

EXTRACTING INFO FROM POLITICAL NEWS THROUGH BIG DATA NETWORK ANALYSIS

Giovanni Giuffrida, Simona Gozzo, Francesco Mazzeo Rinaldi, Venera Tomaselli

1. Integrating Big Data and Network Analysis methods

Recent studies provide convincing evidence that on online data gathering, alongside massive real-world datasets, can give new insights into real-world collective decision-making and can even anticipate future actions.

Business managers, academic researchers, and public policy makers might leverage social media to provide a data-driven analysis of the accounts and decide the best actions in achieving stakeholder goals. A data-driven approach defines a decision-making style focused on collecting, storing and analysing huge amounts of data involved in the decision process in order to extract valuable information for the management oriented to problem solving.

Networks depicting social interactions among actors have been analysed for a long time by a theoretical foundation approach for the application of the *Network Analysis* (NA) methodology (Cronin 2011). Interaction data may be generated by Big Data (BD) techniques in large scale and at low cost providing detailed information about context, content and meaning of social interactions as reported events or surveyed opinions (de Nooy, 2015). As formal method for the relational and structural analysis of decision networks, NA has been applied to analyse decision patterns and structures in municipal politics (Laumann and Pappi, 1976), decision-making and organizational behaviour (Cross and Parker, 2004).

A critical point about BD is the inaccuracy danger (McFarland and McFarland, 2015). Since observational data from social websites online sources are not derived from statistically rigorous designed experiments, they often can hide biases. The sheer size of the data collected can lead to believe to have census data on an overall population whereas it is a very biased sample containing a misrepresentative mixture of subpopulations such as user groups active during the measurement. BD often produces extremely small or zero variances, very large test statistics and then, extremely statistically significant results. Before any analyses on large data sets, the authors propose to employ a community detection data segmentation technique both to identify distinct representative populations at distinct observed activity levels or locations and control for major components of observational data biases.

To improve the accuracy of the results, collecting more data is useful to adjust for the segment skew and perform analyses within each of the segments.

The behavioural sciences have flourished by studying how relatively well-informed individuals and social learning have governed traditional and/or rational behaviour throughout most of human history. In the online age, however, social phenomena can occur with unprecedented scale and unpredictability. Furthermore, individuals have easy access to social connections. The modern ICT has opened new interaction models between individuals, like mobile phone communications and online interactions enabled by the Internet. Such new social exchanges can be accurately monitored for very large systems, including millions of individuals, whose study represents an important chance for social science analysing communities of social networks such as friendship circles, groups of people sharing common interests and/or activities, etc.

The structural attributes of networks can be measured by NA methods, such as size, density, clustering, openness, stability, reachability, and centrality (Wasserman and Faust, 1994) and, then, clustering and categorizing structural patterns (*i.e.*, hierarchies, collaboration networks, and chains). Furthermore, regression models can be specified to test for significant relations between decision structures and performances to assert better and worse performing decision patterns.

Studies integrating BD and NA are very few. Only a study, about a large-scale analysis of the news media coverage of the 2012 US presidential elections campaign (Sudhahar *et al.*, 2015), proposes a BD approach as automatic corpus linguistics methods matched with NA. It is based on information extraction about the key actors and their relations in the media narrative of the US elections, organised as a semantic network graph. By the means of this methodology, political positions are automatically derived from a very large corpus of online news, generating meaningful networks with directed links whose semantic nature is retained and understood. The approach is innovative in order to gain insights in the linguistic analysis of texts by extracting relational data.

