2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

Detecting Influential Communities in Twitter during Brazil Oil Field Auction in 2019

Francis Spiegel Rubin
PPGI/UNIRIO
Rio de Janeiro, Brazil
fran.spiegel@edu.unirio.br

Adriana C.F. Alvim
PPGI/UNIRIO
Rio de Janeiro, Brazil
adriana@uniriotec.br

Rodrigo Pereira dos Santos
PPGI/UNIRIO
Rio de Janeiro, Brazil
rps@uniriotec.br

Carlos Eduardo Ribeiro de Mello
PPGI/UNIRIO
Rio de Janeiro, Brazil
mello@uniriotec.br

***Abstract*—Critical issues in the Oil & Gas business have been discussed in social media, involving policymakers, oil and gas companies, industry individuals, and civic activists. As observed recently, such issues have a critical role in the economic debate and political polarization. In this context, this paper reports on a study for the detection of hidden communities and influential users on social networks through the analysis of Twitter posts during Brazil mega oil fields auction in 2019. We use a methodology focused on the identification of relevant users for the target community and the discovery of effects of interactions between interest groups based on the modeling of the oil and gas social network. Firstly, we applied a standard community detection algorithm called Louvain method. Next, we used the Herfindahl-Hirschman index to identify the most influential users in the network by measuring the distribution of users' influence in a community. As a result, it was possible to visualize the presence of clusters and specific topics of interest shared by the most influential communities. Moreover, it was also possible to distinguish the convergent and divergent opinions that permeated the pre-salt auction and the level of Twitter engagement of key players from the global energy market.**

Keywords—Brazil Oil Auction, Influential Communities, Twitter Analysis, Complex Networks, Community Detection

I. INTRODUCTION

The energy sector is historically considered a conservative sector in an environment of a controversial reputation [1]. In this scenario, the engagement of the oil and gas industry in social media has increased over the past decade mainly to enhance communication and address the needs of different stakeholders in an attempt to gain legitimacy for their long-term prosperity [2]. Given stakeholders' skepticism and overall negative reputation in the oil and gas industry, the public agenda is also being widely publicized and debated in social media as an attempt to engage the main protagonists of the energy sector in a two-way conversation [3, 4]. As key protagonists, we can cite regulatory agencies, oil and gas companies, civil and environmental activists, government, and industry professionals. In this context, the pre-salt mega oil auction featured an opportune time to capture implicit relations from Twitter[1] and to understand the intrinsic features of the oil and gas industry.

In today's socially conscious environment, the relationships and degrees of influence among the protagonists of the oil and gas industry highlight a strategic role of strengthening the promotion of political and economic debates as an attempt to restore public trust with society. Therefore, it is crucial to detect users with similar interests, whose content was mostly shared and commented in each group of protagonists (also known as communities) as well as who are more likely to maximize their opinion and spread their ideas with greater impact than other individuals. Using, for example, Twitter platform, people interact with each other by sending retweets, mentions, and citations. The interactions range from professional connections to common interests for breaking news. Thus, the induced networks of related cyber users become a valuable source for the dissemination of information since they provide the freedom to spread opinion and influence to other cyberspace users.

Based on this scenario, this study seeks to understand the dynamics of the most influential communities in social network platforms through the analysis of Twitter posts collected during the week of the pre-salt oil auction that took place in November 2019 in Rio de Janeiro, Brazil. In the context of complex networks, we aim to model and analyze the existing communication flow in a social phenomena, by connecting a set of nodes, which form a network or graph and identify the communities present in each network.

In our research, we applied a graph-based model to identify the most influential users and to model the leading networks of influence. In the first part of our experimental study, we applied a standard algorithm for community detection, the Louvain method, to identify communities and main interest groups. In the second part, the Herfindahl-Hirschman index was adopted as an approach to identify the most influential users by measuring the distribution and concentration of influence in each detected community.

The results reveal influential demarcated communities and capture implicit relations between influential users assigned to each community. Also, we point out observable differences in the level of engagement among the major oil companies, government, and other protagonists from the oil industry. The Chinese companies were heavily centric, with no interactions captured in social media. In contrast, the Brazilian authorities played an active role in social media, leading by ANP and Petrobras. Our analysis also indicates a set of likeminded users and divergent opinions that permeated the auction towards several discussion issues published in Twitter among the major communities. Furthermore, it was also possible to combine distinct computational techniques to obtain the results of the experimental study.

To guide the achievement of this study, we define the following research question (RQ): *How was the mega auction*

[1] http://www.twitter.com

IEEE/ACM ASONAM 2020, December 7-10, 2020
978-1-7281-1056-1/20/$31.00 ©2020 IEEE

of the so-called "Transfer of Rights" ("Cessão Onerosa" in Portuguese) oil fields reflected on Twitter?

In order to answer our RQ, we elaborate three sub-questions (Sub-Q) as follows:

Sub-Q1: *Within a network emerged from Twitter posts during Brazil mega oil fields auction in 2019, which communities were the most active and had the most interactions?*

Sub-Q2: *Who are the most influential users within the identified communities?*

Sub-Q3: *What were the most convergent and divergent views during the mega oil auction scenario?*

The remainder of this paper is organized as follows: in Section II, we present the background. In Section III, we discuss related work. Section IV brings the methodology. Section V details the results and discussion of the experiments. We also highlight findings and some unexpected results regarding our RQ. Finally, in Section VI, we conclude the paper with final considerations and future work.

II. BACKGROUND

In this section, we summarize the basic definitions and notations used in this work.

