NOT SO #SIMPLE

THE CHALLENGE OF MONITORING PUBLIC POLICIES ON SOCIAL NETWORKS

MARCO AURELIO RUEDIGER
COORDINATOR


CDD – 351

Citation

NOT SO #SIMPLE

THE CHALLENGE OF MONITORING PUBLIC POLICIES ON SOCIAL NETWORKS
NOT SO #SIMPLE

THE CHALLENGE OF MONITORING PUBLIC POLICIES ON SOCIAL NETWORKS

2nd Edition
Rio de Janeiro
FGV DAPP
2018
Founded in 1944, Fundação Getulio Vargas was created with the aim of promoting the social and economic development of Brazil, through the education of qualified administrators in both the public and private spheres. Over time, FGV has expanded its activities to other areas of knowledge, such as social sciences, law, economics, history and, most recently, applied mathematics, which is a benchmark in quality and excellence in all of its eight schools.

Luiz Simões Lopes Building (Main Office)
Address: Praia de Botafogo, 190 – Rio de Janeiro/ RJ
CEP: 22250-900
Tel: +55 (21) 3799-5938 | www.fgv.br

First president and founder
Luiz Simões Lopes

President
Carlos Ivan Simonsen Leal

Vice-Presidentes
Francisco Oswaldo Neves Dornelles (on leave)
Marcos Cintra Cavalcanti de Albuquerque (on leave)
Sergio Franklin Quintella

Director
Marco Aurelio Ruediger

DAPP
(21) 3799-4380
www.dapp.fgv.br | dapp@fgv.br

EDITORIAL BOARD

Coordinator
Marco Aurelio Ruediger

Organizers
Lucas Calil
Tatiana Ruediger

Researchers
Amaro Grassi
Danilo Carvalho Silva
Humberto Ferreira
Pedro Lenhard
Ricardo Faria
Thamyres Dias

Graphic Project
Luis Gomes

Translator
Lucas Maciel Peixoto
# TABLE OF CONTENTS

- **Overview**................................. 6
  - Collective Subjectivities ..................... 8
  - The Linguistic Object.......................... 10
  - Theme Monitor.................................. 12
  - Interactions in the virtual domain: Graph of Retweets. 15
  - Modularity...................................... 16
  - What the data means............................ 18
  - In any language ............................... 20

- **The Political Debate in Brazil**............. 22
  - June Jorney.................................... 23
  - Electoral Competition.......................... 25
  - Protests for the impeachment................... 26

- **Ongoing research**.......................... 28

- **Final Remarks**............................. 30

- **References**.................................. 32
The internet and the social networks represent a dramatic change in the way individuals relate to each other in the beginning of the 21st century. These new paradigms of connectivity and interactivity, which affected the social relations, also opened a lot of room for the interaction between the people and the State and reduced the costs for collective action for social movements. The influence on the internet on politics is verified from issues such as public campaign funding — in the United States, for example — to the organization of large protests in countries of South America, Africa and Asia, such as the Arab Spring, between 2010 and 2012 CASTELLS, 1999, 2015; CHADWICK and HOWARD, 2010; MARGETTS et al., 2016; OATES et al., 2006). Therefore, the new information and communication technologies have the potential to affect the public policies in different moments and under different theoretical positions. They can affect the discussion on the issues that demand the action of the State and alter the classic cycle of public policies, enabling the verification of perceptions about negative and positive external factors of governmental actions in a more immediate way.

This process of reshaping the understanding of the dynamics inherent in the public sphere — in an updated interpretation of the concept established by Habermas (1990) — also represents a challenge for the public administration, which is called upon to respond to the demands of the civil society in a swifter and more engaged way. Transparency, for instance,
as a central concept for carrying out and monitoring public policies, with uninterrupted supervision by the citizens, and feedback – through mechanisms such as the official websites of the institutions and organizations in different social networks. Therefore, the accurate analysis of the public debate is essential for the effective understanding of the interests and narratives of the web.

To monitor the public debate in social networks, FGV/DAPP uses a set of research methodologies and procedures, articulating between the fields of Linguistics, Sociology, Communication, Statistics and Information Technology. This integration domains is necessary so that we can not only extract and organize the large volume of data from the internet, but also qualify and understand the social perception with scientific rigor and respect to the relevance of the social networks – specially Twitter and Facebook – for the Brazilian society.

The social network analysis team at FGV/DAPP is responsible for monitoring the discussions of the society and for carrying out research on any topics regarding public policies – the object that is the essence of the purpose of FGV/DAPP. That is, focusing on debates in the national and international levels, or even in the regional level, about, for example, public security, immigration, health, mobility, education and politics.

