# The 2018 Brazilian presidential run-off: a complex network analysis approach using Twitter data

Juliano E. C. Cruz[a] and Marcos G. Quiles[b]

[a]Applied Computing Postgraduate Program, National Institute for Space Research (INPE), São José dos Campos, Brazil

[b]Federal University of São Paulo (UNIFESP), São José dos Campos, Brazil

**Abstract**

The political role of social networks has increased significantly in recent years. It has shifted from simple data analytics, employed to understand characteristics and desires of social groups, to a crucial digital platform for political campaigns and social influence. Here, we present an exploratory study of social data gathered from Twitter during the second round of the 2018 Brazilian Presidential Election. The study used complex network analysis to identify hidden communication patterns, important features, and key actors within Twitter data. The results were used to better understand how political polarization is embedded into Twitter during presidential campaigns. The dataset was collected in the 2018 Brazilian presidential run-off campaign and was composed of 1 million tweets in several languages, but predominantly in Portuguese. Moreover, the model also provided a straightforward manner to scrutinize the data and infer hypotheses from the obtained patterns. In summary, the number of communities of the network revealed an unbalanced division between the candidates and a lack of moderate behaviors.

**Keywords**: complex network, elections, Twitter, natural language processing, NLP, social network analysis, SNA.

## 1 Introduction

Social media has an important role in political discourse worldwide. A considerable part of the global population uses Internet to read about general and political news. Many people who intend to have a more active role in politics, use social networks, blogs, or other online communities [11]. A smaller group ends up organizing themselves in private communities, in which most members have homogeneous think alike and usually are aligned

to the same party, political view, or ideological currents [11]. Nowadays, one of the most popular social media for politicians is Twitter, which allows both direct and public communication with their supporters or ordinary citizens.

Brazil had become increasingly polarized starting with the 2014 presidential elections, oscillating during four years that followed and reaching a peak during the 2018 campaign. For the first time since the end of the military dictatorship, the far-left candidate won the 2018 election. After many years of centrist and left-wing tenures, the far-right coalition won by putting political "outsiders" on the ticket with a political agenda that appealed to the electorate.

Thus, this paper discusses Twitter data collected during the second round of the 2018 elections to verify the structure and patterns of communication and the shape and features of clusters and communities. Complex network methods, NoSQL databases, and a big data analytics tool were used in the study.

In this first section, background information on political polarization, the Brazilian political scenario up to the 2018 elections, and complex networks are presented. Papers with related themes are analyzed in Sec. 2. Then, the methodology is explained in Sec. 3, where the abstraction model, dataset, and preprocessing process are presented in details. Next, centrality metrics and community detection algorithms employed in this paper are presented. The final section presents the resultant network topology, main nodes, clusters, and communities along with concluding remarks.

### 1.1 Political Polarization

For [18], "polarization generally means expanding extremists and diminishing moderates in a certain distribution or across an ideological spectrum." The study [14] states that this phenomenon is also dynamic: "Polarization is both a state and a process. Polarization as a state refers to the extent to which opinions on an issue are opposed in relation to some theoretical maximum. Polarization as a process refers to the increase in such opposition over time."

Political polarization can be positive or negative and can assume different intensity levels. It can also be divided into elite, mass, or pernicious polarization categories. Elite polarization refers to polarization that occurs among formal political actors, e.g. politicians and political parties, or institutions composed by them [8]. Mass, or societal polarization is trickier to define, because there is no consensus among scholars [18, 8]; but what can be said is that this polarization occurs when the electorate behavior

and opinions regarding issues, policies, and famous figures are explicitly divided along ideological lines [15, 19, 9, 1, 2]. According to [8], pernicious polarization can be described by three points:

- it fuses elite and mass polarization, creating large opposing blocks;
- it is structured around a binary division, splitting the society into two large camps that dominate political life;
- it tends to last beyond the event that caused the polarization.

In their paper, [20] states that pernicious polarization can act in a single political cleavage, i.e. partisan identity, religious versus secular worldview, globalist versus nationalist ideals etc. Usually the description or perception of politics or society is 'us' versus 'them' [8].

Polarization does not appear suddenly; it evolves over time. An echo chamber effect is frequently one of the causes of this evolution. This phenomenon occurs when beliefs, concepts, and opinions are amplified and reinforced by repetition inside a closed system [4]. It is characterized by selective exposure, ideological segregation, and political polarization. People join or stay as members of these groups because they receive information or discuss about subjects that are aligned with their own views and opinions, ignoring completely opposing viewpoints, in what may be characterized as an unconscious employment of confirmation bias [4].

## 1.2 The Brazilian Political Scenario

The second round of the 2018 Brazilian presidential election had candidates from totally opposite political positions.

Jair Bolsonaro, 63 years old, ate the time was a retired army officer who had been a federal congressman for seven consecutive terms. He was the candidate of the Social Liberal Party (Portuguese: *Partido Social Liberal*; PSL), a small and unimpressive conservative party. His agenda was to promote a liberal economy and social conservatism. He was pro-gun and explicitly aligned to the United States, especially to President Trump, and Israel. Due to his beliefs and blunt opinions in interviews and social media postings, he was also considered to be far-right. Even though he had been a politician for a long time, he was considered, and also portrayed himself, as an outsider, mainly because he had been, until then, a second-level politician without any national media exposure. Supposedly, he was far enough from the national political core not to be associated with a series of corruption scandals.

In contrast, the other candidate represented the Workers' Party (Portuguese: *Partido dos Trabalhadores*; PT) which had been in power for sixteen years. Both former presidents where under investigation for corruption, the most recent having been impeached for administration impropriety. But the first, Luis Inácio Lula da Silva, was still extremely popular among his supporters. Only after Lula was barred from running due to corruption and money laundering convictions did PT choose another candidate, just one month before the elections. The 55-years-old, Fernando Haddad, had been mayor of São Paulo, the largest city in the country. An academic, he had also been Minister of Education on Lula and Dilma Rousseff's administration. He ran on an agenda to expand social welfare programs, block privatization of state industries, and maintain civil servants instead of outsourcing government services. His foreign policy agenda was to expand cooperation and commerce with Africa and Latin America.

The widespread corruption scandals occurred within PT's federal administrations tipped the election in favor of the opposition candidate. Bolsonaro was elected by more than 55% of the valid votes. He won in sixteen states, most of which were are among the wealthiest states in the country. Haddad won in just eleven states, which were among the poorest ones. The number of blank (2.14%) and null (7.43%) votes was high, the highest their summed percentage had been in a run-off election since the Brazil's redemocratization [26]. The abstention rate (21.30%) was also the highest for the same period in a presidential run-off election [26]. The high percentage of abstention, blank, and null votes may reflect the reaction of moderates to this pernicious polarized election between far-right and left.

This polarization emerged in the 2014 general elections, where PT won the second round of the presidential election by a tiny margin against the centrist Brazilian Social Democracy Party (Portuguese: *Partido da Social Democracia Brasileira*; PSDB). PT had been governing the country for the last three tenures. In 2013 and 2014 the population's dissatisfaction had increased considerably due to corruption scandals directly involving the most important members of PT in federal and party administrations, resulting in street protests. The most important protests occurred in June, 2013 lasting almost the entire month. They were popular protests that broke out after the city of São Paulo raised public transportation fares. It then quickly spread to other cities around the country throughout almost the whole month of June. In addition, *panelaços* were a way that the population used to show disapproval during presidential TV pronouncements. It consisted in making noise at the window or balcony by banging pots, pans, or other kitchen utensils. Despite this discontent, PT won the 2014 election to its extremely

loyal followers.

In the first year of the new tenure, however, economic number manipulation was uncovered. The administration made the country's economic numbers better than the reality before and during the election campaign. PT's popularity started to decrease rapidly and the federal administration start losing its support among its coalition parties in the parliament (Brazil has a multi party system). Population dissatisfaction also kept increasing. The impeachment of president Dilma Rousseff happened in August, 2016 and the vice-president Michel Temer was sworn in as president. He was a member of another centrist party, the Brazilian Democratic Movement Party (Portuguese: *Partido do Movimento Democrático Brasileiro*; PMDB). The impeachment was another step toward increased polarization.

As president, Temer made some reforms in order to recover the economy. Due to his moderate position, polarization was attenuated temporarily. However, as soon as the 2018 candidates were announced, the polarization returned in force and reached its peak during the presidential run-off.

## 1.3 Complex Networks

The study of complex networks is a relatively new study field, gaining popularity in the early 2000s due to the importance that researchers and research institutes saw in it after a few empirical studies were published [3]. Complex networks are graphs with non-trivial topological architecture and features, which are usually used to model real complex phenomena.

The Seven Brigdes of Köninsberg problem is considered to be the first to be addressed with network theory [3]. This is a puzzle created at Köningsberg, then the prosperous merchant capital of Eastern Prussia, now Kaliningrad in the Russian exclave on the Baltic Sea. It questions if someone could walk across all seven bridges and never cross the same bridge twice. The puzzle remained unsolved until 1736, when Leonard Euler employed a graph for the first time in order to analyze the problem. The conclusion was that there was no solution.

Networks are mainly composed of nodes and links, which in graph theory, are called respectively, vertices and edges. Nodes represent components of a system and links are the direct connections between them. Besides showing how a particular component interacts with other ones, networks can show the big picture of all interactions and allow measurements and properties to be extracted from them [12].

Network theory has been applied in many study fields, such as genetics, social science, physics, linguistics, ecology, climatology, operations research,

computer science, and public health [13]. The main tools for complex network analysis are: network characterization, centrality measurements, connectedness measurements, community detection, robustness analysis, and epidemic modeling.

There are two major types of complex networks, random and scale-free [3]. In random networks, the node degrees do not have a high value variability and the degree distribution follows a Poisson distribution. A national highway network exemplifies random networks, where ordinary cities are connected by a few highways and where the important cites are connected by several highways [3]. Thus, there is no city with no connection to a highway and, also, no city connected to thousands of highways.

On the other hand, scale-free networks are a particular type of network in which a large number of nodes have very few connections and a small number of nodes have a huge number of connections. These highly-connected nodes are called hubs and they are responsible for integrating the low-degree nodes to the network. The network degree distribution follows a power-law distribution. The flight network is an example [3]. The majority of airports are small, having few flights. However, the large airports have a high number of flights.

## 2 Related work

Twitter was founded in 2006 and since then, it has been used for many purposes, like product advertisement, news, propaganda, and social networking for ordinary people. There are research articles that discuss the Twitter usage patterns.

Retweet is the main information dissemination mechanism on Twitter, but it was not known why certain information spread faster than others. Thus, [25] evaluated several features that may affect the propagation of tweets through retweets. This study used the body and contextual data of 74 million tweets in order to identify factors that were significantly associated with the retweet rate. It also created a predictive model, which aimed to estimate a tweet probability of being widely spread. The results showed that links and hashtags are strongly correlated to the retweet rate, as well as, followed user numbers, follower numbers, and account age. The authors also claim that tweet publication frequency did not have any influence on the retweet rate.

Twitter is considered to be different from other social networks. Thus, a study to determine the main social functions of this tool was carried out

in [22]. The paper concludes that Twitter is employed in two distinct ways: for the consumption of information and news, and as a reciprocal social network. Account age is the main feature to infer how the user combines these two forms of usage.

There are also studies regarding exclusively political analysis on Twitter that can be cited. They analyze situations in the United States [11][27], Austria [17], Germany [21], United Kingdom [10], Egypt [6], and Canada [16].

In the first study, [11] investigated how this social network shapes communication and facilitates dialogue between communities of different political orientations. It used retweet and user mention datasets, composed of over 250,000 tweets, that were acquired during the six weeks before the 2010 US elections. Using a combination of algorithms to detect network communities and manually labeled data, the study shows that the retweet network exhibited a highly segregated party structure with an extremely limited connectivity between left- and right-wing users. However, this was not true for the network created by user mentions. It was composed of a single politically heterogeneous group in which ideologically opposed users interacted much more than the retweet network. The explanation for these network architectures is that, in retweets, the user mainly intended to forward information aligned to its ideological group. But, when a user mention was used, it was mainly to attack ideologically opposed groups by citing a particular user. Another study [16] reached the same conclusions as the first one, but using a much smaller dataset of less than 6,000 tweets. The data was obtained during the 2011 Canadian elections.

In [6], the opinion evolution dynamics in Egypt in 2013 was studied. It focused in two political axes: Islamism vs. secularism and military interventionism vs. non-interventionism. The study noticed that, after the power was seized by military forces, the highly polarized tweet postings increased. Before the *coup d'état*, secularists and non-interventionists were more active on Twitter; their activity decreased after the military takeover, giving more voice to Islamists and interventionists. Nevertheless, the study concluded that it was not possible to establish a correlation between being secular and non-interventionist or to be Islamist and pro-interventionist.

Paper [10] analyzes the data from one month before the 2017 British general election. It used over 34 million posts, where 9.6 million are original tweets and 25 million are retweets. The study employed time series analysis regarding relevant news, most cited topics, and most active and popular users. It detected the overwhelming dominance of pro-Labour posts and a disproportionate presence of the Scottish National Party, even though only

Scottish voters could elect this party. The study found that, in this case, social media was just an extension of traditional media. It also claimed that, even though Twitter cannot be used to predict elections due to lack of representativeness, it was a useful tool to access the mood of a particular population niche.

