4 | 3 | 4 | 3 | 0 | 3 |
| RobertoMaroni_ | 4 | 2 | 2 | 2 | 3 | 3 | 4 | 3 | 4 | 3 | 4 | 3 | 4 | 4 | 4 | 4 | 4 | 3 | 3 | 0 |

*Table 3 same as Table 1, with log specification*

Trying a further refinement by setting

$$m_{jk} = \frac{2\sum_{n=1}^{N} e_j^n e_k^n}{\sum_{n=1}^{N} e_j^n + \sum_{n=1}^{N} e_k^n} \log\left(\left|\sum_{n=1}^{N} e_k^n - \sum_{n=1}^{N} e_j^n\right|\right)$$

As in Table 3 does not change much the final result. The most interesting pattern is that, correcting with size mismatch, there is a wider integration between centre-left and centre-right parties.

Generalising the measures used in Table 1 to the whole population, I would like now to check whether there is a tendency to cluster followers with affiliates to the same party. In specific, I now build a new indicator, composed by the mean of $m_{jk}\ where\ party(j) = party(k)$. This will tell us, on average, how similar the following structure of people of the same party is.

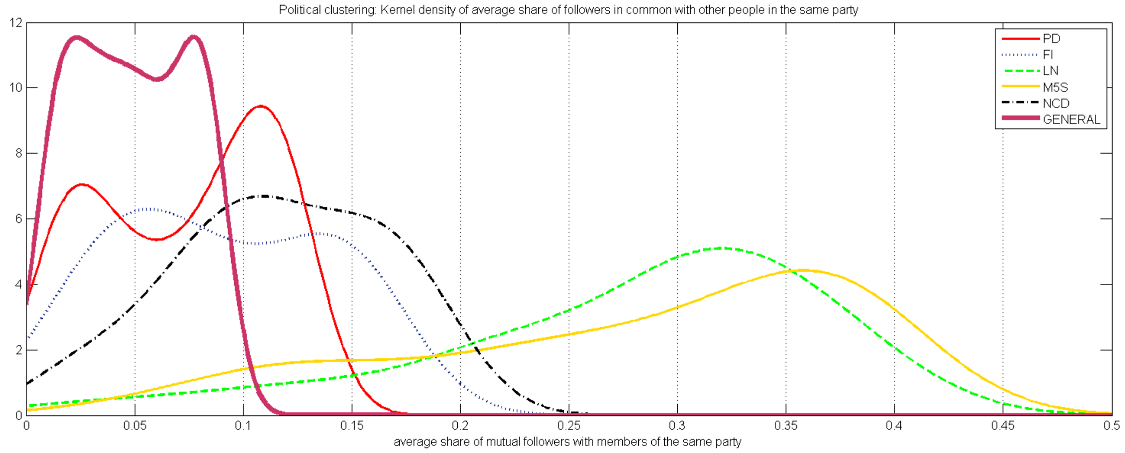


*Figure 9 Kernel density estimator of the average $m_{jk}$ coefficient where both politicians belong to the same party*

This graph confronts the density of this clustering coefficient. Given the baseline, i.e. the average concentration index with respect to the whole population, one can easily notice that by grouping for parties the concentration tends unequivocally to increase. This is in concordance with the assumptions that the estimation strategy makes. We can also point out that the three more central parties (PD-FI and NCD) are the less concentrated. In the case of PD, we have a overrepresentation of small local actors, as many majors are affiliated to this party: thus, we can expect that there will be many politicians uncorrelated to others because many of them are *too local* (e.g. you do not expect a great similarity in the network of majors of small towns very far one to another). However, even if we take into account this over representation of minor actors, it is clear that the two most extreme parties in the spectrum (*Lega Nord* and *Movimento 5 Stelle*) are also those who exhibit a wider clustering. This might suggest the presence of an echo chamber effect that could radicalise voters' opinions (or, on the other hand, it reflects that extremist voters are more likely to be selective in the choice of their newsfeeds). Notice that these graphs are not affected by the estimation of the parameters, as here I take the declared party affiliation, not the estimated one (even though the two pretty much coincide quite often).

Representing the network formed by creating directed edges between politicians where the share of followers of politician *i* that also follow politician *j* can give another visual insight about an eventual clusterization. In particular, if we were to expect a very heterogeneous distribution of followers, where users have in their list of political contacts politicians from all the spectrum, we would not observe any group formation between same party members. Conversely, if strong homophily was true, we would expect distinct groups of politicians of the same party or political ideology. Treating this kind of graph can be particularly untidy. Indeed, having 1004 selected politicians, we could expect as many as one million edges, whose representation would be intractable. Therefore, I restrict the graph to a subset of politicians who have at least 1000 followers (631 of them) and I plot edges only if they share at least 20% of followers. By colouring the nodes based on the estimated political ideology and highlighting groups with a Clauset-Newman-Moore procedure for cluster identification, I can attempt to represent loose political homophily. As it appears at first glance, however, the 20% criterion is not enough to let

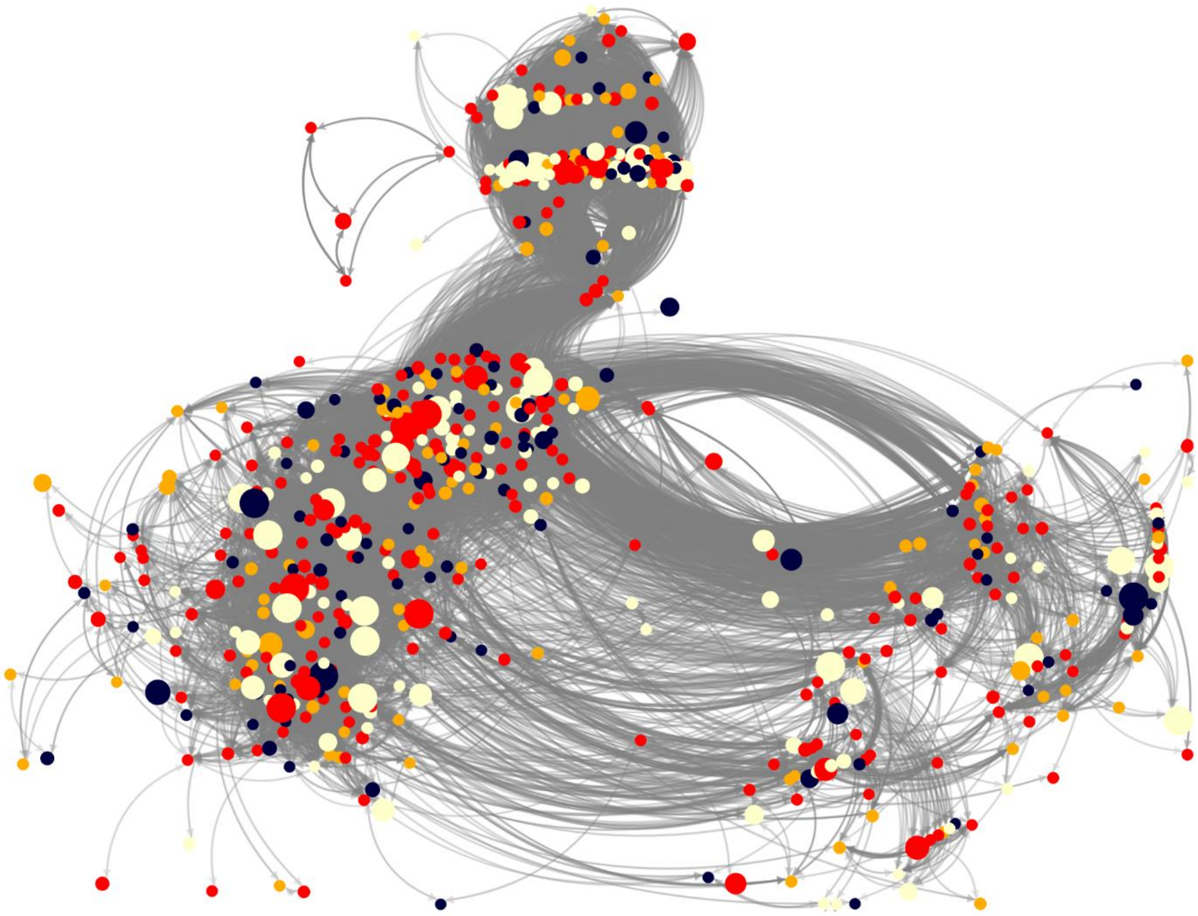


*Figure 10 network graph of loose homophily among followers. Nodes are politicians, colours represent their party affiliation as emerged from the estimation (red=centre-left, yellow=M5S, blue=centre-right, white=others). Edges are directed and they are show only if the origin shares at least 20% followers with the destination. Graph depicted with Harel-Koren algorithm, groups identified with CNM procedure and manually highlighted. 631 nodes and 50.812 edges depicted. No clear political cluster is identifiable. Thicker lines represent stronger homophily. Graph depicted with NodeXL, data elaborated with Matlab*

emerge any political clustering. All groups seem to be heterogeneous in party affiliation, and the high number of nodes (more than 50.000) does not allow to clearly depict anything interesting.

Nonetheless, by reducing the number of nodes and increasing the homophily threshold to 50%, a neat political division appears. This graph is highly unconnected, with 21 components for 249 nodes. The biggest component comprehends 90 nodes only, the second 78 and the third only 10. Therefore, the 50% threshold only selects a very small fragmented network between politicians. However, once nodes are coloured and links are also shown with colours who reflect the *follower* politician's political ideology, the political clustering is blatant. In particular, it is remarkable how few members of different parties are present in the same component. Indeed, the great group of M5S politicians is almost not connected to any other party. The only two political areas that are present in the same group or component are FI and PD, which represent the second biggest component in the southeastern part of the graph. This, again, reinforces the idea that electors of the most extreme party are less likely to follow more moderate candidates. Nonetheless, it is clear that also within FI and PD there is a tendency of grouping together, albeit in a less stringent way than with most extreme parties. With more sophisticated technological tools that could handle a wider amount of data, one could be able to further exploit the continuum of ideologies and try to represent and analyse quantitatively the tendency of clustering, to prove that it increases in political ideology, as suggested by these sketched figures. This would just require a powerful processor that can handle sophisticated procedures, but it can be feasible with the data already collected.