Considering the consolidation of a network society (CASTELLS, 2009), created by the so-called information revolution, the monitoring and analysis of social networks takes on a role of great relevance for better understanding the political behavior of Brazilians (and of the main social actors on the networks), and becomes a productive methodology for political analysis. In this document, we present some already finished studies, explaining in detail the process of data collection and research that we propose as a vector for understanding the public opinion and the social interests highlighted in the vast universe of the internet.

Dealing with large quantities of data is always a challenge, and a rigorous methodology is necessary for such an undertaking, so that we are able to analyze its possible social and political implications. Complex networks such as the ones we analyze have been studied by various fields of knowledge, specifically the identification of communities in computing and in the graph theory (ELHADI; AGAM, 2013; LEE; CUNNINGHAM, 2013; MILKOV et al., 2006). It is a mixture of visual and statistical analysis that use algorithms of the free software Gephi (BASTIAN et al., 2009).

The first challenge after establishing the words that will guide the data collection is trying to detect the different groups. We argue in this study that the groups found in the discussions in social networks can be understood as a virtual layer of the collective subjectivities as envisioned by Domingues (1995, 1999). This argument is based on the empirical finding that clustering algorithms divide a network of retweets in groups of distinct opinions (CONOVER et al., 2011). First, we will explain what collective subjectivities are, and then we will detail the research methodology.
According to Domingues, the concept of collective subjectivity would solve the classic sociological dilemma between action and structure. In his explanation, he makes sure to decentralize the individual as well as the collectivity taken as a unit, focusing on the interaction between the different individuals and subcollectivities (DOMINGUES, 1995; 1999; 2004). Therefore, any and all social systems could be understood as a collective subjectivity. It is important to remember, however, that the collective subjectivities may have variable levels of internal homogeneity, which depend on the interaction of their individuals and subcollectivities in relation to other collective subjectivities. In addition, the collectivities have different levels of intentionality (or centering, as the author puts it), which vary according to the level of identity and organization of the collectivity in question. It is important to emphasize, however, that the causal impact of a given collective subjectivity is not necessarily tied to its level of centering: even subjectivities with a low level of intentionality may exert a great causal impact on the world.

Companies, classes, groups of friends, States, social movements and even families could be understood as collective subjectivities (DOMINGUES, 1998). The concept covers four dimensions of the social relationships: the hermeneutical dimension, the material dimension, the power dimension, and the space-time dimension (DOMINGUES, 1998; 2002). Therefore, all these dimensions affect and outline the collective subjectivities, while also being established by the collective subjectivities themselves in interaction with other collectivities (DOMINGUES, 1995; 1998; 2002).

The main characteristic of these groups is their collective causality, which means that the collective subjectivities interact internally and in relation to the other groups of the real world in such a way that they exert a specific causal effect on the world - intentionally or not. If we assume that the social systems are what actually exists, we must see the causality exerted on them by a given collective subjectivity as being different from the active causality - typical of the weberian tradition, in which the individual action is what counts - or the conditioning causality - the durkheimian one, regarding the effect of society on itself (DOMINGUES, 1999). Collective causality is an attribute of the social systems, and concerns the systems of interaction.

The interesting aspect of this concept is that the collective causality can be manifest (intentional) or latent. Many times, the impacts of the interaction in a specific system are not perceived
by its members (DOMINGUES, 2004, p. 27). However, we must remember that, in some collectivities, there is consciousness related to collective causality itself, even though only in a practical way - in the sense that Giddens distinguishes between practical and discursive consciousness (GIDDENS, 1986). This means that the individuals in question would not be able to discursively express the causal impact they felt or caused. Therefore, Domingues calls involuntary centering the causality exerted by a collectivity whose authors are ignorant of their own collective impact (DOMINGUES, 2004, p.28). Centering, however, depends on the “intrinsic and specific potential of each social system” (DOMINGUES, 2004, p.29) and may not necessarily happen, for it is uncertain.

Two central aspects are responsible both for the changes in the social systems as well as their reproduction and “inertia”: social memory and social creativity. While both are tied to the individual subjectivity and are established by interactive processes, memory is composed of mnemonic elements that result from interactive processes of the collective subjectivities themselves, which crystallize over time in a variable way, depending on the collective subjectivity in question, both materially in symbols and objects as well as in the individual body, brain and subconscious themselves (DOMINGUES, 1998). Such elements, together with the unintended consequences of action, the power and the individual interactive inclinations of each collectivity, help organize the social systems, establishing rules, behaviors and interaction patterns both at the individual and collective level.

However, they are also subject to big or small changes when social creativity is exerted through these interactions. The primary source of creativity is found in the unconscious through what Freud calls primary process, which is the endless sliding of meaning, typical of the unconscious (DOMINGUES, 1998). Afterwards, this process must be articulated with the social reality of the individual by means of the secondary process, in which he tries to match the mental image formed by the id during the primary process with an object of the real world (FREUD, 1900).

