Social Network Analysis and Mining (2020) 10:6
https://doi.org/10.1007/s13278-019-0611-9

ORIGINAL ARTICLE

# Behavioral differences: insights, explanations and comparisons of French and US Twitter usage during elections

Ian Davidson[1] · Antoine Gourru[2] · Julien Velcin[2] · Yue Wu[1]

Received: 7 January 2019 / Revised: 13 September 2019 / Accepted: 29 October 2019


**Abstract**

Social networks and social media have played a key role for observing and influencing how the political landscape takes shape and dynamically shifts. It is especially true in events such as national elections as indicated by earlier studies with Facebook (Williams and Gulati, in: Proceedings of the annual meeting of the American Political Science Association, 2009) and Twitter (Larsson and Moe in New Med Soc 14(5):729–747, 2012). Not surprisingly in an attempt to better understand and simplify these networks, community discovery methods have been used, such as the Louvain method (Blondel et al. in J Stat Mechanics Theory Exp 2008(10):P10008, 2008) to understand elections (Gaumont et al. in PLoS ONE 13(9):e0201879, 2018). However, most community-based studies first simplify the complex Twitter data into a single network based on (for example) follower, retweet or friendship properties. This requires ignoring some information or combining many types of information into a graph, which can mask many insights. In this paper, we explore Twitter data as a time-stamped vertex-labeled graph. The graph structure can be given by a *structural* relation between the users such as retweet, friendship or follower relation, whilst the *behavior* of the individual is given by their posting behavior which is modeled as a time-evolving vertex labels. We explore leveraging existing community discovery methods to find communities using just the structural data and then describe these communities using behavioral data. We explore two complimentary directions: (1) creating a taxonomy of hashtags based on their community usage and (2) efficiently describing the communities expanding our recently published work. We have created two datasets, one each for the French and US elections from which we compare and contrast insights on the usage of hashtags.

## 1 Introduction and motivation

On any given day, more people use social media for news than any other media form. It is slightly more popular than TV and more popular than radio and newspapers combined (Perrin 2015). When we consider individuals aged under thirty, three times as many people use social media daily than those who read news papers daily (Perrin 2015). Social media changes the speed, reach and form of communication. Now anyone can post a message to anyone and they can post often with no cost. Social media also has an interactive aspect like no other media, which means people can comment, re-transmit and even argue in real time. There are no doubt negative aspects such as the lack of fact checking but it is also without a doubt that social media can and does influence the world like no other media source. Considering worldwide organic social justice events such as `#OccupyWallStreet`, `#ArabSpring` and `#JeSuisCharlie`, it is not surprising that they have had an immense impact on established events such as elections, as it has been

IIan Davidson was a visiting Fellow at the French Institute of Advanced Studies (Collegium de Lyon) when this research was completed. All code and data used in this paper are available at www.cs.ucdavis.edu/~davidson/description-clustering/SNAM_code

✉ Ian Davidson
davidson@cs.ucdavis.edu

Antoine Gourru
antoine.gourru@univ-lyon2.fr

Julien Velcin
julien.velcin@univ-lyon2.fr

Yue Wu
yvwu@ucdavis.edu

1 University of California - Davis, Davis, USA

2 Université de Lyon, Lyon 2, ERIC EA 3083, France, Lyon, France

Published online: 14 December 2019

**Table 1** Spending of candidates during the election primary period

| Candidate | Position | Direct contributions | PAC contributions |
|---|---|---|---|
| Donald Trump | Far Right | \$64M | \$2M |
| Ted Cruz | Far Right | \$93M | \$89M |
| Marco Rubio | Right | \$50M | \$57M |
| Jeb Bush | Right | \$34M | \$124M |
| Hillary Clinton | Left | \$174M | \$38M |
| Bernie Sanders | Far Left | \$234M | \$6M |

First number is direct contributions, second number is PAC (political action committee) spending

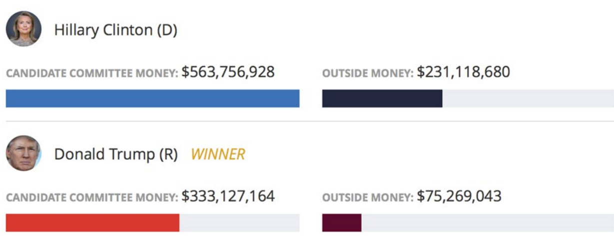



**Fig. 1** Overall contributions to the general election for the USA 2016 election

shown for the 2008 US elections (Johnson and Perlmutter 2013).

