Check for
updates

The interconnectedness of the economic content in the speeches of the US Presidents

Matteo Cinelli[1] · Valerio Ficcadenti[2,3] · Jessica Riccioni[2]

Abstract

The speeches stated by influential politicians can have a decisive impact on the future of a country. In particular, the economic content of such speeches affects the economy of countries and their financial markets. For this reason, we examine a novel dataset containing the economic content of 951 speeches stated by 45 US Presidents from George Washington (April 1789) to Donald Trump (February 2017). In doing so, we use an economic glossary carried out by means of text mining techniques. The goal of our study is to examine the structure of significant interconnections within a network obtained from the economic content of presidential speeches. In such a network, nodes are represented by talks and links by values of cosine similarity, the latter computed using the occurrences of the economic terms in the speeches. The resulting network displays a peculiar structure made up of a core (i.e. a set of highly central and densely connected nodes) and a periphery (i.e. a set of non-central and sparsely connected nodes). The presence of different economic dictionaries employed by the Presidents characterize the core-periphery structure. The Presidents' talks belonging to the network's core share the usage of generic (non-technical) economic locutions like "interest" or "trade". While the use of more technical and less frequent terms characterizes the periphery (e.g. "yield"). Furthermore, the speeches close in time share a common economic dictionary. These results together with the economics glossary usages during the US periods of boom and crisis provide unique insights on the economic content relationships among Presidents' speeches.

Keywords Glossary of economics · Text mining · US Presidents' speeches · Network analysis · Clustering

1 Introduction

Is there a glossary of economic and financial terms whose presence is significant in the US President speeches framework? What is the role of such locutions in the US Presidents public communications? Did the crisis periods affect the choices of mentioning some financial concepts? Is there a group of US Presidents' speeches that can be classified as a cluster

✉ Valerio Ficcadenti
ficcadentivalerio@gmail.com

Extended author information available on the last page of the article

🖄 Springer

because of the usage of specific terminology devoted do describe economics and financial situations?

The US President is one of the most influential people in the world. Therefore his communications have to be considered under many perspectives in order to effectively reach strategic objectives. For such a reason Presidents' talks are carefully calibrated on the basis of the audience, the occasions in which the talks occur and the socio-economic surrounding at the moment in which he is speaking. The entities of US Society change their expectations, and consequently their actions, also on the bases of the informative set carried out by the Presidents' messages. Empirical evidence of these phenomena are given by the effects generated by Trump's announcements on tariffs changes for importing the Chinese goods in the US, or the impact of the Obamacare announcements on the US health-care sector. Sometimes, even just tweets can be influential. For example, in Shaban et al. (2017) the authors have analysed about 130 millions of tweets posted during the 2016 US electoral campaign or the studies of tweets occurred during the 2012 electoral campaign presented in Maldonado and Sierra (2016), and Vargo et al. (2014).

Since Presidents' talks often contain references to the economic and financial situation of the country (e.g. Rule et al. 2015), in this paper we aim at checking if words usually devoted to economics and finance make some speeches closer than others along the years. To do so, we have assembled the huge corpora starting from the written version of Presidents talks, from George Washington (April 1789) to Donald Trump (February 2017). In particular, we have used the same process presented in Ficcadenti et al. (2019) and a summary of it is reported in Sect. 3. The result is a collection of 951 speeches taken from the Miller Center (www.millercenter.org), a Political Research Institution affiliated to the University of Virginia. We have then investigated the network of the US Presidents' speeches where the nodes are talks and the links are the cosine similarity measures obtained from the frequencies of the economic terms stored at speech level.

Using text mining techniques that will be widely presented below, we processed the corpora. Specifically, we proceeded to extract the terms contained in speeches whose meaning could be ascribed to the semantic area of economics and finance. These terms result from the merger of two different glossaries: one deriving from a manual used by prestigious newspapers of the sector as The Economics (see Bishop 2009), and the other is the Wikipedia's glossary of economics see (see Wikipedia contributors 2019a). After excluding the economics locutions absent in the corpora, we have obtained a total of 383 terms. With them, we have estimated the economic content level of each talk by exploiting their absolute and the relative frequencies (see, e.g. Wei et al. 2015; Baker et al. 2016; Tsai and Wang 2017, for comparable processes). The usage of network analysis on the speeches dataset through the employment of the frequencies of occurrences of the aforementioned glossary in the talks allows for the measurement of the degree of connections among Presidents, their speeches and their party affiliations. Hence, the implementation of the network analysis on the speeches of the US Presidents allows to explore the structure of connections and the density of links to understand to which extent different speeches, Presidents and parties are interrelated. The observation of a clustered structure permits us to differentiate the interactions, the information and the implications deriving from mapping the similarities between the speeches collected. Our findings are particularly interesting. Especially because we obtain evidence that the historical periods in which the talks have been delivered are relevant for the clustering system, furthermore we detect a core-periphery structure of speeches based on the economic terms commonly used. This means that there is a category of speeches characterized by the utilization of a subclass of economics terms used to describe more specific situations like relevant economics or financial fact. They

might be speeches delivered during crisis or post crisis period, or peculiar epochs of reforms.

The structure of the paper is the following: in Sect. 2 we present some comparable works; in Sect. 3 we describe the dataset explaining how it was built and how the economic glossary was created. In Sect. 4 the model for the construction of the network is presented. In Sect. 5, we provide the analysis of the network of US President speeches based on their economic content and in Sect. 6 we discuss the results. Conclusions follow.

2 Literature review

Data mining techniques are used by scientists and researchers to manage massive amounts of heterogeneous data with the aim of extracting useful information (see Kocheturov et al. 2019; Ngai et al. 2011; Ravisankar et al. 2011; Malik et al. 2018, for some recent examples.) Relevant sub-fields of data mining are text mining and Natural Language Processing. While the former is related to the analysis of a text, the latter includes natural language comprehension thanks to advanced machine learning techniques (see Felici 1995). Text mining allows to analyzing strings of characters belonging to texts for extracting relevant information like the meaning of sentences or their sentiment. This is particularly useful when a huge amount of documents is involved or when there is the need of quantifying information from qualitative datasets as in Yuan et al. (2018).

Nowadays, textual analysis is employed in a wide variety of studies, for example in medicine (Lee et al. 2019), in tourism management analysis (Cheng and Jin 2019), in design-ing recommendation systems (Ji et al. 2019), in analyzing countries' foreign policies (Cannon et al. 2018), in investigating the blog users' sentiments during rainstorm and waterlogging dis-asters (Wu et al. 2018) or in understanding the potential applications and users of augmented reality tools Li et al. (2018).

Recent contributions like Feuerriegel and Gordon (2018) highlight the importance of the information contained in written documents to analyse the economic paths and to forecast economic and financial variables. Among the latest works, we include Feuerriegel and Gordon (2019) in which text mining techniques are applied for reducing forecast errors of the macroe-conomic indicators by analyzing news. As part of the prediction of market performance, we mention a study of annual reports of more than a thousand of firms (see Balakrishnan et al. 2010), and also the analysis of financial reports through the connection between the words and the financial risk of various companies and banks (see Tsai and Wang 2017; Agarwal et al. 2019). Furthermore, it is worthy to mention the case of Kahveci and Odabaş (2016), in which the textual analysis has been employed to explore the semantic of monetary policy documents from the Federal Reserve Board, the European Central Bank from 2001 to 2015, and the Central Bank of the Republic of Turkey from 2002 to 2015.

In a more general economic and financial framework we can evidence many works that deepen text mining on the analysis of stock markets, stock returns, trading and, more generally, on the ability of transforming qualitative variables into quantitative measures to improve financial forecasts and market predictions (see, e.g. Blasco et al. 2005; Schumaker and Chen 2009; Groth and Muntermann 2011; Huang and Li 2011; Loughran and McDonald 2016; Mishra and Singh 2018). We also mention Antweiler and Frank (2004), Carretta et al. (2011), Hendershott et al. (2015), Peruzzi et al. (2018) as examples of studies of the news impact on the financial institutions and on stock market returns; we refer to Nassirtoussi et al. (2014) for a complete review of text mining applied to sock performance predictions. Within the field of finance, it is also worthy citing some studies in the more specific field of sentiment analysis that

has attracted growing interest in the latest years. Through the analysis of words, punctuation and emoticons collected from social media (such as Twitter), blogs, forums, online reviews, responses to messages from consumers, it is possible to evaluate the positive or negative opinions, their intensity, their emotionalism content and the relevance of the object of analysis with respect to the context (see, e.g. Tetlock 2007; Loughran and McDonald 2011; Price et al. 2012; Garcia 2013; Bao and Datta 2014; Alfaro et al. 2016). A complete methodological review of text mining applications for financial purposes can be found in Kumar and Ravi (2016), it contains relevant studies with sentiment analysis applications. Another interesting review is given by Loughran and McDonald (2016). It contains a summary of text analysis studies in the context of accounting and finance, mentioning pros and cons of each. Our work does not fall in one of the categories listed in Loughran and McDonald (2016); we present a multi-disciplinary study that shares some common points with dictionary-based information extraction methods and documents/authors classification. Our point of view is different because we want to measure the similarities of the Presidents' speeches, Presidents and Presidents' parties affiliations by using a set of terms recognized as meaningful in the economics and finance field. The research here presented does not fall in the category of studies where the sentiments are investigated; therefore we avoid all the cons of using pre-trained classification method where the usage of sentiment dictionaries introduces a certain degree of uncertainty concerning the sentiment classes. On the other hand, we use a pre-determined class of locutions and we check their presence into the speeches. In this way, we select the dimensions of the bag-of-word representation considered meaningful for the analysis of speeches closeness under and economics and finance perspective.