Therefore, we see that, according to Domingues’ concept of collective subjectivities, “social life is woven and constitutes itself as an interactive, multidimensional network, in which individual authors and collectivities casually influence each other” (DOMINGUES, 2002 p. 68). Virtually or face to face, from the moment when the authors find themselves in a specific social situation, they will try to interpret the situation in which they are inserted, in a similar way as explained by Goffman (1974) in his definition of the situation. The way authors perceive each other forms the common ground of interaction both between individuals and between subjectivities (MEAD, 1934; DEWEY, 1927; STONE, 1962; SNOW, 2001; CICOUREL, 1993).

Because of this, it becomes necessary to carry out an analysis that takes into account the different collective subjectivities in relation to one another, attempting to understand the differences in their lifeworlds and also how their interpretations are construed in relation to each other. However, we cannot forget that such collectivities are also formed by individuals, whose constant social interactions and their symbolic interpretations can, on the one hand, maintain the status quo, perpetuating standardized behaviors and symbols, and, on the other hand, transform the symbols through social creativity, giving new meanings to the actions and to reality. The same symbol may emerge from a collectivity and be coopted by another, which employs it in a different meaning from its first use due to the hermeneutic context specific to each collectivity (SWIDLER, 2001; SCHUDSON, 1989; WALZER, 1985; SAHLINS, 1981; HALL, 1973).
Social network research on topics such as security, public health, education and corruption demands a certain degree of complexity in the identification of the object of study and in the reflection on how this object is collected from public debates online. It is different, for example, to filter mentions for a name, a hashtag or a brand, when the words chosen for the query obey a strict . However, how can we repeat the same minimalist principle for something as broad as public security? The word security? And, more importantly: how to make sure that the database analyzed actually corresponds to an acceptable universe − regarding the accuracy and volume − of the debate about the theme on the internet?

Focusing on this, FGV/DAPP pioneered the development of a specific semantic categorization process, which follows the scientific reasoning of discursive semiotics (GREIMAS and COURTÊS, 2011; HJELMSLEV, 1978; GREIMAS, 1986; FONTANILLE, 2008), a field of linguistic studies that is highly important, especially in Europe, and vastly adopted by the main universities in Brazil. Discourse semiotics − whose main historical proponent is Algirdas J. Greimas − understands text as a corpus of analysis that has internal coherence (BARROS, 2005), on which there is an emergence of meaning apprehended by the semantic articulations inherent to the text itself.

In other words: any discursive object, like a painting, a movie scene, a chapter from a novel or a tweet, once selected for study, is considered a text (FIORIN, 2004), and, for the analysis of internet data, the semiotic object becomes a vast corpus, examined in large scale as a coherent, unified set. The referential for coherent identity is the presence, in each post that integrates this corpus, of discursive figures belonging to the semantic field that one wishes to study, which is determined a priori. Once established a hypothesis to search mentions from specific semantic fields, the objects are recovered by the gathering of posts via web.

Take for example public security, a subject that has been extensively studied by FGV/DAPP, and not only in social network researches. Before formulating a hypothetical search based on the set of keywords, it is necessary to understand what comprises public security in Brazil in the social and institutional spheres (or with a selection by predefined states or regions,
considering regional differences). At this stage, the research team carries out a data collection of any concepts or discursive figures that are associated with the abstract notion of security: the police; the police stations; the crimes, such as burglary, robbery or theft; legislative issues, such as drug legalization or the criminal code reform; the authorities; the acronyms; and the public institutions. In order to do so, we carry out an interdisciplinary reflection among the specialists at FGV/DAPP, who delimit the textual constructions that relate to the theme of public security (or any other theme).

This method seeks to iconize (GREIMAS and COURTÉS, 2011, p. 250-251; BARROS, 2005, p. 69) the themes under qualitative bases – as to the accuracy of the posts collected and the relevance of the discursive figures – and quantitative, as to the volume of data obtained in proximity to the estimated complete volume of publications related to the theme in the timeframe (and data source) of the research. Therefore, once the delimitation of the discursive figures that relate to the theme of security is finished, it is necessary to run tests to verify the volume categorized and the assertiveness of the search rule, a set of expressions, sentences and terms organized by logic operators of correlation between semantic fields. At this stage it is possible to remove incorrect expressions, identify ironies and metaphorical expressions and evaluate what must and must not integrate the definitive search rule, with a previous consultation to the universe of data sources used in the research (mainly Twitter). At the end, when a complex rule has been constituted for the research on social networks, the set of texts that fit any attribute of this rule (a reference to the police, or to a crime, or even to a state organization) is seen as having an internal semantic recurrence – which, in semiotics, is known as isotopy (GREIMAS and COURTÉS, 2011, p. 275-278) – in relation to the general theme. This way, it is possible to fragment the search in multiple subthemes, restricting the object to certain categories inside a broader Theme, such as mentions to the police or different crimes.