A. Basic concepts and notation

For modeling social networks, we use the following concepts and notations from the study of graphs [5]. A graph $G = (V, E)$ is simply a way of encoding pairwise relationships among a set of objects: it consists of a collection of nodes V and a collection of edges E, that "joins" two nodes. We represent an edge $e \in E$ as a two-element subset of V: $e = \{u, v\}$ for some $u, v \in V$, where we call u and v the ends of e. Edges in a graph indicate a symmetric relationship between their ends. When we want to encode asymmetric relationships, we use the closely related notion of a directed graph or digraph. A directed graph G' is a pair $G' = (V, E')$ where V is the set of nodes and E' is a set of directed edges, i.e., each $e' \in E'$ is an ordered pair (u, v) and the roles of u and v are not interchangeable. We consider that edge e' leaves node u and enters node v. In this article, we use the notation for ordered pairs $e = (u, v)$ for the undirected graph too. Moreover, we shall consider weighted graphs, which is a graph $G = (V, E)$ with a function w from u to v. We denote the weight of edge $e = (u, v)$ by $w(u, v)$.

The measure of centrality is commonly used to identify relevant nodes within a network. To measure the centrality of graph nodes, we consider the betweenness index metric [6] to represent the degree to which nodes stand between each other. Let $\sigma_{ij} = \sigma_{ji}$ denote the number of shortest paths from $i \in V$ to $j \in V$, where $\sigma_{ii} = 1$ by convention. Let $\sigma_{ij}(v)$ denote the number of shortest paths linking nodes i and j that pass through $v \in V$. The betweenness index $g(v)$ for a node v can be adapted from [6] as:

$$g(v) = \sum_{i \neq v \neq j} \frac{\sigma_{ij(v)}}{\sigma_{ij}} \quad (1)$$

To measure how connected a node is within a graph, we define the node's degree as the number of edges incident on it. In a directed graph, in-degree is the number of incoming edges that enter v, and the out-degree represents the number of outgoing edges that come from v. For undirected graphs, degree is equal to in-degree plus out-degree. We also define the network density as the ratio between the number of existing edges by the number of possible ones contained in a graph.

B. Network influence, community detection and community influence definitions

In this section, we present the definition of influence in a network, community detection, and community influence.

Let the directed graph $G = (V, E)$ be a network in which edge $e = (u, v) \in E$ indicates that node u has some relation with node v and weight $w(u, v)$ indicates the power of the edge $e = (u, v)$ influence. Then, the influence $I(u)$ for node u is defined as the sum of all the edges weight [9]:

$$I(u) = \sum_{e(u,v) \in E} w(u, v) \quad (2)$$

The concept of "community" can be defined as a subset of nodes or vertices that are more densely connected, while the nodes of different communities are sparsely connected [7]. In other words, a community is made up of nodes that often interact with each other to the detriment of others outside the group of interactions. Thus, the objective of community detection of network G is to partition a graph $G = (V, E)$ into k communities (or clusters) $C_1, C_2, \ldots, C_k$ that split V into $V_1, V_2, .., V_k$ components having $\bigcup_{i=1}^{k} V_{i,} = V; V_i \cap V_j = \emptyset$ for $1 \leq i; j \leq k; i \neq j$ and $V_i \neq \emptyset$ for $1 \leq i \leq k$. Each node in V_i is associated with a cluster $C_{i,}$ for $1 \leq i \leq k$ [8].

Community influence is defined as the cumulative influence of all its nodes [9]. The community influence can be defined as:

$$I(C) = \sum_{u \in C} I(u) = \sum_{u \in C} (\sum_{e(u,v) \in E} w(u, v)) \quad (3)$$

The intra-community influence $I_{in}(C)$ is derived from $I(C)$ and it represents the influence that the community nodes have within their community. So, the intra-community influence can be defined as:

$$I_{in}(C) = \sum_{u \in C} I_{in}(u) = \sum_{u \in C} (\sum_{e(u,v) \in E, v \in C} w(u, v)) \quad (4)$$

Another important metric is the node influence share within a community, which is defined as the ratio between the degree of influence for a node i and the intra-community influence. In other words, we can set r_i as the percentage of influence for the leading nodes in a community C_k with $|V_k|$ nodes as:

$$r_i = \frac{I_{in}(i)}{\sum_{j \in V_k} I_{in}(j)} = \frac{\sum_{e(i,j) \in E, j \in C_k} w(i, j)}{I_{in}(C_k)} = \frac{\sum_{e(i,j) \in E, j \in C_k} w(i, j)}{\sum_{i,j \in E, j \in V_k} w(i, j)} \quad (5)$$

C. The Herfindahl-Hirschman Index (HHI)

HHI is commonly used in economics as a measure of concentration and competition among market participants [10]. It is calculated by adding the squares of market shares and was proposed by two economists: Hirschman [11] in 1945 and later published in 1950 by Herfindahl [12] in his doctoral thesis "Concentration in the US Steel Industry". In analogy to the concept of market share, values close to 0% indicate perfect competition between firms and values close to 100% indicate the presence of monopoly.

In a community C_k with $|V_k|$ nodes, we can calculate the HHI for community C_k in terms of the intra-community influence as the squared sum of influence ratios (r_i), for each node $i \in V_k$, where ratios range from $\frac{1}{|Vk|}$ to 1. Higher values indicate concentration on a few influential nodes, whereas

low values represent a more balanced and dispersed community influence. HHI is defined as:

$$HHI(C_k) = \sum_{i \in V_k} r_i^2 = \sum_{i \in V_k} \left(\frac{I_{in}(i)}{\sum_{j\, in\, V_k} I_{in}(j)}\right)^2 \quad (6)$$

Equations (2), (3), (4), (5) and (6) are adapted versions, derived from [9].