Consider, as an example, the impact of social media on the primary US election. Table 1 shows direct contributions by individuals and PACs (political action committees) much of which is spent on advertisement on radio, TV and printed matter. We immediately see an odd result. The least spenders for the Republicans (Donald Trump) and Democrats (Hilary Clinton) were the eventual winners of their respective primaries. In Donald Trump's case, he was outspent 2 as to 1 and even 3 as to 1. It is widely believed that social media allowed them to overcome the lack of paid advertisements in other medias (Enli 2017). This trend carried over even to the main election as shown in Fig. 1 where again Donald Trump was outspent but went on to win the presidency.

It is not believed that Twitter had a similar large-scale impact on the last French elections; however, previous works have clearly shown that these platforms are useful observatory of French politics (Gaumont et al. 2018; Velcin et al. 2014). This motivates us in following the same analysis as in the US election dataset, but performed on a comparable French dataset. It is grounded in the assumption explored in Poblete et al. (2011) that behavioral differences can be observed on Twitter usage in various countries over the world.

### 1.1 Twitter as an observatory of politics

Of the social media platforms, arguably Twitter has had the greatest impact on politics due to many reasons including the large volume of data capable of being generated and the popularity amongst several candidates (i.e., Donald Trump). The Twitter universe consists of accounts which can represent an individual or group each of which can produce many posts that may contain one or more hashtags.

Previous works have shown that hashtags can be useful to capture opinionated messages (Kouloumpis 2011) , which means messages that carry subjective content. Such work has succeeded in recognizing the political leaning of Twitter users (Wong et al. 2016), or, more broadly, to categorize user viewpoints related to various topics (Quraishi et al. 2018). It is believed that hashtags are used in a variety of ways including rumors, i.e., `#CruzSexScandal` (an allegations that a candidate had several extramarital affairs), general support `#ImWithHer` (reference to supporting Hilary Clinton) and general information `#GOPDebate` (the Republican debates). Posts can be propagated to other users a number of ways. All posts by a user are sent to the followers of that users `home feed`. If a person replies to that tweet, it is sent to that person's `home feed` only if she follows the replier, otherwise it is sent to her `notification feed`. If a person retweets another post, it is sent to all her followers. This represents a massive amount of information. In our 8-month period of study for the US election, there were 339,910,403 posts by 3,448,096 individuals using 2,515,421 hashtags. The French dataset is around 60 times less massive with 4,271,444 posts, 64,438 users and 56,239 hashtags.

These data can naturally be modeled as a graph as follows. Each node is of course an account with an edge between nodes indicating some sort of **structural** relationships such as `follower`, `Retweet` or `mention`. This relation indicates potential, actual and probable information propagation, respectively. For each node, we have a series of vertex labels indicating the **behavior** over time. This can be as simple as a vector of how often each hashtag is used or a matrix of usage over time. Thus, this is naturally a complex vertex-labeled graphs of the form $G(V, E, L)$ where $V$ is the set of vertices (user accounts), $E$ is the set of edges and $L$ is a vector of hashtag usage.

### 1.2 Objectives and novelty

Our focus in this paper is to better understand the community structure in this election period data, and we take two main directions:

– Creating a taxonomy of hashtag usage. Figure 7 shows the two-dimensional taxonomy we create (usage vs tra-

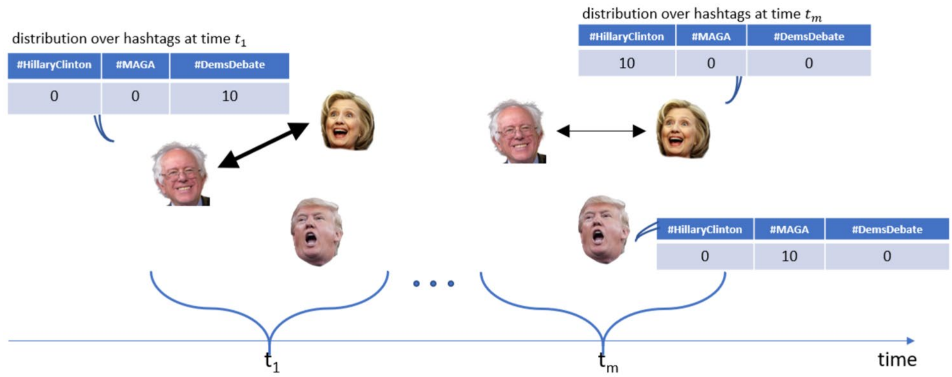



**Fig. 2** How we model Twitter data during elections. The graph structure is given by the retweet relation and the node labels by the users' behavior over time with respect to hashtag usage

jectory) and the distribution of hashtags for the US and French elections across this taxonomy. The full color-coded taxonomy is presented in "Appendix 2".