In the construction of search rules, in order to correctly delimit the corpus, FGV/DAPP, in line with semiotic principles, takes into account the syntax rules inherent to the linguistic system of the language studied, but always paying attention, to take into account, regional variations, grammatical errors and the use of idiomatic expressions and slang, anticipating potential contractions, common in the internet environment, and any colloquial constructions that might integrate the debate. Texts that do not follow the standard variety, from the perspective of Linguistics (DISCINI, 2013), are not distinguished or ignored.
The theme monitor is a tool updated in real time by FGV/DAPP, which shows the volume of mentions of six broad public policy themes to the Brazilian context: security, health, education, mobility, protests and corruption. For three years, the Monitor has been identifying trends, fluctuations and subjects that monopolize, for some time, the discussions on Twitter. Every time the National High School Exam (Exame Nacional do Ensino Médio - Enem) or governmental programs for higher education appear for discussion, the volume of posts on education increases; the same happens with health — in the beginning of the year, for example, there was a higher volume of debate on this subject due to the zika virus; and in general, Brazilians talk mostly about public security, although since the beginning of 2015 there were weeks when the tool collected up to one million posts on corruption.
Since the universe of social network monitoring at FGV/DAPP is not conditioned to any restricted theme or object, we developed a methodology of integration between academic competences, consolidating a research model applicable to any theme, be it broad or narrow, in any language (including searches for brands, acronyms, hashtags or names). And this is possible because the methodology has a linguistic consideration regarding the object of study and the definition of the set of texts that are categorized on the networks.

The graphic above with some research subthemes of public security exemplifies the possibility of broadening or narrowing any scope of analysis, whether the subject is more abstract or more concrete. The debates regarding the prison system, the demilitarization of the police, the Penal Code reform and the reduction of the age of majority compose the broad search on security, and therefore are equivalent when thinking about the general theme, but different when thinking about specific subthemes. That is, a mention to any of the themes is also a mention to public security, but a mention to the overcrowding of prisons in the country is not necessarily a mention to the reduction of the age of majority. This happens because, when contrasting between more abstract or concrete themes, the
concrete elements usually share a semantic field that is more abstract, but integrate other lower scale, different semantic fields. And even a concrete subtheme, such as the prison system, is composed of many discursive figures that can be subject to isolated analyses, such as the situation of the prisoners or employees of the penitentiary system. Important note: what defines the subdivisions of the search is precisely the choice of the theme, the section of the study object and the diversity of semantic figures and fields that, in the field of mentions on social networks, overlay the debate on the subject. In other words: the broader the search, the higher the possibility of subcategorization.

THE TIME PERIOD OF TEXTUAL ANALYSIS

When formulating any search rule, it is important to know that the meaning of words, according to the saussurian notion of value (SAUSSURE, 2006), is narrowed or broadened by the historical situation where the discourse is inserted — the time period. That is, a search rule that is used for a broad, real time search over a number of months must cover specific changes as new figures and expressions become part of the debate. A new bill, for example, impacts the research on a theme, as well as the dismissal or appointment of a minister or an incident highlighted by the press.

A well-known case we can remember happened in March, 2015, when the Federal Attorney General, Rodrigo Janot, published the list of politicians associated to the Operation Car Wash. When searching specifically about the repercussion of this episode — that is, with Janot's report as the sectioned search object — we had to isolate the word “list”, because, according to our verification, it referred only to the theme of corruption during that short period of debate on Twitter. Although the word “list” may refer to dozens of different concepts, it was consolidated in the internet vocabulary, during those 48 hours, as a synonym of the report by the attorney general; therefore, as a discursive figure of Operation Car Wash and of corruption.
A retweet is a way of sharing a message or image, and has great importance in events where information spreads quickly. When we look at the retweet network, we can understand the relationship that was created between different user profiles, which is nothing else than a trace of virtual interaction. In the case of the retweet network, each time user A shares user B’s tweet, the connection between A and B can be understood as a propagation of information that goes from individual A to B (CONOVER et al., 2011; BOYD; GOLDER; LOTAN, 2008). Therefore, by sharing a post from another user, on one hand I endorse what was said there, and on the other hand I repeat and propagate the idea contained in the message to my entire network (MACSKASSY, 2011; BOYD; GOLDER; LOTAN, 2008).

In this process, the individual who shares something is spreading that discursive construction and symbolic articulation, which increases the chances that such symbolic constructions will be internalized by another user. Still, it is important to keep in mind that people, discourses and collectivities are in constant motion, because interactions happen dynamically and not in a fixed way, and must be understood as a process. Because of this, every time we look at a graph, it is essential to keep in mind that it is a circumscribed picture of reality based not only on the search words define for the analysis, but also on the space (in this case, the virtual environment) and time defined by the object scope.