D. The Louvain method

This heuristic method proposed by Blondel et al. [13] uses the concept of Modularity [14, 15] as an objective function to measure the quality of the communities, by measuring the density of links within communities compared to the links mapped between communities. It consists of the interactive repetition of two phases, as illustrated in Fig. 1 with the identification of four communities after step 1, and two communities after step 2. Initially, each node of the graph is a community. For each node in the graph, the gain in modularity is evaluated by removing a node from one community and moving it to the neighboring communities. If modularity improves, the node moves to the new community. The first phase ends when there are no more gains in modularity, which means we reached the maximum local value. The second phase consists of the construction of a new graph, where new communities are shaped by the exit of the grouped nodes of the first phase. The algorithm finishes when there are no more gains of modularity, i.e., when the detection of communities remains stable. Blondel et al. [13] compared the Louvain method with other techniques and concluded that the proposed algorithm achieved significant modularity in less computational time than other heuristic approaches.

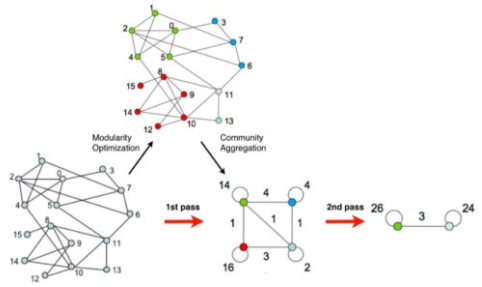

Fig. 1. Illustration of the Louvain method. [12, Fig.1]

E. The Force-Directed method

The Force-Directed algorithm [16] allows the modeling of graphs as a physical system of bodies with forces acting on each other that seek to minimize the energy of the system. In this context, it defines the coordinates of each vertex in space for visualization in 2-dimensions with similar size edges. The intersections are balanced between the pairs of nodes, facilitating successful investigation and interpretation of data. The implementation of this algorithm uses the repulsive forces between the nodes and the forces of attraction between neighboring nodes, and is inspired by the laws of physics of Hooke [17] and Coulomb [18]. This algorithm allowed the visualization of the behavior dynamics revealed by the Louvain communities. A disadvantage of this algorithm is the high computational cost, usually in the order of $O(n^3)$ or $O(n^2)$, being n the number of nodes, which limits its application in graphs with a high number of nodes. We apply this algorithm to present the nodes and edges in a way that groups of nodes with many interconnections are positioned close to one another.

III. Related Work

The problem of community detection has been gaining prominence in the scientific community in recent years and is considered one of the most promising areas of research on social networks [19]. Nevertheless, identifying communities in an arbitrary network is an NP-Hard problem and then is not a computationally trivial task since no polynomial time solution is known to obtain the exact number of elements in a network [20]. In the literature, several studies have been developed to detect communities, applying different techniques [21]: maximum likelihood [22], mathematical programming [23], among others. Most of these techniques are discussed in articles [21, 24, 25].

For example, Granovetter et al. [26] reported that the existing links within communities tend to be stronger, and the links between them tend to be weaker. Java et al. [27] investigated the presence of clusters in interactions between users, enabling the identification of participation percentages for each user concerning the total interactions per cluster. Keyi Shen et al. [28] proposed a hierarchical algorithm to detect community structures by analyzing the number of familiar neighbors as a measure of similarity. In another work, Xinyu Que et al. [29] proposed a parallelization of the Louvain method to detect communities in distributed systems. The work of Xinyu Que et al. achieved significant performance in terms of modularity by adopting hierarchical schemes for the processing of graphs.

It is not rare to see computing environments with insufficient memory size for loading and data processing in their completeness. To guarantee excellent performance in limited computing environments, Yong-Hyuk Kim et al. [30] proposed a variation of the Louvain method based on algorithm partitioning, adding minimization techniques. In turn, Conover et al. [31] investigated the detection of communities using the propagation method proposed by Raghavan et al. [32], by applying a greedy method that interactively associates a classification to each node shared among neighboring nodes. The authors achieved good results in inducing distinct network topologies.

Sluban et al. [9] presented a methodology to identify influential communities from data collected from Twitter. The authors focused on building a network of retweets among users who shared the same interest by extracting the most densely connected communities. The published content was analyzed to infer the common interests in each community. Furthermore, the authors explored Louvain's techniques for community detection and for identifying the most influential users using the Herfindahl-Hirschman index (HHI), with promising results to improve the dynamic flow understanding of interactions and information present in social networks.

In general, the cited works explored distinct community detection techniques to induce networks and served as a basis and motivation for our research. Similarly to the work of Sluban et al. [9], we also applied the Louvain method and the HHI to investigate our data. In contrast, we focus our analysis on the pre-salt mega oil field auction in Brazil based on Twitter publications. Therefore, the detection of communities characterized by influential protagonists from the oil and gas industry captured during the pre-salt oil auction, together with the analysis of their prevalent discussion topic trends, are the main contributions of this research.

IV. METHODOLOGY

This section presents the steps involved in the process of selecting and extracting data, and in applying algorithms to detect users' communities as well as hashtags' communities of significant influence. In the next subsections, we detail the proposed steps of the research: (A) Data Capture, (B) Data Cleaning and Pre-Processing, (C) Community Detection, (D) Viewing of Communities, (E) Identification of Influential Users, and (F) Identification of Prominent Hashtags. Fig. 2 shows the steps and techniques applied in this experimental study.

A. Data Capture

1) Collection period definition

The collected data correspond to the period between November 4th and November 11th, 2019. The day scheduled to host the pre-salt oil mega auction concessions took place on November 6th, in Brazil.