– An in-depth description of the behavior with respect to hashtags of the different communities. For example, Table 3 shows an explanation of communities in terms of hashtag behavior. Figures 16, 17 show the evolution of that behavior month by month.

**Novelty and Previous Work** Our first objective is to better understand the properties of hashtags. Many works studied hashtags diffusion. Those works solely focus on hashtag popularity, or adoption. Wang and Zheng (2014) classify hashtag based on different popularity patterns (single spikes, multi-spikes and fluctuation). Yang and Leskovec (2010) propose a diffusion model of hashtag based on user influence. Romero et al. (2011) propose to define hashtag in terms of stickiness (probability of adoption) and persistence (durability) and show that different topics imply different types of diffusion.

In this work, we focus on the usage of hashtags by communities. We do not study their popularity in frequency of use, but in terms of percent of communities using it. To do this, we created a hashtag taxonomy along two dimensions: (i) the entropy of the hashtag usage across communities and (ii) the stability of that entropy over time. A hashtag is considered to have low entropy if it used predominantly by one community; however, a hashtag's entropy can change over time as we found. Our experiments allow to compare hashtags nature and usage between the USA and France in election time. The results seem to show that Twitter was used mainly for debate in French election, whereas it was intensively used to energize others in the same community in the US election. Following Smith et al. (2014), the US political landscape looks more like a "polarized crowd" than the French landscape that looks more like "community clusters". Appendix in Sect. 2 shows the allocation of the hashtags used in the US and French elections to our taxonomy. It is color-coded so that we can better understand the differences between the countries in the community descriptions we find which we now discuss. It is also related to the previous work that studied the language used in Twitter communities, such as Poblete et al. (2011) and Bryden et al. (2013). However, those works did not use a taxonomy of hashtags, such as ours, to get a better understanding of behavioral differences.

Regarding our second objective, though there exists much work on community structure in social networks, it is limited to just finding communities based on structural connectivity such as the `follower`, `retweet` or `mentions` relation. This limits the usage in other disciplines as the sociological definition of a "community" requires several properties[1] including: (i) interacting people and (ii) members who share *common values, beliefs, or behaviors*. Whilst community discovery methods in social networks address the first point, they usually do not address the second point: We can only guess why those groups of people have been placed in the same community by observing the commonalities between them. To address this limitation, we explore applying our recently published work in the machine learning community on cluster description (Davidson et al. 2018) and we extend it to community explanation.

Given an existing community structure, we show how this method can describe/explain what behavior (using hashtags) individuals in a particular community have used (and when). For example, Tables 2 and 7 explain the behavior (in terms of hashtags) used by the various communities in the USA and France. We can immediately see the amount of general hashtags (colored green and used by most communities) is the overwhelming hashtag type used in the French communities. In contrast, the US communities overwhelmingly use community-speak hashtags (colored red and used by just one community) or intermediate hashtags (colored blue and used by just several communities).

[1] http://www.oxfordbibliographies.com/view/document/obo-9780199756384/obo-9780199756384-0080.xml#firstMatch.

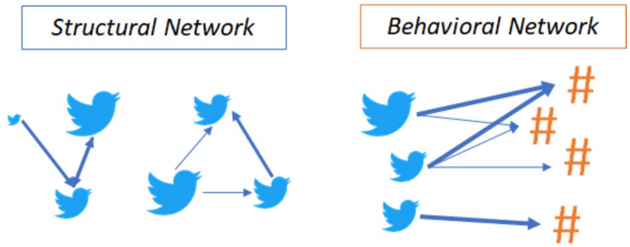



**Fig. 3** Networks we extracted for both primary election datasets. They are weighted and directed

**Organization** We organized the paper as follows: In the next section, we outline how the social network data were generated for our analysis, some general insights about the data and the underlying communities found in the data. In Sect. 3, we overview our methodology for creating a taxonomy of hashtags for the US and French dataset to enhance this understanding and to compare the two elections. In Sect. 4, we overview the methodology on how to create an explanation using the hashtags. In Sect. 5, we apply the methodology in Sects. 3 and 4 to explore several differences between the two countries on Twitter.

## 2 Data collection and community generation

In this section, we address three key issues. First, we explain how our dataset has been set up: the procedure for collecting the data and the design of the complex graph built upon a subset of active accounts. Second, we motivate how we derived the graphs used in our further experiments. In particular, we restrict our analysis to the 10,000 most prolific Twitter users in order to extract a relevant community structure and that we can focus on four main communities for the US dataset and five for the French dataset; we can easily label using well-known accounts. We then spend the remainder of this section to focus on the behavioral aspects (the "posting" activity), as we can access the time-stamped textual content published by the users. By comparing hashtag usage, we demonstrate that the community structure we extracted is relevant. For example, Clinton community behavior is much closer to Sanders community than to Trump community, yet farther than a randomly created community.