The retweet graph is created in the following way: each node represents a twitter profile, and each time user A shares a post from user B, a connection (or edge) between nodes A and B is created. When all the edges and nodes are computed, we can generate a visualization of the network in the free software Gephi which helps us understand the relationship between the different individuals and collective subjectivities.
After entering the data on Gephi, we use the program’s modularity statistics. Modularity enables mapping the different groups in a network (NEWMAN, 2004; 2006). A set/community/cluster of nodes will be classified this way if the connection between the nodes is stronger than its connection to other nodes in the network (VINCENZO, 2008). The algorithm used by the program is based on the Louvain method (BLONDEL et al., 2008). It finds parts of the network that have high modularity in two main steps. The first one optimizes the local modularity, sweeping the entire network with random vertices, and the second one aggregates the community repeating passes interactively until the modularity reaches its highest point. In other words, the network is divided into sets that progressively test all the possible existing combinations of nodes until none of them can belong to any other group (NEWMAN, 2006).

At the end of the process, the algorithm separates the network into sets (GIrVAN; NEWMAN, 2004) called clusters, which will be analyzed as representations of the online interacting collective subjectivities. Each of the sets has more connections between its own nodes than between any of its nodes and nodes of other clusters, which coincides with Newman’s intuition (2004; 2006) that individuals will establish more relationships - represented here by edges - with people inside their own group than with people who are outside of their group.

Each of the subcollectivities of the online collective subjectivity will have its own color so that we can tell them apart in the graph. Each node will belong to a single subcollectivity. After that, we calculate the weighted degree so that we can visually distinguish the nodes that are authorities by the size of their name/label. The degree is a measure that adds up the weights of the edges (adding one for each retweet) and tells us which profiles had the most retweets, that is, who originated the most information that is being spread. We will call these profiles authorities. This measure is relevant in the study of internet phenomena because there is a trend of post massification: first, someone publishes an original message, which can be later republished in an attempt to disseminate the subject or point of view. This way, the
degree, analyzed together with the activity of the profile, indicates the type of participation or engagement that made it relevant.

Now, finally, we can use the ForceAtlas2 algorithm. This algorithm attracts the nodes whose edges are thicker - that is, the points that connect the most -, and repels the nodes that connect the least (GIRVAN; NEWMAN 2004), providing visual information on how the various profiles and groups are related in the network (NOACK, 2009; JACOMY et al., 2014). According to Conover et al. (2011), clustering algorithms such as ForceAtlas2 divide the network of retweets among groups of profiles that express different political views.

The representation we obtain after creating the graphic is informational in a few aspects: the connections between nodes that show a relationships between the profiles, the position of the nodes in the graph, the size of the label of the nodes and their color. We will also look at the authorities of the graph, that is, the nodes that have a higher degree of input. Afterwards, we will combine this result with a qualitative analysis of the main tweets per cluster. This way, we will have a good notion of how each group uses different signifier units, signified units and arguments to publicize specific themes, which will help us identify what unites the profiles inside it - whether there is a predominant ideology in the group, for example. In addition, we will be able to verify whether an interpretation of a subject by a given group is disputed by another group, which will help us understand whether the collective subjectivities in question have interactive inclinations. Interactive inclinations are related to the dimension of the collective subjectivities, which dictate their relationships with others. This, of course, will vary depending on their level of centering, which in turn varies according to the internal level of identification.

What leads me to state that a cluster can be seen as a collective subjectivity is the high level of common sharing that the individuals inside the same cluster have (NEWMAN, 2004; 2006), which indicates a high level of interaction among its members. This means that, if I am inside cluster 'x', I will probably share posts or news that other people in my cluster will also share. This behavior indicates that we somehow think in a similar way, that there is a certain type of logic we share and that makes it so that the same content 'x' seems interesting and relevant to the point where we feel impelled to share it, and, in doing so, publicize it (BOYD; GOLDER; LOTAN, 2008). To a certain extent, this also means that there is a shared logic inside the groups, because their individuals share the same understanding of certain symbols and values.

Since people interact in a more or less organic form in the network, the creation of interest groups happens due to an involuntary centering, as proposed by Domingues (2004). Therefore, considering the view of Dowbor and Szwako (2013), where social movements are seen as performances, we can interpret a retweet as the moment when someone from the public starts to chant the repertoire along with the actors. New repertoires may emerge in this transition from passiveness into activeness, which will be assimilated as these repertoires find actors who will join the chant.
The building and categorization methodology used by FGV/DAPP is important because it prevents the arbitrary or limited collection of data about a theme, based on a small set of words and hashtags, without considering the conceptual perspective of text, and, above all, the semantic possibilities regarding a certain subject. Therefore, linguistic reflection about the study object in social networks is essential. A word is more than simply a word; it is, above all, a discursive figure, concrete or abstract, embedded of a meaning enclosed by the other linguistic constructions it connects to. Because of this, the search rules used by FGV/DAPP take into account associations and restrictions between terms and expressions.