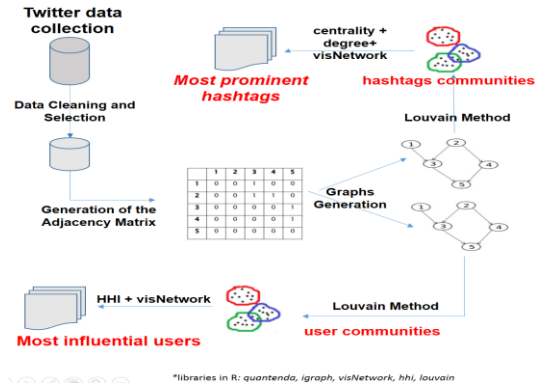

Fig. 2. Proposed steps and techniques for the research

2) Identification of tweets (Twitter publications) corresponding to the pre-salt oil auction

Portuguese terms related to the context of the oil and gas industry were selected to enable the analysis of the publications referring to the oil auction. The selected terms served as input parameters for filtering the tweets. To collect as many tweets as possible, in the context of the oil auction, the potential variations of punctuation, accentuation, and spelling of the selected terms were considered. We defined the following filters by manually sampling the most frequent mentions of auction-related search terms on Twitter: "Cessão Onerosa", "ANP", "Pré-Sal", "LeilãodoPréSal", "MegaLeilão", "Petróleo", and "Petrobras".

3) Data capture execution

To retrieve content from Twitter, we run the Python TwitterScraper[2] library. It allows the application of filters by selected terms and date range. The program scans Twitter publications based on the input parameters. The data collected resulted in a total of 50,521 tweets. We reduced the dataset to 50,381 tweets published by 30,158 distinct accounts after the removal of duplicate tweets.

B. Data Cleaning and Pre-processing

In the cleaning process, we removed from the tweets collection punctuation, accents, articles, and prepositions, except for those marked with '#' and '@'. These markers identify the hashtags used in each tweet and the presence of interactions between separate accounts. The detection of stopwords in Portuguese/Spanish was also applied to discard terms of no relevance or little meaning for this research.

In the following, we define the steps to construct our *User's Network*. From the dataset with 50,381 tweets, we selected the users mentioned in at least one tweet, which results in the set of *Mentioned Users* with 3,212 users mentioned in 7,983 tweets. Next, we define the *Users' Network* as a network given by a weighted directed graph $G = (V, E)$ containing an edge $e = (u, v)$ if user v is mentioned by user u in, at least, a specific tweet. That is, user v is marked with the '@' character in, at least, one tweet by user u. Then, we select users from set Mentioned Users that are mentioned the most (degree above 8) and delete loops. The result set E has a cardinality $|E| = 778$. Duplicate edges are then combined and added up together to compute the weight of the new edge, meaning the number of tweets that user v is mentioned by user u. Then, to extract the backbone of the network, so the underlying structure emerges more clearly, we remove edges with weights equal to one and delete isolated nodes. The final set V has a cardinality $|V| = 108$ and set E has a cardinality $|E| = 141$. To illustrate, consider a small Mentioned Users Set instance with four tweets (t1="oi @chevron", t2="oi @petrobras", t3="oi @chevron @petrobras", t4="oi @petrobras" posted by four distinct author (a1="@petrobras", a2="@chevron", a3="@cnodc", a4="@anp"). For this instance, the set with the two Most Mentioned Users is: @petrobras cited in three tweets (t2, t3, t4), and @chevron cited in two tweets (t1, t3). Fig. 3 shows the illustration of the Users' Network for this small instance.

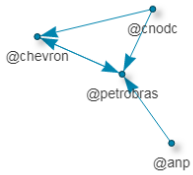

Fig. 3. Illustration for the small instance. @petrobras has in-degree equals to 3, @chevron has in-degree equals to 2, @cnodc and @anp has in-degree equal to 0. Edges weight equal to 1.

Fig. 4 shows the Users' Network with 108 users and 141 edges. It shows the most frequent mentions and aims to reveal users with similar interests and unexpected relations without examining the actual content. We used the Influence Metric (2) to calculate the degree of influence for each mentioned user.

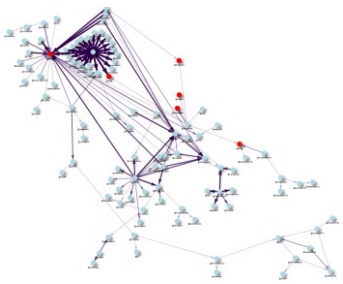

Fig. 4. Users' Network with 108 users and 141 edges – oil companies colored in 'red'

Then, we define our *Hashtags' Network*. The dataset with 50,381 tweets contained a set of 5,135 *Distinct Hashtags*. Let *Most Popular Hashtags* be the set with the 110 most popular hashtags from the set of Distinct Hashtags. The Hashtags' Network is a network given by an undirected weighted graph $G = (V, E)$ containing an edge $e = (u, v)$ if hashtag u occurs, at least, in a specific tweet than hashtag v, meaning a co-occurrence of both hashtags. Set V consists of all hashtags

[2] https://pypi.org/project/twitterscraper/0.2.7/

from set Most Popular Hashtags with degree above zero and has a cardinality $|V| = 106$. The weight $w(u, v)$ of edge e means the number of tweets that hashtag u and v have in common, i.e., the frequency of co-occurrences. Fig.5 shows the Hashtags' Network. It contains 106 hashtags and 833 edges and aims to reveal distinct views published on Twitter during the oil auction scenario. Then, we remove spurious connections with weights equal to one and isolated nodes, resulting in 102 hashtags and 590 edges. We derived the popularity of each hashtag according to the number of tweets that hashtag u and v have in common.

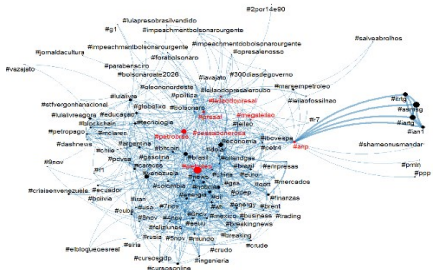

Fig. 5. Hashtags' Network with 106 hashtags and 833 edges – hashtags with filters "Cessão Onerosa", "ANP", "Pré-Sal", "LeilãodoPréSal", "MegaLeilão", "Petróleo" and "Petrobras" are colored in 'red'

C. Community Detection

We applied the Louvain method to detect the communities in the Hashtags Network and within the Users' Network by calling the *cluster_louvain* function from the *igraph* R library. We named the *Hashtags Communities* to represent the communities identified in the Hashtags' Network and the *Users Communities* to represent the communities identified in the Users' Network. To calculate the measurements of modularity and network density, we called the *modularity* and *edge_density* functions. Next, for Users' Network communities, we calculated the community degree of influence according to the intra-community influence metric (equation (4)).