### 2.1 Twitter data collected

Our US dataset contains a subset of all the tweets published during the Republican primary election, between the December 30, 2015, and the August 18, 2016. We used the names of politicians[2] involved in the election at this time as query for the Twitter Stream API.[3] Researchers in political science selected these names. We therefore have all the tweets that contain at least one of the words shown in footnote. Similarly, the French dataset contains all tweets published during the French "Les républicains" primary election, between June 21, 2016, and November 4, 2016, containing politician names.[4]

The network is complex as for each user we have their tweet behavior over time that we represent as a time-stamped, vertex-labeled graph as shown in Fig. 2. This means we have not only who retweeted who (the graph structure) but also the hashtags used in those tweet and the time/date of the tweet. Alternatively, we can view the data as a fused regular and bipartite graph as now described. For each dataset, we extract two networks: the retweet network, as it has been shown to grasp political polarization (Conover et al. 2011), and the bipartite relation between hashtags and users. The retweet network is a directed weighted graph, an edge (a,b) connects node a to b if a retweets b and has the number of retweets as value. The user/hashtags network is a weighted bipartite graph, with edge (a,b) being the number of time user a used hashtag b. Edges of both networks are time-stamped. We provide an illustration of the networks in Fig. 3. We recall that both the number of users and of hashtags are huge: for the US dataset, we have 3,448,096 individuals and 2,515,421 hashtags, and for the French dataset, we have 64,438 users and 56,239 hashtags.

### 2.2 Creating the network and extracting communities

Our first goal is to extract the community structure from the structural information. As mentioned earlier, several previous works showed that the retweet network is well suited to grasp politically orientated communities (Conover et al. 2011; Gaumont et al. 2018). To find communities within this retweet, we use the notion of modularity (Girvan and Newman 2002), which is broadly used. Consider $A$ the adjacency matrix, $m$ the volume of the graph, and $deg(i)$ the degree of the $i$-th vertex. Then, modularity $M$ is defined as:

$$M = \frac{1}{2m} \sum_{i,j} \delta_{i,j} \left( A_{i,j} - \frac{deg(i)deg(j)}{2m} \right)$$

[2] bush,carson,christie,cruz,fiorina,gilmore,graham,huckabee, kasich, pataki,paul,rubio,santorum,trump.

[3] https://github.com/AdrienGuille/TweetStreamer.

[4] copé,coppé,fillon,kosciusko-morizet,nkm,lefebvre,le maire,mariton, morano,myard,poisson,sarkozy,sarkosi,sarkosy ,sarko,wauquiez,guaino,aliot-marie,allio-marie.

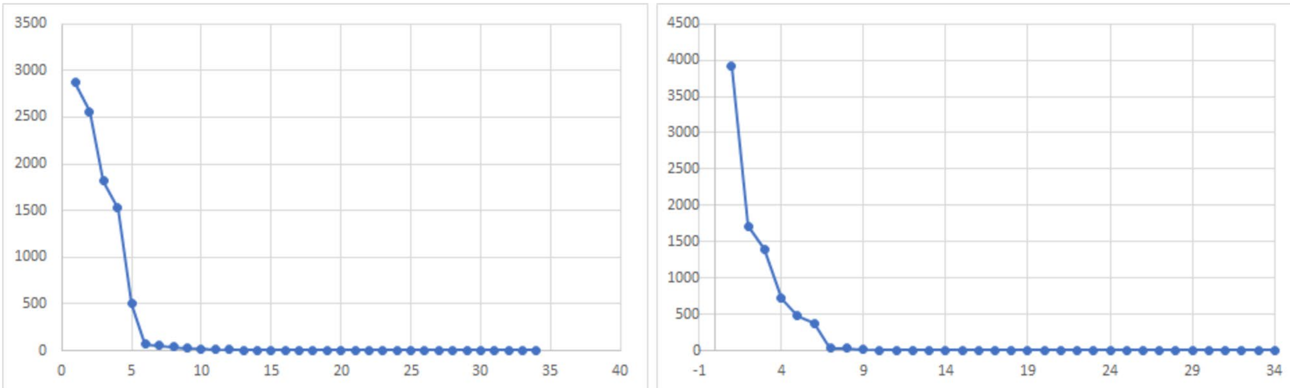



**Fig. 4** Number of vertices per communities (left: USA, right: France). The vertices are concentrated in six communities for USA and seven for France