Text mining and natural language processing methods are often employed with network analysis. For example, in Chae (2015), the authors proposed an analytical framework for analyzing tweets about supply chains and further developing insights into the potential role of Twitter for supply chains practices and researches. Their approach combines three methodologies: descriptive analytics, text mining and sentiment analysis, and network analysis throughout network visualization and metrics. The list of papers to be mentioned could be endless, here we have reported few of the most recent studies employing text mining methods and network analysis for economics and finance applications.

In the context of the US Presidents' communications studies, there are not many works devoted to the analysis of such a wide Presidents' speeches corpus with a combination of text mining and network analysis approaches strictly comparable to our methodological combination. For example, in Light (2014), the author combined text mining and network analysis to analyse the Presidents' Inaugural Addresses, but he has used the Stanford POS-Tagger and a different similarity measure. Hence, we use some excellent studies as reference for validating our approach. From a methodological point of view, we are in line with (Bail 2016). In such a work, the author presents remarkable results in the social field combining text mining and network analysis to study how advocacy organizations stimulate conversation on social media. Despite the differences in objectives pursued, our research and Bail (2016) share the utilization of the frequencies of the words to measure text similarity. Then, both studies involve the usage of communities detection algorithms to capture clusters in the data. In our study, advocacy organizations can be compared to the Presidents, and the speeches can be compared to the posts published by the advocacy organizations. Differently from Bail (2016), we computed the cosine similarity to evaluate the distance between the speeches, while Bail has used the co-presence of terms within the posts on Facebook of advocacy organizations to build a bipartite affiliation network. Moreover, to make our network we have used a prefixed list of locutions whose frequencies of occurrence contribute to the creation of the similarity matrix (as briefly mentioned before, we consider the union of two glossaries of economics -
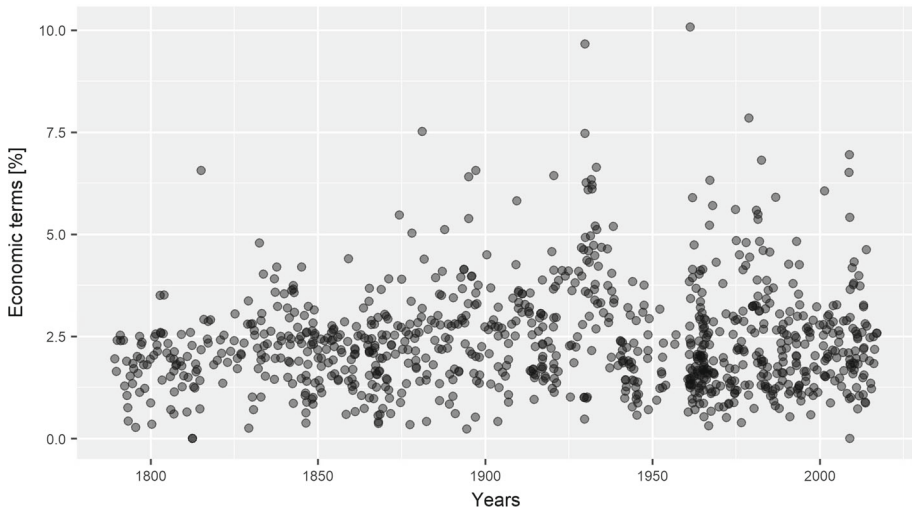
Fig. 1 Relative frequency of all the economic terms per speech along the years

the one reported in Bishop (2009) and Wikipedia contributors 2019a's—as a tool to explore the speeches proximity).

In our analysis, the main underlying assumption is given by the idea that a set of words related to economics and finance, whose presence is quite stable during the years (see Fig. 1), can be the core of a common message straighten by the Presidents in their talks. So, we address the meaningfulness of those terms concerning Presidents' speeches connections, to understand the structure of relationships based on such terminology. Furthermore, we look for the presence of a group of speeches based on the presence of such a glossary.

Another relevant study that inspired our research is Rule et al. (2015), where the authors have performed a wider analysis of the US President speeches but on a sub-sample of our dataset. They have considered the State of the Union Speeches (SoU hereafter) occurred between 1790 and 2014, to investigate changes in topics along the years. Rule at al. have created a set of semantic classes by means of the co-occurrence approaches (Callon et al. 1991, for further info on the method), therefore the authors have generated the classes from the speeches (endogenously), and then they have carefully labelled these group on the bases of the main points treated.

The presence of semantic trends like "Domestic Policy", "Foreign Policy" and "Political Economy" is a prerequisite for our analysis. Hence, we start from the idea that the economics and finance are relevant arguments of the political debate, therefore we quantify the relevance of their dictionary in making Presidents' talks closer, and consequently in creating clusters. Alternatively, from a different perspective, we aim at describing the ability of a class of terms attributable to economics and finance fields to explain the closeness of the speeches, the Presidents and/or their parties affiliations along the years.

Finally, after the determination of the semantic classes, the authors of Rule et al. (2015) used a TF-IDF (term frequency—inverse document frequency) approach to compute the cosine similarity between speeches. They employed a multi-semantic dictionary made by 1000 words whose frequencies are the drivers of the speeches' differences. This procedure is comparable with our even if we decided to use the words' frequencies instead the TF-IDF

because we have already selected the terms that matter the most for our study/topic; therefore we do not need an additional weighting scheme.

Another paper that identifies the words related to economics and finance as remarkable for their ability to divide the media sentiment during the electoral campaign of 2012 is Sudhahar et al. (2015). Almost all the words reported in it and related to economics and finance are part of our glossary. The same can be said for the words used in Schonhardt-Bailey et al. (2012) to identify economics related part of Reagan's speeches.

In Bernauer and Bräuninger (2009), the authors used the wordscore approach to measure the intra-party heterogeneity of preferences within parliamentary parties in the German Bundestag during 2002–2005. Wordscore presented in Laver et al. (2003) has many common points with our approach, even if it is designed to capture texts' political positions. As for almost all the text mining approach for content analysis, wordscore imposes a priori assumptions about the algorithm training sources. Indeed, it is based on the relative frequencies of words of a pre-selection of corpora that belong to a set of political classes. Therefore, the selection criteria of the set of texts as well as the classes, have to be based on assumptions or on selections suggested by other relevant studies (as stated in Laver et al. 2003). Finally, after that a "political" score class-based is assigned to the words appearing into the reference text, the so called "virgin" texts are addressed (in Bernauer and Bräuninger 2009 the "virgin" texts are those under investigation, whose political orientation is unknown). Basically, the classification of a virgin text depends on the probability of meeting words with relative frequencies similar to those that appear into one or more reference texts belonging to a certain class. The more the presence of the virgin text words is similar to that appearing in a set of reference texts belonging to a specific class, the more likely the virgin text can be classified as part of that class.

In the operational research field, one of the works analyzing US Presidents political activities is Cochran et al. (2014). The authors have addressed the US elections using the registered voters' behaviours to determine the best candidate's communication strategy (in term of political positioning) to get their attention. More in general, operational research studies contain many references to the methods employed to pursue objectives in line with the present study. For example, Alfaro et al. (2016) reports an application of sentiment analysis and opinion mining on comments posted in an organizational and administrative affairs weblog. Similarly to our study, in Alfaro et al. (2016), the texts are represented in a document-term matrix and cosine similarity is used to obtain a similarity matrix. But, differently from us, the authors performed the analysis using machine learning techniques as support vector machines and k-Nearest neighbours. In Oliva et al. (2018), the authors have developed a novel approach to identify groups of decision makers by means of the coherence of their opinions. They presented the new network based approach by analyzing the situation after the 2012 election.

Our study differs from the rest of the literature mainly for the peculiar text mining-network based approach employed. The combination of the two techniques allows for an exploration of the data from a topological prospective but keeping into consideration the micro-relationships deriving from the economics words employed. The vector representation of the texts along the economics and finance dimensions (given by the glossary terms frequencies) manifests a great ability in explaining the speeches proximity. Therefore, the analysis of the network highlights clusters and we conclude interpreting information about economics and financial turbulence occurred during US history. Furthermore, the original dataset employed makes this work a unicum for extracting insights about US economy and its risks.

3 Data

The speeches of US Presidents under analysis in this paper are 951. They have been stated in a period that spans from 1789 to 2017. The transcripts are downloaded from the Miller Center database (https://millercenter.org/) in June 2017 with a procedure described in Ficcadenti et al. (2019) and summarized in the next subsection. For each speech we know its date, the name of the speaker and his party affiliation.

Furthermore, we employ the list of locutions resulting from the union between the glossary of economics resulting in Bishop (2009), and the one presented in Wikipedia contributors (2019a). The result of this union is appropriately prepared for the analysis object of this paper as we will see in the next section.