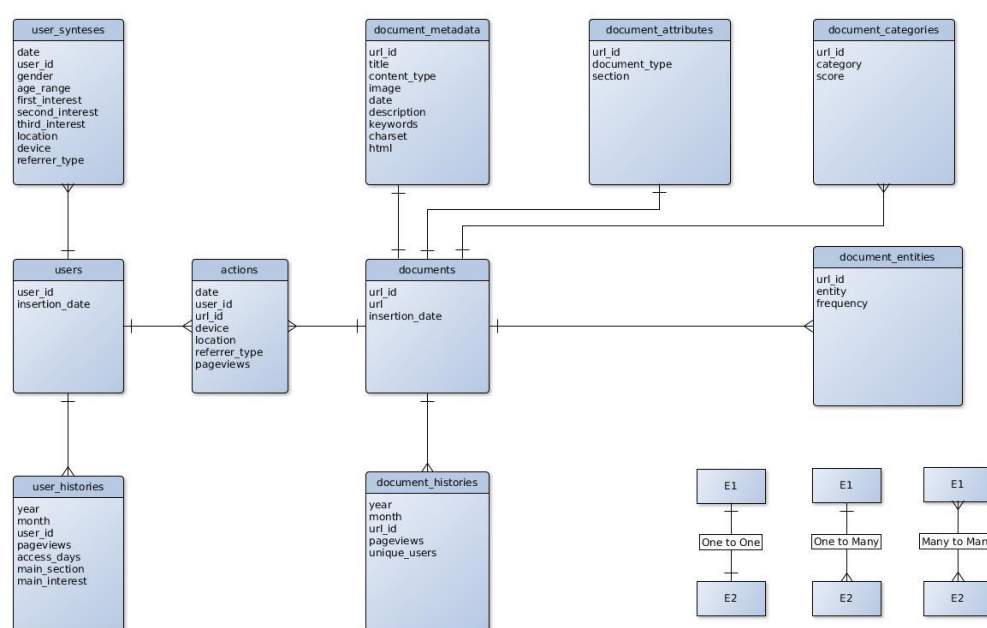
The main aim of this study is to improve knowledge on BD techniques combined with NA methods in social research in order to test the efficacy of the methodological integration between NA and BD approaches. As a first step, the present study describes the audience analysis system developed for extracting valuable information from the user traffic on selected articles about the Italian Senate reform using the advanced search feature of one of the most diffused online newspaper. Then, a very large database is performed. In the second step, the study shows how NA can be integrated with BD tools detecting online news.

2. The Big Data model

The audience model here employed is a generic data model that has been already applied to other online newspapers (Mazzeo Rinaldi, *et al.*, 2017).

The following picture (Figure 1) shows the entity-relationship (ER) diagram of our data model.

Figure 1 – Entity-relationship diagram



This is a Relational models composed of Tables and relationships among those. Tables are collections of rows, where each row contains several fields arranged in columns. Lines connecting tables indicate the kind of relationship linking them, which can be “one to many”, “many to many”, or “one to one”. “A one to many” relationship happens when one row in one table corresponds to many rows in the second table. For instance, a user performs many actions, giving rise to “a one to many” relationship. “A many to many” relationship happens when many rows from one table correspond to many row in the second table. The following sections describe in detail each table with their fields, and how derived fields are computed.

Users - The table *users* contains one row for each user. Each user is identified by a *user_id*, which is a 33 character hexadecimal string stored in a cookie in the user's browser. Then, for each user, we store the first date in which the user was seen by the system.

Documents - The table *documents* contains one row for each document. A document is an *html* document accessible from a given *url*. Each document is uniquely identified by a *document_id*, which contains the hash of the *url* obtained using the md5 algorithm. As for users, we store the first date that a document is seen by the system.

Actions - The table *actions* contains the actions performed by users. For each given day, this table memorizes the number of *pageviews* that a user identified by a *user_id* views a document identified by a *document_id* from a given *device*, *location*, and *referrer_type*. A *device* gives us information on whether, for instance, the user connected from a desktop, a laptop, a mobile phone, etc. A *location*, which is derived from the *ip address* of the user, gives us information on the geographic location from which the user connected from. Finally *referrer types* give us information on what the user visited before landing on the current page.

Document attributes and document metadata - These tables contain information that are related to the documents. What we store is the type of the document, which tells us whether the document was a homepage, a section page, an article, a gallery, or a video. Moreover, each document is classified using an algorithm, which analyses the *url* of the document, as belonging to a certain section of the newspaper.

Document categories - The table *document_categories* contains a textual classification for each document in the database. For each document, we use a text classifier that analyses the text extracted from the *html* of the document, and associates it to one or more categories. Thus, each content may generate more than one row in the *document_categories* table. The classifier assigns to each category of the document a score, which indicates the level of confidence to which a the category belongs to the document.