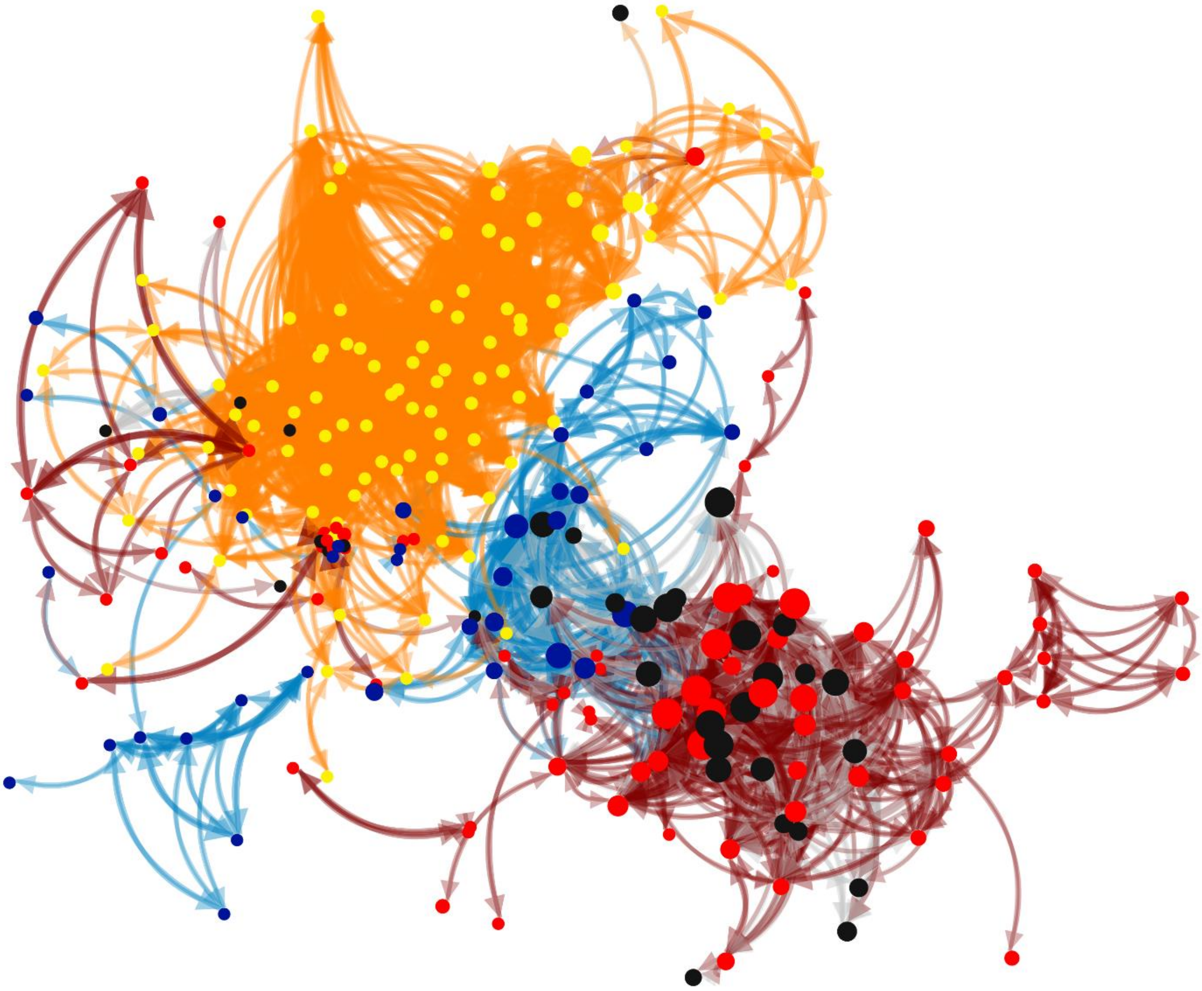


*Figure 11 network graph of strong homophily among followers. Nodes are politicians, colours represent their party affiliation as emerged from the estimation (red=centre-left, yellow=M5S, blue=centre-right, black=others). Edges are directed and they are show only if the origin shares at least 50% followers with the destination. Their colour reflects the origin's identity. Their size and transparency reflect the number of shared followers (darker and thicker = more common followers). Graph depicted with Harel-Koren algorithm, groups identified with CNM procedure and manually highlighted. 249 nodes and 5.498 edges depicted. Graph shows 21 components, strongly politically identified. Graph depicted with NodeXL, data elaborated with Matlab and Visual Basic.*

### Users' interactions

Let us now analyse the network created between users. To allow an easier visualisation of the problem, I restrict the dataset to 10.000 randomly chosen *informative* users. However, I dispose of the edge list for the whole 6 million population. By arbitrarily reducing the database, we will lose interesting properties about the connectedness of the whole population. This will artificially increase the average geodesic distance between two random users, as the *bridge* between them might exist but have been cut out from the graph because excluded from the subsample. On the other hand, the amount of data provided by the full adjacency matrix is in the order of $10^{13}$ relationships. This cannot be handled by a standard computer and would need a more powerful server. Indeed, also the $10{,}000 \times 10{,}000$ matrix I will be using is extremely heavy to be stored for the RAM of a common laptop. Let alone the technical problems, also the interpretation and visualisation of a network graph with 6 millions nodes will be highly intractable. One can, however, simplify the problem reducing the amount of data with an edge list, thanks to the relative rareness of connections.

For the following analysis I will use UCInet 6.0, NodeXL and Patek, according to the most fit program for perform the tasks I need.

Let us start with some descriptive statistics of the whole network, which cannot be represented in a comprehensible way having more than 200.000 edges.

| **Graph Type** | **Directed** |
|---|---|
| | |
| **Vertices** | 8778 |
| | |
| **Edges With Duplicates** | 219234 |
| **Total Edges** | 219234 |
| | |
| **Reciprocated Vertex Pair Ratio** | 0,084586615 |
| **Reciprocated Edge Ratio** | 0,155979456 |
| | |
| **Connected Components** | 1 |
| **Single-Vertex Connected Components** | 0 |
| **Maximum Vertices in a Connected Component** | 8778 |
| **Maximum Edges in a Connected Component** | 219234 |
| | |
| **Maximum Geodesic Distance (Diameter)** | 8 |
| **Average Geodesic Distance** | 3,272852 |
| | |
| **Graph Density** | 0,001422775 |

*Table 4 Descriptive statistics of the userFollowUser network, from NodeXL*

First, we can notice that out of ten thousands users, only 8778 follow or are followed by some other component of the subgroup. This is surely due to the low density in the link creation (0,001422775) but also to the fact that many individuals have a private account on Twitter, therefore the Tweetinvi API would result with an empty list of followers. This can be taken account of by a refinement of the data collection process, but it does not impact the network analysis.

The network exhibits a single giant component where all members are connected. Therefore, information can flow though the whole network, disregarding the political identity of users. The average geodesic is 3,27 that means that on average on can reach any other user within three steps. Thus, albeit not very dense, the connectedness of users is high. The diameter, i.e. the maximum distance between two nodes, is of eight steps. It is also interesting to notice that on average only 15% of following relationships are symmetrical.

We try to picture the full graph, hiding all members that are followed by less than 40 other users so to make more comprehensible the graph. What we notice at first glance is the strong interconnectedness of users. It is not spottable any area isolated or pseudo isolated member. This graph is depicted with a

Harel-Koren algorithm that by definition tries to separate different communities. Thus, I proceed with a cluster analysis through the implementation of a Clauset-Newman-Moore (2004, 2008) approach to find groups. This allows to detect main communities in large networks and visualize them graphically. 31 groups are hence identified, but the three biggest encompass almost the whole population, and are the only who own some actor with *indegree* higher than 40. Among these actors, we try to distinguish them in terms of betweenness centrality, where bigger dots have higher values, and eigenvector centrality, where less transparent dots have higher values. However, the sample is pretty homogenous (this is surely due to the fact that we only select the most connected guys by definition, excluding smaller ones to make the graph possible to plot).