3.1 Data mining and pre-processing

We provide a description of the process realized to make the dataset ready for the analysis.

The first step consists in implementing a web scraping routine to collect the speeches transcripts from the Miller Center website www.millercenter.org. Furthermore, it is also devoted to code the solutions for managing some common errors occurring at this point due to website inconsistency. For example, we met missed download due to misreported transcripts on the web pages or speeches doubly transcribed into the same page. It implicates memorization of records containing blanks or doubly repeated transcripts.

Secondly, we assessed all the stored transcripts checking for typos. For example, there are missed blanks between words or there are misuses of punctuation creating strings that do not make sense. Therefore, we coded functional solutions to manage them by employing regular expression and removing the punctuation. Once we got rid of these, we used the Hunspell dictionary (see Ooms 2017) to check the remaining peculiar words. Indeed, some terms in the speeches are part of ancient English and they are not present in the current Hunspell's dictionary, therefore we have ensured about their existence online and, when they did not result to be typos, we decided to keep them.

We have taken into consideration just the statements that the Presidents have planned to delivery without any external influences provided by journalists or audience interventions. For example, it could happen that the topic of a press conference radically changes after a tricky question from a journalist. This constitutes a deviation from the idea that the President had in mind when he thought about the press conference. For this reason, in the third step, we have eliminated all the transcripts parts that come after the journalists' questions.

The process here summarized in three steps is actually wider, we prefer to refer the reader to Ficcadenti et al. (2019), where the process is described with additional details. The refined dataset is made by 951 speeches transcripts containing words stated by all the 45 US Presidents.

As introduced in Sect. 1, we utilized two sets of locutions related to economics and finance to map the speeches in their vector space. In order to do that, we did the following operation:

$$T = S \cup W$$

where S is the set containing The Economist's glossary of economics whose elements are listed in Bishop (2009), W contains the terms listed in Wikipedia contributors (2019a) and T is the set resulting from the union of the first two. It includes the salient terms employed in a context where economics and finance are treated at a scientific or journalistic level. It comprises names and surnames of relevant economists, but we have decided to replace those

bigrams with just the surnames assuming that a President more likely refers to the surname when he wants to speak about a person, especially in official talks. This assumption does not affect the frequencies of occurrence of the bigrams (names and surnames), on the contrary, it allows for a more careful accounting of occurrences. Furthermore, the terms divided by a hyphen (-) are treated as divided, so as bigrams, as well as single words in the hyphened form because there could be different transcription versions [e.g. "Most-favoured nation" has been accounted as "Most favoured nation" as well, see Bishop (2009) for the definition]. Furthermore, all the locutions originally reported in singular form were pluralized in order to take in consideration both the types; the acronyms have been treated in their extended forms, e.g. "OECD" became "Organization for Economic Cooperation and Development".

At this point, the number of elements in T is almost doubled with respect to what is resulted from the original union; we grouped the resulting content in different classes on the bases of the number of words that composes the locutions. For example, the word "Tax" has been put into the group of terms made by single words, while the locution "Organization for Economic Cooperation and Development" falls in the set where the expressions are made by six words. We adopted this subdivision because we want to count the occurrences of the terms in the speeches transcripts without losing or double counting any of them. But, for example, the word "Tax" is contained in locutions like "Tax rate" or "Tax avoidance". Therefore, if one looks for "Tax" into the speeches transcripts before of looking for "Tax rate", he/she will introduce a bias because the frequency of occurrence of "Tax" will include the frequency of occurrence of "Tax rate". To avoid this, we have grouped the locutions on the bases of their length and we got seven groups (indeed, the longest elements in T are made by seven words and the shortest are made by one word). Then, we have looked for the presence of such expressions in the US Presidents speeches firstly accounting for the frequencies of the longest expressions and eliminating them after that. The absolute frequencies of the $T's$ elements are stored in a table and are used to compute the respective relative frequencies at speech level (absolute frequencies dived by speech length for each one). The locutions that do not occur at least once are eliminated and the occurrences of locutions for plural and singular versions have been summed. Finally, the number of elements in T results to be 383. Therefore, the speeches are mapped in a vector space made by 383 dimensions.

The sum of the relative frequencies of $T's$ elements for each transcript can be considered as a proxy of the economics and finance content of the speeches. It represents the proportion of a talk devoted to the aforementioned locutions. This measure has a small bias by definition because the frequencies of the economic words might not have one as an upper bound but the effect of this tiny distortion does not affect the key features of the analysis. In Fig. 1 each dot exhibits the summed relative frequencies of all the terms belonging to T in the respective speech. It is possible to see that some points fall on zero, they represent three talks where none of the $T's$ locutions occurred:

- George W.Bush, Final Press Conference—12/01/2009 (Miller Center 2019a);
- James Madison, Proclamation of Day of Fasting and Prayer—09/07/1812 (Miller Center 2019c);
- James Madison, Proclamation of a State of War with Great Britain—19/06/1812 (Miller Center 2019b).

These speeches have been removed.

Finally, we organize the collected occurrences in a matrix **I** having 948 rows (one for each speech) and 383 columns (one for each term of the Glossary of Economics terms resulting in T). It contains zero when the locution j does not occur into the speech i, while it contains the relative frequencies of occurrence if the locution j appears in the speech i.

4 The network model

We describe the system of the US Presidents' speeches as a weighted complete network $G = (V, E)$, where the V is the set of the nodes and E collects the edges. The cardinality of V is n while the cardinality of E is m.

Each node is a speech. The connection between two speeches $i, j \in V$ is weighted by the cosine similarity w_{ij}:

$$w_{ij} = \frac{\boldsymbol{x}_i \cdot \boldsymbol{x}_j}{||\boldsymbol{x}_i|| \cdot ||\boldsymbol{x}_j||}$$

where $\boldsymbol{x}_k$ represents the vector of absolute frequencies of the economic words in the speech k, for each $k = 1, \ldots, n$. High values of w_{ij} indicate high similarity in terms of economic content among a couple of speeches, whilst low values of w_{ij} indicate low similarity. The weights w_{ij} are collected in a matrix $\mathbf{W}$ of dimension 948 by 948. We assume that $w_{kk} = 0$, for each k, so that loops are not allowed.

In order to analyse the network, as common practice in networks of correlations such as financial networks (Namaki et al. 2011) and brain networks (Bullmore and Sporns 2009), we perform a thresholding of links which are not relevant in terms of strength of association between two speeches.

The thresholding of similarity matrices for filtering out relevant connections employs various methods, somewhat more principled than the use of an arbitrary threshold, whose choice depends on the considered raw data and on eventual information regarding their structure and composition. Such methods are based either on the analysis of the statistical significance of the weights or on the detection of an eventual hierarchical structure of the data. In the first case thresholds derive from analytical arguments (for instance when the underlying data are Gaussian time series), or from permutation tests. While in the second case, methods searching for minimum spanning trees are employed (Battiston et al. 2010). For building the network in the case of presidential speeches, we meet a lack of an underlying data structure such as a set of time series and we do not have any (a priori) signals regarding a hierarchical organization of the data. For this reason, to assess the similarity scores contained in $\mathbf{W}$, we statistically test the values w_{ij} by means of permutation tests. We perform a random reshuffle of the elements in each row of the document-term frequency matrix $\mathbf{I}$. Such a reshuffling associates each speech to a randomized set of economic words keeping the distributions of words frequencies and the amount of economic words contained in each speech. Using the reshuffling procedure, a set I_{rand} of 1000 instances of the matrix $\mathbf{I}$ is generated. For each instance in I_{rand}, the corresponding matrices $\mathbf{W}$ of cosine values are computed by performing $\binom{n}{2}$ pairwise comparisons among randomized speeches. The resulting set of 1000 cosine matrices is called W_{rand}.

Each matrix in W_{rand} is indicated as $\mathbf{W}_h$, where $h = 1, \ldots, 1000$. Called w'_{ij} the binary vector of size W_{rand} whose entries are computed by applying the condition $w'_{ijh} = \begin{cases} 1 & \text{if } w_{ijh} \geq w_{ij} \\ 0 & \text{otherwise} \end{cases}$ to the elements w_{ijh} of the matrix $\mathbf{W}_h$, we compute the probability that two randomized speeches are more similar to each other than two real speeches as: $p_{ij} \frac{\sum_{h=1}^{|W_{rand}|} w'_{ijh}}{|W_{rand}|}$. If such a probability results greater than a threshold value τ, i.e. if $p_{ij} \geq \tau$, the corresponding entry, $w_{ij} \in \mathbf{W}$, is discarded since it is not considered statistically significant. The threshold is set to $\tau = \frac{\alpha}{\binom{n}{2}}$ whereas $\alpha = 0.001$ and the coefficient $\binom{n}{2}^{-1}$ is the Bonferroni correction for multiple comparisons (Miller 1981), that in our case is represented by the number of tested links. The network resulting from such a procedure, in which only
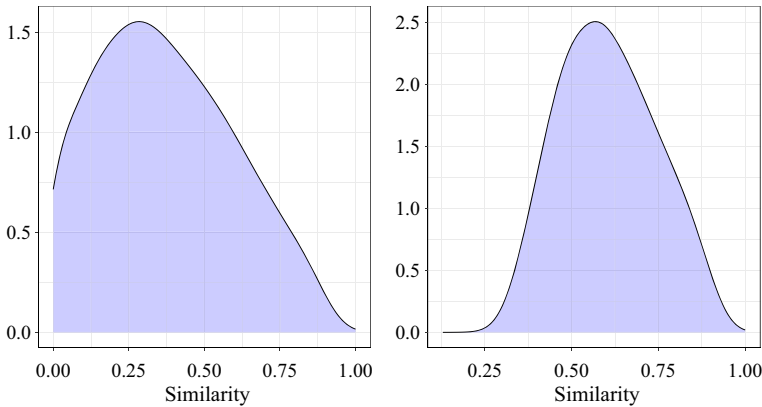
Fig. 2 Probability density function (pdf) of cosine similarity of economically relevant words in presidential speeches. Left: pdf of the $\binom{n}{2}$ cosine similarity values obtained via pairwise comparisons. Right: pdf of statistically significant cosine similarity values. (Color figure online)