Document entities - The table *document_entities* contains semantic entities extracted by analyzing the *html* of each document. For each document, we use a text annotator that analyses the text extracted from the *html* of the document, and annotates this text with one or more entities. The text annotator used is Tagme (<http://tagme.di.unipi.it/>), a tool that is able to identify meaningful short phrases in an unstructured text and link them to a pertinent Wikipedia page. An entity is therefore a Wikipedia page associated to a short phrase included in the text extracted from the *html* document.

Document histories - The table *document_histories* contains statistics relative to the fruition of each content by the users in a given month. We store the *pageviews*

generated by the content in the given month and the number of *unique visitors* that have generated at least one pageview for the content in the given month.

User histories - The table *user_histories* contains statistics for each user relative to their consumption of content, computed for a given month of navigation. We compute the number of *pageview* generated by the user, the number of days that the user has accessed the site. We also compute the section most visited by the user.

3. Some results from the Big Data model

In this study, 18 articles about the Senate reform are identified by hand using the advanced search feature of the online newspaper. We searched using the query "riforma Senato", and set a time frame from January 1st, 2014 to December 31th, 2014. The query returned 47 articles, of which 5 were removed because they contained only pictures, and 24 were removed because they were too general and discussed the reform only marginally. We were thus left with 18 articles about the senate reform, whose effective timeframe of publication was between March 12th, 2014 until August 8th, 2014¹. All these articles were in the newspaper section about politics. The articles in the section about politics published in the same timeframe were in total 1,788. We then measured the number of pageviews and the number of comments of all articles. The 18 specific articles collected 886,898 pageviews and 2,461 comments, whereas the 1,788 articles in the politics section collected 32,774,270 pageviews and 14,108 comments.

We analysed these statistics in order to ascertain the level of interest of the public with respect to the senate reform, relative to the level of interest of the public with respect to general politics. We then computed: the average number of pageviews per article, the average number of comments per article, and the probability of commenting. The average number of pageviews per article indicates the interest of a general reader in the topic. The 18 specific articles have 49,272 average pageviews, whereas the 1,788 general articles have 18,330 average pageviews. The average number of comments per articles indicates how much readers are engaged with the topic. The 18 specific articles have 137 average comments, whereas the 1,788 general articles have 8 average comments.

Finally, the probability of commenting estimates the likelihood that a comment is written upon reading an article. The 18 specific articles have 0.28% probability of commenting, whereas the 1,788 general articles have 0.04% probability of commenting. Clearly, all these measures indicate that the public is very interested

¹ See for more details Giuffrida et al., (2016).

on the topic of the Senate reform. To check the hypothesis about the Senate reform's relevance, we compare NA results obtained from two data samples: the first is composed by 18 news about the Senate's reform and the second includes 18 random news selected from political section of the newspaper.

Next, we analysed the sentiment of each comment, classifying each comment with a score ranging from -1 to +1. Furthermore, we extracted from each of the 18 articles the *semantic entities* contained in the text of the articles. Both the sentiment analysis and the entity extraction were made using a public API service provided by <https://dandelion.eu/>. It was found out that the most popular entities, that is, the entities that received the most of the comments, were: "Senato", "Governo", "Parlamento", "Senatore", and "Matteo Renzi". These entities received, respectively, an average sentiment score of -0.426. Despite the negative sentiment score, this is not to be immediately interpreted as an implication that commenters were against the proposed bill. This is because, generally comments in a newspaper tend to be negative (among others: Reis et al., 2015), therefore the average sentiment score of the top entities must be compared with the average sentiment score of the comments in the general politics section, which was computed to be equal to -0.38. Therefore it can be inferred that commenters have a slightly negative sentiment (0.04), representing 8% of the standard deviation measured on comments in 0.5, thus insignificant.

4. Results from network analysis application

This study analyse links among subjects reading the same news of newspaper web site. Each link (line in the graph) represents the relationship between two users or nodes that are connected by the line. This kind of data structure is here employed to process data by means of *Gephi*, a software for visualization, analysis and data mining of graphs. When you want to analyse a graph, the nodes are the reference objects and the links among nodes are the units of analysis. The two graphs obtained on this basis represent links between readers of the two samples selected among the registered users, removing repetitive and recursive ties. The graphs are obtained applying the algorithm Force Atlas 2.