As for the 2017 German federal elections, [21] measured the election dynamics using Twitter data, from a dataset of more than 39 million tweets, with a little more than 130 thousand users. The study analyzed how the party Alternative for Germany (German: *Alternative für Deutschland*; AfD), a far-right party, also known as alt-right due to its opposition to classical conservatism, was able to take control of a large number of parliament seats. Initially, the study performed community detection to then identify how each cluster interacted with others and also to determine the most relevant themes for each cluster. In the dataset, there was significant content in English, which was identified as the American alt-right support to AfD. Likely due to language difference, the support did not convert into interconnection between those groups. The community topics were mostly about immigration and race. Bot detection was also performed, which found 11% of users as non-regular users.

Study [17] is about the 2016 Austrian presidential elections. It used more than 300 thousand tweets that were written in German or English. The dataset analysis was performed using complex network analysis, sentiment analysis, bot detection, and tweet temporal analysis. Three relevant findings were found. The main finding was that the election winner, Alexander van der Bellen, was considerably more popular and influential on Twitter than his opponent, Norbert Hofer. There was a clear sentiment polarization regarding the followers of the candidates. In addition, hashtags carrying a negative connotation were predominantly associated with Hofer. The second finding was that just a few bots were detected in the dataset. Finally, the users who supported van der Bellen were the ones responsible for spreading misinformation about him.

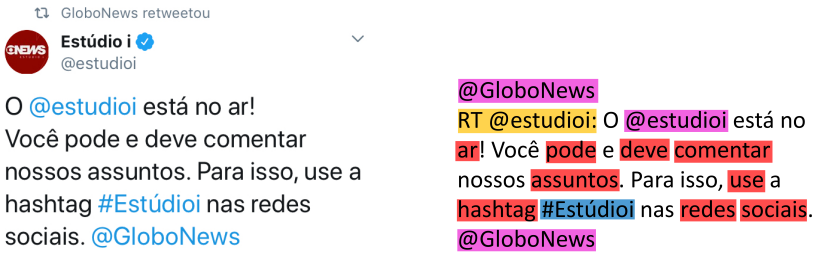
Paper [27] reports on the 2016 American presidential election, where the main candidates were Hillary Clinton and Donald Trump. The dataset had almost 2.9 million tweets, which were acquired during the elections. One of the objectives was to evaluate how accurately tweets represented the public opinion. It found that sentiment and topics expressed on Twitter could be a good proxy for public opinion. Another finding was that little original content was created by users. They normally retweeted opinions and the user-to-user communication rate was quite low. Finally, the last relevant finding was that sentiments generated by Donald Trump discourses

were more optimistic and positive than those generated Hillary's, directly reflecting the comment sentiments of Twitter users at the time.

# 3 Methodology

## 3.1 The Abstraction Model

The proposed abstraction model is able to keep all the information extracted from tweets by merging different nodes and relationship types into a unique network. In addition, it also allows node and relationship types to be filtered out in particular queries. The node types are: users, hashtags, retweets and words, called stems. The relationships are: copresence and authorship. Thus, in a scenario in which it is necessary to filter out only the users with a copresense relationship, it can be easily done with this approach. Fig. 1 shows what a tweet looks like initially (a) and, in (b), the colored particles are the ones which are selected to become nodes in the preprocessing phase, as detailed in Section 3.2, . Some words are not tagged due to stop word the removal procedure.




(a) (b)

Figure 1: (a) Original tweet and (b) extracted text with color tag representing the node types, where pink highlights users, yellow-retweets, red-words, and blue-hashtags

The example shown in Fig. 1 (a) is a retweet of a TV news account. It was chosen for having all of the node types in just one tweet. The translation is: "@estudioi is on air! You can and should comment our stories. Therefore, use the hashtag #Estúdioi on social networks. @GloboNews"

Networks would not exist with only nodes. Therefore, connections are also a really important feature. The copresense relationship in Fig. 2 (a) refers to the relationship created when the terms appear in the same tweet

body. It is not directional because the terms are together and there is no action between them. The authorship relationship in Fig. 2 (b) refers to the act of a user writing a tweet. In this case, the relationship is directional and goes from the user to tweet body terms. Both edge types are weighted, where it means the number of times that particular relationship happened.

(a) Copresence relationships

(b) Authorship relationships

Figure 2: Node relationship types: (a) represents terms that appear in the same tweet and (b) represents a user that wrote a tweet with certain terms. This example refers to the Fig. 1(b)

### 3.2 Dataset and Preprocessing

The data used in this study was acquired during the second round of the 2018 Brazilian presidential election, from October 8th to 27th. It has 1 million public tweets with user name and tweet body. For a faster and more dynamic analysis, an abstraction model, described in Section 3.1, was employed. Several processing approaches were employed to design the model some of which are related to Natural Language Processing (NLP).

The following steps were performed in the preprocessing phase:

- (A) Relational database read
- (B) Language detection
- (C) Translation to Portuguese when applicable
- (D) Tweet segmentation in hashtags, user mentions, author users, retweets and words (regular text)
- (E) Stop word removal and word stemming

- (F) Insertion in MongoDB and data transformation
- (G) Insertion in Neo4j.

Step A to F (MongoDB insertion) was performed in parallel tweetwise. In step A, the tweets were retrieved from a relational database where they had been stored when collected from Twitter.

There was a considerable amount of tweets in other languages. So, language detection was needed for each tweet, which was done in step B. About 91.7% (917,311) of the tweets were in Portuguese, 5.5% (55,169) in English, 1.7% (17,293) in Spanish and 0.4% (4,015) in French. The other languages totaled only 6,212 occurrences, which corresponds to about 0.6%, comprising more than 15 languages. There is no guarantee that all language detection was correct. However, this misclassification would not have a significant impact, mainly because these tweets represent such a small fraction of the data (0.6%) and also because their contents were usually non-textual, e.g. emoticons.

In step C, non-Portuguese tweets were automatically translate to Portuguese to avoid language clusters and create a network that focused on meaning.

Step D performed the tweet segmentation, where tweet body words were separated into five classes or, what we call, tweet particles: hashtag, author user, user mention, retweet, and text. This segmentation procedure is very important for Step G, which needs to know what type of tweet particle it is handling, so it can than create the right type of node and relationship. Due to the fact that each term type has a well-defined pattern, simple regular expressions were used to perform such segmentation.

Step E seeks to avoid words that do not have intrinsic meaning and to avoid meaning duplicity by performing stop word removal and stemming, respectively. Stop words are words with a high occurrence frequency that have no, or very little, value in natural language processing in terms semantics. They usually belong to the grammatical categories of pronouns, articles, prepositions, and modal or auxiliary verbs. Stemming performs word morphological root extraction. The idea is not to have more than one node for the variation of the same word. For example, “votes”, “vote”, “voting”, and “voted” are represented by their root, “vot”.

In step F, the tweets were stored in the MongoDB database, where the document properties are the five tweet particles obtained in segmentation performed by step D. MongoDB is a document-oriented NoSQL (non-relational) database that also has the Map Reduce framework embedded in it. The objective of this task was to transform how data were organized.

Such unique particle became the entry key and the copresent terms became the entry properties. MapReduce was employed for authorship relationships and Spark was used for copresence relationships. Spark is an analytics engine for big data and was needed because the standard MongoDB standard document size was not able to handle intermediate processing steps for copresense relationships.

In step G, all nodes and relationships among them were inserted in Neo4j. Neo4j is a native graph-oriented NoSQL database, which has its own query language and a graphical user interface for data visualization. This tool was chosen because it can handle a large amount of data and has metrics and algorithms already implemented, allowing fast, dynamic and intuitive data analysis and processing.

The resultant network had 468,643 nodes and 3,854,159 connections. Regarding node types, there are 22,123 hashtags, 40,790 retweets, 66,831 stems, and 338,899 users. There were 1,676,120 authorship relationships and 2,178,039 copresense relationships.

### 3.3 Analysis

After the network was generated, further processing needed to be performed in order to make analysis feasible. Topology and features of the network, clusters, and communities, were analyzed. The methods employed are presented in Sec. 4. The basic metrics and further results are shown Sec. 5.

## 4 Methods

The methods employed for further network analysis are explained in this section.

### 4.1 Centrality Metrics

Centrality metrics measure how important a node is to the network. There are several approaches, but the ones used in this paper are approaches that consider the influence beyond the first connection layer. The methods are eigenvector centrality and PageRank.

#### 4.1.1 Eigenvector centrality

Eigenvector centrality was proposed in 1986 [5]. It is the first centrality metric to consider the transitive importance of a node in a graph rather

than just its direct connections. Thus, relationships with high scoring nodes can be said to contribute more to the scoring of a particular node than connections to low scoring nodes.

The $v$ node centrality is given by:

$$x_v = \frac{1}{\lambda} \sum_{t \in M(v)} x_t = \frac{1}{\lambda} \sum_{t \in G} a_{v,t} x_t, \tag{1}$$

where $a_{v,t}$ is the connection between the node $v$ and $t$, $M(v)$ is a neighbor set of $v$, $\lambda$ is a constant, and $G$ is the graph.

### 4.1.2 PageRank

PageRank was initially created to rank websites in Google search [23]. The score is based on the inbound link quantity and quality, as shown in Eq. 2. It also relies on the assumption that an Internet user can get bored after several clicks, going then to a random page. It can be understood as a Markov chain, where states are pages and transitions are links, all of which have equally transition likelihood. Thus, if the method reaches a page with no outbound link, it will randomly choose a page to continue the process. PageRank pragmatically considers that pages without outbound links are connected to all pages in the network. Therefore, scores found for this particular page are divided equally among all other pages. This residual transition probability is typically set to 0.85. The value is estimated by averaging how often users use their bookmark list to go to a new page.

$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}, \tag{2}$$

where $p_1, p_2, ..., p_N$ are pages in analysis, $M(p_i)$ is the page set linked to $p_i$, $L(p_j)$ is the number of outbound links on page $p_j$, and $N$ is the total number of pages.

## 4.2 Community Detection

The objective of community detection methods is to find clusters in the network. Contrary to clustering methods that group samples in terms of their features, community detection only uses nodes and theirs relationships. This study employs label propagation and Louvain methods.

### 4.2.1 Label Propagation

The Label Propagation algorithm is a fast algorithm for finding communities in graphs, which was proposed by [24]. It detects communities using only the network structure and does not require a predefined goal function or prior community information. However, it has an interesting feature that allows to use initial labels to narrow down the final solution, making it a semi-supervised mode.

The algorithm assumes that a single label can easily become dominant in a densely connected group of nodes, which is unlikely to happen in a poorly connected region. At the end, nodes with the same label belong to the same community.

In Label Propagation:

- each node is initialized with a unique label;
- following a specified order. Each node updates its label regarding the majority of its neighbors, if there is a tie at the first place of the ranking, the label is chosen randomly among the tied nodes; this step is repeated until the stop criterion is met.

The algorithm's name comes from the fact that some labels propagate through the network during the iterative process of label updating. Densely connected groups of nodes quickly reach consensus on a single label during the iterations. So, only a few labels will remain at the end. The stop criterion is met when every node is labeled with the majority label of its neighbors, or when the execution reaches the maximum number of iterations defined by the user.

### 4.2.2 The Louvain Algorithm

Proposed by [7] the Louvain algorithm performs hierarchical community detection. It is based on modularity and is one of the fastest algorithms, also performing well on very large graphs [7]. The basic idea consists in optimizing the communities' modularity and then, aggregating the community nodes. The modularity score quantifies the assignment quality of a community to a node by comparing how densely connected that community has become compared to a scenario in which it is a random network.

In the first step of this method has two-step iterative process, each node is individually analyzed, removing it from its current community and inserting it in its neighboring communities. The modularity change (Eq. 3) is calculated for each insertion. If none of these attempts has a positive

outcome, the node remains in its current cluster. However, if any of the attempts are positive, then the node is assigned to the community that yielded the highest value. The stop criterion is met when there are no new assignments.

$$\Delta Q = \left[\frac{\Sigma_{in} + 2k_{i,in}}{2m} - \left(\frac{\Sigma_{tot} + k_i}{2m}\right)^2\right] - \left[\frac{\Sigma_{in}}{2m} - \left(\frac{\Sigma_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right], \quad (3)$$

where $\Sigma_{in}$ is the sum of all the weights of the links inside the community $i$ is moving into, $\Sigma_{tot}$ is the sum of all the weights of the links to nodes in the community $i$ is being inserted into, $k_i$ is the weighted degree of $i$, $k_{i,in}$ is the sum of the link weights between $i$ and other nodes in the community that $i$ is being inserted into, and $m$ is the sum of the weights of all links in the network.

The second step aggregates nodes in order to create a network of communities. After the first step, communities might be composed of nodes and inner and outer connections. The second step transforms each community into a single node. The connections are also merged, where the inner connections become a self-loop. The number of the connections that were merged is the weight of the new links.