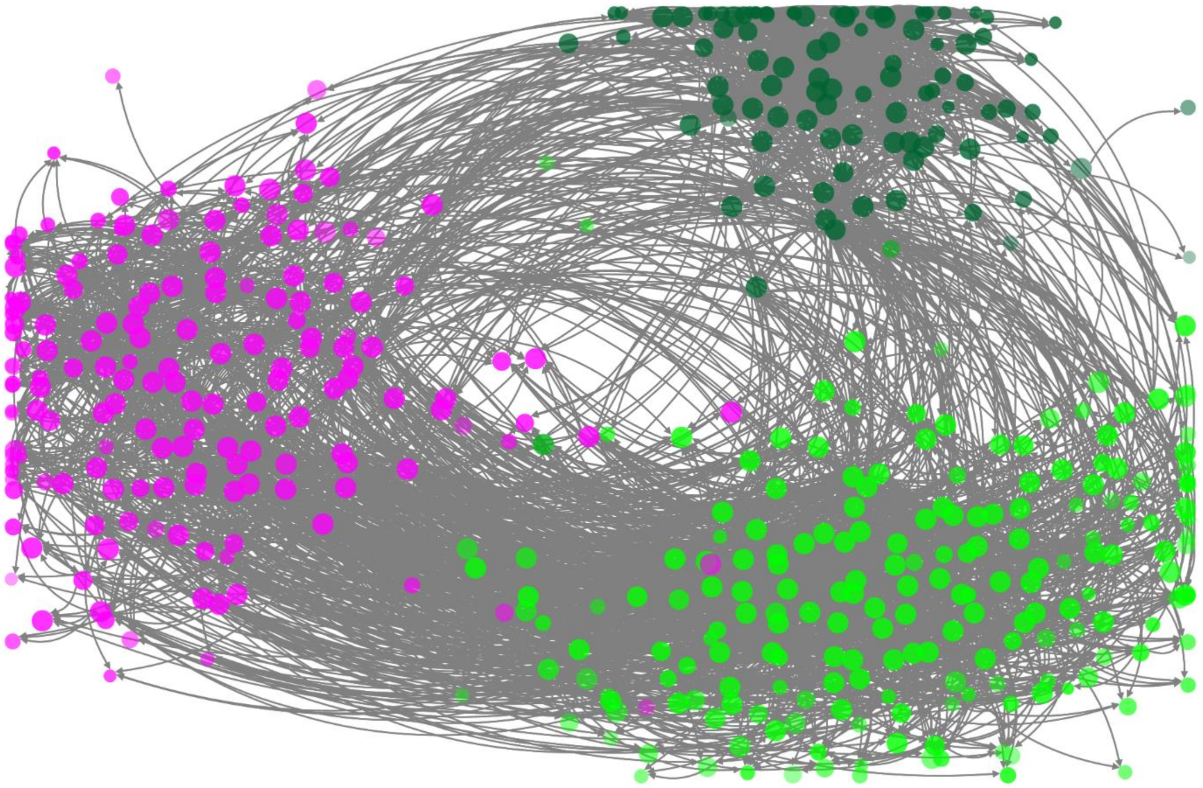
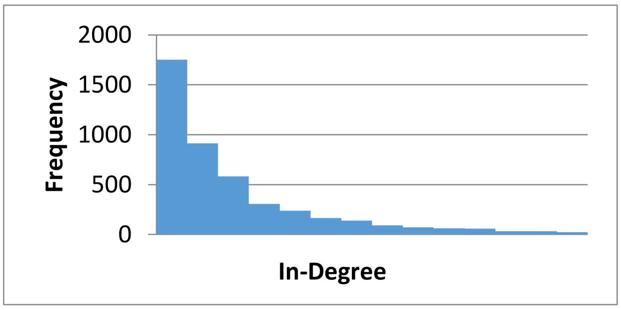
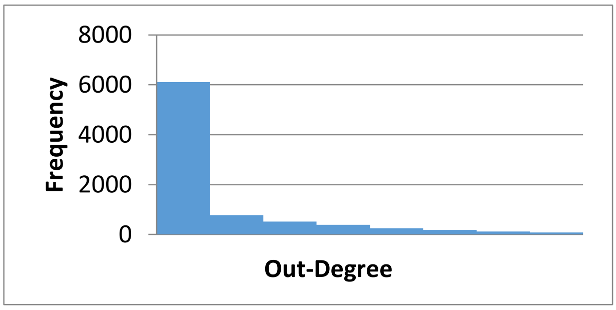


*Figure 12 Network graph with Harel-Koren algorithm, modified to highlight the main groups. Only indegree>40 showed. Colors represent main groups, transparence reflects eigencentrality and size betweenness centrality. NodeXL.*

Here follow some other descriptive statistics about the network:

| | | | |
|---|---|---|---|
| Maximum In-Degree | 219 | Maximum Out-Degree | 388 |
| Average In-Degree | 12,488 | Average Out-Degree | 12,488 |
| Median In-Degree | 6,000 | Median Out-Degree | 0,000 |





| | | | |
|---|---|---|---|
| Max Betweenness Centrality | 1885491,114 | Max Eigenvector Centrality | 2,582e$^{-3}$ |
| Avg Betweenness Centrality | 19952,094 | Average Eigenvector Centrality | 0,114e$^{-3}$ |
| Median Betweenness C. | 2240,677 | Median Eigenvector Centrality | 0,046e$^{-3}$ |

| | |
|---|---|
| Maximum Clustering Coefficient | 1,000 |
| Average Clustering Coefficient | 0,079 |
| Median Clustering Coefficient | 0,042 |

It is striking the low level of clustering coefficient, as a priori one would have expected a greater level in homophily among users. The average number of connections is roughly around 12.5, which is plausible considering that we are using a subset of the informative users subset. Therefore, considered that users in the subgroup are screened by having at least 150 followers, the fact that only a dozen of them lies within this matrix tells us about how limited is the analysis of the partial network.

It is, now, interesting to work integrating the network data with the politicization indexes computed before. We want to see if people of the same party exhibit characteristics that are different from the whole population. I will use the same way of labelling users from the same area as I used in the previous sections.

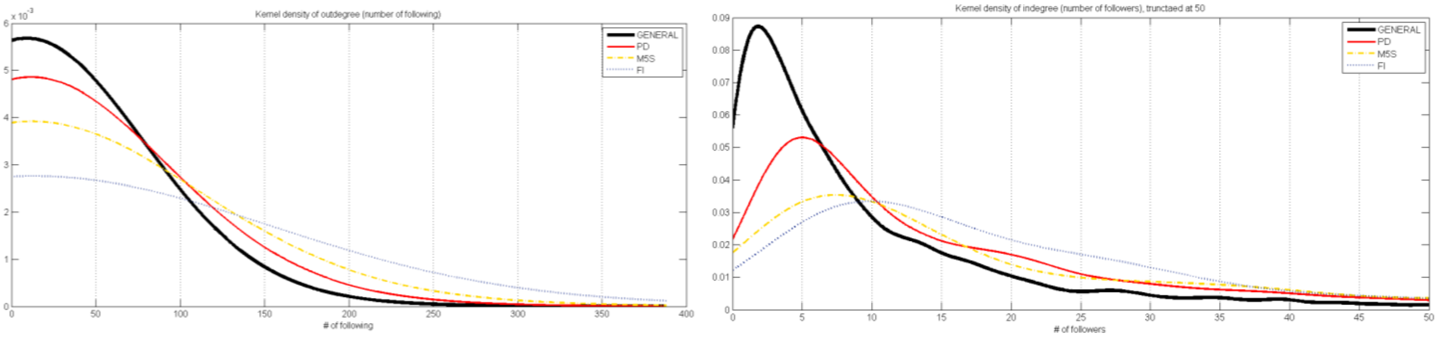


*Figure 13.1-2 kernel density of outdegree (11.1) and indegree (11.2) per party Vs. general population*

These two graphs represent the *outdegree* and the *indegree* level of users. It is surprising to notice that, despite the total number of *followers* and *following* (i.e. the sum of all indegree and outdegree) should be equal, the distribution of the two are radically different. While the *following* relationship sees a well-dispersed distribution, meaning that users tend to be more or less active in their political interests, the distribution of the number of followers is much more concentrated in the first ten percentiles. This reflects the fact that few, popular, users will attract a wide amount of followers within the matrix. In terms of

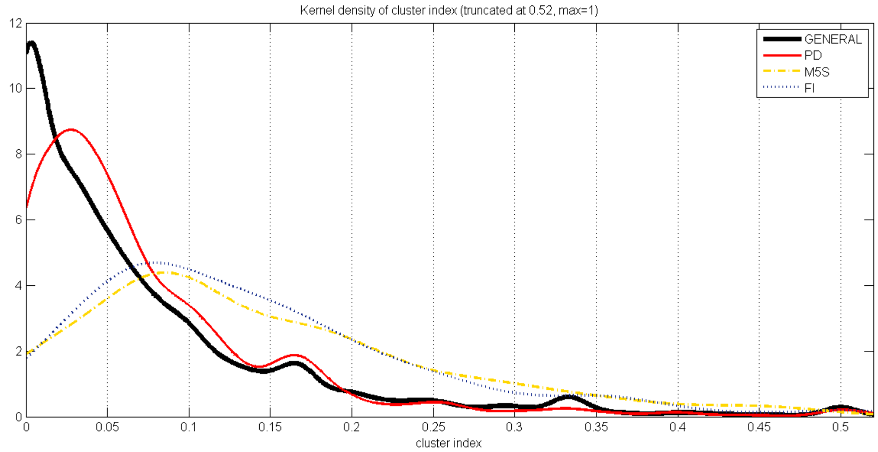


*Figure 14 Kernel density of the cluster index per party Vs. general population*

standard deviation, the *indegree* has a std of 19.11, whereas the *outdegree* hits 26.73. The asymmetry between these relationships is a fascinating phenomenon that, however, does not seem to tell an insightful story about different parties' structure. If anything, it can be noticed that the centre-right party tends to have more active users both in terms of following and followers.

More intriguing is the analysis of the distribution of the cluster index. Indeed, we notice there a clear different trend between the two most extreme parties (M5S and FI) and the most central one (PD). PD voters exhibit a trend closer to the whole of the population, sensibly lower than that of the other parties. This give a hint on the fact that more radical party also have a higher possibility of echo chambers. However, this graph does not tell about political preferences of followed people, but it barely reflects a higher presence of clusters among the *extremist* population.

Therefore, I have selected a random sample of one hundred users for each of the parties and plotted their followers/following relationships. This will allow a neater representation that allows to identify better what we have discussed though the kernel density analysis. The graph is obtained with a Harel-Koren algorithm, allowing for curvy edges. Arrows point in the direction of the *followed* actor. One can count as many as 1710 relationships, 45% of whom are reciprocal. This is already three times bigger that in the whole population, suggesting that highly politicized actors, i.e. those whose estimator falls within a party's boundaries, might tend to stick more together. The density, indeed, grows from 0,0014 in the general population to 0,0237 here. In particular, we over represent extremists, as we now have selected 200 persons from the tails of the distribution. However, the average geodesic remains around 3.17, meaning that the increased number of relationships is not evenly distributed among the sample. The distribution of *outdegrees* and *indegrees* changes sharply, too, with a great decrease of the values around zero.