Force Atlas 2 is an improvement of Force Atlas, an algorithm that displays net with a 'scale free' (such as internet) and 'small world' (relations among people) topology. Namely, this algorithm is used to show graphs with high connectivity or with nodes easily reachable, making the graph more readable and offering a clear interpretation of connections among nodes. However, Force Atlas works on 1-10.000 nodes while Force Atlas 2, under equal conditions, works on larger graphs and limiting the dispersion. The application of Force Atlas 2 highlights "islands" of

more connected nodes. In these areas of the graph there are readers who often view the same news.

The first descriptive measure about the structure of our networks is the graph diameter, or the largest geodesic distance (computed as number of edges among nodes): 4 for the random graph and 3 for the Senate reform's graph, while the average geodesic distance is less than two items connect pair of nodes. This means that no actor is more distant than three/four steps from any other actor (close to the small world structure), i.e. there are at most three/four articles connecting all readers within the graphs. Diameter of a network points out how many steps are necessary to get from one side of it to the other. Although both graphs are compact networks, the second shows a shorter distance among readers and this despite the greater number of nodes. The opposite of diameter is the geodesic path, that is the shortest path through the network. The average short path is 1,82 for the graph about political articles and 1,67 for graph about Senate reform.

Gephi has also enabled to distinguish the news with greater impact from the less relevant. Considering the substantial number of ties generated for each link on the same article, the graphical output distinguish 7 of the 18 articles as the most impacted while other items are clustered into a single category.

Another tool to detect the size of connection within the network is the eccentricity, a measure of how far each actor is from the furthest other. It is similar to the diameter (computed on the whole network) but referred to each node. Therefore, it measures how far, at most, is each node from every other node. This procedure shows information about a relative measure (referred to each node), while the Diameter is referred to the whole net.

Having regard to the structure of the data, a greater eccentricity value indicates the presence of readers who have read the same articles. The link represents, in fact, the visualization of a single article. The modal eccentricity for both graphs is 3, with 1.110 (random) and 1.000 (Senate) nodes with this value. This means that, for both graphs, most of the nodes have a maximum distance of 3 links (or visualized news). The second most common value is 4 for the random graph (500 nodes) and 2 for the graph on the Senate reform. This means that readers who select news about Senate reform are more connected, with lower maximum distances. Overall, the readers of articles about Senate reform have lower dispersion, focusing on fewer articles, displayed by more readers. This information allows policy makers to identify the most important news for the public opinion, which thickens the debate.

However, the most common tools applied to study the structure of a graph are the centrality measures. There are different measures of centrality. The most intuitive is the *degree* centrality computed on the number of ties, where the actor with the most ties is the most important. The definition of centrality can be

extended from the node level to the whole graph. Considering the data structure, *degree* increases the more the more the node consults articles viewed by other nodes. The centralization index shows the spread of shared information among nodes. Articles about Senate reform have a higher co-occurrence of consultations. It is, therefore, probably a topic on which focuses the attention of web users (*attractor*), even comparing the data with the graph referred to policy-relevant information (assuming a relative specialized interest of the average user). Looking at the diameter, the information is even more interesting when you consider that - in both cases - the attention of users is polarizes on 3-4 of 18 selected articles (almost three items explain more than 50% of contacts). These results are obtained analysing data both by mean of Gephi and Nodexl, another software useful for network analysis on big relational data. A second measure of centrality is the *closeness* centrality, applied when an actor is considered important if he/she is relatively close to all other actors. The *closeness* measures the average distance between a node and all other nodes and it is based on the inverse of the distance of each actor to every other actor in the network. The distribution of *closeness* indices is similar to the *degree*. Graph about the Senate Reform is more compact. The readers have more similar profiles considering both direct and indirect edges (or shared news). Another centrality measure is the *betweenness*. This measure is useful to define models based on communication flow and increases the more the more an actor can control communication flows within the net, playing as a broker. *Betweenness* centrality counts the number of shortest paths between *i* and *k* that actor *j* resides on. Considering the data structure for the whole net, *betweenness* centrality does not detect useful information because the links do not represent actual relationships but shared information. So users cannot effectively “convey” information. However, the structure of this measure points out that a low number of “hubs” underpins the network structure. Both analysed graphs present high level of *degree* and low *closeness* and *betweenness*. This means that the net is composed form “clusters” of readers and that each node has redundant connections (that is, there is a great sharing of information in the network).