Louvain is a hierarchical method. Thus, it tries to go a level further every iteration, merging communities whenever possible. The overall stop criterion is met when an iteration does not result in any reassignments.

# 5 Results

## 5.1 Network basic features

This section presents the statistical and structural profile of the resulting network, an estimated view of which can be seen in Fig. 3. The network has four node types, where 338,899 nodes are unique users; 66,831 nodes are unique stems; 40,790 nodes are unique retweets; and 22,123 nodes are unique hashtags. Thus totaling 468,643 nodes and 3,854,159 edges, where 2,178,039 edges are copresence relationships and 1,676,120 edges are authorship relationships. From the total user nodes (338,899), 52% (176,079) were users that wrote tweets and 48% (162,820) were users that only were mentioned in tweets without any active role in the dataset. Some of the author users (3,390) were also mentioned. Notable almost 32% (149,095) of all nodes did not have any connections which was due to the fact that some tweets did not have any text body, but only a link or a picture.

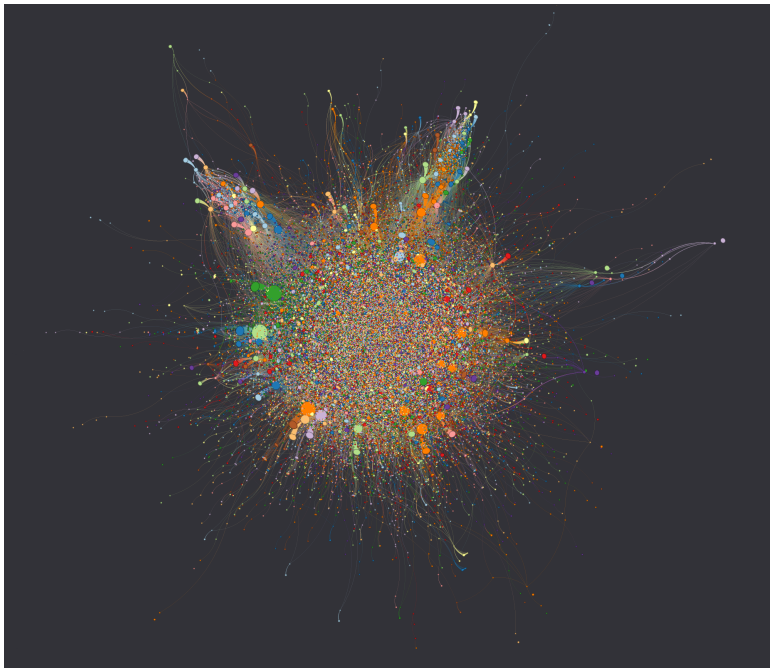


Figure 3: An estimated view of the network, where circle size is proportional to node degree. This figure was generated employing only a third of all connected nodes.

Table 1 shows a comparison among the entire network, subnetworks, and the network segmented by node types according to average degree, average connection weight per node, and average clustering coefficient. It is useful to have in mind the practical significance of these metrics. The higher the degree, the higher the diversity of tweet particles that the node is exposed to. Thus, if a node has low degree, it is possible to say that it was used in a tweet or tweets with little text. The higher the average weight, the higher the probability that the node was used in replicated or similar tweets. The clustering coefficient tells how connected the neighbors (among them) of a particular node are; where 1 means they are fully connected and 0 means they are not connected.

In our study, the network had an average degree of 2842.0 and an average edge weight of 3.65 per node. The highest degree was 42616 (é), shown in Table 2. Considering only the authorship edges of the whole network, the average degree was slightly lower (2345.3). But, when the network is analyzed regarding only copresence edges, the average degree drops signifi-

cantly. The possible explanation for this is the fact that a high amount of different users (176,079) mentioned other nodes only once on average. On the other hand, mentioned users appeared in tweets with a lower particle variety but with a higher repetition, 6.09 on average.

As shown in Table 1, the average clustering coefficient of the entire network was a little bit more than 0.16. That means, on average, the neighborhood of each node was connected to 16% of the total possible connections. Due to its characteristic of only connecting the author and the written particles, the network of only authorship relationships averaged an extremely low clustering coefficient (0.0007). When the analysis is segmented by the node type, most of them average close to the overall average (0.1618), except for retweets that had a slightly higher average (0.2269).

Table 1: Average degree, average weight per node, and average clustering coefficient for the entire network and for different node types.

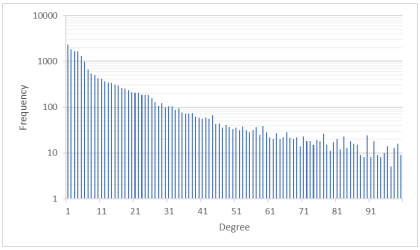
| Node Type | Population | Avg. Degree | Avg. Weight | Avg. Clustering Coef. |
|---|---|---|---|---|
| all | 468,643 | 2842.0 | 3.65 | 0.1618 |
| all (authorship) | 236,563 | 2345.3 | 1.00 | 0.0007 |
| all (copresence) | 139,055 | 1746.9 | 6.09 | 0.0455 |
| hashtags | 22,123 | 6145.0 | 8.59 | 0.1303 |
| stems | 66,831 | 1823.4 | 3.94 | 0.1682 |
| retweets | 40,790 | 789.0 | 8.10 | 0.2269 |
| users | 338,899 | 259.1 | 1.55 | 0.1548 |

Segmenting by node type and ranking by average degree, hashtags (6145.0) rank first, stems (1823.4) second, retweets (789.0) third, and users (259.1) last. The plausible explanation for the node order is the appearance frequency in tweets. For example, hashtags are normally used in more unique tweets than the other nodes; but they are also the ones which appear more repeatedly (8.59 times on average) with the same particles. Hashtags index keywords or topics, which may explain why they were more popular than other types. Stems are the morphological roots of important words of the tweet text. Thus, a word group related to elections together with "general" words are used to create a text with a message which is grammatically comprehensible. So, the words related to elections tend to be more repeated in tweets, which did not happen much with the other words. In addition most stems without a context were neutral, i.e. they did not support or attack any political view, and as such, were used by both political groups. The repetition of appearance of a stem with other particles in a tweet was lower, on average, than half of the average weight of hashtag and retweet nodes. The political neutrality of stems, did not occur to the other node types, they carried either implicit or explicit support or opposition to a political view or candidate. A tweet from a prominent user tended to be retweeted several

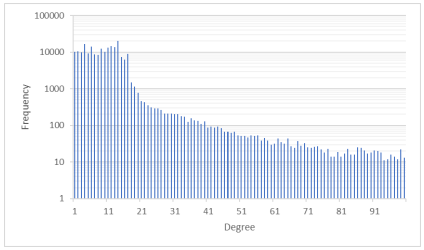
times, which made the particles appear exactly as the original text several times, thus increasing the relationship weight.

The user node is the last one in the rank, which on average, appears with less varied particles (259.1) and have less repetition (1.55) than the other nodes. The average weight is highly impacted by the fact that nodes with an authorship relationship have the average weight close to 1. That happens because more than 72% of user nodes have no repetition of tweet particle they use. The user node is a special case due to its authorship relationship, where this node type is mandatorily connected to the rear part of a directed edge.

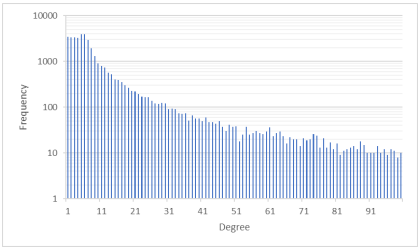
As for degree distribution, Fig. 4 shows a non-cumulative semi-log plot for each node type. For the sake of visualization, degree was cropped in the range between 1 and 100. In general, the plots have a similar shape and approach a magnitude of 10 in the end. Users, Fig. 4(b), and stems, Fig. 4(d), begin the graph with the magnitude of 10000, while hashtags and retweets begin at 1000. User nodes have a significant plateau at the beginning, ending at degree 17. For retweets, Fig. 4(c), there is also a plateau; it is much smaller ending at degree 7.



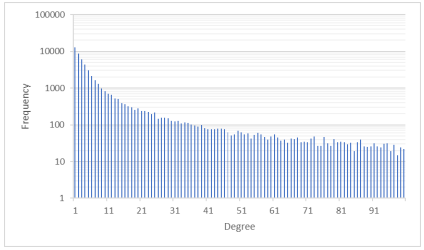

(a) hashtags




(b) users




(c) retweets

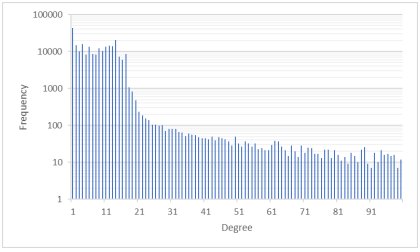



(d) stems

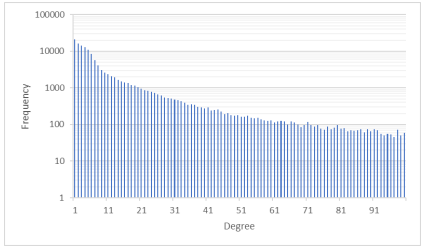
Figure 4: Degree distribution for different node types.

Fig. 5 shows the entire network and also segments the distribution according to connection type. Fig. 5(a) shows only the authorship relationship.

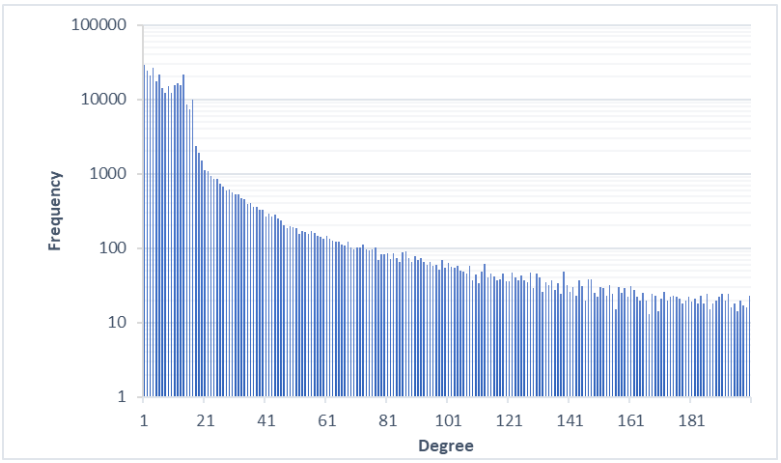
It has a similar shape to the degree distribution of user nodes in Fig. 4(b) due to the important role that node type has in this relationship. On the other hand, the copresence relationship shown in Fig. 5(b), has a shape similar to the hashtag, retweet, and stem distribution in Fig. 4. Finally, Fig. 5(c) is a combination of both subnetworks, where the copresence subnetwork shape is more preponderant, but the authorship relationship contributes with a secondary peak, which is visible from degree 12 to 17.




(a) only authorship relationship




(b) only copresence relationship




(c) network

Figure 5: Degree distribution for each relationships and for the entire network.

The average weight per node is shown in Fig. 6. Given that the weight represents how many times a relationship happens between two nodes, the average weight per node of the entire network is 3.65 (Table 1), which means that a node interacts with another one 3.65 times on average. 87% (296,354) of all nodes have a rounded average weight between 1 and 2.

In Table 2, the fifteen nodes with the highest degree are shown. Note

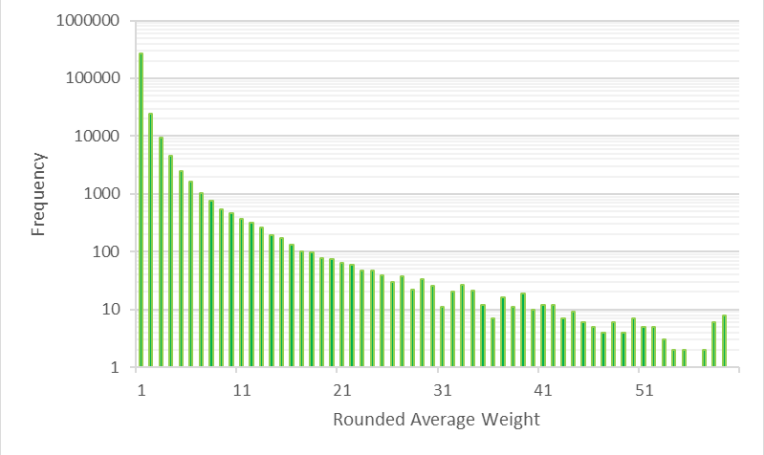



Figure 6: The distribution of average connection weight per node.

that the majority are hashtags. There are also five stems and one user. The degree ranges from around 17,000 to more than 42,000. All but four of the nodes have an average weight lower than 10 and the majority are under 6. The exceptions are three hashtags related to the candidates (#haddad13, #bolsonaropresidente, and #bolsonarosim) and one that is a candidate account (@jairbolsonaro).