Some words has to be spent about the use of colours, arrows and shapes. Colours reflect the party affiliation; the size of the marker is, instead, the *betweenness* indicator; darker, less transparent, markers are those with higher eingecentrality and squared markers are for those with a clustering index above 0.18, the median of the sample. The groups identification is done through a Wakita-Tsurumi (2007) procedure. Three main groups are identified, and their presence is straightforward in the graph. A fourth, smaller one, is represented by those dots between the big component and the two smaller ones (half of the picture, roughly). I use for this graph people from the whole centre-right spectrum, not only strictly FI, to have a more comparable average absolute value of radicalisation with respect to *Movimento Cinque Stelle*.

It is particularly stunning the high political clusterisation that emerges mainly for electors of FI and M5S. Indeed, at a first glance, one can see that right-wing voters are highly interdependent and almost cover the totality of the two smallest major groups. Similarly, M5S electors are strongly connected one another and surrounded by PD ones, who act as a bridge between the groups. The centrality of this party acts as a connection between the right wing and M5S who would not have otherwise any contact. Nonetheless, PD electors do not all have a high *betweenness* coefficient. While many of them do, especially those connecting the groups, the pivotal groups seems to be the smallest FI one in the upper

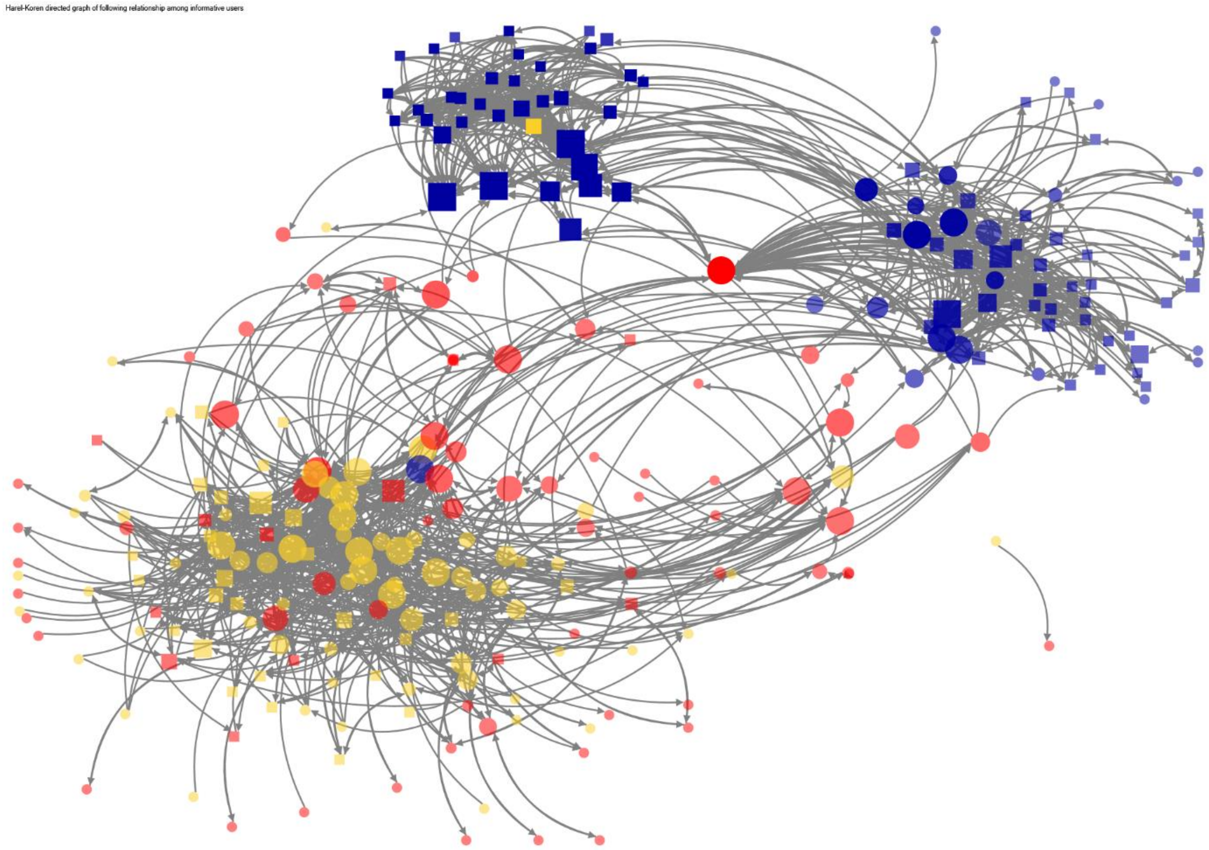

Harel-Koren directed graph of following relationship among informative users

*Figure 15 network of 100 partisan users for each of the main parties, groups are manually highlighted after the algorithmic identification. Colours represent parties (yellow=M5S, red= PD, blue=FI); size of the marker is betweenness index, transparency eigencentrality; squares represent the most clustered actors. Arrows point towards followed users. Graph obtained with NodeXL*

part of the graph. This small, highly clustered, group detains many key players in terms of *betweenness* as well as those with a highest eigencentrality. The most central player, however, is the PD voter that spots in the middle of the graph, as he has the highest values for most indicators of centrality. It is interesting to notice that he, contrary to most other pivotal players, has very few *outdegrees* and many *indegrees*. Therefore, he seems not to receive as many information as what he shares. This kind of players are the most important in models of convergence of beliefs, as they become the *trendsetters*: they do not change their opinion much, but they influence a lot of people, while staying in the middle of the population.

Selecting other random samples, I had qualitatively similar graphs. Hence, this peculiar relationship between parties seems to hold. More extreme people tend to be the most selective in their relationships, clustering more with similar actors, while centre ones are more diverse in taste. However, the connectedness between the two parties on the left of the political spectrum is higher than with right-wing voters. I would like to point out that these results are insightful about the Twitter debate, which is heavily important in Italian politics, but far from being representative of the whole population. In particular, a caveat comes from the very heterogeneous electorate of M5S: analysis of electoral fluxes after the elections pointed out that a good share of voter of the party come from extreme and moderate left, but a considerable and increasing amount of them is made by right-wing anti-system voters. Although the ideological and functional position of the M5S is radically at the opposite of that of *Forza Italia*, one cannot fail to see a high, if not higher, attempt to keep the distances also from the centre-left. This analysis, thus, does not reflect a political closeness between the two parties, but instead a relative degree of clustering in the electorates, at least on Twitter.

It is also important to discuss the role of the estimation strategy that we have adopted to identify voters of each party. Indeed, the model relies on the structure, so clustering and the other indicators of the shape of the network will influence the estimators. However, I have used only a matrix of the type *user follows politician*, so the relationships between users are not taken into account during the MCMC procedure. Therefore, if one could argue that it is not the case that more extreme users are more clustered but the other way round in the politician analysis, now this argument does no longer hold. These is no a priori reason, then, for these clusters to emerge, if not because of an underlying tendency of users to prefer people with closer ideology. This result is in line with a wide literature, as discussed in the literature review, but it is one of the few, to the best of my knowledge, to perform an analysis at the user level, while most papers focus on *user follows politician* approaches.

As Barberà (2015) points out, clustering algorithm could bias the results into finding an excessive polarisation, due only to the mathematical process behind the identification of groups. Therefore, I try also other approaches to assess the level of homophily of users.

Thus, I would like to explore further the results obtained at the network analysis level, with a different apporach. Let me build now another indicator of homophily, provided by the share of followed actors who are identified within the same party, as in Curarrini et al. (2009).

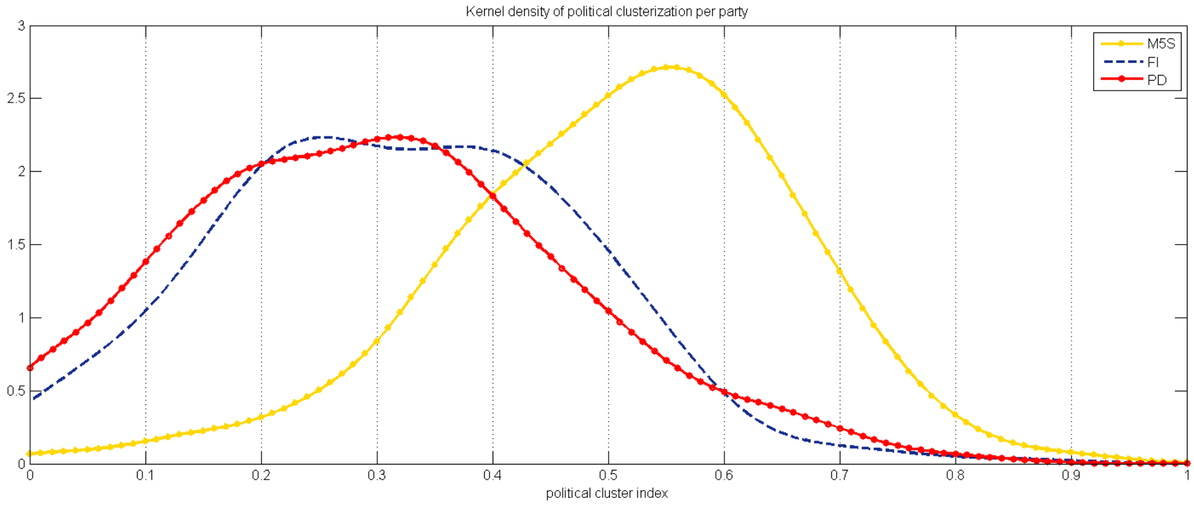


*Figure 16 political homophily: share of followed users who are of the same party as the follower, kernel density for the three main parties*

This graph depicts an interesting phenomenon. Here the definition of *clustering* is not what usually intended in network literature, but it is more a synonym of political homogeneity. Now *Forza Italia* results to be less clustered as one would have guessed with the network graph. Nonetheless, it is remarkable that the right wing spectrum is more fractionalised in different parties. In the network graph, I consider everyone on the right, while here I restrict the sample to those within one standard deviation from the mean FI parliamentary. Moreover, this graph measures political *homophily* which can be correlated with clustering, but does not have to.