statistically significant links are kept, is sometimes called Bonferroni network (Tumminello et al. 2011). Additionally, it is worth noting that even if the Bonferroni correction is very conservative the resulting network G_τ is not that different in terms of density, than the network G_α that one would obtain without the Bonferroni correction. Indeed, recalling that the network density $d = \frac{2m}{n(n-1)}$, we have $d_\tau = 0.342$ and $d_\alpha = 0.4$. The empirical distributions of cosine similarity values G (without any corrections) and G_α are reported in Fig. 2.

The resulting network, after the removal of six nodes disconnected from the largest connected component, has $n = 942$ nodes, $m = 153677$ links and it contains only significant relationships of similarity. We investigate the structure of such a network in order to understand if the associations among presidential speeches outline a peculiar network structure driven by the presence of the terms belonging to the glossary of economics and finance.

5 Results

We firstly analyse the degree and the strength distribution of the network of presidential speeches reported in Fig. 3. We observe that both empirical distributions display bimodality, meaning that the nodes could be reasonably partitioned in two different groups characterised by low/high degree and strength. Another important aspect that we take into account is the clustering of the network, i.e. the cohesiveness of triplets of nodes, by means of both global and local clustering coefficients. The global clustering coefficient $C \in [0, 1]$ measures the ratio of closed triangles to connected triples (i.e. subgraphs with three nodes and two or three links). Such a measure of global clustering can be considered both in the case of weighted and unweighted networks (Opsahl and Panzarasa 2009). The unweighted version of the clustering coefficient can be expressed in the following way:

$$C = \frac{3 \cdot n_\Delta}{n_\wedge} \tag{1}$$

where n_Δ and $n_\wedge$ are respectively the number of triangles and connected triples. We find that the global clustering coefficient is $C = 0.718$ indicating a relatively high cohesiveness of nodes, surely related to its high density. It also suggests the presence of clusters (communities) in the speeches network. Such an aspect can be further investigated by considering the local
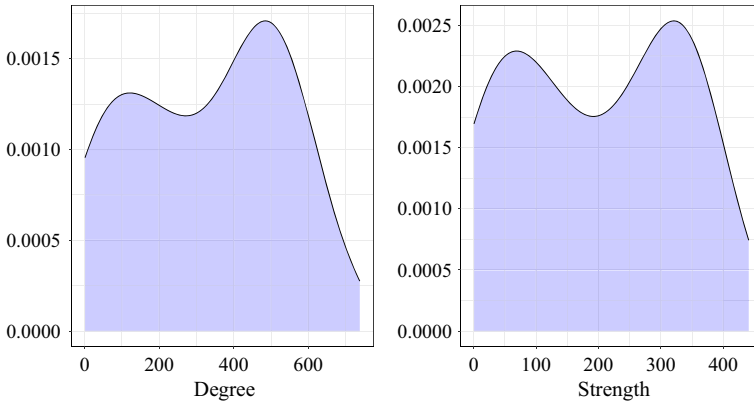
Fig. 3 Distribution of degree (left) and strength (right) of the network of presidential speeches. Both distribution display bimodality. (Color figure online)

weighted clustering coefficient (see Barrat et al. 2004) reported in Eq. 2, that considers the local cohesiveness of each node, combining the topological information with the weights distribution.[1]

$$c_i = \frac{1}{s_i(k_i - 1)} \sum_{jh} \frac{(w_{ij} + w_{ih})}{2} a_{ij} a_{ih} a_{jh} \tag{2}$$

In Eq. 2, w_{ij} is the cosine similarity between speeches i and j (as defined in Sect. 4), $s_i = \sum_j w_{ij}$ is the strength of the node i and $k_i = \sum_j a_{ij}$ is the degree of the node i. The matrix **A**, whose elements can be referred as a_{ij}, is the binary (unweighted) version of the matrix **W**. The local weighted clustering coefficient $c_i \in [0, 1]$ groups the structure of the neighbourhood of each node (in terms of connected triplets) with the intensity of connections in the neighbourhood, expressed in terms of links weight.

By plotting the complementary cumulative distribution function (ccdf) of the local weighted clustering coefficient i.e. $P(x > X) = 1 - P(x \leq X)$, displayed in Fig. 4, we observe how a relatively high proportion of the nodes in the network displays a high value of c_i thus indicating the presence of remarkably clustered neighbourhoods. Such an evidence can be associated with a peculiar arrangement of the links weights for two main reasons. First, the average local clustering coefficient in its unweighted form (obtained setting $w_{ij} = cost$ in Eq. 2) is slightly lower, yet similarly distributed as shown in the inset of Fig. 4, than the weighted local clustering coefficient, namely $\overline{c}(w_{ij} = cost) = 0.74$ while $\overline{c} = 0.75$. Second, the ccdf of the weighted clustering coefficient for the actual network is right-shifted with respect to the ccdf curve of a null distribution associated to such a clustering coefficient obtained from 100 networks with the same topology, but reshuffled edge weights. Beyond clustering, another interesting quantity is represented by the assortativity coefficient $r \in [-1, 1]$ (Newman 2003) that measures to which extent similar nodes tend to be interconnected relatively to the expected proportion of links under a null model called configuration model. The assortativity coefficient can be computed in the case of scalar node attributes. It can be structural (e.g. the nodes degree) or non-structural (e.g. the year of a certain speech) as in Eq. 3,

$$r = \frac{\sum_{ij} \left(a_{ij} - \frac{k_i k_j}{2m}\right) x_i x_j}{\sum_{ij} \left(a_{ij} x_i^2 - \frac{k_i k_j}{2m} x_i x_j\right)} \tag{3}$$

---

[1] It is worth noting the alternatives presented in Onnela et al. (2005) as well as the extended versions of the clustering coefficient in Fagiolo (2007), Clemente and Grassi (2018).
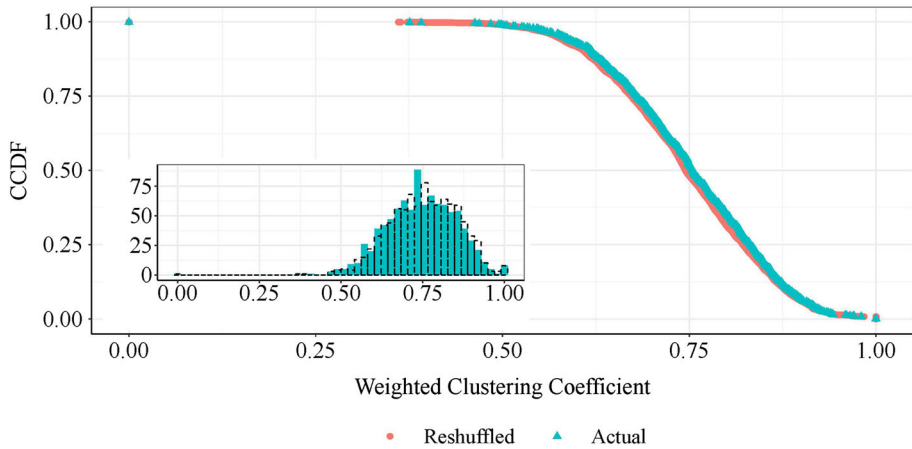
Fig. 4 Complementary cumulative distribution function of the weighted local clustering coefficient for the actual network of presidential speeches and for the an ensemble of 100 networks with reshuffled weights and same topology of the actual one. The inset reports the distribution of the local clustering coefficient in its weighted (blue histogram) and unweighted version. (Color figure online)

where $\mathbf{x}$ is the n-sized vector of scalar features whose elements are x_i and x_j, m represents the number of nodes, k_i is the degree of the node i and a_{ij} are the elements of the matrix $\mathbf{A}$, binary version of the matrix $\mathbf{W}$. The assortativity can also be computed in the case of categorical nodes' attributes (e.g. the political affiliation of the president who gave the speech) as in Eq. 4

$$r = \frac{\sum_{ij} \left(a_{ij} - \frac{k_i k_j}{2m}\right) \delta(f_i, f_j)}{2m - \sum_{ij} \frac{k_i k_j}{2m} \delta(f_i, f_j)} \tag{4}$$

where $\mathbf{f}$ is the n-sized vector of categorical features and $\delta(f_i, f_j)$ is the Kronecker delta function of the elements f_i and f_j which are components of $\mathbf{f}$. Therefore, the assortativity coefficient, similarly to a network-based version of Pearson's correlation coefficient (Noldus and Van Mieghem 2015), provides us with a value that quantifies the tendency of similar nodes to be interconnected. When we observe a positive value of r we say that the network is assortative, meaning that similar nodes are interconnected while when we observe a negative value of r we say that the network is disassortative meaning that diverse nodes are interconnected. Interestingly, we could observe very different values of r depending on the attribute that we take into account.