Table 2: Top 15 regarding degree.

| Node | Degree | Average Weight |
|---|---|---|
| é | 42616 | 4.93 |
| #elenao | 35035 | 4.51 |
| #haddad13 | 31400 | 19.51 |
| #bolsonaro | 26420 | 6.83 |
| #elenunca | 23349 | 4.83 |
| #bolsonaropresidente | 22805 | 10.61 |
| #bolsonarosim | 22616 | 12.83 |
| #bolsonaro17 | 22196 | 8.01 |
| brasil | 22185 | 5.44 |
| @jairbolsonaro | 18995 | 11.61 |
| #haddadpresidente | 18972 | 8.19 |
| pra | 18916 | 4.76 |
| bolsonar | 18489 | 5.79 |
| #elenão | 18218 | 4.21 |
| vot | 16928 | 4.83 |

Table 3 shows the top 15 refer to hashtags with the highest degree. All of the hashtags shown refer to one of the two candidates. There is no hashtag about a neutral subject among them. Seven support Haddad: #elenao, #haddad13, #elenunca, #haddadpresidente, #elenão, #haddadsim, and #haddad. Eight support Bolsonaro: #bolsonaro, #bolsonaropresidente, #bolsonarosim, #bolsonaro17, #elesim, #b17, #marketeirosdojair,

and #bolsonaro2018. Only four in the ranking have the average weight over 10: #haddad13 (19.51), #bolsonarosim (12.83), #marketeirosdojair (11.09), and #bolsonaropresidente (10.61). As seen in Table 3, there is no correlation between degree and average weight. This might suggest that depends on the group of people who are using the hashtag or the theme the hashtag is being associated. Hashtags have the highest degrees because of their function in Twitter, to index keywords or topics. Table 3 reflect this; the majority of node degrees are higher than shown in Table 4, 5, and 6.

Table 3: Top 15 hashtags regarding degree.

| Node | Degree | Average Weight |
|---|---|---|
| #elenao | 35035 | 4.51 |
| #haddad13 | 31400 | 19.51 |
| #bolsonaro | 26420 | 6.83 |
| #elenunca | 23349 | 4.83 |
| #bolsonaropresidente | 22805 | 10.61 |
| #bolsonarosim | 22616 | 12.83 |
| #bolsonaro17 | 22196 | 8.01 |
| #haddadpresidente | 18972 | 8.19 |
| #elenão | 18218 | 4.21 |
| #elesim | 14716 | 4.85 |
| #haddadsim | 11539 | 6.42 |
| #b17 | 10427 | 6.04 |
| #haddad | 9981 | 6.39 |
| #marketeirosdojair | 9808 | 11.09 |
| #bolsonaro2018 | 9557 | 7.28 |

According to Table 1, stems have the second greatest degree average. This is also true for the top 15, shown in Table 4. The average weight per node has lower values compared to hashtags, as shown in Table 1. Eleven are under 5 and none of them are over 7.

Table 4: Top 15 stems regarding degree.

| Node | Degree | Average Weight |
|---|---|---|
| é | 42616 | 4.93 |
| brasil | 22185 | 5.44 |
| pra | 18916 | 4.76 |
| bolsonar | 18489 | 5.79 |
| vot | 16928 | 4.83 |
| haddad | 16918 | 6.55 |
| faz | 16649 | 2.96 |
| tod | 16226 | 4.34 |
| pov | 14258 | 5.22 |
| democrac | 13654 | 4.56 |
| ser | 13445 | 4.20 |
| hoj | 12735 | 4.39 |
| favor | 12634 | 1.69 |
| vai | 11479 | 4.59 |
| tá | 11191 | 3.30 |

As for retweets and user nodes, Table 1 shows them in the last positions regarding degree. For the average weight per node, retweets are ranked second, on average, but users are ranked last. Table 5 and Table 6 show that

the top 15 retweet and user nodes have similar degree occurrences. Even though retweets are second in terms of average weight shown in Table 1, this is not true for the top 15. Retweet occurrences have higher values than hashtags, in Table 3. This may be due to the fact that, is the fact that when someone retweets, the text normally remains unchanged, increasing the weight each time a retweet occurs.

In Table 5 there are eight retweets supporting Bolsonaro (rt @jairbolsonaro, rt @wallyssonrg2, rt @glauberrosa, rt @mniederauerm, rt @vs_ferreira, rt @francischini_, rt @nizmycuba, and rt @dompeddropedro) and seven supporting Haddad (rt @haddad_fernando, rt @akayona_, rt @samyy_l, rt @yaasxxc rt, @xicosa, rt @jeantissociall, and rt @scalvint).

Table 5: Top 15 retweets regarding degree.

| Node | Degree | Average Weight |
|---|---|---|
| rt @jairbolsonaro | 9799 | 12.01 |
| rt @haddad_fernando | 8849 | 16.72 |
| rt @wallyssonrg2 | 7905 | 23.48 |
| rt @glauberrosa | 7829 | 19.63 |
| rt @akayona_ | 6242 | 5.82 |
| rt @samyy_l | 6209 | 12.23 |
| rt @yaasxxc | 4399 | 9.81 |
| rt @mniederauerm | 4350 | 7.21 |
| rt @xicosa | 4199 | 8.52 |
| rt @jeantissociall | 4161 | 2.58 |
| rt @vs_ferreira | 3839 | 20.46 |
| rt @francischini_ | 3782 | 10.24 |
| rt @nizmycuba | 3754 | 8.30 |
| rt @scalvint | 3397 | 8.39 |
| rt @dompeddropedro | 3151 | 5.90 |

As for average weight, users are slightly higher than stems. Author nodes, that were not mentioned in tweet bodies may have influenced these averages significantly, giving users the advantage as seen in Table 1.

Among the users, five supported Haddad: @haddad_fernando, @manueladavila, @guilhermeboulos, @ptbrasil, and @lulaoficial. Five supported Bolsonaro: @jairbolsonaro, @lobaoeletrico, @carlosbolsonaro, @bolsonarosp, and @roxmo. And five users were neutral, including press media (@folha, @claudioedantas, and @estadaopolitica), digital platform (@youtube), and a judicial branch institution (@tsejusbr).

The two accounts of the candidates, @jairbolsonaro and @haddad_fernando, were the first two in the ranking of users and also in retweets, as shown in Table 5. Of note, Bolsonaro had more relationship repetitions than Haddad as user, but fewer retweets. In both tables, Table 5 and 6, Bolsonaro has higher degree scores than Haddad.

Fig. 7 presents the cumulative distribution of the entire network and also its subnetworks: copresence and authorship relationships. The dashed

Table 6: Top 15 users regarding degree.

| Node | Degree | Average Weight |
|---|---|---|
| @jairbolsonaro | 18995 | 11.61 |
| @haddad_fernando | 14612 | 9.44 |
| @lobaoeletrico | 7053 | 8.78 |
| @manueladavila | 5417 | 6.52 |
| @folha | 5221 | 5.00 |
| @guilhermeboulos | 4750 | 10.09 |
| @youtube | 4695 | 5.56 |
| @tsejusbr | 4559 | 5.96 |
| @carlosbolsonaro | 4528 | 6.19 |
| @ptbrasil | 3797 | 3.60 |
| @bolsonarosp | 3501 | 4.68 |
| @lulaoficial | 3499 | 3.70 |
| @claudioedantas | 3169 | 5.98 |
| @estadaopolitica | 2993 | 3.28 |
| @roxmo | 2896 | 3.28 |

lines show examples of power-law and Poisson distributions. Note that the distribution of the overall network and authorship subnetwork show a similar behavior. User nodes are more than 72% of all nodes and almost 51% of user nodes are only authors not mentioned in tweets, which means they are not present in the copresence subnetwork. Therefore, author users may be the reason for the similar behavior of the overall network and authorship subnetwork. Finally, both overall network and authorship subnetwork distributions do not seem to follow either Poisson or power-law distribution when comparing the networks and the example lines. However, the copresence subnetwork does seem to follow a power-law distribution.

## 5.2 Authorship relations

The analysis of the most mentioned or retweeted users enables the estimation of main influencers in the network. A user who got retweeted or mentioned by several different users is considered to be more popular than a user who got retweeted or mentioned the same number of times but from fewer different users. Thus, considering the same total weight sum, more connections count more than fewer connections with higher weights. This section will cover the authorship relationships between user-user and user-retweet.

In Fig. 8, the twenty most cited users (blue, red, and gray) and the main users who mentioned them (light blue) were selected. The most cited ones were @jairbolsonaro, which is the main hub of the cluster at the left side of the figure, and @haddad_fernando who is also the main hub but at cluster on the right side. Other users were significantly mentioned but less relevant: @manueladavila, @rogerwaters, @ceosoares, @lulaoficial, @ptbrasil, @guilhermeboulos, and @cirogomes supporting Haddad; @lobaoeletrico, @jairbolsonarohttps, @carlosbolsonaro, @bolsonarosp, @flaviobolsonaro, @danilo-

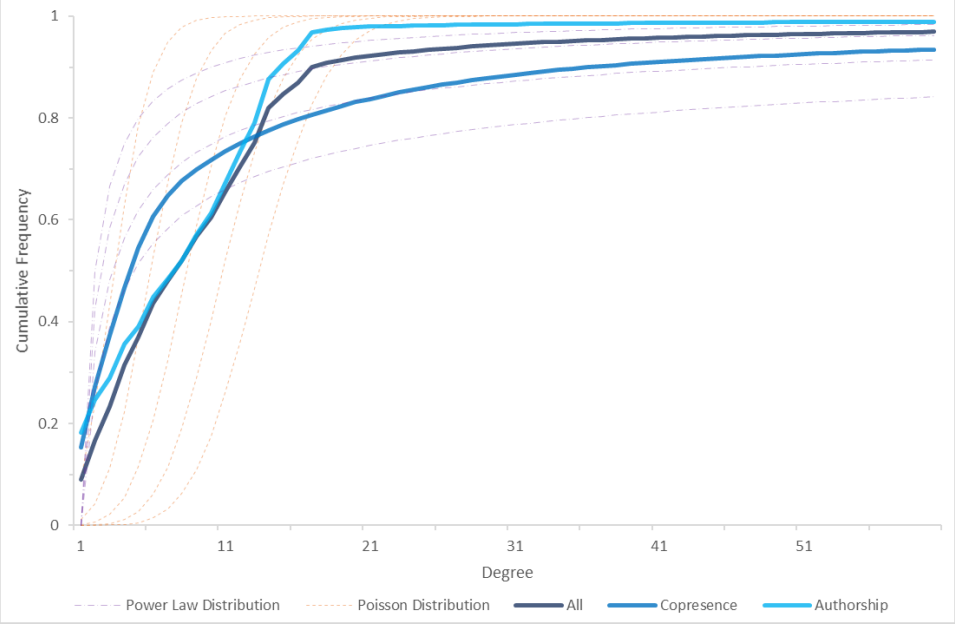



Figure 7: Cumulative distribution functions of the network and its subnetworks (solid lines), example of power law distributions (dash-dot lines), and example of Poisson distributions (dashed lines).

gentili, and @lsentoes supporting Bolsonaro; and @youtube , @milenio, @tsejusbr, and @folha as neutral. The "user" @jairbolsonarohttps is the result of mentions of the twitter account of Bolsonaro followed by a truncated link but without a space separating them. @tsejusbr is the only neutral node that is not a media institution, but a judicial branch institution. The correspondent degree for each node can be seen in Table 7.

In Fig. 8, there are clearly three major clusters with several satellite clusters around them. The major ones are composed of a cluster which is related to Bolsonaro, another one which is related to Haddad, and finally, a central one that communicates mutually with both sides.

The difference between Table 6 and Table 7 is that the latter only considers inbound authorship connections. In copresence connections, nodes present in many long tweets tend to have a larger degree than nodes present in shorter tweets. Copresence degree, practically speaking, means how many different tweet particles a certain node appeared together in tweets. On the other hand, authorship degree means how many different users wrote a certain particle. That said, almost all nodes in Table 6 are present in Table 7,

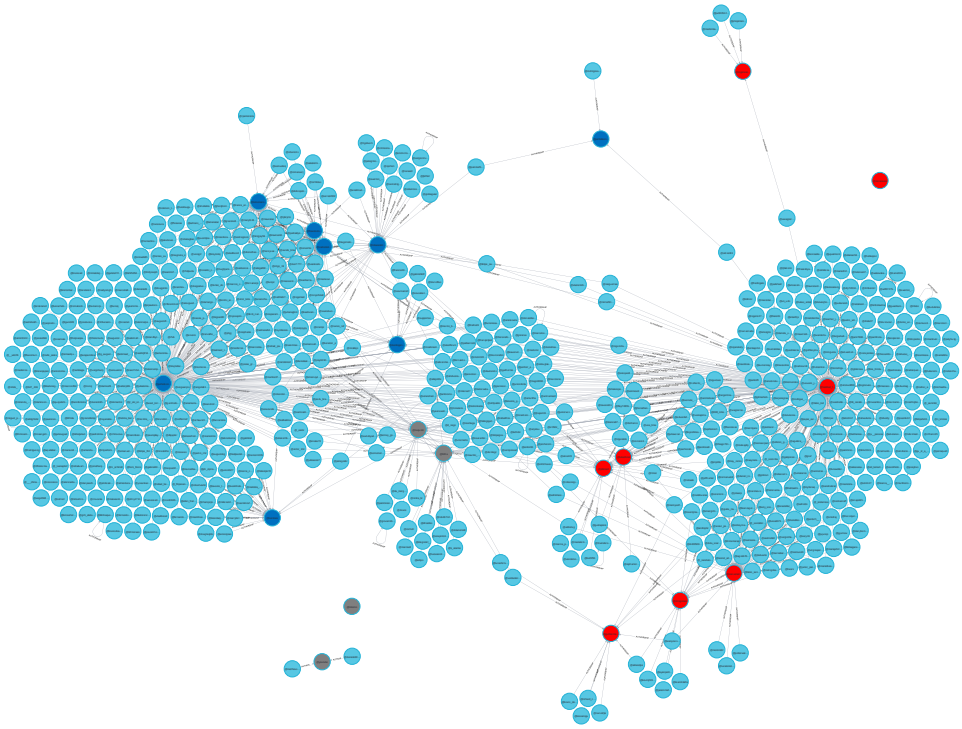


Figure 8: The twenty most mentioned users and the main users who mentioned them (light blue). The color of mentioned users represent which side they were related to. Blue is for nodes supporting Bolsonaro, red Haddad and gray neutral.

although in different positions, except for @claudioedantas, @estadaopolitica, and @roxmo.