However, the striking evidence coming from this graph is the high homophily index of M5S. This was already an intuition when, seeing the distribution of users, I pointed out a small bulge of people around the extreme-left values. However, I had not information at the time about their interconnectedness. Here I show that more than half of M5S voters has more of half of his friends with the same ideology. Even more puzzling, very few of them do not have friends of the same party, at a very significantly different rate from other politicised actors.

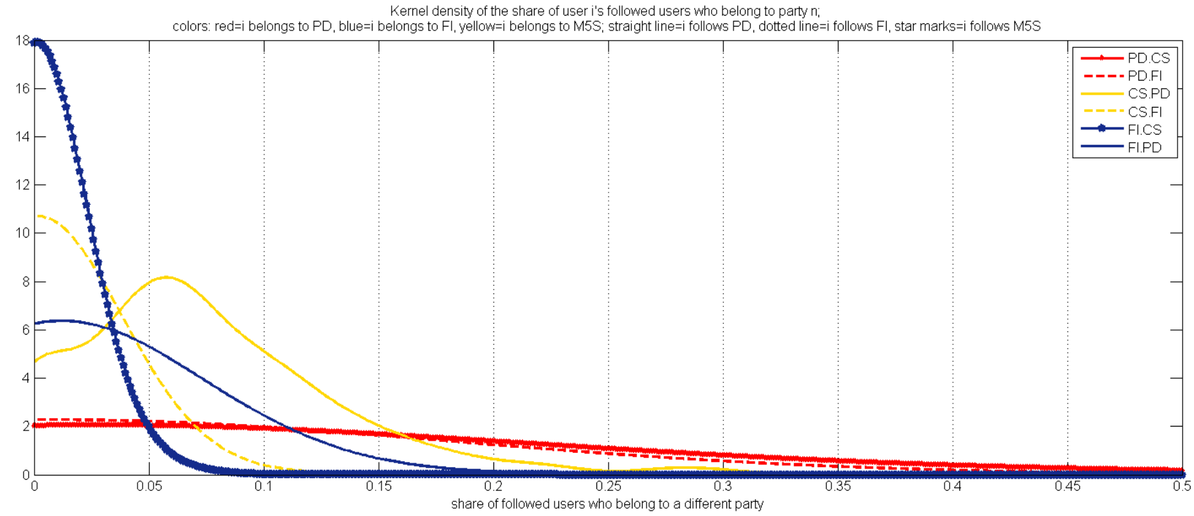


*Figure 17 share of followed users that belong to a party different from that of the follower. Colours represent follower's party, dotted line represent followed=FI, star-marked=M5S, plain=PD*

To check this insight, I built also a graph representing the political heterogeneity of one's network. In particular, I have computed for each politicised user the share of followed people who belong to the other two parties. The results that are depicted in the figure confirms the insight brought from the analysis of the previous graphs. The PD users are those who tend to have more diversified platforms, with a considerable amount of people having a great share of differently polarised users. Remarkably, this heterogeneity does not distinguish the two other parties. PD users confirm again their median role between the other actors.

The highest rate of zero matches is, instead, given by FI-M5S and M5S-FI, confirming again the lack of communication among them in the network. M5S voters have a discretely higher rate of followed accounts from centre-left with respect to FI. While PD voters follow people from all the distribution, they are not followed back as much as they do. Nevertheless, it emerges once more the relative closeness with M5S, albeit from the political indicator PD is equidistant from the centre-right and M5S.

### TWEETS AND DIFFUSION OF CONTENTS

Let us now analyse the political communication and try to assess the differences between the different parties. To do so, I have downloaded a database of over 250.000 tweets sent by major politicians during the last months. Due to time limits to retrieve the information, I could not expand the database as much as I wanted to, including also smaller actors. Indeed, with the database I have constructed, it would be possible to trace almost the whole history of political communication in Italy up to the micro level, as we have all data about regional and municipal authorities. However, the Twitter API query limits do not allow such a research in feasible times without having more computers downloading the data at once. Indeed, with the right amount of data, one could geo-localize the research, stratify it for different ages and sex and get a full roadmap of the determinants of content diffusion. However, for the time being, we can have a more modest attempt to look for major areas of discussion and see how they interact with political ideology, as estimated by the model.

In particular, I focus on some of the most frequent words, using a SQL procedure to count and list the most repeated entries. To avoid content injection, I focus on words with a single meaning, looking for all derivations (in Italian, for instance, masculine and feminine adjectives differ) and words that have the same semantical root. This allows to identify some commonly used words and compare the choice of wording and thematic from different parties. With a wider dataset, one could exploit the fact that all politicians are on a continuum of political identity, however here I restrict the analysis into 11 categories, provided by 0.5 increments in $\varphi_j$ starting from -3.

|  | Key Words (ITA) | M5S |  |  | PD |  |  | NCD-FI |  |  | FI-LEGA |  | Key Words (ENG) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  |  | -3 | -2,5 | -2 | -1,5 | -1 | -0,5 | 0 | 0,5 | 1 | 1,5 | 2 |  |
| corruption | casta | **0%** | **0%** | 0% | **0%** | 0% | 0% | 0% | 0% | 0% | 0% | 0% | caste |
|  | corrott* | **4%** | **3%** | 0% | 0% | 1% | 1% | 1% | **1%** | 1% | 0% | 1% | corrupted |
|  | dimissioni | **1%** | 1% | **2%** | **1%** | 0% | 1% | 0% | 1% | 0% | 0% | 0% | resign |
|  | scandalo | **1%** | **1%** | 0% | **0%** | 0% | 0% | 0% | 0% | 0% | 0% | 0% | scandal |
|  | sprec* | 0% | 0% | 0% | 0% | 0% | 0% | **1%** | **1%** | 0% | 0% | **1%** | waste |
|  | vergogna | **1%** | 0% | 1% | 1% | **1%** | 0% | 0% | 0% | 1% | 1% | **1%** | shame |
| negative | basta | 1% | 1% | **3%** | 1% | 2% | 2% | 2% | 2% | 2% | **2%** | **5%** | enough |
|  | contro | 5% | 3% | 6% | 5% | **9%** | 5% | 6% | **6%** | 6% | 5% | **9%** | against |
|  | senza | 3% | 3% | 4% | 3% | 4% | 4% | 4% | **4%** | **4%** | 4% | **8%** | without |
| positive | grazie | 4% | 3% | 7% | 3% | **11%** | 9% | **10%** | **14%** | 5% | 7% | 5% | thanks |
|  | proposta | **5%** | 3% | **3%** | **6%** | 3% | 3% | 3% | 2% | 2% | 3% | 0% | proposal |
| economics | crisi | 0% | 0% | 1% | 0% | 1% | **2%** | **3%** | **3%** | 2% | 1% | 2% | crisis |
|  | IMU+TASI+tasse | 2% | 2% | **6%** | 2% | 1% | 1% | 2% | 2% | 4% | **4%** | **5%** | taxes (name of some) |
|  | impresa | **0%** | **0%** | **0%** | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | firm |
|  | lavoro | 3% | 2% | 7% | 7% | 6% | **10%** | **12%** | 6% | 4% | 6% | **8%** | jobs |
|  | ripresa | 0% | 0% | 0% | 0% | 0% | 0% | **0%** | 0% | **0%** | 0% | **1%** | growth |
| institutions | cittadin* | 7% | **12%** | 6% | **11%** | 3% | 4% | 4% | 2% | 3% | 2% | **8%** | citizens |
|  | governo | 8% | 6% | 7% | 9% | 8% | 8% | 8% | **9%** | **10%** | 9% | **16%** | government |
|  | legge | 5% | **9%** | 2% | **9%** | 5% | 5% | 4% | **6%** | 4% | 4% | 2% | law |
|  | politica | 2% | 3% | 6% | 4% | 6% | 6% | **6%** | **6%** | **7%** | 5% | 2% | politics |
|  | presidente | 4% | **7%** | **6%** | 5% | 3% | 5% | 5% | 4% | 4% | 4% | **7%** | president |
| leaders | berlusconi | 3% | 3% | **12%** | 5% | 1% | 2% | 2% | 3% | **8%** | **11%** | 0% | berlusconi |
|  | grillo | **2%** | 1% | **2%** | **4%** | 1% | 2% | 1% | 2% | 1% | 1% | 0% | grillo |
|  | pd | **19%** | **15%** | 6% | 9% | 9% | **13%** | 8% | 4% | 4% | 4% | 1% | pd |
|  | renzi | 9% | **10%** | 2% | 7% | 9% | 4% | 3% | 3% | **12%** | **10%** | 1% | renzi |
| EU | euro | **4%** | **4%** | 3% | 3% | 1% | 2% | 3% | 3% | 2% | 1% | **5%** | euro |
|  | europa | 1% | 1% | 2% | 1% | 3% | **5%** | **6%** | **5%** | 3% | 4% | 1% | europe |
| nationalism | clandestin* | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | **1%** | **0%** | **0%** | illegal immigrant |
|  | immigrati | 0% | 0% | 0% | 0% | **2%** | 1% | 0% | 1% | **3%** | 2% | **5%** | immigrants |
|  | italia | 4% | 5% | 6% | 2% | 6% | 5% | 5% | **6%** | **6%** | **8%** | 5% | italy |
|  | marò | 0% | 0% | 0% | 0% | **4%** | 0% | 0% | 0% | **1%** | **1%** | 0% | marines |

*Table 5 list of key words grouped by semantical cluster. For each, the share of occurrences within the whole list of key words is shown. Green coloured cells represent, for each column, the five most used words; yellow coloured are the following top 5. In bold, for each row, the three political categories who used the word the most.*

This table reports what a word *n*'s share of a certain type of user's key words. Words are grouped for semantical similarities and politicians are grouped for party affiliation (very roughly). Green elements represent, for each column, the 5 most frequent words, while yellow one are the following 5 most repeated. In each row, bold numbers represent the category who use most frequently the corresponding word. I use a percentage measure because, obviously, the relative size of the categories differs sensibly. For instance, in the -0.5/0 range we have 23.656 repetition of key words, in 1.5/2 15.682 but in -1.5/-1 only 406. This is due to the uneven distribution of politicians in the scale.