A simple search about public security would return a vague result: how can we study the debate about the theme starting only from the words “security” and “police”, for example?

With the database collected from this complex process of categorization, FGV/DAPP produces different analyses and visualizations, such as graphs, word clouds, graphics and historical timelines – each visualization with its own specific functions, depending on the objectives of the research.

**Graphs**

Map of a network of profiles that interact about a given subject or profile. The graphs enable the identification of the groups and subgroups that communicate the most with each other (the clusters), the most relevant influencers and the people who operate as “bridges” between these two distinct groups. In studies about political issues, these bridges are usually the press, because it produces the news that reverberate collectively.
The graph above is about the repercussion on Twitter about the voting for the process of impeachment against then-President Dilma Rousseff, at the Chamber of Deputies. In the graph — presented here without the identification of the actors who interact in the debate — it is possible to see, in blue and on the left side, the profiles whose discourse is favorable to opening the impeachment against Dilma, and in red and on the right side, the profiles against the process. As previously explained, the bigger the node representing the actor in the discussion, the higher the relevance they exert on the debate, since the diameter of the circle is proportional to the amount of retweets that this actor obtained.

In addition, similarly to what we explained before, the more central a profile is in the positioning of the graph, the higher the reach of what they publish and the broader the network of profiles with which they interact; in this case, the orange cluster is composed of profiles of the press, situated closer to the blue cluster, and of actors from political groups opposed to the impeachment, but who are adversaries of the Dilma administration in the political sphere.

It is also important to highlight that we used the linguistic process of categorization by search rules for the Twitter data collection that composes the graph, elaborating an extensive search bracket with hundreds of words and lexical constructions to accurately capture the large volume of debate in April 17, 2016, when the Chamber decided to continue Dilma’s impeachment process. This process is the same for any visualization, analysis or section of study, because the object, as previously mentioned, is the same: the text.
Likewise, FGV/DAPP carries out studies of social networks in different languages, with research dedicated to international sociopolitical contexts — we have even carried out several analyses for the British newspaper Financial Times, among which are our interpretations of the protests in Argentina and Venezuela, in Spanish, and about the general elections in the United Kingdom in 2015, with the reelection of David Cameron as prime minister. The logical operators used to build the linguistic construction of the search cover any western language, evidently respecting the differences in syntax among languages, but with an identical study process to the one used in searches in Portuguese, with the collection of words and expressions, verification of accuracy and specific refinement.

Consider the case study we carried out over Brexit, in June, 2016, regarding the repercussion, in France, Germany and the United Kingdom, of the plebiscite that culminated with the decision to eventually have Great Britain leave the European Union. In that research, we selected a two-week period for data collection to search for particular expressions and hashtags and for the thematic debate in each country. We verified that the migratory issue was greatly highlighted, specially in Germany and France, where people were debating the potential arrival of more refugees, while the economic consequences obtained a larger volume of debate among the British, specially among profiles against the United Kingdom leaving the European Union.

By a textual perspective of the political context, the main words and expressions about Brexit were usually associated with the prominent figures related to the subject — such as, in the case of the first wordcloud demonstrated below, Jo Cox, a member of parliament who was against Brexit and was murdered a few days before the plebiscite, and Nigel Farage, leader of the nationalist party Ukip and one of the main advocates of United Kingdom's exit (and more mentioned, on Twitter, by opponents than by allies). The second wordcloud highlights the name of German chancellor Angela Merkel.

To elaborate the clouds presented below, we initially ordered the profiles in a graph, identifying the clusters associated with each party or position regarding Brexit. Afterwards, the textual content of the posts from each cluster was isolated and applied to a generator that works by word count, which produces the cloud. However, as previously presented, punctuation signs, prepositions and pronouns always appear with a large volume of repetitions, and are usually excluded from the final visualization.
Word Clouds

They identify the words and expressions with the highest amount of repetitions in the textual collection for a given subject. FGV/DAPP, however, uses a cleaning method on the clouds before displaying them, favoring nouns, adjectives, adverbs and any constructions of specific value for better understanding the subject, removing repetitions, articles, numbers, prepositions and pronouns (with a few exceptions, if the research theme demands). With the clouds, it is possible to clearly visualize the main discursive figures that integrate the debate.
The combination between the methodology for network analysis developed by FGV/DAPP and the various forms of data visualization presented here also enables a detailed monitoring of the political behavior of Brazilians on the internet — and of the interconnections between this debate and the social perception about public policies, essential for decision making by the public administration. The Theme Monitor, for example, carries out a daily collection of themes since 2014, such as health, education and security, observing their articulation with a complex and vast political debate.