Further measures are the modularity and the clustering coefficient. These tools permit to analyse the structural behaviour of graphs, showing the presence of groups overlapped within the whole net. The modularity is 0,366 for links among articles of political news and 0,185 for news about the Senate reform, while both graph have the same clustering coefficient (0,84). Modularity was designed to measure the strength at which a network is divided into clusters (groups or communities): networks with high modularity have dense connections between the nodes within clusters but sparse connections between nodes in different communities. So, data about Senate reform are more compact compared to the graph about general political issues. This latter shows separated sub-groups and a

weak community structure. As further confirmation, we applied the Clauset-Newman-Moore algorithm to obtain clusters grouping the graph's vertices by means of Nodexl. The first graph (about political articles) shows intersections among 7 collapsed groups, while the second shows only 4 clusters.

Each cluster has a specific relational structure and it presents specificities if you compare to the structural characteristics of the whole network. Graph about general political news is (as expected) more heterogeneous. It stands 7 different sub-groups identified simply by analysing the relationship dynamic (without any reference about the content of each article). Besides, there are strong structural differences among groups (Table 1).

Table 1 – Degree, Betweenness, Clustering for groups (Political News)

Political news	Degree	Betweenness	Clustering
TOT (averages)	379,04	756,27	0,84
GROUP 1	402,25	1353,45	0,73
GROUP 2	355,98	371,68	0,91
GROUP 3	484,29	1063,77	0,79
GROUP 4	274,57	102,57	0,96
GROUP 5	315,26	824,47	0,85
GROUP 6	406,00	0,00	1,00
GROUP 7	252,63	115,68	0,96

First, second and thirist groups are the most important, with highest *degree* (more articles viewed) and *betweenness* (more brokers) centrality measures. What distinguishes, however, the most relevant groups from the others is the index, particularly high, of *betweenness*. This configuration indicates that - within some groups - there are nodes with a greater heterogeneity in cultural consumption and therefore able to “connect” with very different contexts and arguments, acting as a via (or broker) on other thematic groups. Groups of this type have a more varied thematically configuration and are more likely to consist of an average user informed, which will tend to select most likely the main information on lines differentiated issues. More specialized are, however, groups that are formed in the second graph Table 2).

The most informed readers, with the highest *betweenness* index, make the first group. This group is the “attractor”, structure of the graph depending by it. The second group is internally compact but it is not particularly important for the network structure. Third group shows the highest *degree* value. This means that nodes are directly and strongly connected each other (see often the same news).

Forth group has the highest clustering coefficient and this means that nodes tend to form “triangles”: transitivity of information spread specifically within this group is strong.

Table 2 – *Degree, Betweenness, Clustering for collapsed groups (Senate Reform)*

Senate's reform	Degree	Betweenness	Clustering
TOT (averages)	517,21	515,99	0,85
GROUP 1	411,98	748,35	0,81
GROUP 2	605,00	184,02	0,91
GROUP 3	626,41	568,31	0,80
GROUP 4	412,20	113,50	0,98

5 Conclusion

BD are today pretty much ubiquitous and increasingly collected every day in real-time from many different users' device. Applications for BD analysis are getting more and more widespread. In this paper we proposed a Network Analysis approach over a Big database collected from readers' interaction with a political topic on a large online Italian newspaper. In particular, we isolated a certain number of published articles on the topic of interest and we analyzed the readers' collective behavior towards this subset against all articles in the more general political section. We also report on some basic statistics to show how readers are in general interested to the topic at hand, and we did this by comparing some quantitative measurements among the two subsets above discussed.

We collected historical data on a recent constitutional reform on the Italian Senate. This has raised a pretty active debate in Italy and this is well reflected in our measurements. As a matter of fact, both basic statistics and the proposed Network analysis models prove the validity of such approaches to measure overall readers' sentiment. We believe that this type of analysis may turn into a very valuable tool for policy makers in general as they can measure in real-time readers' reaction to new law proposals. This can surely help the policy maker in shaping the new proposal in order to maximize citizens' consensus. As future work, we will be trying to collect additional data from our logs, such as user's demographic information and behavioral data in order to expand our analysis models. We will be also testing additional models and we hope to be able to work with real policymakers in order to shape such models at best.