The Presidents' speeches network displays a value of degree assortativity $r_{degree} = 0.04$ indicating no particular mixing to degree while it displays a value of $r_{strength} = 0.15$ meaning that nodes with high strength tend to be connected to other high strength nodes while low strength nodes tend to be connected to other low strength ones. Additionally, we observe $r_{date} = 0.177$, $r_{party} = 0.009$ and $r_{president} = 0.008$ meaning that we observe noticeable associations for what concerns the economics terms presence, among speeches close in time, while we do not observe any particular association among speeches given by Presidents affiliated to the same party or even by the same President. Beyond aspects concerning nodes' attributes, the combination of a high clustering coefficient and a positive strength assortativity coefficient let us room for further investigations related to the presence of communities (Newman and Girvan 2004) or other higher order framework such as core-periphery structure
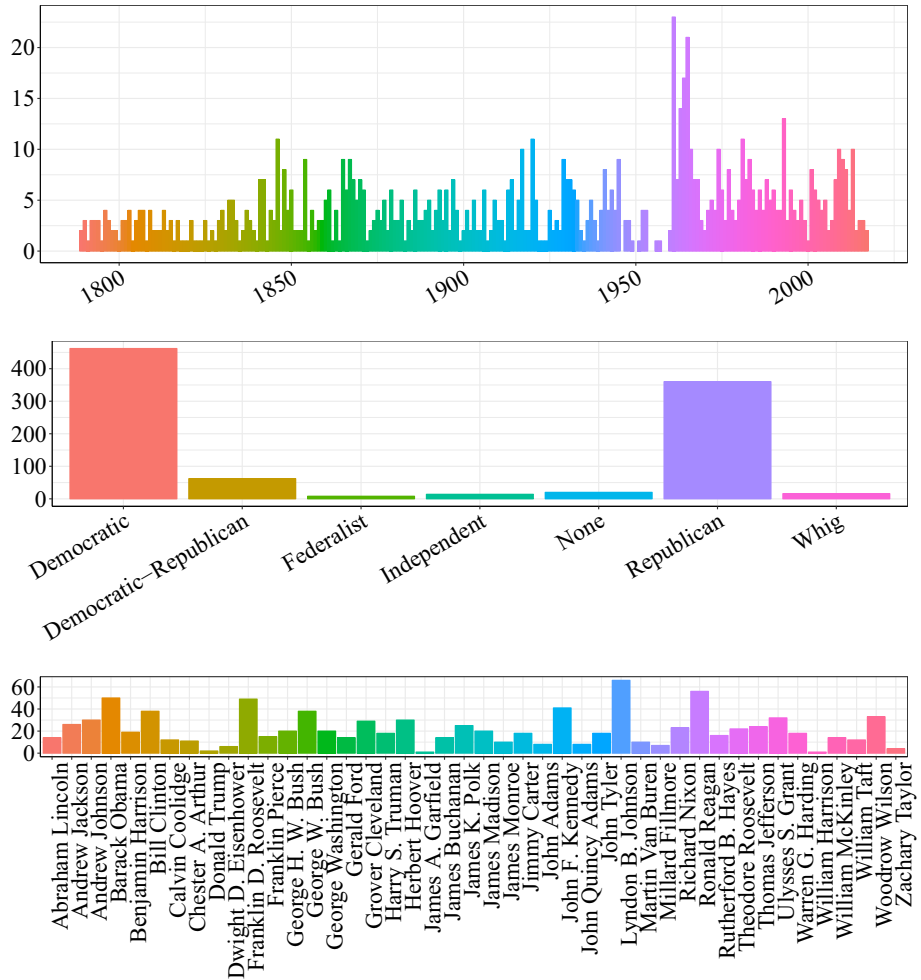
Fig. 5 Bar charts of date, party and president's name related to each speech in the network. (Color figure online)

(Borgatti and Everett 2000). When the network is divided into communities, we observe subgraphs whose nodes have a higher probability to be linked to the nodes of the subgraph than to any other nodes of the network. When we observe a core-periphery structure, the network topology allows for the partitioning into a set of central and densely connected nodes (the core) and a set of non-central and sparsely connected nodes (the periphery).

We retrieve the community structure of the network by means of community detection algorithms such as Walktrap (WT), Pons and Latapy (2005); label propagation (LP), Ragha-van et al. (2007), spectral partitioning (LE), Newman (2006); Louvain (L) algorithm, Blondel et al. (2008); Infomap (IM), Rosvall and Bergstrom (2008) and hierarchical aggregation (FG), Clauset et al. (2004). We compare the obtained partitions with the Adjusted Rand index ARI $\in [-1, 1]$ (Hubert and Arabie 1985), which measures the agreement of such partitions in terms of assignment of nodes into communities, known that ARI = 1 indi-
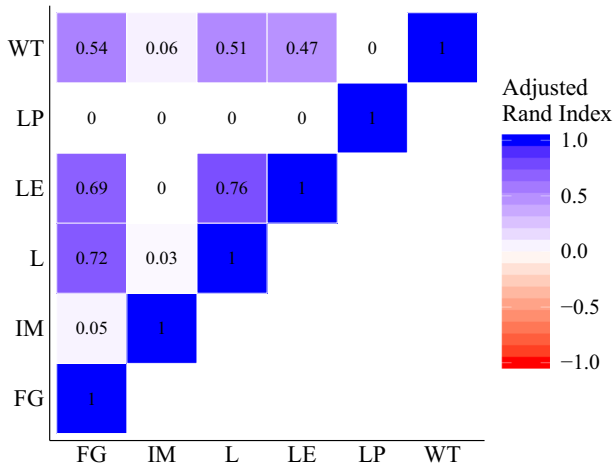
Fig. 6 Values of the Adjusted Rand Index for six different methods of community detection. (Color figure online)

cates perfect agreement and ARI $= -1$ indicates perfect disagreement. The inconsistency among different values of ARI displayed in Fig. 6 indicates that there is no overall agreement among the different methods. Therefore, it results in hard to interpret the outcomes. Such a disagreement is also due to the fact that not all the algorithms share the same objective function. For instance IM maximises the so-called map equation (Rosvall and Bergstrom 2008) while other algorithms maximise modularity Q, a quality function that is analogous to the numerator of Eq. 4. Additionally, certain community detection algorithms are normally applied to sparse networks, therefore our outcome may suffer because it comes from the analysis of dense networks. The evidence from Fig. 6, together with the high clustering indicators, assortativity and the slightly bimodal degree and strength distributions suggest the presence of another higher order structure that can further characterize the structure of the considered network. In particular, a positive value of strength assortativity with a noticeable degree of heterogeneity suggests the presence of a core-periphery structure called rich-club (Zhou and Mondragón 2004). A rich-club is observed in a network when the nodes with the highest degree are tightly interconnected in order to form a dense subgraph. The presence of a rich-club is measured through the rich-club coefficient $\phi(k) \in [0, 1]$ which measures the density of the subgraph made of nodes with degree $d > k$. The concept of rich-club can be easily extended to measures beyond degree and to weighted networks using appropriate null models for each of the cases, e.g. Opsahl et al. (2008), Cinelli et al. (2018); Cinelli (2019). In this one, we assess rich-club ordering in the case of node strength and we measure the density of connections among nodes with the highest strength $\phi(s)$. The value of $\phi(s)$ is compared against its average value $\bar{\phi}(s)$ across an ensemble of 100 networks with the same topology but reshuffled edges weights. When the ratio $\phi(s)_{norm} = \frac{\phi(s)}{\bar{\phi}(s)} > 1$ the network is said to display rich-club ordering. It is worth noting that the null model that we are taking into account preserves the topology while reshuffling the edges weights; it means that the nodes strength distribution is not preserved across randomized networks. For this reason, as explained in Cinelli (2019), we rank the nodes by increasing strength in each of the considered networks and we measure the density of connections among nodes whose rank is higher than a value $p \in [1, n]$. Accordingly to the explanation given for $\phi(s)$ we computed the index $\phi(p)$ and its normalized version $\phi_{norm}(p)$.
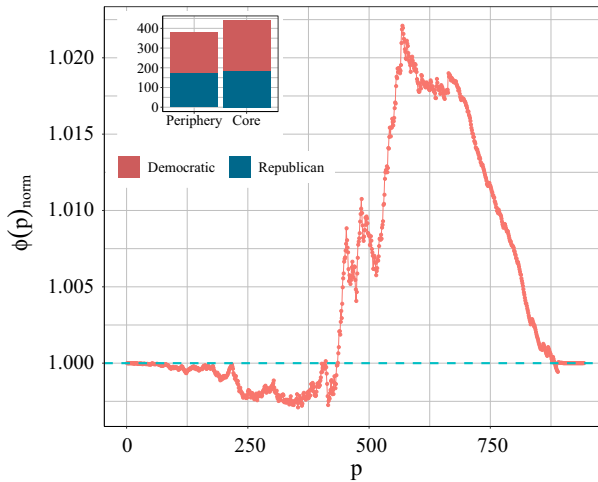
Fig. 7 Normalized rich-club coefficient of the presidential speeches network. Rich-club ordering is observed for values of $\phi(p)_{norm} > 1$. The inset reports the number of speeches inside and outside the rich-club considering also considering the affiliation of the speaker. For the sake of representation only affiliations to Democratic and Republicans have been reported. (Color figure online)

By computing the rich-club coefficient for each value of p, we obtain a curve of the density of the considered subgraph which equals one when such graph is complete.