The most used words are undoubtedly those in the *institution* category, followed by *leaders*. These are very neutral words, and it is thus comprehensible that they attract the most users. However, it is interesting that the words *Renzi* (PD's secretary and Italian Prime Minister) and *PD* are more used by M5S politicians than by centre-left ones. Conversely, *Grillo*, M5S's leader, is more spoken of among PD members. On the other hand, *Berlusconi* is mostly quoted on the centre-right.

Categories that seem to be fairly unpopular, as the *corruption* one, show still that this vocabulary belongs more to M5S than to other parties. This is not beyond belief, as the party has always had an history of criticising illegal activities. The two most *emotional* categories, *negative* and *positive*, are particularly common within the extreme right (the first) and centre-right (the second), with *grazie* (*thank you*) being one of the most used words in that part of the spectrum.

Astonishingly, words related to the economics are not particularly popular. Only *lavoro* (jobs/work/employment) hits many repetition, especially for parties in the government at the moment (PD and NCD), while *tasse* (here I aggregate the data about the word that means *taxes* and the name of some tax-related vocabulary i.e. *IMU, TASI, Equitalia*) is more common among people in the minority.

Europe-related words are also pretty low in volume, even though in the dataset we have also the period of last European elections (May 2014). A caveat come here because these words have often been manipulated or used mostly in hashtags, and thus not detected by the algorithm I used. For instance, #noEuro is a clearly Europe-related word, but it would not be detected as different from the word *euro* alone. More sophisticated version of the same algorithm could, however, provide more complete results.

Finally, nationalism results to be clearly a domain of the centre-right and extreme right. Particularly interesting is the case of the word *marò* that is almost used by right-people only (even though some high percentage is exhibited in the lowly populated -1/-0.5 interval). This word is related to two Italian marines actually hold in prison in India and that became a symbol for right-wing parties. Almost half of the tweets containing this word in the dataset belong to the same politician, Giorgia Meloni (*Fratelli d'Italia*).

Aggregating the query into the macro-areas can help to visualise better the communication pattern. Indeed, it points out that the two main parties at the opposition, M5S and FI (NCD is so small that it is almost completely absorbed by FI when aggregated), are also the most polarised in terms of political debate, i.e. they focus on some core area while discussing a few about others. The table depicts in darker colours the party that talks more about a certain subject, while in light one that who does it less.

| | **M5S** | **PD** | **NCD-FI** | **FI-LEGA** |
|---|---|---|---|---|
| **Corr** | **7%** | 3% | 3% | 2% |
| **Neg** | 9% | 11% | **12%** | 11% |
| **Pos** | 8% | **12%** | 11% | 10% |
| **Eco** | 6% | 13% | **14%** | 12% |
| **Inst** | **29%** | 28% | 27% | 24% |
| **Lead** | **31%** | 21% | 18% | 25% |
| **EU** | 5% | 6% | **7%** | 5% |
| **Nat** | 4% | 6% | 8% | **11%** |

*Table 6 percentage of keywords belonging to a particular field aggregated by party-area estimated from the MCMC model. Dark-highlighted cells represent, for each row, the party who adopted more that language. Light-highlighted ones are for those who did it less.*

These two minority parties seem to compensate one another: the most debated topics of the first are the least of the latter and viceversa. This difference is not very pronounced, however. But it still could tell an interesting story about how parties at the two ends of the majority government tend to polarise their choice of topics, focusing on some political empty niche or attacking the government from different angles. M5S targets a lot more about corruption and institutions that the right, while FI will focus on economics and Europe, being a more business-oriented party. The extremely high value of M5S in the *Leaders* section also denotes the strong personalisation of this party's strategy. Almost on fifth of their key words is represented by the word *PD*, which is object of a continuous attack.

Conversely, the party at the government seems to have a more fairly distributed share of key words, denoting a more institutionalised communication as well as the attempt to defend himself from the critics arriving from both sides. With the whole panel of data that could be retrieved from Twitter one could also see the trend in time of communication strategies. For instance, one could check whether parties in positions of government (crossing national with also regional and local data) change their wording and the subjects of their communication so to adapt to their new role. Moreover, a detailed study of how words acquire a political meaning can be interesting. New terms often appear to denote older phenomena. Often a sort of re-branding of complex issues is done by making the word that defines it with a greater or a smaller political and emotional connotation. The way language evolves and is shaped by political actors can now be followed up to the micro-level, using the interconnected data of users, with their political ideology estimators.

This content analysis can, indeed, only give a flavour of what one could do with more advanced hardware resources, with whom circumvent the API query limits. Given the list of users, it is possible to map their language to their followed politicians and to those in the same part of the spectrum, so to eventually find a correspondence between the language used by candidates and that of followers. This semantic analysis of political debate could be extended so to see which words can trigger the most effective impact. In particular, we can focus on users with high eigencentrality within the users' network, to find the best way to elicit them to share a candidate's tweet to their audience. This algorithmic approach can potentially implement a micro-based political marketing campaign that can identify the key players, given their position, their ideology and that of their followers. Moreover, one could see what topics can provoke a higher contagion through those people who are less politicised but connected with some politicised actor.

### FURTHER DEVELOPMENTS

The results provided so far in this research are promising but limited. Where many hints of a correlation between polarisation and homophily were found, nothing systematic can be claimed about the causal relationship between the two. Exploiting the peculiarity on the Italian political system that, contrary to the American one, exhibits a great party fragmentation and the institutionalisation of extremist movements, I could identify an increasing degree of clustering between users as the political dimension diverges from the centre. However, this is just a snapshot, whereas a complete research would need a video.

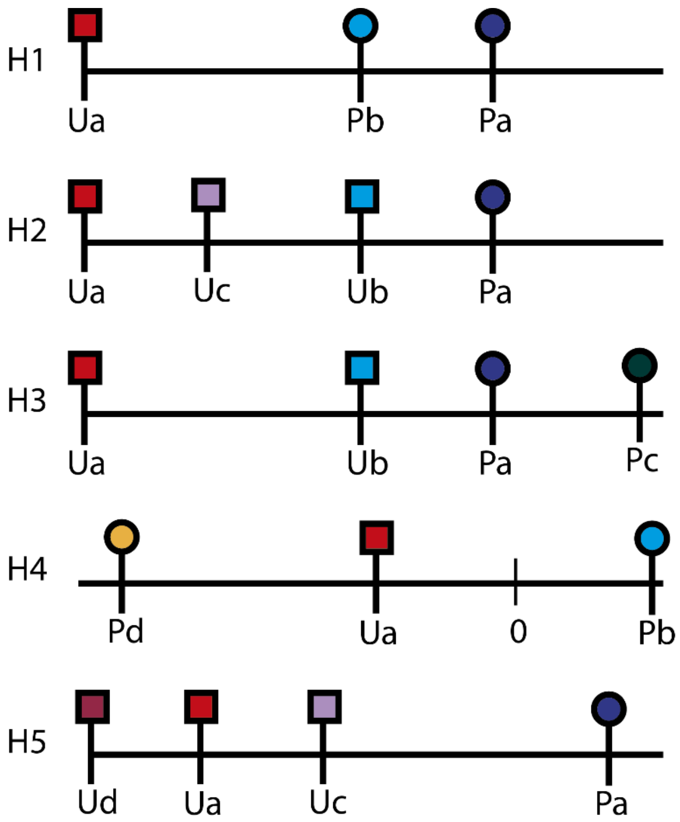
Therefore, I plan to extend my research to a dynamic setting, by creating a panel data about the following relationships. Unfortunately, Twitter does not allow to trace back to the date of creation of the link, therefore the only way to do it is to scan the existing network for a long time and use differences to trace changes in the structure. This way, I could compute the estimation of political ideology for users and politicians through time, tracking any sign of convergence or centrifugal forces. Nonetheless, I fear that the method used for the estimation could be not sensitive enough to capture slight changes. Thus, it can be a starting point but it has to be used in a cautious way.

Once a good measure of political ideology is established, finding a good way to catch the link between echo-chambers and radicalisation is still challenging. One could exploit exogenous sources of radicalisation or in heterogeneity to see what is the chain of reaction of the network. If, for instance, after a shocking event as Charlie Hebdo, one could have an upraise in extremisms, controlling if this also fosters an increase in segregation could reinforce the idea of selective choice of media being more intense in magnitude with radicalisation. However, the way to capture how the exposition to same-minded could increase one's extremisms remains murky.