with $\delta_{i,j} = 1$ if i and j are in the same community, 0 if they are not. Community detection becomes an optimization problem of finding the partition that minimizes this measure. Clearly, computing this for each partition of the graph is highly combinatorial and intractable. The most popular modularity optimization method is the Louvain method (Blondel et al. 2008). In this method, each node is at first randomly associated with a community, then for each node $i$, the modularity increases when moving $i$ to the community of each of its neighbor (i.e., the gain of adding it to $j's$ community minus the gain of removing it from its current community). If all gains are negative, $i$ is not moved. The method assumes convergence when no modularity gain can be obtained. This method is known to be more suitable to large networks, and robust, even if it obviously leads to a locally optimal solution (Blondel et al. 2008).

**Community extraction in our dataset** We show in Sect. 1 that we grasp a coherent community structure with regard to the full network when we consider high-impact users only. We keep the 10,000 users with the higher in-degree. We perform a Louvain community detection (Blondel et al. 2008) on the retweet network.[5]

For the French dataset, we obtain 1261 communities, and for USA, we obtain 442 communities. As shown by the user distribution over communities in Fig. 4, the vertices are concentrated in five communities for the USA and six for France. We therefore aggregate all the nodes that are associated with the smallest communities to a virtual global community we called the "Lambda" community. For the rest of this section, we will deal with these six communities for USA and seven for France (Fig. 5).

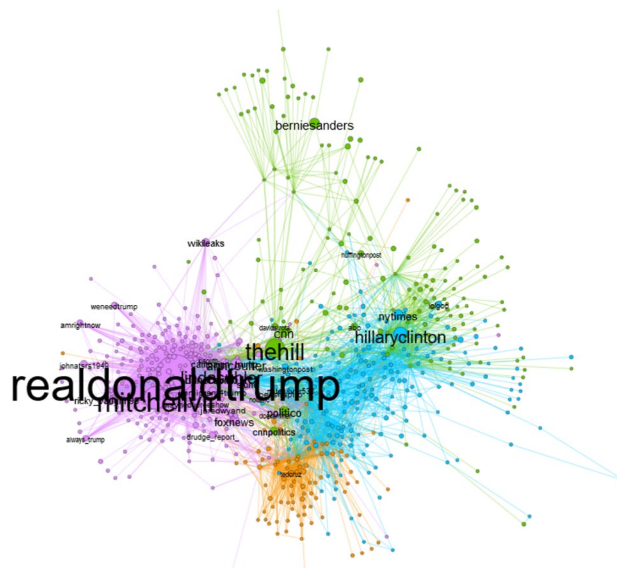



**Fig. 5** Network of the US dataset. Colors correspond to the communities automatically found by the Louvain algorithm (color figure online)

As Conover et al. 2011; Gaumont et al. 2018; Aragón et al. 2013, we find politically polarized communities : pro-Clinton, pro-Trump, pro-Cruz and pro-Sanders for the US dataset, pro-NKM-Juppé, pro-Sarkozy, pro-Fillon, pro-FN and left-wing accounts|media for the French dataset. This is not surprising as people supporting the same candidates most often retweet each other than supporters of other candidates. We label the community according to most central users, as further explained in Sect. 1. We perform a deeper statistical analysis of the communities in the same section for the interested readers.

[5] We performed preliminary experiments for comparing different clustering algorithms from the literature. We found that the Louvain method used by Gaumont et al. (2018) in a similar context leads to the fastest and most robust solution. Furthermore, a thorough comparison with many state-of-the-art clustering algorithms draws the same conclusion in Yang et al. (2016).