From Fig. 7 we observe that the network of presidential speeches displays a rich-club ordering and that it enters in the so-called rich-club regime (i.e. $\phi(p)_{norm} > 1$) for a value of $p \sim 430$. Additionally, by plotting the amount of speeches given by Presidents affiliated to the two main parties in the US, we note how the network nodes, both inside and outside the rich-club, display a balanced proportion of political affiliations, meaning that the partition of speeches tends to resemble the economic content more than other political aspects.

6 Discussion

Our analysis shows the relevance of the terms belonging to the glossary of economics and finance in connecting the Presidents' speeches between each other. We employ a text mining-network analysis approach to measure the similarity of the speeches by using the aforementioned glossary.

In Sect. 5 we note high levels of clustering coefficients at local and global level that indicate the capacity of the speeches of being organized in communities. The relationships between speeches are quite complete thanks to the presence of many triangular schemes. This aspect confirms the idea that the analysed glossary of economics is recurrent in the political debate and it constitutes a significant component of the connection between Presidents' talks. Looking at Fig. 1, one can find further confirmations of this fact by noticing the high density of the points in certain periods as well as the quite stable presence of the economics locutions along the years. The talks interconnections and the presence of local triangles bridged with other closed triangular relationships are indications of the use of a shared terminology within local communities of speeches. It might be related to the type of phenomena faced by the speakers during different socio-economic periods (e.g. recession periods characterized by high unemployment rate or crises leaded by particular sector of the economy).

The assortativity analysis allows for the determination of the features that can be relevant in justifying the relationships between talks. The structural assortativity, measured by using the strength of the nodes, reveals that speeches characterized by close levels of similarity tend to be closer. This offers further hints on the existence of sets of talks that share the utilization of some economics terms picked from our glossary. They might be the State of the Union or the Inaugural Addresses for example. During these type of talks, the Presidents usually devote part of their speech in commenting the economic situation of the country, sometimes involving a more specific/technical analysis of the indicators (see Rule et al. 2015; Light 2014). We check for the influence of the political parties affiliations of the Presidents, the years in which the speeches have been stated and the speakers as discriminant features to identify the clusters. The strongest contribution is provided by the years of the talks, indeed the assortativity for that case manifests noticeable relationships between talks belonging to the same years. Namely, public communications that are close in time tend to be similar for what concerns the employed economics terms. As an example, the devastating crisis of 1929 was the starting point of a turbulent period of high instability and the US economy and financial conditions have been at the center of the debate for long time. Indeed, during the years 1929 - 1933 the Presidents have spoken a lot about the crises (see Fig. 9), therefore we expect speeches occurred at that time to be closer with respect to those stated during war periods. Another example is given by the more recent financial crisis started in 2008 (see Fig. 9).

The low level of assortativity that emerges when we test the contribution of Presidents (so the relevance of the speakers in clustering the speeches) and their affiliations to parties are partially justified by a low level of politicization in the usage of economics and financial terminology. Furthermore, the Presidents usually utilize the political rhetoric to accentuate the government difficulties coming from exogenous factor to exalt their achievements or to manifest awareness of the country challenges. This is particularly true for Presidents that have faced crises as well as for the Presidents that had the responsibility of leading the US after a tough period. This phenomenon generates reference to the economy and finance in the US Presidents' speeches.

In addition, if one looks at changes of governments occurred during or immediately after the toughest US crises, e.g. *Great Depression* and the *Great Recession*, it is possible find additional justifications for the low levels of assortativity by Presidents and their affiliations. Indeed, these two crises have involved the efforts of more than one government and President. Specifically, the *Great Depression* occurred during the Republican Presidency of Hoover has produced aftermath during the successive Presidency of the Democrat Roosevelt. While the *Great Recession* has first involved the Republican Bush and then the Democrat Obama. It means that, during the period in which there are picks of economics words usages (crisis or booms periods for example), economics and finance are addressed by different Presidents and therefore different parties. Furthermore, it confirms that during a mandate, the way of using economics terms is common among Presidents and parties.

Certain events have such a disruptive power that the whole society pays attention to them. Therefore, in speaking about these facts, the parties division does not matter; the public debate embeds such events regardless of the storytellers' identity. Moreover, the President is the most important political landmark in the US, therefore each time something relevant happens in the country, he has a sort of institutional obligation to speak about it, regardless his political affiliation or his sensibility. Consequently, the lexicon used by different Presidents does not change much, especially considering that our glossary is made up of a terminology useful to identify concepts (nouns for example) more than sentiments or adjectives associated to them.

Fig. 8 Wordcloud for the rich-club (left) and the rest of the network (right) obtained considering two partitions of the matrix I. The size of the words is proportional to their frequency. In the top panels we consider the entire set of words while in the bottom panels we cut out words with very high frequencies. We observe that while the main economic concepts are common to both the rich-club and the rest of the network, less frequent concepts differ between the two groups. (Color figure online)

The talks particularly devoted to economics and finance are probably characterized by the contribution of technicians and ghostwriters with high knowledge of the economy. It means that the Presidents attitude are not so manifested in these cases. In addition, as said, the glossary here used does not allow to capture contributions coming from ideologies or different visions of the economic system. This is expected to be an important feature for discriminating the speeches on the bases of the parties for example.

Ultimately, the assortativity by party can be affected by the unbalanced proportion of speaker affiliations to different political parties (we have more speeches stated by democrats than republicans, see Fig. 5—mid). While the distributions of speeches by President and by dates are relatively homogeneous, as displayed in Fig. 5 (bottom and top). However, as suggested in Newman (2003), we have to carefully consider these results given that in networks with many attributes, disassortative mixing tends to resemble a random assignment of node attributes since, in presence of several attributes, then random mixing will most often pair unlike nodes.

The analysed network results to be relatively dense despite the threshold applied to links for avoiding scarce statistical significant links. Such density value reflect the idea of having a glossary of economics particularly popular among the Presidents. But, at the same time, it may create noise when community detection algorithms are applied (see the disagreement between methodologies in Fig. 6).

The research of a core-periphery structure grounds on the positive and remarkable assortativity based on nodes' strength and on the bimodal distribution of the nodes' degree and strength. Therefore, we look for a core of strongly interconnected speeches in which there is a regular presence of common terms referred to economics and finance. Figure 7 confirms the presence of a rich-club (when $\phi(p)_{norm} > 1$) for $p \sim 430$. Consequently, we can state that the discriminant between the two groups should be reflected by the presence of two sets of economics terms, which drive the differences of similarity regime. They can be represented by words clouds as displayed in Fig. 8. The two clouds at the top of the figure show similar words occurrences (dimension of the font), therefore the nodes' strength within the core and the periphery has to be driven by the presence of locutions reported in the two clouds shown at the bottom of Fig. 8. Indeed, the bottom clouds contain the words less frequently occurred within the two groups. Namely, the core and the periphery share a common set of words and the occurrence of more marginal terms conditions the belonging to the core or to the periphery of a speech.

Concluding, the network results to be divided into a core of speeches where nodes with the highest strength are connected to each other and a periphery with the opposite characteristic. This outcome implicates the aforementioned discriminatory behaviour of some terms' frequency distribution. So, given a core of words regularly present among the speeches, the dimensions of the vector space devoted to bring/eject a speech in/from the rich-club community is provided by a relatively high presence of terms characterizing a contingent situation in which the talk has been stated. For example, the word "Manufacturing" has a notable presence in the set made by speeches belonging to the rich-club but, in contrast, it disappears from the rest of the network (bottom Right cloud of Fig. 8). On the other hand, the bottom blue cloud shows the presence of "Innovation" in the peripheral network but not in the core (bottom left cloud of Fig. 8).