Retweets were also analyzed. Fig. 9 shows the twenty retweet nodes with the greatest degree in terms of the authorship relationship, as well the main users who retweeted them. There are no connections among the clusters, in which the core is a retweet node surrounded by users. Actually, there is no user that retweeted more than one node from the top 20. Given that the one-million dataset used in this study is a small sample from the real world data, the lack of user retweets does not necessarily imply that this relationship is non-existent in the real world. The most plausible scenario is a very low occurrence. The lack of connectivity among these clusters may be a manifestation of the echo chamber, similar to the behavior reported by

Table 7: Top 20 most mentioned users.

| Node | Degree |
|---|---|
| @jairbolsonaro | 5089 |
| @haddad_fernando | 3653 |
| @youtube | 643 |
| @lobaoeletrico | 541 |
| @manueladavila | 485 |
| @milenio | 347 |
| @tsejusbr | 324 |
| @rogerwaters | 311 |
| @ceosoares | 306 |
| @folha | 295 |
| @lulaoficial | 246 |
| @jairbolsonarohttps | 238 |
| @carlosbolsonaro | 217 |
| @bolsonarosp | 198 |
| @flaviobolsonaro | 192 |
| @ptbrasil | 189 |
| @danilogentili | 170 |
| @guilhermeboulos | 154 |
| @cirogomes | 126 |
| @lsentoes | 122 |

[11].

In Table 8, most retweets are pro Haddad (16). Three are pro Bolsonaro and one neutral from a press media account. The nodes supporting Haddad are: rt @haddad_fernando, rt @samyy_l, rt @akayona_, rt @jeantissociall, rt @yaasxxc, rt @guilhermeboulos, rt @duabrazil, rt @manueladavila, rt @delucca rt @alugms, rt @charlottelawr, rt @j_livres, rt @o_ave_conas, rt @emersonanomia, rt @ivonepita, and rt @lulaoficial. The nodes supporting Bolsonaro are: rt @jairbolsonaro, rt @carlosbolsonaro, and rt @lobaoeletrico. The only neutral one is rt @dw_espanol.

Table 8: Top 20 most retweeted users.

| Node | Degree |
|---|---|
| rt @jairbolsonaro | 8345 |
| rt @haddad_fernando | 7690 |
| rt @samyy_l | 5586 |
| rt @akayona_ | 4628 |
| rt @jeantissociall | 4061 |
| rt @yaasxxc | 3904 |
| rt @guilhermeboulos | 2541 |
| rt @duabrazil | 1796 |
| rt @carlosbolsonaro | 1386 |
| rt @manueladavila | 1164 |
| rt @delucca | 939 |
| rt @alugms | 830 |
| rt @charlottelawr | 793 |
| rt @j_livres | 693 |
| rt @lobaoeletrico | 684 |
| rt @o_ave_conas | 615 |
| rt @emersonanomia | 584 |
| rt @ivonepita | 523 |
| rt @lulaoficial | 482 |
| rt @dw_espanol | 480 |

Except for two nodes (rt @xicosa and rt @scalvint), all the nodes ab-

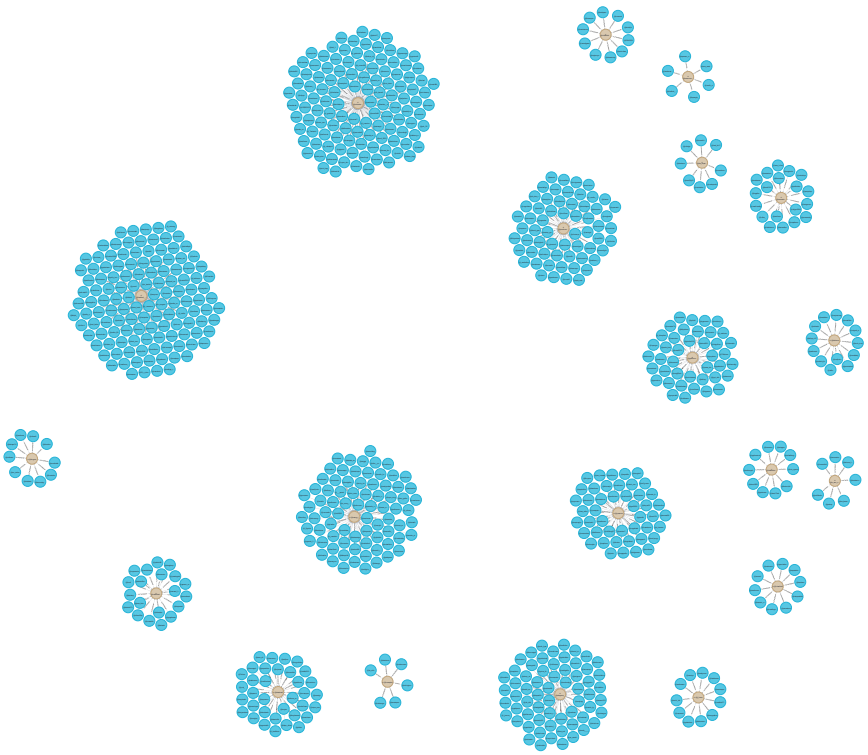


Figure 9: Retweet nodes with highest authorship degrees (beige) and the main users (light blue) who wrote them.

sent in Table8 in comparison to Table5 are from Bolsonaro supporters (rt @wallyssonrg2, rt @glauberrosa, rt @mniederauerm, rt @vsferreira, rt @francischini_, rt @nizmycuba, and rt @dompeddropedro). They represent almost half of Table 5.

### 5.3 Top 50

In this subsection, the top 50 regarding centrality are presented. Centrality metrics employed, Sec. 4.1, were chosen because they consider not only the direct influence, but also other neighborhood levels. Moreover, the difference between their outcomes can be analyzed.

Fig. 10 was obtained employing eigenvector centrality. It has 771 connections and is composed of fifteen users, twelve hashtags, and twenty-three stems. There is no retweet node. Within user nodes, four supported Bolsonaro (@lobaoeletrico, @bolsonarosp, @flaviobolsonaro, and @jairbol-

sonaro), four supported Haddad (@manueladavila, @lulaoficial, @haddad_fernando, and @zehdeabreu), and seven were neutral, of which four were press media (@folha, @uol, @globonews, and @estadaopolitica), two were judicial branch institutions (@mpf_pgr and @tsejusbr), and one was a soccer team arena (@allianzparque). The arena was mentioned several times due to a concert of Roger Waters there. Roger Waters was publicly against Bolsonaro.

Within hashtags, eight supported Bolsonaro (#suasticafake, #elesim17, #17neles, #ptjamais, #folhafakenews, #mito, #ibopefake, and #cagueiproibope), two supported Haddad (#elenunca and #vote13), and two were neutral (#brazil and #eleições2018). Words by themselves do not support any side; they need context and other tweet particles. The stems were: "é", "brasil", "bolsonar","pt", "pra", "haddad", "fal", "faz", "presid", "dia", "pod", "vot", "contr", "agor", "pov", "diz", "q", "vai", "apoi", "quer", "candidat", "deu", and "qu".

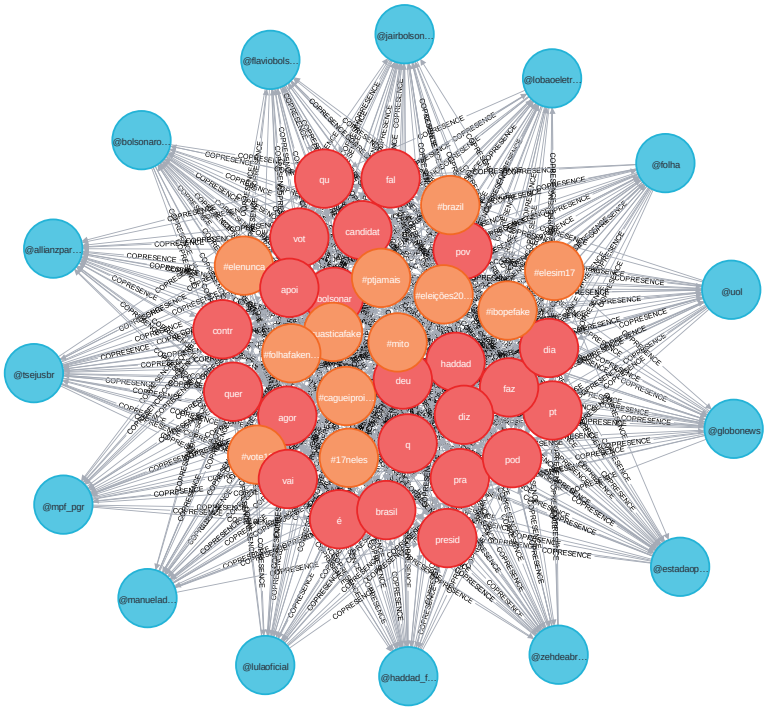



Figure 10: Top 50 using eigenvector centrality.

The ten highest scores were: @estadaopolitica (345.3), é (303.1), brasil (290.8), bolsonar (271.6), @lobaoeletrico (265.9), @jairbolsonaro (254.9), @haddad_fernando (249.2), pt (249.0), pra (246.7), #elenunca (244.7). In

terms of values, the nodes ranged from more than 345 (highest) to just under 180 (lowest). Somehow, the three node types were distributed evenly throughout the list. For further information, all the scores are in Table 10 in Appendix, Sec. 7.

Fig. 11 was obtained employing PageRank. It has 723 connections and is composed of nineteen hashtags, nineteen stems, nine retweets, and three users. Within user nodes, two supported Bolsonaro (@lobaoeletrico and @jairbolsonaro) and one supported Haddad (@haddad_fernando). As for hashtags, eleven supported Bolsonaro (#bolsonarosim, #bolsonaropresidente, #bolsonaro, #bolsonaro17, #marketeirosdojair, #elesim, #euvotobolsonaro, #haddadéfefeca, #bolsonaro2018, #ptnão, and #b17) and eight supported Haddad (#haddad13, #elenao, #elenão, #haddadpresidente, #elenunca, #haddadsim, #haddad, and #viravoto). Within retweets, five supported Bolsonaro (rt @lobaoeletrico, rt @glauberrosa, rt @jairbolsonaro, rt @wallyssonrg2, and rt @vs_ferreira) and four supported Haddad (rt @haddad, rt @samyy_l, rt @_vitroia, and rt @akayona_). The stems were: "é", "brasil", "haddad", "bolsonar", "pra", "vot", "tod", "xa0","xa0♥", "russ", "ser", "pt", "fal", "democrac", "dia", "morr", "faz", "hoj", and "vai".

Note that there are links between some users and the retweet with the same username; because someone retweeted the user post and complemented it by mentioning the original user, perhaps adding some more text. This happened 17 times between @lobaoeletrico and its retweet (@lobaoeletrico), 126 times between @haddad_fernando and its retweet (rt @haddad_fernando), and 134 times between @jairbolsonaro and its retweet (rt @jairbolsonaro).

The ten highest scores were: #haddad13 (4616.0), #bolsonarosim (2884.4), #bolsonaropresidente (2366.1), #elenao (2246.6), é (2155.7), @jairbolsonaro (1930.5), #bolsonaro (1890.2), #bolsonaro17 (1801.7), #haddadpresidente (1531.2), rt @wallyssonrg2 (1492.4). The scores of all top 50 range from more than 495 (lowest) to more than 4616 (highest). Stems appear more at the bottom of the ranking; the other node types are more evenly distributed. For further information, all the scores of this top 50 are in Table 11 in Appendix, Sec. 7.