Using graph analysis, it is possible to monitor the evolution of the political debate on the networks for an extensive period of time, and identify changes in the forms of interaction and in the main actors connected to this discussion; as the political agenda of the country changes, the dynamics of discussion also tend to change. Simply as an example of the analytical possibilities of this methodology — but far from exhausting its applications —, we present below four graphs produced between 2013 and 2016 from Twitter data.
The first visualization is the result of monitoring the network during the entire month of June, 2013, focusing on the discussion about the demonstrations who took the streets of Brazil. The movement started due to the increase in the public transport fare and quickly gained new objectives, such as demands for more investment on areas such as health and education. Although, in the heat of the moment, it was difficult to identify the different political groups involved in the protests, the method described here was efficient to carry out this task.
We can see that there were several different subcollectivities during June, 2013, each of them represented by a color. The subcollectivity in yellow encompasses several segments of the traditional press and is central in the graph because it is used by all groups. Is is responsible for approximately 13% of the mentions.

On the left side, in pink, we find profiles that are more aligned to the left wing, including the profile of deputy Jean Wyllys. The pink subcollectivity is composed by the demonstrators who called out the first protests, and was responsible for approximately 5% of the posts. In red, also on the left side of the graph, we find the subcollectivity identified as a group supporting the Labor Party (PT) and Dilma, whose most popular hashtags were #ApoioDilma (support Dilma), #TamoJuntoDilma (we are together, Dilma) and #Brasil13 (Dilma and PT’s campaign number).

On the bottom of the graph, we can identify the blue cluster, with right wing ideological orientation. Among its authorities are BlogdoNoblat, PMESP and PastorMalafaia. Although the most popular hashtags referred to the protests in general - such as #vemprarua (come to the street) and #mudabrasil (change Brazil) -, another one that appeared in the collectivity was #ForaDilma (out Dilma), revealing the strong anti-PT attitude of the group. On the top of the graph, we can see the green cluster, the biggest subcollectivity of the period, encompassing almost 29% of the mentions. In this group, many of the most popular tweets were posted by virtual celebrities and youtubers such as Felipe Neto, PeceSiqueira, Rafinha Bastos and even the fictional Irmã Zuleide. The subcollectivity seems to fluctuate between the spheres of the right and left winds, depending on the theme being discussed. PeceSiqueira, for example, has declared himself to be left wing, while FelipeNeto seems to have a more libertarian bias, in addition to being anti-PT. However, neither of them have any type of partisan fidelity, which could explain how they ended up in the same group.

Perhaps they appeal to the masses who have recently joined the political sphere, frustrated with the institutional political system we have today. This would also explain the strong rejection to politics and the media. However, what seems to be a consensus in the group is the national pride, which is also reflected in the most popular hashtags, such as mudabrasil, changebrazil, ogiganteacordou (the giant has awaken) and verásqueumfilhoteunãofogeàluta (a part of the national anthem). In addition, other consensual points in the group are the criticism towards police misconduct and the traditional media, seen as biased and strongly rebuked for trying to delegitimize the protests.
ELECTORAL COMPETITION

The second example is related to the repercussion of the second round of the 2014 elections on the network, using data collected over 24 hours on October 24, the eve of the election day. In the graph, there is a distinct polarization between two opposing political groups, reflecting the electoral dynamic. These groups are organized in two clusters: the first one (in blue, on the left), formed by profiles who supported the candidacy of senator Aécio Neves (PSDB), and the second one (in red, on the right), formed by profiles supporting the reelection of then president Dilma Rousseff (PT) — both positioned around profiles who exert the biggest influence inside each cluster. At the center of the graph, and therefore with higher reach in different clusters, there are profiles connected to press organizations, with an important articulation role.
The third graph, in turn, represents the public debate on Twitter about the protests for the impeachment held on March 15, 2015, in several states. In this case, the cluster of the press is dissipated, integrating itself with the one favorable to maintaining the Dilma administration — in red — or the one supporting her removal — in blue. This last one occupies a larger and more central space in the visualization, which denotes a more hegemonic position in the debate, demonstrating a significant adherence to the ideas of the movement by the main actors who participated in the discussion.
The last graph, published below without the labels identifying the authors represents the mentions on Twitter to the protests of March, 2016, which gathered more than 3 million people in all of Brazil around a mostly pro-impeachment agenda. We can see the emergence of a big cluster, in yellow, taking the spotlight in the graph. In this group are profiles favorable to the impeachment of then president Dilma and to the investigations of Operation Car Wash.

The polarization verified in previous moments gives room — considering the crisis of trust of the Brazilian people regarding politicians and politics, related to the growing number of corruption reports —, to a strong sense among the population of not identifying with either of the opposing groups, similarly to what happened in 2013, when the popular movements were detached, in the field of public debate of the internet, from the traditional political articulations.