The core-periphery structure can be interpreted under the prospective shown in Fig. 9. Namely, by comparing the percentage of economics words present into the speeches stated during recession periods versus the presence of such words in the talk delivered during the rest of the time (see Wikipedia contributors 2019b, for the classification of crisis period in the US history). Figure 9 clearly shows two regimes with a change between 1940 and 1960. The number of crisis has diminished during the last century, therefore the number of speeches delivered during the recession has slowed down. Furthermore, by looking at the last crisis (2008 - *Great Depression*), it is possible to notice few big red points. The fact that during the last crisis the Presidents have used fewer economics terminology than before can be attributed to the need of modern leaders of avoiding media identification with turbulent periods to maintain low association with unpleasant events. However, the red points manifest lower presence of economics words and seems to be located in the past, while the blue points are more concentrated during recent years and they seem to have a higher mean. This has certainty conditioned the similarity, contributing to the creation of a core-periphery structure.
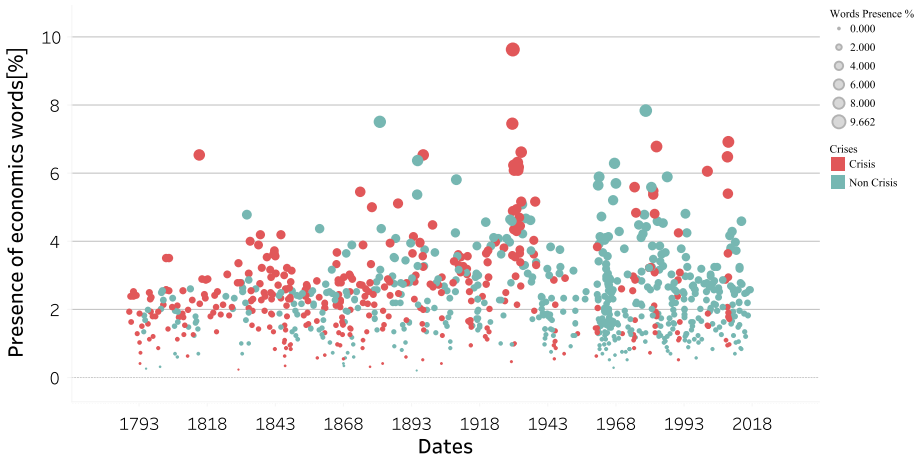
Fig. 9 Percentage of the economics terms per speech along the years divided by period of recession. Such a division has been taken from the Wikipedia's list of crises, see Wikipedia contributors (2019b). The red indicates that the speech has been stated during a period of recession, while blue indicates a non recession period. (Color figure online)

7 Conclusions

In this study we have designed a procedure made by a combination of text mining and network analysis techniques. The tools employed to perform the analysis are well established in the literature (see Sect. 2) but their combinations and the application in this context constitutes - at best of our knowledge - a unique case. The designed approach is general and we believe that the usage of it in other fields can provide promising results.

Here we focused on economics and finance topics that are recurring in political discussions. Indeed, Politicians have a preponderant role in deciding the fiscal and monetary policy; their actions impact the economy of the country as well as the financial sector. The Presidents' public communications related to economics and finance are designed to reach certain listeners and to spread different type of messages on the bases of contingent situations, the audience and political objectives to be reached. Anyway, the 45 Presidents have referred to the economy and financial sector at least once in their political life.

This study highlights the relevance of the economic jargon employed by the US Presidents since the foundation of the country. Indeed we have found a quite stable presence of the terms belonging to the glossary of economics in the talks; see Fig. 1 for a visual inspection of this fact.

The network analysis approach allows to detect communities of talks characterized by the economics terms employed, hence on the bases of their economic content similarity. From such an investigation three main results are derived:

- The US Presidents speeches share the use of a core dictionary referred to economics and finance, see Fig. 8 to see some examples.
- The speeches clusterization based on the cosine similarity network lead to a core-periphery structure. Namely, the 948 analysed speeches are divided in two sets, one made by stronger connected speeches and another with lighter edges.
- The words that lead the division in a core-periphery structure are those peculiar of certain events, namely the terms used to explain some local phenomena, see the differences between the locutions present in the bottom clouds of Fig. 8.

Concluding, the analysis of Fig. 9 and the comments to the network clustering indicators provide clues about the potential explanations for such a core-periphery structure. Indeed, we consider plausible to hypothesize that the core-periphery setup is linked to the presence of two regimes: one for the speeches stated in the older critical periods and another made by speeches more recently stated during non-recession periods.

These results throw the basis for further researches. For example, the causes for the utilization of certain terms might be further investigated and they might confirm the connections with events like crises. The timing for the presence of some terms can be taken into consideration as well. Indeed, the combination of times and terminology resulted to be the main elements related to a core-periphery structure of the network. Finally, the sentiments associated with the terms of the economics glossary can be object of study. We believe that in this way it would be possible to determine the impact of the Presidents' party affiliation on the network.

References

Agarwal, A., Gupta, A., Kumar, A., & Tamilselvam, S. G. (2019). Learning risk culture of banks using news analytics. *European Journal of Operational Research*, *277*, 770–783.

Alfaro, C., Cano-Montero, J., Gómez, J., Moguerza, J. M., & Ortega, F. (2016). A multi-stage method for content classification and opinion mining on weblog comments. *Annals of Operations Research*, *236*, 197–213.

Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards. *The Journal of Finance*, *59*, 1259–1294.

Bail, C. A. (2016). Combining natural language processing and network analysis to examine how advocacy organizations stimulate conversation on social media. *Proceedings of the National Academy of Sciences*, *113*, 11823–11828.

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. *The Quarterly Journal of Economics*, *131*, 1593–1636.

Balakrishnan, R., Qiu, X. Y., & Srinivasan, P. (2010). On the predictive ability of narrative disclosures in annual reports. *European Journal of Operational Research*, *202*, 789–801.

Bao, Y., & Datta, A. (2014). Simultaneously discovering and quantifying risk types from textual risk disclosures. *Management Science*, *60*, 1371–1391.

Barrat, A., Barthelemy, M., Pastor-Satorras, R., & Vespignani, A. (2004). The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences*, *101*, 3747–3752.

Battiston, S., Glattfelder, J. B., Garlaschelli, D., Lillo, F., & Caldarelli, G. (2010). The structure of financial networks. In *Network science* (pp. 131–163). Springer.

Bernauer, J., & Bräuninger, T. (2009). Intra-party preference heterogeneity and faction membership in the 15th German Bundestag: A computational text analysis of parliamentary speeches. *German Politics*, *18*, 385–402.

Bishop, M. (2009). *Essential economics: an A to Z guide* (Vol. 22). Hoboken: Wiley.

Blasco, N., Corredor, P., Del Rio, C., & Santamaría, R. (2005). Bad news and Dow Jones make the Spanish stocks go round. *European Journal of Operational Research*, *163*, 253–275.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *10*, 10008.

Borgatti, S. P., & Everett, M. G. (2000). Models of core/periphery structures. *Social Networks*, *21*, 375–395.

Bullmore, E., & Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, *10*, 186.

Callon, M., Courtial, J. P., & Laville, F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemisitry. *Scientometrics*, *22*, 155–205.

Cannon, B. J., Nakayama, M., Sasaki, D., & Rossiter, A. (2018). Shifting policies in conflict arenas: A cosine similarity and text mining analysis of Turkey's Syria policy, 2012–2016. *Journal of Strategic Security*, *11*, 1–19.

Carretta, A., Farina, V., Martelli, D., Fiordelisi, F., & Schwizer, P. (2011). The impact of corporate governance press news on stock market returns. *European Financial Management*, *17*, 100–119.

Chae, B. K. (2015). Insights from hashtag #supplychain and Twitter analytics: Considering Twitter and Twitter data for supply chain practice and research. *International Journal of Production Economics*, *165*, 247–259.

Cheng, M., & Jin, X. (2019). What do Airbnb users care about? An analysis of online review comments. *International Journal of Hospitality Management*, *76*, 58–70.

Cinelli, M. (2019). Generalized rich-club ordering in networks. *Journal of Complex Networks*. https://doi.org/10.1093/comnet/cnz002.

Cinelli, M., Ferraro, G., & Iovanella, A. (2018). Rich-club ordering and the dyadic effect: Two interrelated phenomena. *Physica A: Statistical Mechanics and its Applications*, *490*, 808–818.

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

Clemente, G. P., & Grassi, R. (2018). Directed clustering in weighted networks: A new perspective. *Chaos, Solitons & Fractals*, *107*, 26–38.

Cochran, J. J., Curry, D. J., Radhakrishnan, R., & Pinnell, J. (2014). Political engineering: optimizing a US Presidential candidate's platform. *Annals of Operations Research*, *215*, 63–87.

Fagiolo, G. (2007). Clustering in complex directed networks. *Physical Review E*, *76*, 026107.

Felici, G. (1995). Talking to Sibilla: An approach to context dependent natural language comprehension. *European Journal of Operational Research*, *85*, 263–281.

Feuerriegel, S., & Gordon, J. (2018). Long-term stock index forecasting based on text mining of regulatory disclosures. *Decision Support Systems*, *112*, 88–97.

Feuerriegel, S., & Gordon, J. (2019). News-based forecasts of macroeconomic indicators: A semantic path model for interpretable predictions. *European Journal of Operational Research*, *272*, 162–175.

Ficcadenti, V., Cerqueti, R., & Ausloos, M. (2019). A joint text mining-rank size investigation of the rhetoric structures of the US Presidents' speeches. *Expert Systems with Applications*, *123*, 127–142.

Garcia, D. (2013). Sentiment during recessions. *The Journal of Finance*, *68*, 1267–1300.

Groth, S. S., & Muntermann, J. (2011). An intraday market risk management approach based on textual analysis. *Decision Support Systems*, *50*, 680–691.

Hendershott, T., Livdan, D., & Schürhoff, N. (2015). Are institutions informed about news? *Journal of Financial Economics*, *117*, 249–287.