Hence, I would like to dedicate more attention to the retweets chain. Again, once a panel data is constructed, I would focus on the determinants of information propagation. Given that friends' retweets are shown to increase the likelihood of one's retweeting by a large amount of studies, to the best of my knowledge nobody has focused on the political closeness of these intermediary retweeters between the politician and the user. In particular, I would like to test the intuition that friends can act as a foot-in-the-door technique. My hypothesis is that the same message has a different value according to who express it and, in particular, how ideologically far the actors are.

This little diagram can help to sketch the hypothesis I would like to test with the suitable data. The line represents political spectrum, squares are users, dots are politicians and the retweeting of a politicians tweet takes the notation of $RT_{U_i}^{P_j}$

$$H1: P\left(RT_{U_a}^{P_a}\right) < P\left(RT_{U_a}^{P_b}\right)$$



This is a baseline, easy to test, hypothesis that simply implies that, ceteribus paribus, the likelihood of retweeting someone who is ideologically closer to the user is higher than that of further actors. Once one can control for factors as the popularity of the politician, *tweet-specific* characteristics (e.g. if that tweet is linked to some particular event or has a not politically charged content), it is possible to assess whether ideological closeness fosters endorsement. This would be an additional evidence of selective exposure: not only people follow less extremists, but even when they do follow heterogeneous politicians, they are less likely to interact and endorse those who are further.

$$H2.1\colon P\left(RT_{U_a}^{P_a}|RT_{U_{i\neq a}}^{P_a}\right) > P\left(RT_{U_a}^{P_a}\right)$$

$$H2.2\colon P\left(RT_{U_a}^{P_a}|RT_{U_c}^{P_a}\right) > P\left(RT_{U_a}^{P_a}|RT_{U_b}^{P_a}\right)$$

These two hypothesis start to investigate the role of *user* friends into message diffusion. The first one says that the probability of retweeting is higher if a followed user who is not a politician has already retweeted the message. Again, this has to be cleaned by all noise factors, but the baseline idea is that receiving the message also from another source is *in se* a factor that can increase the likelihood to share the message. Moreover, *H2.2* exploits the ideology estimators to state whether, given a certain distance between the politician and user $U_a$, having a closer friend who retweets the message increases the likelihood of retweeting. This would go in the direction of the *food-in-the-door* theory I was discussing above. If a person who is close to me shares a message, I trust him more and are more tempted to adopt his ideas. Surely, here correcting for unobservable is even more insidious. The problem of endogeneity is incredibly high. However, I am confident that with the right amount of micro-level data one can find ways to clean the effect so to assess the targeted mechanism only. In particular, we can exploit geographical position as a source of non-politicised heterogeneity, trying to find a solid way to identify the effects of closeness of diffusion.

$$H3\colon P\left(RT_{U_a}^{P_a}|RT_{U_b}^{P_a}\right) - P\left(RT_{U_a}^{P_a}\right) < P\left(RT_{U_a}^{P_c}|RT_{U_b}^{P_c}\right) - P\left(RT_{U_a}^{P_c}\right)$$

This hypothesis says that, given a certain distance between $U_a$ and $U_b$, the effect of $U_b$'s retweet of a politician $P_j$ is greater the further the politician is from the retweeter. This is to say that if a friend, who I know to be centre-left oriented, suddently endorses an extreme right candidate this gives me more information about the candidate's quality than it would have done if he endorsed a closer politician.

To make a concrete example: if Fox News endorses a Democratic candidate over a Republic one, there must be something exception about the candidate and, therefore, one should be more likely to endorse him. Thus, it is not only the distance between me and the candidate and that between me and the retweeter that matter: also the distance between politician and endorser is crucial.

$$H4: P\left(RT_{U_a}^{P_d}\right) > P\left(RT_{U_a}^{P_b}\right) \; if \; |\varphi_d - \theta_a| = |\varphi_b - \theta_a| \wedge \varphi_d\theta_a > 0 \wedge \; \varphi_b\theta_a < 0$$

This hypothesis says that the model should not be exactly symmetrical: the political distance matters in an asymmetric way towards the spectrum. In particular, I suppose that one would be more likely to endorse a candidate on the same part of the ideological spectrum more than an equally distant one on the other side. Centre-right voters should be more likely to retweet a right-wing candidate than a centre-left one. This hypothesis relies on a bipolar ideology where the main attacks are between the two sides of the spectrum and mainly relies on a centrifugal assumption. However, recent political phenomena see an increase of the bitterness within the same part of the spectrum, with extremism attacking their closest rival, i.e. centrist of the same part of the spectrum. If this was to be true, we could reverse the sign of the inequality and see if any centripetal force makes moderates stick together against extremisms.

$$H5: P\left(RT_{U_a}^{P_a}|RT_{U_c}^{P_a}\right) < P\left(RT_{U_a}^{P_a}|RT_{U_d}^{P_a}\right)$$

This hypothesis is a slightly modified version of *H3*, now including fact that, given a certain distance between me and the endorser, I would be more likely to retweet when I am between the endorser and the politician. The mechanism behind this hypothesis is similar to that aforementioned when discussing *H3*: if a person so different from the politician shows her support, this provides me more information on the candidate's quality.

Moreover, it could be interesting to see if the effect of retweet, given a certain distance between the endorser and the politician, is higher when the user is between the two of them, or outside the segment on the endorser's side or on the politician's one. This idea can be developed further to see whether some endorsement are potentially negative. If, for instance, the retweeter is further from me than the candidate, this might be a disincentive me to endorse the candidate myself.

All these hypothesis have a very interesting role in the debate about selective exposure and the spread of opinions. Indeed, few studies are capable to classify in a detailed manner how the medium (endorser friend) between the receiver (users) and the message (politician) influences the diffusion. This is like adding a further dimension on the debate about segregation: not only we find homophily, but among the followed actors the closest ones play a more pivotal role. The findings could have an important echo in the network theory, starting to treat nodes in a less anonymous way. Enriching the models that already encompass ideological closeness with an empirical-oriented framework can provide general and insightful results.

On the top of the academic interest, these findings could be exploited in a wide range of situations. Micro-level marketing is already paying an increasing attention of the way to identify those actors that can make a content go viral. Therefore, finding the class of people that can infect a population in a way that does not only rely on the structure of the network, i.e. for instance using eigencentrality, but also on the relative position on the topic can identify new strategy to gather new potential customers. What works for politics might very well be adopted to other fields: a classical music lover might be more tempted to listen to Rihanna if another melomaniac suggests him to do so. Although the generalisability of these findings could be arguable, there is no doubt that strong results in the dynamics of political persuasion would be of first interest for social media political marketing. This could also have very strong political economics implication: instead of a theory of median voter, one could elaborate a theory of *key* player, where policies are biased towards those agents that can best persuade other people on how to vote. If, for instance, all grandsons could persuade their grandparents to vote for any party, the optimal policy should not increase pensions but lower tuition fees. Thus, finding the key players for each party and targeting one's agenda towards them could re-shape a vast literature. This would be another, ambitious, step that could follow the research on networks.

Finally, I would like to pursue a deeper content-analysis of tweets. My goal would be to investigate how some words are politically charged, rank them in an ideological space and see whether the language used by politician varies along the spectrum, if it matches their followers' way to communicate and if it can shape it. In his essays about the micro-physics of power, Michel Foucault (e.g. Foucault, 1977) warns the reader about the link between power and knowledge, embodied and shaped by language. One's cognitive and linguistic acts are the results of a structure of hierarchical relationships, strongly interconnected. If political language could impose new words and concepts, rebranding for political convenience complex phenomena, then we would establish a new link between power and the way reality is framed through language. Whereas this might seem a purely speculative topic, the concrete implications that would follow are extremely relevant. Big data could help provide an algorithmic way to design political communication, strategically focusing on the words and the thematic areas that can impact the network profile and its attributes in a favourable way for the candidate.

## CONCLUSIONS

This work started from two easy questions: *how comes that mutually exclusive truths can survive together*? *Why people diverge in their beliefs even when communication is cheap and frequent?* Bayesian theories are often too rosy in predicting people's convergence to a single point, whereas reality is full of examples of centrifugal beliefs.

Thus, I started to work on models of opinion dynamics to find good candidates for these failures. The echo-chamber effect seemed to be the most promising one. Indeed, the double-counting of a small like-minded community can very well sustain otherwise bound to disappear beliefs. The great role of clustering and homophily in Bayesian failures became an inspiration for a more general enquiry about ideologies. If sociophysics can predict hoaxes to propagate, why cannot we apply similar insights to other fields. Politics was to me the most interesting one to investigate.

Indeed, the great diffusion of extremist movements, with often claims that are completely incongruous with moderate parties and *official* truth, is a socially relevant phenomenon that deserves wider research. For the greatness of the possible data resources, I chose Twitter as the core of my research. Indeed, the political dimension of this medium is quite elevated and, albeit non representative of the whole population, it can provide a good sample to test the models about ideas' dynamics. Moreover, the literature fails to reach a consensus about the topic of social networks and polarisation, so there is still room for new findings.

Hence, I have incurred a challenging data collection process, so to have enough data to start a work on the topic. I built a database with the accounts of all elected politicians in Italy during the last five years up to the municipal level. Then, I collected the full list of their followings and of the followers' followers. These two wide adjacency matrix, of the type *user follows politician* and *user follows user* are the skeleton of my analysis.

I have used the first to replicate a model that exploits the assumption of homophily to estimate political ideology of politicians and of their followers. The positive feature of this model is the ability to provide a continuum of measures, which can be exploited once a panel data was retrieved to assess the radicalisation of both candidates and users. This model gave satisfying results, even though the process should be refined, possibly with more advanced hardware supports. Nonetheless, we had good indicators that could serve for a preliminary descriptive analysis of the political network.

Then, with the estimated data, I could start to investigate the presence of homophily between candidates' followers. I find some evidence of within-party clustering, which is particularly strong for more extreme parties. Interestingly enough, there seems to be a correlation between homophily and extremism. However, it is not possible to say whether the first causes the second or the other way round.