D. Viewing of Communities

The communities found by the Louvain method in the previous stage represent a division power of the primary influencing users during the pre-salt oil auction week. To highlight the most relevant relationships captured among individuals and organizations, and the main themes discussed in Twitter, it was essential to select an algorithm that allowed us to visualize the Louvain communities. The Burnes-Hut Force-Directed Layout algorithm grows only as $O(nlogn)$ [33] and allows the visualization of the internal and external relations between community nodes, giving a clear understanding of the communities. For this step, we used the *visNetwork* library in R. An advantage of this library is the simplicity and interactivity in a real problem modeling, as well as in the understanding of the dynamics of the communities.

E. Identification of Influential Users

To detect the most influential users within each User Community, we applied the node influence share metric (equation (5)). To assess whether the detected communities have a balanced distribution among their users, or whether the behavior of these networks characterizes a monopoly where few users concentrate the most interactions, the calculation of HHI was applied (equation (6)). We inputted the selected users and the shares of influence calculated for each user, as defined in equation (5). Then, we selected the *hhi* library in R to calculate HHI for each community, according to equation (6). The hhi is calculated by squaring each user's percentage share, expressed as decimal, and then summing over each user's squared share. An index close to zero indicates nearly perfect competition with no dominant users, and close to one indicates monopoly. The presence of distinct groups supporting or criticizing the auction allowed the identification of niches within the detected communities. Finally, we display the most influential users, complementing the users' community analysis.

F. Identification of Prominent Hashtags

To identify the most prominent hashtags, we apply the betweenness index (equation (1)) to compute the centrality for the hashtags. The identification of distinct central nodes within the hashtags community allowed the identification of the main convergent and divergent views published in Twitter during the pre-salt mega oil auction week. In order to answer Sub-Q3, we complemented our analysis by highlighting the most significant hashtags for each community according to its centrality and degree. We report on the results in the following section.

V. Results and Discussion

In this section, we describe the results and report analysis and discussion.

A. Users Network

This network summed 108 vertices and 141 edges, corresponding to 1% density concerning the total possible edges. We observed that 84% of the vertices' degrees range between 0 and 5, with few users retaining a high centralizing power within their interactions. Fig. 6 shows the distribution of vertices degrees.

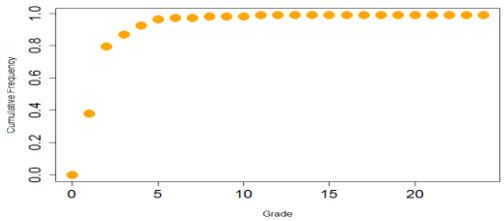

Fig. 6. Frequency per degree – Users' Network

To answer Sub-Q1, we detected the presence of seven communities. We labeled them "C1", "C2", …, "C7". Each detected community was assigned to a different color by the Louvain method. Fig. 7 shows a different approach than the graph structure from Fig. 4 by applying the Burnes-Hut forced-directed algorithm. We can easily visualize the communities with nodes acting as charged particles that repel each other, and links acting as springs that pull related nodes together [34].

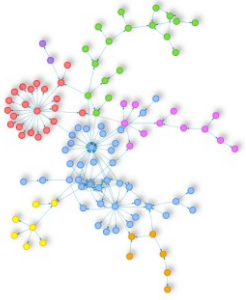

Fig. 7. Visualization of the seven demarcated Users' Communities

Table I indicates the following measures for each community C: "#v": the number of vertices; "#e": the number of edges; "nd": network density; and "ici": intra-community influence. The following measures supported the answer to Sub-Q2.

TABLE I. USERS' COMMUNITIES MEASURES

C	#v	#e	nd	ici	hhi	color
1	41	66	0.04	912	0,26	Blue
2	2	1	0.5	28	1	Purple
3	6	5	0.16	257	0,40	Orange
4	22	21	0.04	435	0,07	Red
5	7	6	0.14	53	0,24	Yellow
6	18	19	0.06	350	0,17	Green
7	12	11	0.08	161	0,15	Purple

Legend: C – community label; #v – number of vertices; #e – number of edges; nd – network density; ici – intra-community influence; hhi – hhi metric; color – color designated.

Then, we observed dense connections within the communities and sparse connections between communities. From 141 edges, 15% connected vertices from different communities, and 85% connected vertices from the same community. Communities "C1", "C4" and "C6" had the highest number of nodes. Communities "C1" and "C6" had the lowest network density ratio. Communities "C1" and "C4" had the highest intra-community influence. Communities "C2", "C3, "C5" had the highest density ratio, with users interacting actively. Community "C2" presented disconnected behavior and little interaction with users from other communities.

We could also observe the protagonism of users in the function of connecting distinct communities, also known as separating vertices, whose removal would disconnect the rest of the graph in independent pieces. Table II shows a sample of vertices with this behavior.