Some nodes of both top 50 where not in the tables of Sec. 5.1. The tables show only the direct influence (degree) of a node. They do not take into account the influence of other relationship levels, whereas the metrics employed in this section also take into account the neighborhood. Thus, a node with a lower degree can be more influential than a higher-degree node, if it has $n$-level neighbors with higher influence.

Both methods share the most important nodes for the electoral scenario,

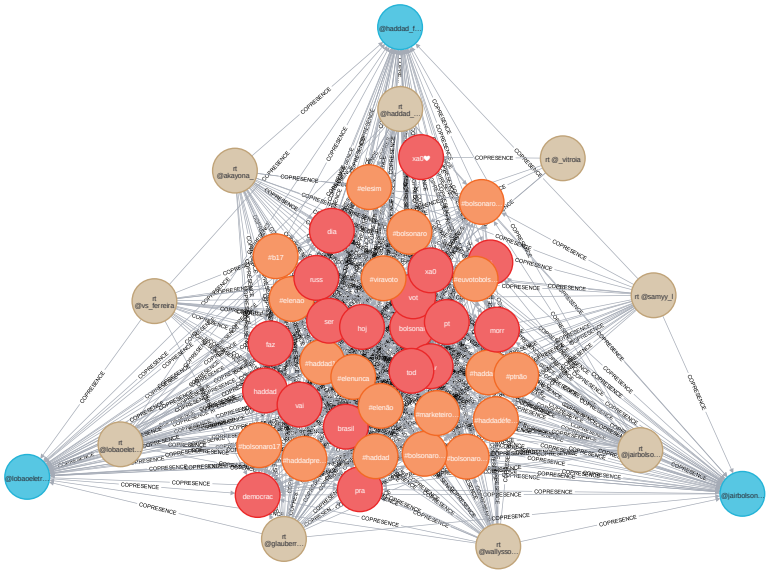


Figure 11: Top 50 using PageRank.

but the other nodes differ slightly. Eigenvector has no retweet node and also has a high amount of stems and users. On the other hand, PageRank has fewer user nodes and stems, but more retweets and hashtags.

## 5.4 Communities

In order to detect communities, Louvain and Label Propagation algorithms were used, see Sec. 4.2. The Label Propagation result was not satisfactory. About 58% of the nodes were labeled to the same community and the size of each remaining community did not exceed 136 members (0.03%). It found 193,893 clusters in total. As it did not reach a reasonable performance level, no further analysis was carried out.

On the other hand, the Louvain algorithm performed well. It detected 150,574 communities, the 23 largest of which represented 67.7% of all nodes. In the top 23, the largest community had 43,415 nodes and the smallest 84. In the remaining communities outside top 23, the size varied from 22 to 1 node. Communities with only one node represented more than 32% (150,079) of all nodes, which was quite near of the number of nodes that did not have any connections (149,095), as shown in Sec. 5.1.

Fig. 12 presents the size of each cluster and which candidate they are related to. The cluster labeling was done, analyzing the top 50 nodes accord-

ing to PageRank. The analysis mainly considered the political orientation of hashtags, users, and retweets. Notably, the first three communities are Bolsonaro-related and the occurrence behavior of next ones seems not to have a defined pattern. There are nine clusters (1, 2, 3, 7, 11, 14, 16, 17, and 23) related to Bolsonaro, totaling 166,220 nodes. There are six clusters (4, 6, 8, 9, 13, and 18) related to Haddad, totaling 93,759 nodes. Community 9, besides being related to Haddad, has a strong thematic about Manuela d'Ávila, the vice presidential candidate of Haddad. There are eight neutral clusters (5, 10, 12, 15, 19, 20, 21, and 22) which represent 57,288 nodes in total. Most of them are made up of discussions between both political sides. Communities 5 and 21 are two neutral clusters worth mentioning. The first community, besides the clash between opposite political sides, is also centered on Ciro Gomes, a left-wing candidate from the first round. The second community, besides the clash, is also centered on Roger Waters, a famous singer who was performing seven concerts in Brazil, most of them during the second round of elections. He was publicly against Jair Bolsonaro, at the time.

There are also some clusters that have lower political debate. In Community 15, the theme is about death and personal relationships with people of the opposite political side. Community 22 has Arabic and Hindi characters.

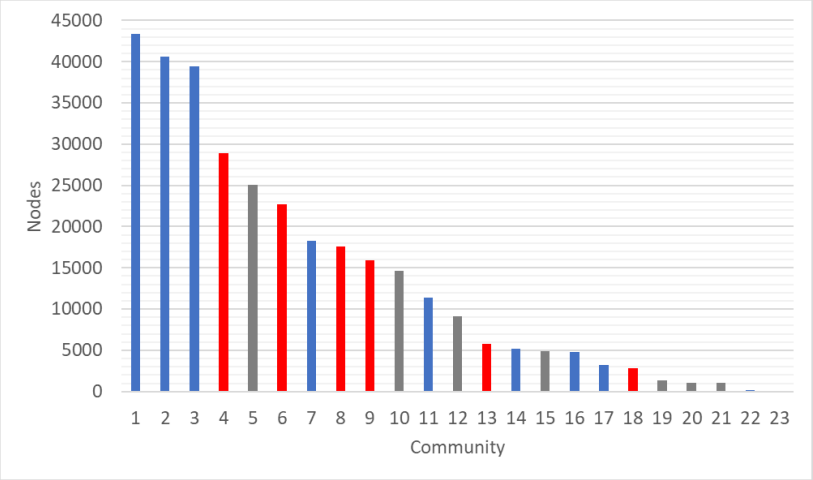



Figure 12: Largest communities obtained with Louvain algorithm. Blue are Bolsonaro-related, red-Haddad-related, and gray-neutrals.

Fig. 13 shows the highest PageRank score for each community, as detailed in Table 9. Except for the Community 13, there is a certain trend of

decreasing PageRank whenever the community size decreases.

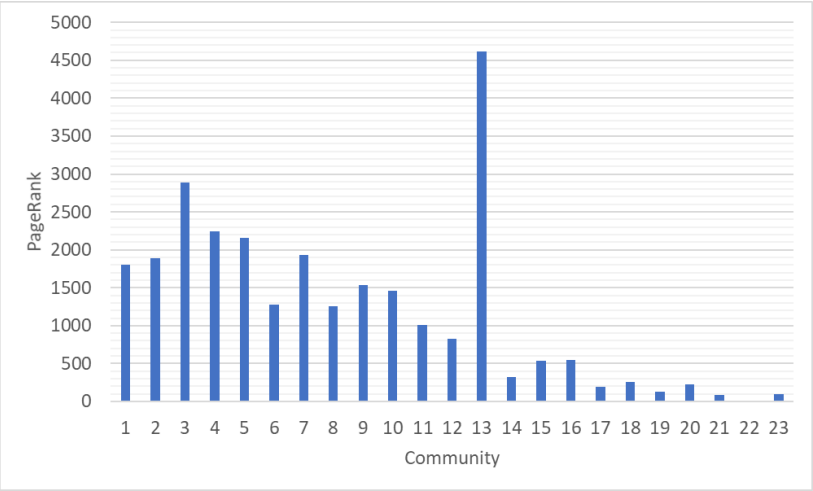



Figure 13: Highest PageRank scores for each community.

Analyzing the node type of main communities in Table 9: nine are hashtag, six are stems, five are retweets, and three are users. Apparently, there is no correlation among community label and main node type, nor among community size and main node type.

Stems by themselves are neutral. In order to give further information complementing Table 9, the second and third main nodes are going to be listed when the first one is a stem. For Community 5 we have: haddad (1043.8), #haddadéfefeca (813.6); for Community 12: #haddad (615.2), #atentadofakedopt (241.0); for Community 15: rt @jeantissociall (462.5), solt (327.1); for Community 16: #bolsonaropresidente17 (305.7), #nordeste17 (299.2); for Community 17: né (128.0), #haddadnao (87.7); and for Community 23: #brasilacimadetudo (54.5), #17paraobemdobrasil (42.0).

Fig. 14 shows the PageRank score mean and standard deviation for every community, showing that most nodes in every community have a very low PageRank score. Except for a few cases, the average ranges from around 0.5 to around 1.5. The standard deviation begins near one and goes towards 0. The outlier is Community 13, mainly because it has the node with the highest PageRank score, #haddad13 (4616.0).

Fig. 15 and Fig. 16 further analyze the communities inner structure. The charts present the same information but in different ways, allowing complementary analysis. Fig. 15 shows the parts to a whole and Fig. 16 the

Table 9: Highest node PageRank scores for each community.

| Community | Node | PageRank |
|---|---|---|
| 1 | #bolsonaro17 | 1801.7 |
| 2 | #bolsonaro | 1890.2 |
| 3 | #bolsonarosim | 2884.4 |
| 4 | #elenao | 2246.6 |
| 5 | é | 2155.7 |
| 6 | rt @haddad_fernando | 1281.1 |
| 7 | @jairbolsonaro | 1930.5 |
| 8 | @haddad_fernando | 1255.0 |
| 9 | #haddadpresidente | 1531.2 |
| 10 | #elenunca | 1463.1 |
| 11 | rt @jairbolsonaro | 1007.3 |
| 12 | vot | 825.1 |
| 13 | #haddad13 | 4616.0 |
| 14 | rt @francischini_ | 320.3 |
| 15 | morr | 533.8 |
| 16 | dia | 543.7 |
| 17 | vc | 195.4 |
| 18 | rt @gleisi | 259.1 |
| 19 | rt @crissrc | 127.8 |
| 20 | @tsejusbr | 228.8 |
| 21 | #rogerwaters | 81.9 |
| 22 | #true_prophecies | 9.3 |
| 23 | “u2060” | 100.4 |

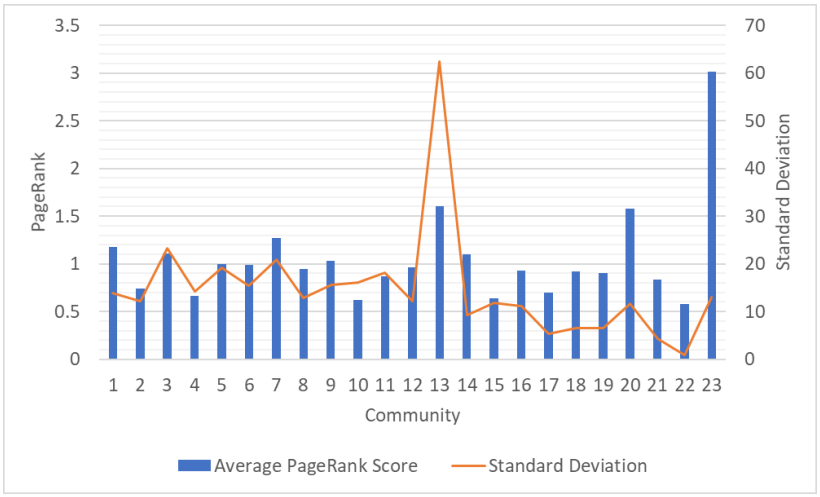



Figure 14: PageRank score average and standard deviation for each community.

node type percentage individually for each community. Note that, in almost all communities, user nodes represent at least half of the entire population. Two communities, 11 and 15, have user nodes representing more than 90%. User nodes do not have the highest ratios in only two cases, where are stems. The other node type ratios do not surpass 20%.

User nodes range from 26.5% (Community 22) to 96.5% (Community

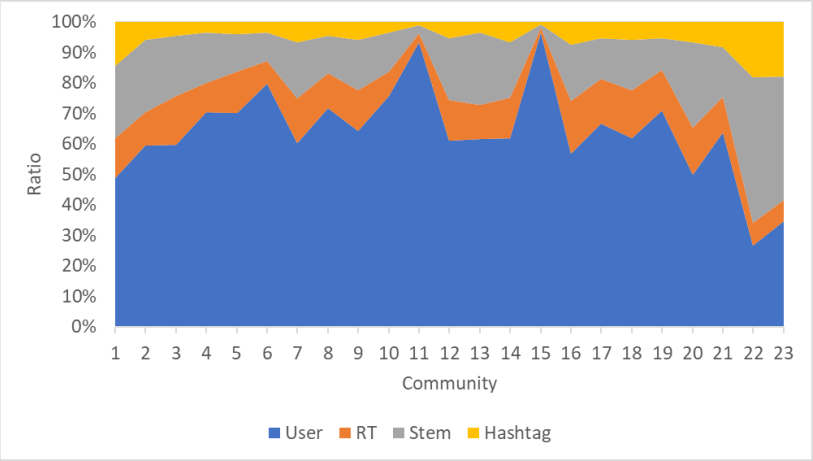



Figure 15: Node-type ratio for each community, parts to the whole.

15). Retweet node ranges from 1.6% (Community 15) to 17.5% (Community 16). Stems range from 1.1% (Community 15) to 47.7% (Community 22). Hashtags range from 0.7%(Community 15) to 18.2%(Community 22). All ratios for every community can be found in Table 12 in Appendix, Sec. 7.