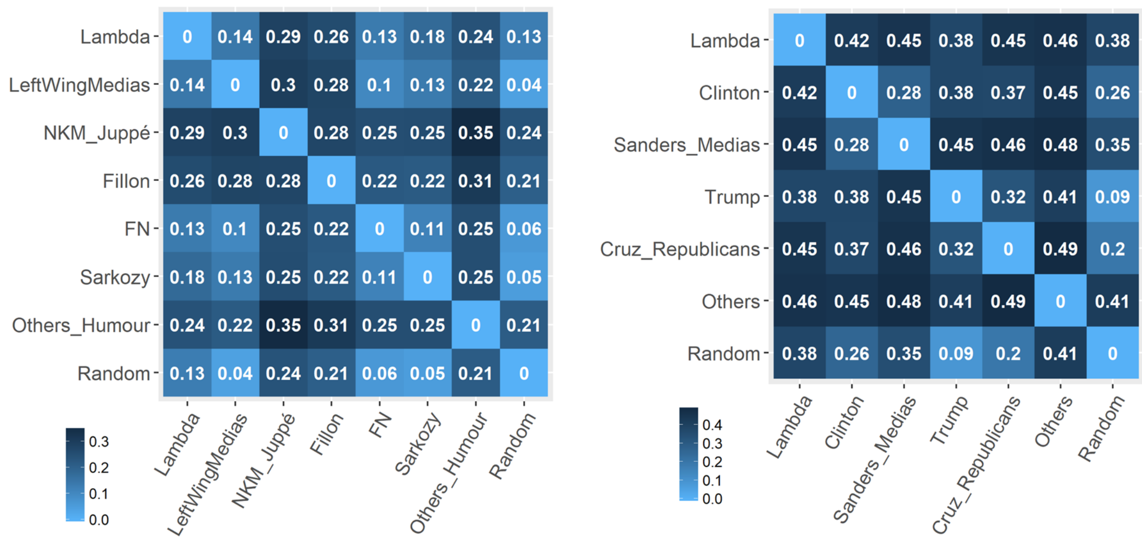


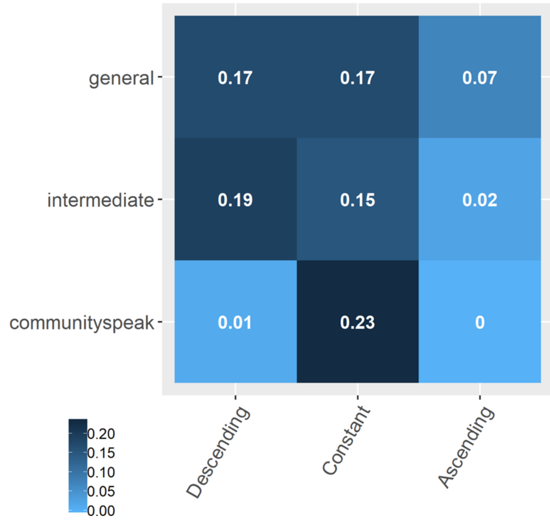

**Fig. 6** JSD (normalized). Every communities are more distant to the other communities than the random community, except for Clinton/Sanders and surprisingly Lambda/Trump and Others/Trump. The lower the more coherent the behavior between the communities

**Hashtag selection** For interpretability's sake, we restrict our analysis to a smaller number of hashtags. As shown earlier, the number of hashtags is enormous and many of them are not interesting to study as they are rarely used. Therefore, we wish to keep popular hashtags but in an unbiased manner. For example, the Trump community intensive use of hashtag will bias any selection measure based on frequency. To circumvent this issue, we take the intersection of two sets of hashtags. First, we extract the 200 most frequently used hashtags, and next the 200 hashtags most used by the 10,000 users (that is for each hashtag we compute the proportion of the 10,000 users that used it at least one time). The intersection of those subsets gives 136 hashtags for the US dataset and 178 for the French dataset which are both popular and "consensual," meaning that they are used often and by most people.

## 2.3 A first comparison of communities using hashtags' use

Now that we have found a limited set of potentially interesting communities based on structural properties, we would like to check whether they are meaningful in terms of behavior. Here, we show that the communities found use language consistency with their political ideologies. A solution would be to compare these communities with a ground truth associating each account to a political opinion. Because of the network size, it is difficult to find such available datasets, and even when we find one of them [see, for instance, the work of Fraisier et al. (2018) on the last French election] the annotation does not encompass all the Twitter accounts.

Instead, we propose to use the behavioral information to bypass this problem in estimating whether the communities show a clear difference in their usage of hashtags. We note by $h$ the number of hashtags, $c$ the number of communities, $t$ the number of time periods. We aggregate the behavioral adjacency matrix by community, to obtain a $c \times h$ matrix $Y$ containing the count of hashtags used by the communities on the whole period. When normalizing this matrix by row, we obtain what we call the communities' distribution over hashtags, noted $Y_n$. We can therefore compute a Jensen–Shannon divergence (JSD) kernel $J$ which is $c \times c$. Each entry of this kernel provides a measure of behavioral distance between communities: the more two communities use common hashtags, the lower the JSD between them. To help the reader to interpret these results, we also provide the JSD with a random community. Users of this community are drawn randomly using a uniform distribution. We draw a number of users equivalent to the mean size of communities. We repeat this measure of JSD on 100 different random samples for each communities to obtain a baseline. We present the results in Fig. 6, where we normalized the JSD divergence using $1 - exp(J_{i,j})$. Every community is more distant to the other communities than the random community in the French dataset and in the US dataset except for Clinton/Sanders, which seems coherent, but more surprisingly between Lambda/Trump and Others/Trump. This shows that the obtained communities provide coherent groups of users in terms of hashtag usage.