TABLE II. USERS CONNECTING DIFFERENT COMMUNITIES

Users	Communities
@lulaoficial	1, 6, 7
@anp	1, 5
@petrobras	1, 4, 6, 7

As noted, Community "C1" showed the greatest intra-community influence based on the users' influence metric for this network. This community also had the highest number of vertices, and the highest share of federal government users. Given the importance of Community "C1", we generated a subgraph Community $C1' = (V', E')$ from Community $C1 = (V, E)$, where V' equal to V and edge $e(u', v') \in E'$ exists only if u' and $v' \in V'$. Next, we conducted a new application of the Louvain method exclusively on the generated subgraph for Community C1', to deepen the analysis on the role of the most relevant community of the graph. By exploding this subgraph into new communities, we could observe the characteristics of the most representative niches in the generated subgraphs. We segmented the 41 users from Community C1' into six new communities, which we called *niches*, according to Table III. Fig. 8 illustrates the niches identified for Community C1'.

In addition to the analysis, we could observe a strong link between the current government committee (Ministry of Mines and Energy, Ministry of the Environment and Ministry of Communications) in Niches 1.2 and 1.3, and Petrobras, the mega bidding round winner in Niche 1.4. It was also interesting to observe the representativeness of Niches 1.6 and 1.2, with the activity of opposition users to the current federal government, environmental activists and the Brasilian Army.

TABLE III. TOTAL OF VERTICES – COMMUNITY C1'

N1.1 (red)	N1.2 (blue)	N1.3 (pink)	N1.4 (yellow)	N1.5 (purple)	N1.6 (green)
3	12	4	10	5	7

Legend: N - Niche

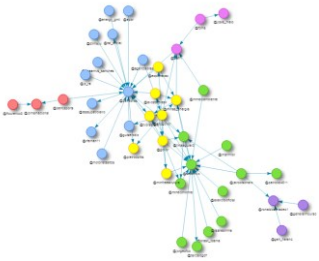

Fig. 8. Niches from Community C1'

Table IV shows the two most influential users and their percentage of representativeness in each niche for C1'.

TABLE IV. MOST INFLUENTIAL USERS – COMMUNITY C1'

Niches	Users	% Representativeness
Niche 1.1	@jornalnacional @bemjsports	40.0 40.0
Niche 1.2	@anpgovbr @govbr	21.0 14.0
Niche 1.3	@paf11 @folha	50.0 20.0
Niche 1.4	@petrobras @epbr	68.0 3.0
Niche 1.5	@jairbolsonaro @ronaldosenadeo	40.0 10.0
Niche 1.6	@secomvc @brasil_ibama	71.0 4.0

As we noticed, Communities "C1" and "C4" presented the highest degree of equal share among its users, followed by Communities "C7", "C6". The absence of "monopoly" in Community "C1", the largest community of the network, indicated a natural characteristic of interactivity among the leading users. There was no exacerbated concern among the main protagonists of the mega auction to play a "monopolistic" role in Twitter. This characteristic corroborates, to some extent, the conservative behavior expected for the oil and gas sector. By detailing the participation of the users from the captured data, we could observe the prominent presence of government agencies and Petrobras, as the holder of the most mentioned interactions. However, concerning Communities "C2" and "C3", we see a steady trend of concentration in their interactions.

It was also interesting to observe the activity of former presidential candidates in "C2" (e.g. @cirogomes), cryptomarket users in "C4" (e.g. @criptomoedas, @bitcoin, @blockchain), the national press in "C6" (e.g. @g1, @portalr7, @uol), and the international press in "C5" (e.g. @lemondefr, @telegraaf). We also noticed the presence of activists and opposition against the auction in "C4" (e.g. @ptbrasil, @lulaoficial, @lulalivre) and "C6" (e.g. @fup_brasil, @cut_brasil). Although their representativeness was not relevant to the total number of interactions, the

movement of these users was organized, concentrated and of high representation in communities. The national oil regulatory agency from Brazil (ANP) also played an essential role in connecting distinct users, and its content was exhaustively explored by Community "C1". We also noted the presence of mentioned tweets between the oil and gas companies that did not participate in the auction, e.g. PDVSA and Venezuelan government users in Community "C4", and Pemex in Community "C7".

Moreover, we observed that among the thirteen companies qualified to participate in the mega bidding round[3], only Ecopetrol, ExxonMobil, Shell and Chevron presented activity records in Twitter, gathered in Community "C7", besides Petrobras in "C1". There was no Twitter record activity by the Chinese state-owned companies CNOOC and CNODC, even though they were the only foreign representatives that submited bids during the auction to form a consortium. This behavior suggests that Chinese companies remain knit to party-state strategic interests and policies, despite their enthusiastic demand for hydrocarbons resources [35].

B. *Hashtags Network*

This network summed 102 vertices and 590 edges with 11% of network density and asymmetrical distribution of vertices degrees. We can observe a minority of high degree vertices in the distribution chart, as shown in Fig. 9.

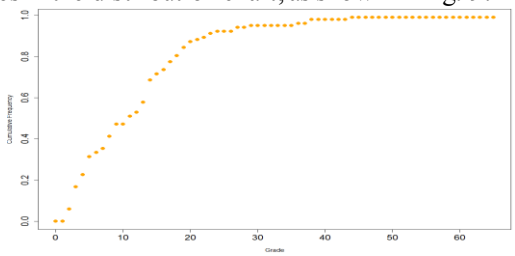

Fig. 9. Frequency per degree – Hashtags Network

After the application of the Louvain method, we detected the presence of seven hashtags communities and modularity of 61%. 160 edges out of 590 edges (27%) connected vertices from distinct communities, and 430 edges (73%) connected vertices from the same community. Table V shows the hashtags community measures. To answer Sub-Q3, we generated the most relevant hashtags according to the degrees and the betweenness index.