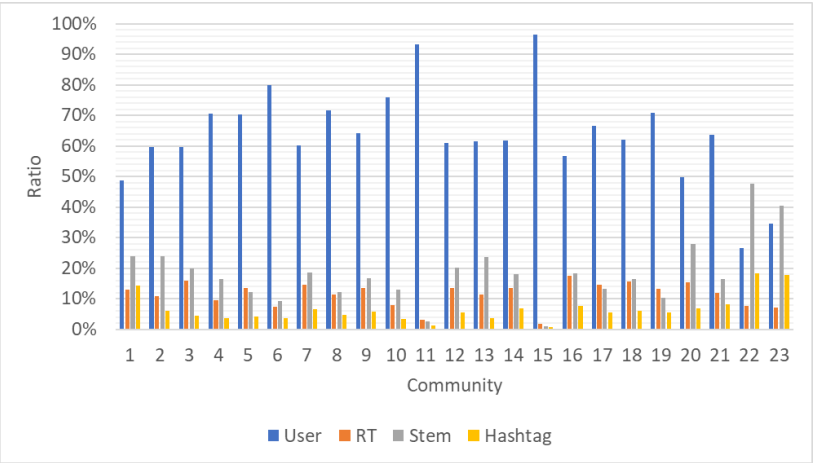



Figure 16: Node-type ratio for each community.

Even though user nodes are the great majority in most communities, they have the lowest PageRank scores on average for almost all communities, as shown in Fig. 16. This happens because just a tiny group is mentioned in tweets, while the vast majority performs only the author role. In all com-

munities, the highest PageRank score is either hashtag or stem. Hashtags are the highest in fourteen communities and stems in nine.

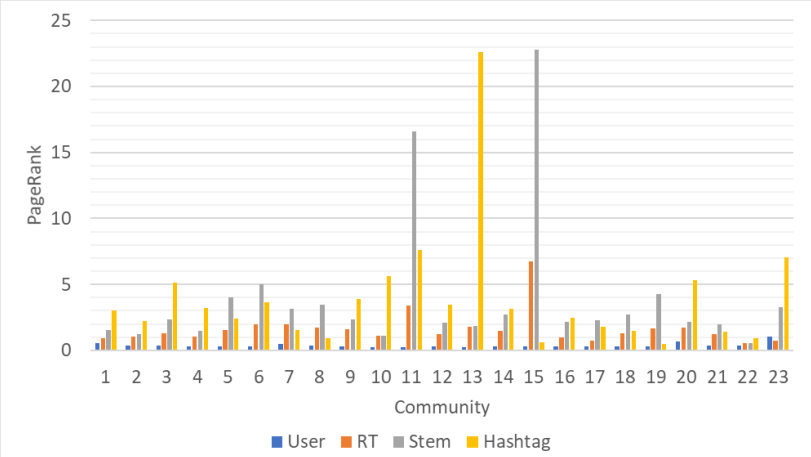



Figure 17: PageRank average per node type in each community.

PageRank was chosen because it generated a top 50 composed of all node types, which did not happen with eigenvector centrality, see Sec. 5.3. Having a more diverse node population, theoretically, helps in order to have a better inference about community theme. Thus, from visually analyzing the top 30 of every community in Fig. 18, Fig. 19, and Fig. 20, it is possible to note that only two communities (19 and 11) do not have all node types present. Also, most of the top 30 have stems as the majority nodes, except for Community 1. Stems do not have the highest PageRank scores on average (Fig. 17), but at the top 30 they are the majority in most communities. Another notable trend is that, the smaller the community is with the top 30, the more likely it is to be less connected.

# 6 Conclusions

The objective of this paper was to discuss Twitter data from the Brazilian 2018 presidential run-off election. The study to employed tools to facilitate storage, processing, and analysis of a considerable amount of data.

From a initial dataset of 1 million tweets, the study obtained 468,643 unique nodes after the preprocessing phase. Given that, theoretically assuming there is no empty tweet, there would be an average of 2.13 term repetitions within the corpus. Hashtags nodes had the highest average de-

gree (6145.0) followed by stems (1823.4), retweets (789.0), and finally users (259.1). Due to the function of indexing keywords and topics, hashtags would naturally have a higher average degree, because the same hashtag appears in several different tweets. A similar situation occurs with stem nodes, although in a lower intensity, because a single node has a generic meaning, with only a semantic function within a tweet. Retweets were the third in ranking probably because they are widely used to disseminate information in the exact way the tweet was initially published. Thus, there is little variability over time of terms that appear together. In other words, it is much more about repetition than variation of content and context. Users are ranked last. There is no apparent reason for that.

The employed tweet abstraction model made the network analysis easier and much more dynamic. Just one network it was necessary. Analysis was performed by queries that allowed selecting multiple types of nodes and connections.

Fig. 7 shows the degree distribution of the entire network and its subnetworks. The copresense subnetwork seems to follow a power-law distribution, while the entire network and authorship subnetwork do not seem to follow either power-law or Poisson distribution.

Fig. 8 shows the twenty most mentioned users and some of the most important users that mentioned them. The hubs coincide with the most prominent personalities and organizations during the period of the political campaign. The main users are @jairbolsonaro and @haddad_fernando, both presidential candidates, and secondarily there are users orbiting them. The fact that there is just a medium-size group of users connected to both candidates may suggest that the majority of users might get focused exclusively on attacking or supporting one of the sides, with regard to user mention. Besides mentioning, the other ways the interactions between electorate and candidates can happen is with hashtag and retweet.

The claim that users who support opposite parties or politicians hardly interact through retweets [11] was confirmed by the analysis shown in Fig. 9. In addition, no user retweeted more than one node from the top 20.

The results from Label Propagation algorithm were not satisfactory and consequently, were not analyzed. However, the Louvain algorithm produced results. There were 150,574 communities in total, but only 23 were analyzed. This small group concentrated more than 67% of all nodes, where the largest community had 43,415 nodes and the smallest just 84. The communities with just one node totaled 150,079. A top 30 according to PageRank metric was generated for each community. Most of the top 30 had stems for their majority of nodes. Their main node was hashtag (9 communities), fol-

lowed by stems (6 communities). Frequently research creates more questions than answers. Further studies should examine the communities in greater depth, to study main user roles and influences, hashtag and stem themes, etc. They also should probe the relationship among communities in order to understand the relationship structure and to examine whether neutral communities orbit polarized groups or form an independent major group.

Finally, polarization normally means expanding extremists and diminishing moderates. Moderate numbers diminish, but do not disappear completely. Therefore, except for Community 15, one could not find explicitly moderate users or moderate behavior in the data analyzed. It is impossible to say if moderates ignored the second round election completely or supported one of the sides; there is no data to establish by what ratio either happened nor can it be ascertained what role moderates had in communities, or even if they were present in the major ones.

## References

[1] Abramowitz, A. A. and Saunders, K. L., Is polarization a myth?, The Journal of Politics 70(2):542-55, 2018.

[2] Abramowitz, A., The disappearing center: engaged citizens, polarization, and American democracy, 2011, Yale University Press.

[3] Barabási, A. L., Network Science, 2016, Cambridge University Press.

[4] Barberá, P., Jost, J.T., Nagler, J., Tucker, J.A., Bonneau, R., Tweeting from left to right: Is online political communication more than an echo chamber? Psychological Science 26.10:1531-1542, 2015.

[5] Bonacich, P., Power and centrality: a family of measures, American journal of sociology, 92(5):1170-1182, 1987.

[6] Borge-Holthoefer, J., Magdy, W., and Darwish, K., Weber, I., Content and network dynamics behind Egyptian political polarization on Twitter, Proceedings of ACM Conference on Computer Supported Cooperative Work & Social Computing,700-711, 2015.

[7] Blondel, V. D., Guillaume, J., Lambiotte, R., Lefebvre, E., Fast unfolding of communities in large networks, Journal of Statistical Mechanics: Theory and Experiment, 2008(10):P10008, 2008.

[8] Carothers, T., O'Donohue, A., Democracies Divided: The Global Challenge of Political Polarization, 2019, Brookings Institution Press.

[9] Claassen, B. and Highton, B., Policy polarization among party elites and the significance of political awareness in the mass public. Political Research Quarterly, 62(3):538-551, 2009.

[10] Cram, L., Llewellyn, C., Hill, R., Magdy, W., UK General Election 2017: a Twitter Analysis, arXiv:1706.02271, 2017.

[11] Conover, M. D., Ratkiewicz, J., and Francisco, M., Gonçalves, B., Menczer, F. and Flammini, A., Political polarization on Twitter, AAAI Conference on Weblogs and Social Media, 2011.

[12] Costa, L. F., Rodrigues, F. A., Travieso, G. and Boas, P. R. V., Characterization of complex networks: a survey of measurements, Advances in Physics, 56:167-242, 2007.

[13] Costa, L. F., Oliveira, O. N., Travieso, G., Rodrigues, F. A., Boas, P. R. V., Antiqueira, L., Viana, M. P. and Rocha, L. E. C. Analyzing and Modeling Real-World Phenomena with Complex Networks: A Survey of Applications, arXiv:0711.3199v3, 2010.

[14] DiMaggio, P., Evans, J., Bryson, B., Have American's social attitudes become more polarized?, American Journal of Sociology 102.3: 690-755, 1996.

[15] Fiorina, M. P., Abrams, S. and Pope, J. C., Culture war? The myth of a polarized America, 2nd Edition, 2006, Pearson Longman.

[16] Gruzd, A. and Roy, J., Investigating political polarization on Twitter: A Canadian perspective, Policy & Internet 6(1):28-45, 2014.

[17] Kušen, E., Strembeck, M., Politics, sentiments, and misinformation: An analysis of the Twitter discussion on the 2016 Austrian Presidential Elections, Online Social Networks and Media 5:37-50, 2018.

[18] Lee, J. M., "The political consequences of elite and mass polarization." PhD thesis, University of Iowa, 2012. https://doi.org/10.17077/etd.vmg54vi9

[19] McCarty, N. and Poole, K. T. and Rosenthal, H., Polarized America: The dance of ideology and unequal riches, 2016, MIT Press.

[20] McCoy, J., and Somer, M., Toward a Theory of Pernicious Polarization and How It Harms Democracies: Comparative Evidence and Possible Remedies, Annals of the American Academy of Political and Social Science 681.1:234-271, 2019.

[21] Morstatter, F., Shao, Y., Galstyan, A., Karunasekera, S., From alt-right to alt-rechts: Twitter analysis of the 2017 German federal election, Proceedings of the The Web Conference, 621-628, 2018.

[22] Myers, S. A., Sharma, A., Gupta, P., Lin, J., Information network or social network?: The structure of the twitter follow graph, Proceedings of International Conference on World Wide Web, 493-498, 2014.

[23] Page, L., Brin, S., Motwani, R., Winograd, T., The pagerank citation ranking: Bringing order to the web, Stanford InfoLab, 1999.

[24] Raghavan, U. N., Albert, R., Kumara, S., Near linear time algorithm to detect community structures in large-scale networks, Physical Review E 76(3):036106, 2007.

[25] Suh, B., Hong, L., Pirolli, P., Chi, E. H., Multifractal analysis of low-latitude geomagnetic fluctuations, IEEE International Conference on Social Computing, 177-184, 2010.

[26] Superior Tribunal Federal, Brazilian Electoral Data Repository, http://www.tse.jus.br/hotsites/pesquisas-eleitorais/index.html

[27] Yaqub, U., Chun, S. A., Atluri, V., Vaidya, J., Analysis of political discourse on twitter in the context of the 2016 US presidential elections, Government Information Quarterly, 34(4):613-626, 2017.

# 7 Appendix

Table 10: Top 50 for eigenvector score.