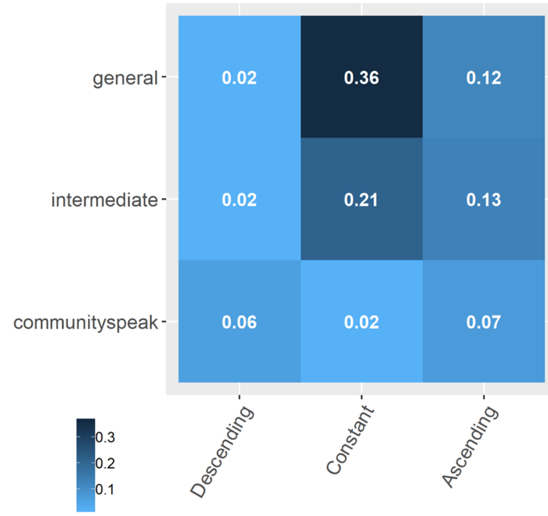




**Fig. 7** Distribution of the hashtags in our full taxonomy subclasses. On the left the results for the US dataset, on the right for the French dataset. The hashtags are mainly community-speak and constant in the US dataset, while they are general and constant in the French dataset. This taxonomy reveals insights on the difference of usage of Twitter between the two countries

## 3 Toward a simple taxonomy of hashtag usage

In this section, we outline a simple taxonomy of hashtags for the political usage of social media. We then map both the French and US election hashtags onto this taxonomy which allows us to compare the behavior of both elections directly for the first time. The taxonomy we create has two dimensions: (1) *Who* uses the hashtags and (2) the *life cycle* of the hashtag. The former dimension addresses whether the hashtag is used exclusively by one community or it is widely used by many communities. The later dimension addresses whether this usage is constant or changes over time. For example, does a hashtag transition from being used exclusively by one community to be used by many communities or vice versa or something else.

To obtain the first dimension, we first use the whole time period and find three clusters of hashtags: hashtags that are used exclusively by one community (something we call **community-speak**), hashtags used by few communities (intermediate hashtags) and general hashtags used by all the communities. Next to obtain the second dimension, we measure how this changes over time. We find, without supervision, different usage patterns, or typical life cycles. For the US dataset, four patterns arise : two constant patterns (hashtags constantly used by one or by a lot of communities), an ascending pattern (one-to-many, meaning that it is used by a few communities, then by many) and a descending pattern (many-to-one, meaning that it is used by many communities initially then almost exclusively by one community). For the French dataset, results are comparable. We find five patterns: two constant patterns (but for many and almost every communities) and a non-monotonous ascending pattern. We simplified both countries dynamics into three evolutionary behaviors: "constant," "ascending" and "descending" in Fig. 7. For each country, we present the frequency of hashtags for each combination of our full taxonomy in Fig. 7.

In this section, we discarded the "Others" and "Lambda" communities for clarity of interpretation and focus on the four main communities corresponding to the pro-Clinton, pro-Trump, pro-Sanders and pro-Cruz groups for the US dataset. We proceed similarly for the French dataset, by ignoring Lambda and Others.

### 3.1 Dimension #1: hashtag usage by community

Firstly, we would like to build a classification of hashtag usage by community. One behavior we noticed is that some hashtags are only used by one community and not the others (e.g., `#MAGA`). These hashtags can be viewed as a kind of "community-speak" or markers that clearly assess the belonging of a member to the community and similar types of hashtag usage could be found. This raises two questions: 1) How do we quantify the usage of a hashtag? and 2) How do we come up with categories of hashtag usage?

We address the former question using the notion of entropy. Recall that there are $m$ hashtags and $c$ communities. For each hashtag, over the entire election period, we compute the distribution of usage over the main communities discovered in the previous sections. This is done by normalizing our $c \times h$ matrix $Y$ of hashtag use by the column sum. We obtain a set $\{y_j\}_{j=1}^{h}$ of normalized column vectors, with