TABLE V. HASHTAGS' COMMUNITY MEASURES

C	#v	#e	nd	sd	h (bi)	h (d)
1	37	161	0,1	606	#presal (78) #venezuela (66) #argentina (40)	#petrobras (55) #brasil (54) #venezuela (48)
2	8	14	0,2	53	#anp (4)	#anp (32)
3	3	3	0,5	14	#leilaofossilnao (0) #marsempetroleo (0) #salveabrolhos (0)	#leilaofossilnao (6) #marsempetroleo (6)
4	41	237	0,1	826	#oil (98) #noticias (66) #colombia (47)	#petroleo (72) #oil (45) #gas (34)
5	5	7	0,3	72	#cuba (1)	#eeuu (24) #cuba (19)
6	6	7	0,2	72	#politica (1) #300diasdegoverno (0)	#economia (33) #politica (15)
7	2	1	0.5	9	#lulapresobrasilvendido (0) #opresalenosso (0)	#opresalenosso (6) #lulapresobrasilvendido (7)

Legend: C – community; #v – number of vertices; #e – number of edges; nd – network density; sd – summed degree; h (bi) – top hashtags per betweenness index; bi – betweenness index; h (d) – top hashtags per degree; d – degree.

[3] https://www.energy-pedia.com/news/brazil/brazil-approves-13-companies-to-participate-in-the-6th-production-sharing-round-177754

Communities "C1" and "C4" have concentrated most of the connections, representing key themes in the collection of tweets. The calculus of the betweenness index allowed us to identify central hashtags that are mentioned in the context of a large number of other themes in each community. Fig. 10 displays the communities detected from the hashtags' network.

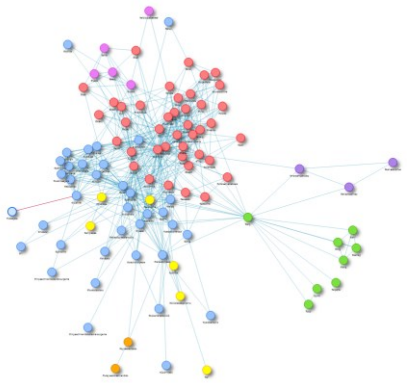

Fig. 10. Communities C1 (blue), C2 (green), C3 (purple), C4 (red), C5 (pink), C6 (yellow), C7 (orange) – Hashtags' Network

Based on the betweenness index and degrees obtained, publications were concentrated mainly on #petroleo, #oil, #presal, #brasil, and #petrobras hashtags. A characteristic observed in the tweets is the outstanding presence of comments correlating the hashtags of different countries, such as #brasil, #venezuela, #argentina, #china, #rusia, #iran, #usa, #siria, #bolivia, #colombia, #equador, #mexico, and #chile. This behavior demonstrates the diversity of debates of political and economic interest, confirming the importance of the auction oil event in a global context. We also observed the presence of the opposition with the role of criticizing the auction and the current Brazilian government. The opposition hashtags were driven by the hashtags #leilaodopresaleroubo, #impeachmentdobolsonarourgente, #lulalivre, #lulapresobrasilvendido, and #opresalenosso. Moreover, we noticed the presence of environmental activists raising a voice against the auction with the hashtags #leilaofossilnao, #salveabrolhos, and #marsempetroleo. This research also revealed an unexpected presence of hashtags related to the cryptocurrencies market (#blockchain and #bitcoin) linked to the energy sector.

VI. CONCLUSION

The last decade has been characterized by an increase in online activity among the key players from the global energy market [1]. Despite the billions of dollars in turnover, the global oil market still exhibits shy and conservative behavior in social network platforms when compared to other sectors of the economy [36] - a likely symptom of a highly regulated industry. Not all foreign companies qualified for the pre-salt oil auction have manifested themselves in social network platforms, notably the Chinese state-owned oil companies. This singularity suggests the presence of a natural political and strategic link between the Chinese state apparatus and Chinese NOCs [37], in an environment marked by overseas expansion and international alliances with major Western oil companies.

We observed that the online demonstration of social activity is still considered a taboo in the energy sector, giving limited space for daring and informality, in contrast to other economic sectors. However, this characteristic did not hinder the realization of this research. It was possible to verify that

the activities in Twitter somehow echoed the desires that guided the week of the pre-salt auction of the "Transfer of Rights" area, reflecting the optimism and engagement of the main protagonists from the oil and gas industry.

In this work, we identified the main groups of influence and the most relevant hashtags, revealing distinct trends and points of view, answering the research questions. As expected, it was possible to identify the presence of environmental activists and groups opposed to the current federal Brazilian government. The oil and gas sector proved to be an excellent thermometer for testing the proposed computational methods and for demarcating communities with well-defined and easily identifiable protagonists. The Louvain technique, the calculation of the Herfindahl-Hirschman index and the betweenness index applied to the hashtags' communities made it possible to reach the results for the data sample. To improve the qualitative results, we can add the external influence a community has on the rest of the system as well as combine new community detection techniques. As a future work perspective, we intend to add the Leiden algorithm [38] to improve the robustness of the analysis and to deepen the understanding of the proposed phenomena.

REFERENCES

[1] Abitbol, A., Meeks, J., Cummins, R.G., "Does Oil and Goodwill Mix?: Examining the Oil and Gas Industry's Impact on Stakeholder Engagement on Facebook", Environmental Communication, 13:2, 192-208 (2019). DOI: 10.1080/17524032.2018.1546751.

[2] Du, S., Vieira, E.T., "Striving for Legitimacy Through Corporate Social Responsibility: Insights from Oil Companies", J Bus Ethics 110, 413–427 (2012). DOI: 10.1007/s10551-012-1490-4.

[3] Woolfson, C., Beck, M. (2005) "Corporate social responsibility failures in the oil industry". New York: Baywood Publishing.

[4] Korschun, D., Du, S., "How virtual corporate social responsibility dialogs generate value: A framework and propositions". Journal of Business Research. 66. 1494–1504 (2013). DOI: 10.1016/j.jbusres.2012.09.011.

[5] Kleinberg, J. , Tardos, E., "Algorithm Design". Addison Wesley, 2005.

[6] Freeman, Linton (1977). "A set of measures of centrality based upon betweenness". Sociometry 40: 35–41.