The monitoring of the political debate on the internet is established as one of the most important lines of action of the social network analysis team at FGV/DAPP. As previously pointed out, the use of this methodology allows us to monitor the variations of the behavior models on the most relevant social networks for the Brazilian society — specially Facebook and Twitter —, and relate them to the changes in the political agenda. The analysis of these historical series also helps to anticipate political behaviors based on patterns of interaction. This application was successfully used by FGV/DAPP to estimate ahead of time the dimensions of the protests of March 13, 2016, pictured above, through the observation of their repercussion on the networks in previous dates.
The methodology briefly presented here is the result of a research project that began in 2013 and is going through constant updates and improvements, in line with the innovations in the processing of large volumes of data and in the software used for the analyses. The proposal of FGV/DAPP is precisely to solve problems in the analysis of social networks regarding qualitative and quantitative results, such as restricting the search terms, the inconsistency of the sectioned object and the lack of a linguistic reflection — and Linguistics is still not present enough, in Brazil and internationally, as an integral part of the process for analyzing social networks.
One of the current stages of development of research on social networks by FGV/DAPP is building a better structure for sentiment analysis, without the — already widespread — use of uniform word bases which attribute positive or negative values. Based on the reasoning that guides the search rules, FGV/DAPP is elaborating search keys on social networks, attributing a specific opinion value for each rule, theme and time period — taking into account what is actually positive or negative for each theme or subtheme. Because of this, a search about public security evaluates as positive a highlight about the reduction of crimes in a certain place and as negative an increase in the feeling of unsafety or posts that emphasize the fear of day-to-day criminal activity.

Fluctuations on the social perception about any subject do not obey a fixed system of words that are predefined as positive or negative — it is easy to think about this principle by observing the difference between attributing an esthetic value to a work of art or person (attractive or unattractive, for example) and attributing an opinion about the economy of a country, where the value is conditioned to a variation of quantity (whether the GDP has increased or decreased) or of pace (whether productivity has accelerated or decelerated).

The evaluation of sentiment in social networks already integrates part of the analyses carried out by FGV/DAPP, and today it consists the main focus of research by the team, with the expansion of the database with positive and negative categorizations to cover the hundreds of themes and subthemes already regularly analyzed. Currently, FGV/DAPP has a semantic dictionary with more than 3,500 search rules, in seven languages, which are used and updated on demand. All of them are subject to the same complex process of categorization, evaluation, cleaning and verification that makes our studies on social networks coherent, methodologically reliable and correct regarding the identification of volume, content and reach of the debate about public policies on the internet.
The use of qualitative and quantitative methodologies, together with the linguistic understanding of the textual volume sectioned from social networks, ensures the accuracy of the study corpus and the reliability of the search results, with the dialogue between Sociology, Information and Communication Technologies, Linguistics and Social Communication.

The methodology employed by FGV/DAPP considers the networks produced by the traces of interaction in the online environment, such as shared posts, mentions, likes and connections found on social networks as texts under the foundation of semiotics, considering the strategies of textual production, the figures of discourse and the identification of the effects of meaning that are present, recurring and coherent on the database selected as a corpus.
APPLIED LINGUISTICS

The methodology employed by FGV/DAPP considers the publications on social networks as texts under the foundation of semiotics, considering the strategies of textual production, the figures of discourse and the identification of the effects of meaning that are present, recurring and coherent on the database selected as a corpus.

VISUALIZATION OF DATA

The displaying of graphics, graphs and word clouds illustrates, with accuracy and simplicity, the public debate on the social networks, allowing for a wide reach of search results, and consequently the understanding of important subjects in Brazil and in other countries by the society. The integration between a strict research methodology and design tools is an essential part of FGV/DAPP, which is engaged on making complex and important questions accessible to the society.

DIFFICULTY IS POSITIVE

The use of complex structures and rules with hundreds of expressions and syntactic articulations enables the study of abstract themes in debates on social networks and the subdivision in specific categories, expanding the social perception about public policies.

FREQUENT UPDATES

All the categories and themes researched by FGV/DAPP go through a constant process of updating, modification and verification regarding the accuracy of the identified texts. Therefore, changes in lexicon and shifts related to political and administrative changes or to unforeseen events are duly incorporated to the categorization, which respects the time period of the search on social networks.
REFERENCES


ELHADI, H.; AGAM, G. Structure and Attributes Community Detection: Comparative Analysis of Composite, Ensemble and Selection Methods. Em: SNA KDD 2013 International Workshop on Social Network Mining and Analysis held in
conjunction with ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Agosto, 2013.


MILKOV, E, COHEN, W, and NG, A. Contextual Search and Name Disambiguation in Email using Graphs. In SIGIR. 2006.