Huang, K. W., & Li, Z. (2011). A multilabel text classification algorithm for labeling risk factors in SEC form 10-K. *ACM Transactions on Management Information Systems (TMIS)*, *2*, 18.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*, 193–218.

Ji, Z., Pi, H., Wei, W., Xiong, B., Woźniak, M., & Damasevicius, R. (2019). Recommendation based on review texts and social communities: A hybrid model. *IEEE Access*, *7*, 40416–40427. https://doi.org/10.1109/ACCESS.2019.2897586.

Kahveci, E., & Odabaş, A. (2016). Central banks' communication strategy and content analysis of monetary policy statements: The case of Fed, ECB and CBRT. *Procedia-Social and Behavioral Sciences*, *235*, 618–629.

Kocheturov, A., Pardalos, P. M., & Karakitsiou, A. (2019). Massive datasets and machine learning for computational biomedicine: Trends and challenges. *Annals of Operations Research*, *276*, 5–34.

Kumar, B. S., & Ravi, V. (2016). A survey of the applications of text mining in financial domain. *Knowledge-Based Systems*, *114*, 128–147.

Laver, M., Benoit, K., & Garry, J. (2003). Extracting policy positions from political texts using words as data. *American Political Science Review*, *97*, 311–331.

Lee, W. Y., Bachtiar, M., Choo, C. C. S., & Lee, C. G. (2019). Comprehensive review of hepatitis B Virus-associated hepatocellular carcinoma research through text mining and big data analytics. *Biological Reviews*, *94*, 353–367.

Li, H., Gupta, A., Zhang, J., & Flor, N. (2018). Who will use augmented reality? An integrated approach based on text analytics and field survey. *European Journal of Operational Research*. https://doi.org/10.1016/j.ejor.2018.10.019.

Light, R. (2014). From words to networks and back: Digital text, computational social science, and the case of presidential inaugural addresses. *Social Currents*, *1*, 111–129.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, *66*, 35–65.

Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, *54*, 1187–1230.

Maldonado, M., & Sierra, V. (2016). Twitter predicting the 2012 US Presidential Election?: Lessons learned from an unconscious value co-creation platform. *Journal of Organizational and End User Computing (JOEUC)*, *28*, 10–30.

Malik, M. M., Abdallah, S., & Ala'raj, M. (2018). Data mining and predictive analytics applications for the delivery of healthcare services: A systematic literature review. *Annals of Operations Research*, *270*, 287–312.

Miller, R. G. (1981). Normal univariate techniques. In *Simultaneous statistical inference, chapter 2* (pp. 37–108). Springer.

Miller Center (2019a). Final press conference. Retrieved May 25, 2019 from https://millercenter.org/the-presidency/presidential-speeches/january-12-2009-final-press-conference.

Miller Center (2019b). Proclamation of a State of War with Great Britain. Retrieved May 25, 2019 from https://millercenter.org/the-presidency/presidential-speeches/july-9-1812-proclamation-day-fasting-and-prayer.

Miller Center (2019c). Proclamation of day of fasting and prayer. Retrieved May 25, 2019 from https://millercenter.org/the-presidency/presidential-speeches/july-9-1812-proclamation-day-fasting-and-prayer.

Mishra, N., & Singh, A. (2018). Use of Twitter data for waste minimisation in beef supply chain. *Annals of Operations Research*, *270*, 337–359.

Namaki, A., Shirazi, A. H., Raei, R., & Jafari, G. R. (2011). Network analysis of a financial market based on genuine correlation and threshold method. *Physica A: Statistical Mechanics and its Applications*, *390*, 3835–3841.

Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications*, *41*, 7653–7670.

Newman, M. E. (2003). Mixing patterns in networks. *Physical Review E*, *67*, 026126.

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

Newman, M. E., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, *69*, 026113.

Ngai, E. W., Hu, Y., Wong, Y., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, *50*, 559–569.

Noldus, R., & Van Mieghem, P. (2015). Assortativity in complex networks. *Journal of Complex Networks*, *3*, 507.

Oliva, G., Scala, A., Setola, R., & Dell'Olmo, P. (2018). Opinion-based optimal group formation. *Omega*. https://doi.org/10.1016/j.omega.2018.10.008.

Onnela, J. P., Saramäki, J., Kertész, J., & Kaski, K. (2005). Intensity and coherence of motifs in weighted complex networks. *Physical Review E*, *71*, 065103.

Ooms, J. (2017). *hunspell: High-Performance Stemmer, Tokenizer, and Spell Checker*. https://CRAN.R-project.org/package=hunspell r package version 2.9.

Opsahl, T., Colizza, V., Panzarasa, P., & Ramasco, J. J. (2008). Prominence and control: The weighted rich-club effect. *Physical Review Letters*, *101*, 168702.

Opsahl, T., & Panzarasa, P. (2009). Clustering in weighted networks. *Social Networks*, *31*, 155–163.

Peruzzi, A., Zollo, F., Quattrociocchi, W., & Scala, A. (2018). How news may affect markets' complex structure: The case of cambridge analytica. *Entropy*, *20*, 765.

Pons, P., & Latapy, M. (2005). Computing communities in large networks using random walks. In P. Yolum, T. Güngör, F. Gürgen, & C. Özturan (Eds.), *Computer and information sciences—ISCIS 2005* (pp. 284–293). Springer.

Price, S. M., Doran, J. S., Peterson, D. R., & Bliss, B. A. (2012). Earnings conference calls and stock returns: The incremental informativeness of textual tone. *Journal of Banking & Finance*, *36*, 992–1011.

Raghavan, U. N., Albert, R., & Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, *76*, 036106.

Ravisankar, P., Ravi, V., Rao, G. R., & Bose, I. (2011). Detection of financial statement fraud and feature selection using data mining techniques. *Decision Support Systems*, *50*, 491–500.

Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, *105*, 1118–1123.

Rule, A., Cointet, J. P., & Bearman, P. S. (2015). Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790–2014. *Proceedings of the National Academy of Sciences*, *112*, 10837–10844.

Schonhardt-Bailey, C., Yager, E., & Lahlou, S. (2012). Yes, Ronald Reagan's rhetoric was unique-but statistically, how unique? *Presidential Studies Quarterly*, *42*, 482–513.

Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The AZFin text system. *ACM Transactions on Information Systems (TOIS)*, *27*, 12.

Shaban, T. A., Hexter, L., & Choi, J. D. (2017). Event Analysis on the 2016 U.S. Presidential Election using social media. In *International conference on social informatics* (pp. 201–217). Springer.

Sudhahar, S., Veltri, G. A., & Cristianini, N. (2015). Automated analysis of the us presidential elections using big data and network analysis. *Big Data & Society*, *2*. https://doi.org/10.1177/2053951715572916.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, *62*, 1139–1168.

Tsai, M. F., & Wang, C. J. (2017). On the risk prediction and analysis of soft information in finance reports. *European Journal of Operational Research*, *257*, 243–250.

Tumminello, M., Miccichè, S., Lillo, F., Piilo, J., & Mantegna, R. N. (2011). Statistically validated networks in bipartite complex systems. *PLoS ONE*, *6*, 1–11.

Vargo, C. J., Guo, L., McCombs, M., & Shaw, D. L. (2014). Network issue agendas on Twitter during the 2012 US presidential election. *Journal of Communication*, *64*, 296–316.

Wei, Y. M., Mi, Z. F., & Huang, Z. (2015). Climate policy modeling: an online SCI-E and SSCI based literature review. *Omega*, *57*, 70–84.

Wikipedia contributors (2019a). *Glossary of economics—Wikipedia, the free encyclopedia*. Retrieved May 25, 2000, from https://en.wikipedia.org/w/index.php?title=Glossary_of_economics&oldid=898838737.

Wikipedia contributors (2019b). *List of recessions in the united states—Wikipedia, the free encyclopedia*. Retrieved May 25, 2000, from https://en.wikipedia.org/w/index.php?title=List_of_recessions_in_the_United_States&oldid=891896869.

Wu, X., Cao, Y., Xiao, Y., & Guo, J. (2018). Finding of urban rainstorm and waterlogging disasters based on microblogging data and the location-routing problem model of urban emergency logistics. *Annals of Operations Research*. https://doi.org/10.1007/s10479-018-2904-1.

Yuan, H., Xu, W., Li, Q., & Lau, R. (2018). Topic sentiment mining for sales performance prediction in e-commerce. *Annals of Operations Research*, *270*, 553–576.

Zhou, S., & Mondragón, R. J. (2004). The rich-club phenomenon in the internet topology. *IEEE Communications Letters*, *8*, 180–182.

Affiliations

Matteo Cinelli[1] · Valerio Ficcadenti[2,3] · Jessica Riccioni[2]

Matteo Cinelli
matteo.cinelli@infn.roma1.it

Jessica Riccioni
j.riccioni@studenti.unimc.it

[1] Applico Lab, CNR-ISC, Via dei Taurini 19, 00185 Rome, Italy

[2] Department of Economics and Law, University of Macerata, Via Giovanni Mario Crescimbeni 14, 62100 Macerata, Italy

[3] London Centre for Business and Entrepreneurship Research, London South Bank University, 103 Borough Road, London SE10AA, United Kingdom