By analysing the network structure at the users' level, I obtain again many hints of the presence of a echo-chamber system for the most extreme parties. Studying the groups identified with the canonical algorithms of network theory, one find a strikingly good correspondence between those and party affiliation. In particular, the network graph I have obtained by the adjacency matrix of a subset of users reflects a structure that matches very well the political description of electoral fluxes. Hence, both the estimation of ideology method and the insight of segregation seem to be confirmed, with the appealing result of more radical groups being also more homophile. All results seem to point to this direction.

Then, I start to draft a content analysis of tweets, to see if there is any correspondence between radicalisation in ideology and in language. The extent of this study is very limited, but it gives a flavour of a promising route to follow. The sketchy evidence of thematic specialisation of parties, with radical minority parties concentrating most of their efforts on few topics and the majority one being more various in its campaign tells an interesting story that should be expanded.

Finally, I list some of the possible developments that this work might have. The hope is to extend it to enrich the debate on how ideas spread though linked people. The insight that factors other than the position within the network should be taken into account is finding a growing space in literature, but a consensus on how one should model this is still far from being reached. A correlation between extremism, hoaxes and segregation has been found by many studies, but no causal relationship could be established clearly.

Exploiting an innovative way to measure political ideas of all users, we can go beyond the simple analysis of content virality and focus on a more specific issue. All media between the source of the political content and its final recipient can be now politically identified, in the hope to find that the message intermediator is not neutral but influences the receiver through his (perceived) ideology. Adopting epidemiologic models to social science can be a starting point, but it has important limits. I can catch malaria from any of my neighbours, but I cannot *catch* socialism in the same way. Thus, a new theory of homophily, selective information and diffusion of beliefs has to be elaborated.

## BIBLIOGRAPHY

Abramson, G. (2001). Mathematical modelling of the spread of infectious diseases, notes from lectures given at PANDA, UNM

Acemoglu, D., Dahleh, M. A., Lobel, I., & Ozdaglar, A. (2011). Bayesian learning in social networks. *The Review of Economic Studies*, *78*(4), 1201-1236.

Acemoglu, D., Ozdaglar, A., & ParandehGheibi, A. (2010). Spread of (mis) information in social networks. *Games and Economic Behavior*, *70*(2), 194-227.

Amblard, F., & Deffuant, G. (2004). The role of network topology on extremism propagation with the relative agreement opinion dynamics. *Physica A: Statistical Mechanics and its Applications*, *343*, 725-738.

Andrieu, C., De Freitas, N., Doucet, A., & Jordan, M. I. (2003). An introduction to MCMC for machine learning. *Machine learning*, *50*(1-2), 5-43.

Bailey NTJ (1957) The mathematical theory of epidemics. London: Charles Griffin.

Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. (2012, April). The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web* (pp. 519-528). ACM.

Bala, V., & Goyal, S. (1998). Learning from neighbours. *The review of economic studies*, *65*(3), 595-621.

Barberà, P. (2014). How Social Media Reduces Mass Political Polarization. Evidence from Germany, Spain, and the U.S., available at j.mp/BarberaPolarization

Barberá, P. (2015). Birds of the Same Feather Tweet Together. *Bayesian Ideal Point Estimation Using Twitter Data*.

Bloch, M. (1924). *Les rois thaumaturges*. Istra.

Boyd, S., Ghosh, A., Prabhakar, B., & Shah, D. (2006). Randomized gossip algorithms. *Information Theory, IEEE Transactions on*, *52*(6), 2508-2530.

Brundidge, J. (2010). Encountering “difference” in the contemporary public sphere: The contribution of the Internet to the heterogeneity of political discussion networks. *Journal of Communication*, *60*(4), 680-700.

Bryant, J., & Miron, D. (2004). Theory and research in mass communication.*Journal of communication*, *54*(4), 662-704.

Campante, F. R., Durante, R., & Sobbrio, F. (2013). *Politics 2.0: The multifaceted effect of broadband internet on political participation* (No. w19029). National Bureau of Economic Research.

Clauset, A., Moore, C., & Newman, M. E. (2008). Hierarchical structure and the prediction of missing links in networks. *Nature*, *453*(7191), 98-101.

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

Colleoni, E., Rozza, A., & Arvidsson, A. (2014). Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *Journal of Communication*, *64*(2), 317-332.

Conover, M. D., Gonçalves, B., Flammini, A., & Menczer, F. (2012). Partisan asymmetries in online political activity. *EPJ Data Science*, *1*(1), 1-19.

Conover, M. D., Gonçalves, B., Ratkiewicz, J., Flammini, A., & Menczer, F. (2011, October). Predicting the political alignment of Twitter users. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on* (pp. 192-199). IEEE.

Conover, M., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., & Flammini, A. (2011, July). Political polarization on Twitter. In *ICWSM*.

Currarini, S., Jackson, M. O., & Pin, P. (2009). An economic model of friendship: Homophily, minorities, and segregation. *Econometrica*, *77*(4), 1003-1045.

De Choudhury, M. (2011, October). Tie formation on Twitter: Homophily and structure of egocentric networks. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on* (pp. 465-470). IEEE.

DeGroot, M. H. (1974). Reaching a consensus. *Journal of the American Statistical Association*, *69*(345), 118-121.

DeMarzo, P. M., Zwiebel, J., & Vayanos, D. (2001). Persuasion bias, social influence, and uni-dimensional opinions. *Social Influence, and Uni-Dimensional Opinions (November 2001). MIT Sloan Working Paper*, (4339-01).

Falck, O., Gold, R., & Heblich, S. (2014). E-lections: Voting behavior and the internet. *The American Economic Review*, *104*(7), 2238-2265.

Flaxman, S., Goel, S., & Rao, J. M. (2013). Ideological segregation and the effects of social media on news consumption. *Available at SSRN 2363701*.

Foucault, M. (1977). *Discipline and punish: The birth of the prison*. Vintage.

Galam, S. (2003). Modelling rumors: the no plane Pentagon French hoax case.*Physica A: Statistical Mechanics and Its Applications*, *320*, 571-580.

Galam, S. (2008). Sociophysics: A review of Galam models. *International Journal of Modern Physics C*, *19*(03), 409-440.

Gale, D., & Kariv, S. (2003). Bayesian learning in social networks. *Games and Economic Behavior*, *45*(2), 329-346.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). Introducing markov chain monte carlo. *Markov chain Monte Carlo in practice*, *1*, 19.

Golub, B., & Jackson, M. O. (2010). Naive learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics*, 112-149.

Halberstam, Y., & Knight, B. (2013). Are Social Media more Social than Media? Measuring Ideological Homophily and Segregation on Twitter.

Halberstam, Y., & Knight, B. (2014). *Homophily, Group Size, and the Diffusion of Political Information in Social Networks: Evidence from Twitter* (No. w20681). National Bureau of Economic Research.

Hegselmann, R., & Krause, U. (2002). Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of Artificial Societies and Social Simulation*, *5*(3).

Hethcote, H. W. (2000). The mathematics of infectious diseases. *SIAM review*,*42*(4), 599-653.

Himelboim, I., McCreery, S., & Smith, M. (2013). Birds of a feather tweet together: integrating network and content analyses to examine cross-ideology exposure on Twitter. *Journal of Computer-Mediated Communication*, *18*(2), 40-60.

Kim, H., Son, I., & Lee, D. (2012). The Viral Effect of Online Social Network on New Products Promotion. Available at http://jiisonline.evehost.co.kr/files/DLA/7-18-2.pdf

Kohut, A., Doherty, C., Dimock, M., & Keeter, S. (2012). In changing news landscape, even television is vulnerable. *Pew Center for the People and the Press*.

Krause, U. (2000). A discrete nonlinear and non-autonomous model of consensus formation. *Communications in difference equations*, 227-236.

Kretzschmar, M., & Morris, M. (1996). Measures of concurrency in networks and the spread of infectious disease. *Mathematical biosciences*, *133*(2), 165-195.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010, April). What is Twitter, a social network or a news media?. In *Proceedings of the 19th international conference on World wide web* (pp. 591-600). ACM.

Lazarsfeld, P. F., Berelson, B., & Gaudet, H. (1968). The peoples choice: how the voter makes up his mind in a presidential campaign.

Meadows, M., & Cliff, D. (2012). Reexamining the relative agreement model of opinion dynamics. *Journal of Artificial Societies and Social Simulation*, *15*(4), 4.

Messing, S., & Westwood, S. J. (2012). Selective exposure in the age of social media: Endorsements trump partisan source affiliation when selecting news online. *Communication Research*, 0093650212466406.

Newman, M. E. (2002). Spread of epidemic disease on networks. *Physical review E*, *66*(1), 016128.

Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. Penguin UK.

Smith, A. (2013). Civic engagement in the digital age. *Pew Internet*.

Stroud, N. J. (2010). Polarization and partisan selective exposure. *Journal of Communication*, *60*(3), 556-576.

Sunstein, C. R. (2009). *Republic. com 2.0*. Princeton University Press.

Wakita, K., & Tsurumi, T. (2007, May). Finding community structure in mega-scale social networks:[extended abstract]. In *Proceedings of the 16th international conference on World Wide Web* (pp. 1275-1276). ACM.

Weisbuch, G., Deffuant, G., & Amblard, F. (2005). Persuasion dynamics.*Physica A: Statistical Mechanics and its Applications*, *353*, 555-575.

Wu, F., & Huberman, B. A. (2004). Social structure and opinion formation.*arXiv preprint cond-mat/0407252*.

Yardi, S., & Boyd, D. (2010). Dynamic debates: An analysis of group polarization over time on Twitter. *Bulletin of Science, Technology & Society*,*30*(5), 316-327.