References

- CRONIN B. 2011. A window on emergent European social network analysis. *Procedia Social and Behavioral Sciences*, Vol. 10, pp. 1-4.
- CROSS R., PARKER A. 2004. *The Hidden Power of Social Networks: Understanding How Work Really Gets Done in Organizations*. Boston: Harvard Business School Press.
- DE NOOY. W 2015. Structure from interactive events. *Big Data & Society*, Vol. 1-4, doi: 10.1177/2053951715603732.
- GIUFFRIDA G., MAZZEO RINALDI F., ZARBA C. 2016. Big data e news online: possibilità e limiti per la ricerca sociale. In AGNOLI S. PARRA SAIANI P. (Eds.), *Sulle tracce di big data. Questioni di metodo e percorsi di ricerca*, numero monografico di *Sociologia e ricerca sociale*, anno XXXVII, n. 109, Milano FrancoAngeli.
- LAUMANN E. O., PAPPI F. U. 1976. *Networks of Collective Action: A Perspective on Community Influence Systems*. New York: Academic Press.
- MAZZEO RINALDI F., GIUFFRIDA G., NEGRETE T. 2017. Real-time monitoring and evaluation - Emerging news as predictive process using Big Data based approach. In: PETERSSON G.K., BREUL J.D. (Eds) *Cyber Society, Big Data, and Evaluation*. vol. 24, New Brunswick, N.J.: Transaction.
- McFARLAND D. A., McFARLAND H. R. 2015. Big Data and the danger of being precisely inaccurate. *Big Data & Society*, Vol. 1-4, doi: 10.1177/2053951715602495.
- PRESIDENZA DEL CONSIGLIO DEI MINISTRI 2014. Constitutional review bill, Available at: <http://www.governo.it/governoinforma/documenti/CONSTITUTIONAL%20REVIEW%20BILL.pdf>
- REIS J., BENEVENUTO F., OLMO P., PRATES R., KWAK H., AN J. 2015. Breaking the News: First Impressions Matter on Online News, *Proceedings of the Ninth International AAAI Conference on Web and Social Media*, Oxford: Oxford University, pp. 357-366.
- SUDHAHAR S., VELTRI G. A., CRISTIANINI N. 2015. Automated analysis of the US presidential elections using Big Data and network analysis. *Big Data & Society*, Vol. 1-28, doi: 10.1177/2053951715572916.

SUMMARY

Extracting info from political news through Big Data Network Analysis

Recent years have witnessed extraordinary growth in globally generated data. The advent of automation of many of our activities has resulted in large volumes of routinely generated data about what we search for on-line, what we read, what we buy, and the list goes on. The automatic extraction of all this information, together with innovative data mining and predictive analytics techniques, combined with the analysis of relational patterns, can represent an innovative opportunity in supporting decision-making. Responding to a rising interest to understand and improve actionable analytics-driven decision patterns, this paper proposes the use of network analysis as a method both to foster the structured planning, to develop the implementation of decision-making actions and also measure effectiveness, efficiency, and quality of analytics decision-making networks from an empirical methodological perspective.

The main aim of this study is to improve knowledge on big data techniques combined with network analysis methods in social research. In particular, descriptive measurements and clustering methods of *Network Analysis* are here employed in order to define relational structures within a big data set. A Big Data tool is developed to collect info from user interactions with published news and comments about a case study related to a recent Italian constitutional review bill with important political implications. It has been extensively reported online and it produced a lot of interest from online readers. Here, one of the most popular Italian national newspapers is the main news source for the data collection of this analysis. Network analysis on the big data set is then used to extract and analyse essential information from a vast dataset, assuming that data is not only large, but also meaningful. By research outcomes, network analysis is suitable for big data to obtain informative data on the basis of relational structures among many and many comments used to measure people sentiment about online news in order to detect relevant items for policy-makers.

Giovanni GIUFFRIDA, Department of Political and Social Sciences, University of Catania, giovanni.giuffrida@gmail.com

Simona GOZZO, Department of Political and Social Sciences, University of Catania, simonagozzo@yahoo.it

Francesco MAZZEO RINALDI, Department of Political and Social Sciences, University of Catania; KTH, Royal Institute of Technology, School of Architecture and the Built Environment, Stockholm, fmazzeo@unict.it

Venera TOMASELLI, Department of Political and Social Sciences, University of Catania, tomavene@unict.it