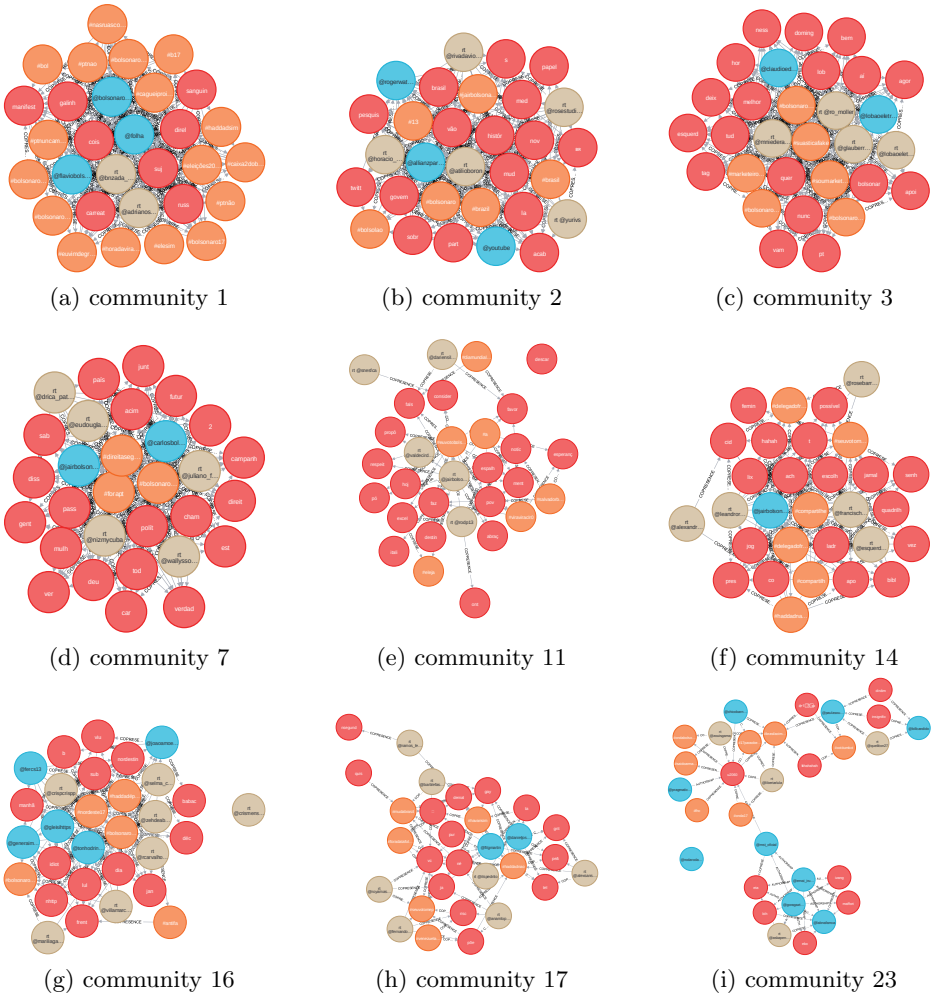
| Node | Eigenvector score |
|---|---|
| @estadaopolitica | 345.3 |
| é | 303.1 |
| brasil | 290.8 |
| bolsonar | 271.6 |
| @lobaoeletrico | 265.9 |
| @jairbolsonaro | 255.0 |
| @haddad_fernando | 249.2 |
| pt | 249.0 |
| pra | 246.7 |
| #elenunca | 244.7 |
| @zehdeabreu | 243.4 |
| @allianzparque | 235.9 |
| haddad | 232.1 |
| fal | 229.0 |
| faz | 228.0 |
| @globonews | 227.4 |
| presid | 222.2 |
| dia | 219.4 |
| #suasticafake | 219.3 |
| @tsejusbr | 219.3 |
| @mpf_pgr | 219.2 |
| @bolsonarosp | 216.6 |
| pod | 210.0 |
| vot | 208.3 |
| contr | 199.6 |
| agor | 196.3 |
| #elesim17 | 194.2 |
| pov | 194.0 |
| #ibopefake | 192.9 |
| #vote13 | 192.7 |
| diz | 192.6 |
| @manueladavila | 192.2 |
| @flaviobolsonaro | 190.3 |
| q | 189.3 |
| #cagueiproibope | 189.1 |
| #brazil | 188.5 |
| vai | 188.0 |
| #17neles | 187.0 |
| apoi | 187.0 |
| @folha | 186.8 |
| quer | 185.9 |
| candidat | 185.4 |
| #ptjamais | 183.6 |
| deu | 183.5 |
| @uol | 182.8 |
| #folhafakenews | 182.7 |
| @lulaoficial | 180.3 |
| #mito | 179.9 |
| qu | 179.8 |
| #eleições2018 | 179.6 |

Table 11: Top 50 for PageRank score.

| Node | PageRank score |
|---|---|
| #haddad13 | 4616.0 |
| #bolsonarosim | 2884.4 |
| #bolsonaropresidente | 2366.1 |
| #elenao | 2246.6 |
| é | 2155.7 |
| @jairbolsonaro | 1930.5 |
| #bolsonaro | 1890.2 |
| #bolsonaro17 | 1801.7 |
| #haddadpresidente | 1531.2 |
| rt @wallyssonrg2 | 1492.4 |
| #elenunca | 1463.2 |
| rt @haddad_fernando | 1281.1 |
| rt @glauberrosa | 1261.1 |
| @haddad_fernando | 1255.0 |
| brasil | 1230.1 |
| haddad | 1043.8 |
| rt @jairbolsonaro | 1007.3 |
| bolsonar | 995.3 |
| #elenão | 973.0 |
| #marketeirosdojair | 933.0 |
| pra | 872.7 |
| #elesim | 864.5 |
| vot | 825.1 |
| #euvotobolsonaro | 814.7 |
| #haddadéfefeca | 813.6 |
| rt @lobaoeletrico | 775.8 |
| #haddadsim | 730.0 |
| pov | 719.9 |
| rt @samyy_l | 697.1 |
| tod | 683.8 |
| #bolsonaro2018 | 653.3 |
| xa0 | 649.7 |
| rt @vs_ferreira | 641.5 |
| #ptnão | 637.0 |
| rt @_vitroia | 624.6 |
| xa0 | 624.4 |
| #haddad | 615.2 |
| #b17 | 606.2 |
| #viravoto | 604.5 |
| russ | 602.0 |
| ser | 598.6 |
| pt | 597.1 |
| democrac | 595.3 |
| dia | 543.7 |
| @lobaoeletrico | 540.6 |
| morr | 533.8 |
| faz | 529.2 |
| hoj | 526.7 |
| vai | 517.8 |
| rt @akayona_ | 495.2 |

Table 12: Node type ratios per community.

| Community | User | RT | Text | Hashtags |
|---|---|---|---|---|
| 1 | 48.7% | 13.1% | 23.9% | 14.4% |
| 2 | 59.5% | 10.8% | 23.8% | 5.9% |
| 3 | 59.7% | 16.0% | 19.8% | 4.5% |
| 4 | 70.5% | 9.6% | 16.4% | 3.6% |
| 5 | 70.2% | 13.5% | 12.2% | 4.1% |
| 6 | 79.8% | 7.5% | 9.3% | 3.5% |
| 7 | 60.3% | 14.6% | 18.5% | 6.7% |
| 8 | 71.7% | 11.5% | 12.2% | 4.6% |
| 9 | 64.2% | 13.4% | 16.6% | 5.8% |
| 10 | 75.8% | 7.8% | 13.0% | 3.4% |
| 11 | 93.2% | 3.1% | 2.6% | 1.2% |
| 12 | 61.0% | 13.4% | 20.2% | 5.4% |
| 13 | 61.5% | 11.4% | 23.5% | 3.6% |
| 14 | 61.7% | 13.4% | 18.1% | 6.7% |
| 15 | 96.5% | 1.6% | 1.1% | 0.7% |
| 16 | 56.6% | 17.5% | 18.3% | 7.5% |
| 17 | 66.7% | 14.6% | 13.3% | 5.5% |
| 18 | 62.0% | 15.6% | 16.4% | 6.0% |
| 19 | 70.8% | 13.3% | 10.4% | 5.4% |
| 20 | 49.8% | 15.4% | 28.0% | 6.7% |
| 21 | 63.7% | 11.8% | 16.3% | 8.2% |
| 22 | 26.5% | 7.6% | 47.7% | 18.2% |
| 23 | 34.5% | 7.1% | 40.5% | 17.9% |

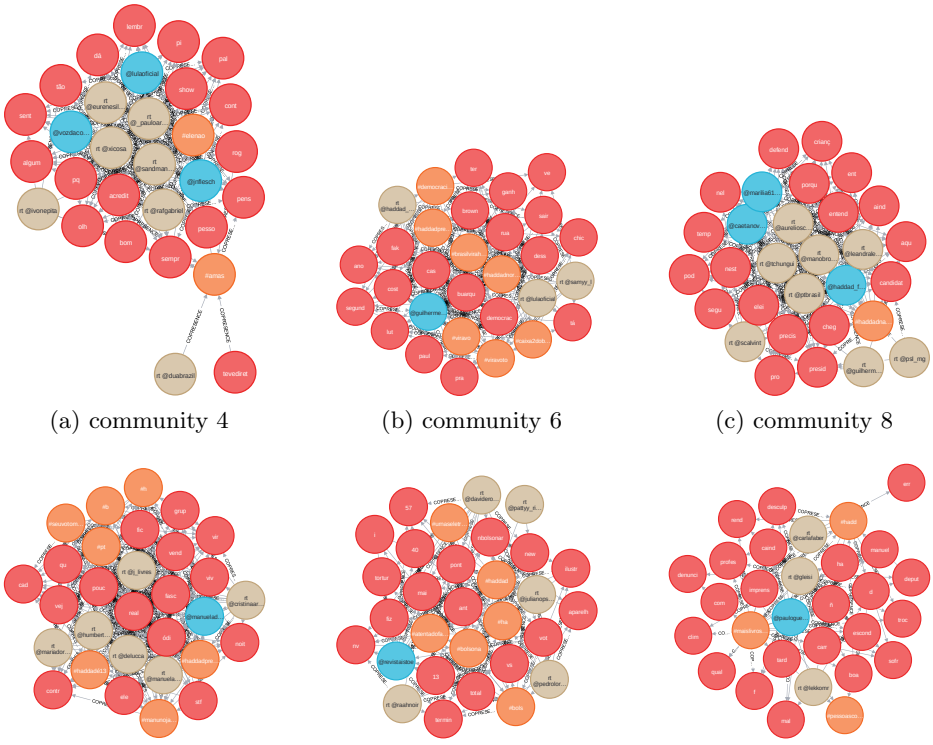


(a) community 1 (b) community 2 (c) community 3

(d) community 7 (e) community 11 (f) community 14

(g) community 16 (h) community 17 (i) community 23

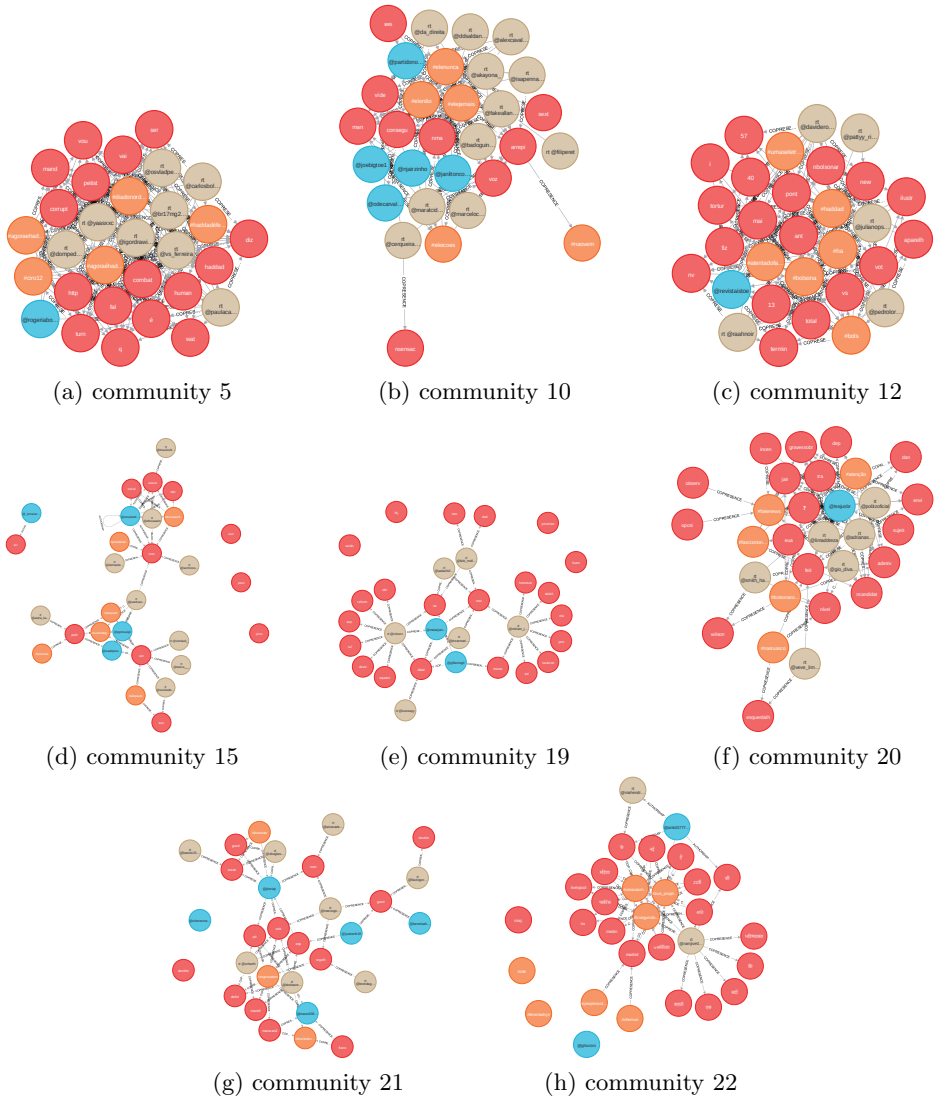
Figure 18: PageRank top 30 for Bolsonaro-related communities.



(a) community 4 (b) community 6 (c) community 8

(d) community 9 (e) community 12 (f) community 18

Figure 19: PageRank top 30 for Haddad-related communities.



(a) community 5 (b) community 10 (c) community 12

(d) community 15 (e) community 19 (f) community 20

(g) community 21 (h) community 22

Figure 20: PageRank top 30 for neutral communities.