[7] Fortunato, S., Hric, D., "Community detection in networks: A user guide". Physics Reports. 659 (2016). DOI: 10.1016/j.physrep.2016.09.002.

[8] Köhler, V., Fampa, M., Araújo, O., "Mixed-Integer Linear Programming Formulations for the Software Clustering Problem". Comput Optim Appl 55, 113–135 (2013). DOI: 10.1007/s10589-012-9512-9

[9] Sluban, B., Smailović, J., Battiston, S. et al., "Sentiment leaning of influential communities in social networks", Compu Social Networls 2, 9 (2015). DOI: 10.1186/s40649-015-0016-5.

[10] Werden, G.J., "Using the Herfindahl–Hirschman index". In: Phlips, L. (ed.) Applied Industrial Economics, pp. 368–374. Cambridge University Press, Cambridge, UK (1998).

[11] Hirschman, A.O., "National power and structure of foreign trade". Berkeley, CA: University of California Press, 1945.

[12] Herfindahl, O. C., "Concentration in the steel industry", Ph.D. dissertation, Columbia University, 1950.

[13] 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). DOI:10.1088/1742-5468/2008/10/P10008.

[14] Girvan, M., Newman, M.E.J., "Community structure in social and biological networks", Proc. Natl. Acad. Sci. USA **99**, 7821–7826 (2002).

[15] Newman, M.E.J, "Modularity and community structure in networks". Proc Natl. Acad. Sci. U. S. A. 103(23), 8577–8582 (2006).

[16] Kobourov, S. G. "Force-Directed Drawing Algorithms". 2004.

[17] Hooke, R., "Lectures de Potentia Restitutiva, or of Spring Explaining the Power of Springing Bodies", London, 1678.

[18] Jackson, J. D. Classical Electrodynamics (3rd ed.). New York: Wiley. ISBN 978-0-471-30932-1. OCLC 318176085, 1999.

[19] Bedi, P., Sharma, C., "Community detection in social networks", Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 6. n/a-n/a (2012). DOI: 10.1002/widm.1178.

[20] Brandes, U., Delling, D., Gaertler, M., Gorke, R., Hoefer, M., Nikoloski, Z., Wagner, D. "On modularity clustering". IEEE transactions on knowledge and data engineering, 20(2), 172-188 (2007).

[21] Fortunato, "Community detection in graphs". arXiv:0906.0612 (2009)

[22] Clauset, A., Moore, C., Newman, M. E. J., "Hierarchical structure and the prediction of missing links in networks", Nature, 453 (2008), pp. 98–101.

[23] Agarwal, G., Kempe, D., "Modularity-maximizing network communities using mathematical programming", The European Physical Journal B, 66 (2008), pp. 409–41.

[24] Danon, L.,Diaz-Guilera, A., Duch, J., Arenas, A., "Comparing community structure identification", Journal of Statistical Mechanics: Theory and Experiment, P09008 (2005).

[25] Schaeffer, S. E., "Graph clustering", Computer Science Review, 1 (2007), pp. 27–6.

[26] Granovetter, M., "The strength of weak ties", The American Journal of Sociology, 78 (1973), pp. 1360–1380.

[27] Java, A., Song, X., Finin, T., Tseng, B., "Why we twitter: understanding microblogging usage and communities", In: Proceedings of the Joint 9th WebKDD and 1st SNA-KDD Workshop on Web Mining and Social Network Analysis, pp. 56–65, August 2007.

[28] Shen, K., Song, L., Yang, X., Zhang, W., "A Hierarchical Diffusion Algorithm for Community Detection in Social Networks", In: 2010 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, Huangshan, 2010, pp. 276-283. DOI: 0.1109/CyberC.2010.57

[29] Que, X., Checconi, F., Petrini, F., Wang, T., Yu, W., "Lightning-fast community detection in social media: A scalable implementation of the louvain algorithm", Department of Computer Science and Software Engineering Auburn University Tech. Rep. AU-CSSE-PASL/13-TR01, 2013.

[30] Kim, Y.H., Seo, S., Ha, Y.H., Lim, S., Yoon, Y., "Two applications of clustering techniques to twitter: community detection and issue extraction". Discret Dyn Nat Soc 2013:1–9 (2013).

[31] Conover, M., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., Flammini, A., "Political polarization on twitter". In: Proc. Fifth Intl. Conf. on Weblogs and Social Media (ICWSM). AAAI, Palo Alto, California (2011).

[32] Raghavan, U. N., Albert, R., Kumara, S., "Near linear time algorithm to detect community structures in large-scale networks", Physical Review E, vol. 76, no. 3, p. 036106, 2007.

[33] Barnes, J., Hut, P.A., "hierarchical *O(N log N)* force-calculation algorithm". *Nature* **324,** 446–449 (1986). DOI: https://doi.org/10.1038/324446a0

[34] Bostock, M., Ogievetsky, V., Heer. J., "D^3 DataDriven Documents". IEEE Transactions on Visualization and Computer Graphics, vol. 17, no. 12, pp. 2301-2309 (2011).

[35] de Graaff, N. A., "Global networks and the two faces of Chinese national oil companies". *Perspectives on Global Development and Technology*, *13*(5-6), 539-563 (2014). DOI: 10.1163/15691497-12341317

[36] 2020 Social Media Industry Benchmark Report. https://www.rivaliq.com/blog/social-media-industry-benchmarkreport/#title-all-industry.

[37] Wu, W., "Chinese Oil Enterprises in Latin America: Corporate Social Responsibility". Springer (2018).

[38] Traag, V.A., Waltman, L., van Eck, N.J., "From Louvain to Leiden: guaranteeing well-connected communities". Sci Rep 9, 5233 (2019). DOI: 10.1038/s41598-019-4169.