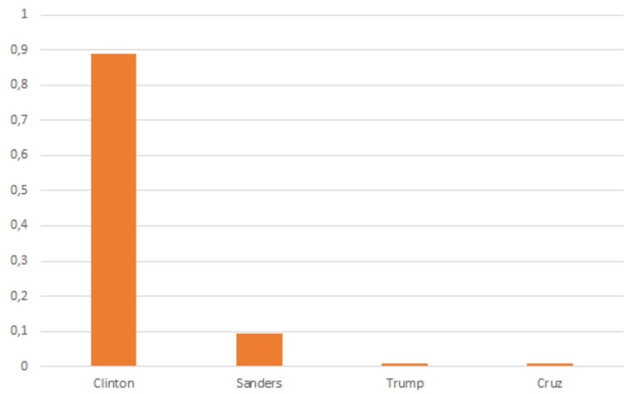



**Fig. 8** Distribution over the four biggest communities of the hashtag `#ImWithHer`

$y_j \in \mathbb{R}^c$. We can compute entropy for each of those vectors. We recall that the entropy is defined, for a specific hashtag distribution $y_j$, by:

$$H(y_j) = -\sum_{d=1}^{c} y_{j,d} \log y_{j,d} \tag{1}$$

A small number of communities using a hashtag leads to a low entropy. On the other hand, hashtags with high entropy tend to be popular, to deal with general topics, i.e., transversal topics. For example, the distribution over communities for the hashtag "#ImWithHer", provided in Fig. 8, has a relatively small entropy (0.4, 16th smallest entropy on our dataset). It was mainly used by the pro-Clinton community (almost 90% of the usage).

To these measures of entropy, we can apply classical Gaussian mixture model (GMM) to cluster the hashtags.[6] Using model selection through BIC criterion, we obtain three clusters, as illustrated in Fig. 9. One of these three clusters is composed of hashtags that are very close to one community (i.e., the community-speak hashtags). The other two are hashtags used by either a few number of communities ("intermediate" hashtags) or by many communities ("general" hashtags). Interestingly, the results are similar to the French dataset as shown in the right-hand figure. We obtain three clusters when using an automatic model selection. The fitted Gaussians are highly close in shape, even if the "community-speak" cluster has higher variance in the French dataset. Hence, we use a common hashtag usage classification for both datasets.

**Visualization of Hashtag Usage By Community.** We propose a visualization to explore the community usage of hashtags. We constructed the plot in this way: we project $c$ community as points on the corner of a c-gon. With $X$ being a $c \times 2$ matrix of coordinates, and $Y_n$ the community usage of hashtags (normalized by rows), we project the hashtags as the barycenter weighted by the community usage. Their coordinate becomes $Y_c$, a $m \times 2$ matrix, explicitly defined as $Y_c = Y_n^t X$. The result is presented in Fig. 10 for the US dataset. Hashtags with colors different than yellow are community-speak hashtags (clearly stuck to each community), and their color corresponds to the community they are associated with. The yellow dots are general hashtags, and gray ones are intermediate hashtags. As we can see, `#Trump` is a general hashtag, meaning that it is used by all the communities, whilst examples of the community-speak hashtags are `#Cruz2016`, `#ImWithHer`, `#NeverHillary` and `#MAGA`. Intermediate hashtags are `#NeverTrump`, used mainly by the Cruz community and the Clinton community, or `#DemsInPhilly`, specific to Clinton and Sanders communities (Fig. 11).

A three-way classification of our hashtags based on entropy seems relevant, but it does not give any hint on the dynamic of the hashtags. Yet, a plausible hypothesis is that "community-speak" hashtags' profile will stay stable in time, as, for example, the `#MAGA` usage will not evolve/change over time. To briefly evaluate this, we compute how many times the community-speak hashtags found on the complete period move to another cluster when we consider time. If a hashtag changes from one cluster to another, that means its entropy significantly changes. For each time period (here, on a monthly basis), we cluster each hashtags in the community-speak cluster or non-community-speak (i.e., the two other clusters) using its entropy at the considered time period. We count how many times community-speak hashtags are not clustered as community-speak hashtag ("a change"). Those hashtags changed nature 14.45% of the time, but this ratio only concerns 43.75% of the hashtags, meaning that 56.25% of the community-speak hashtags stayed in the community-speak cluster for the US dataset. For the French dataset, the percent of change is slightly higher with 22.6%, and it concerns 65.4% of the community-speak hashtags. This observation empirically suggests that "community-speak" hashtags follow a specific dynamic pattern. We will see that other kind of profiles can be automatically found in the next section.

### 3.2 Dimension #2: hashtag life cycles in terms of usage

Given the different usages of hashtags across communities (e.g., the "community-speak" profile), we wish to build a further understanding of hashtags based on the hashtags dynamic usage, that is, how does their usage evolve over time? This time, we would like to use the timestamps of hashtags in the clustering process. As in the previous

[6] To do so, we used the Mclust package (Fraley and Raftery 2006).

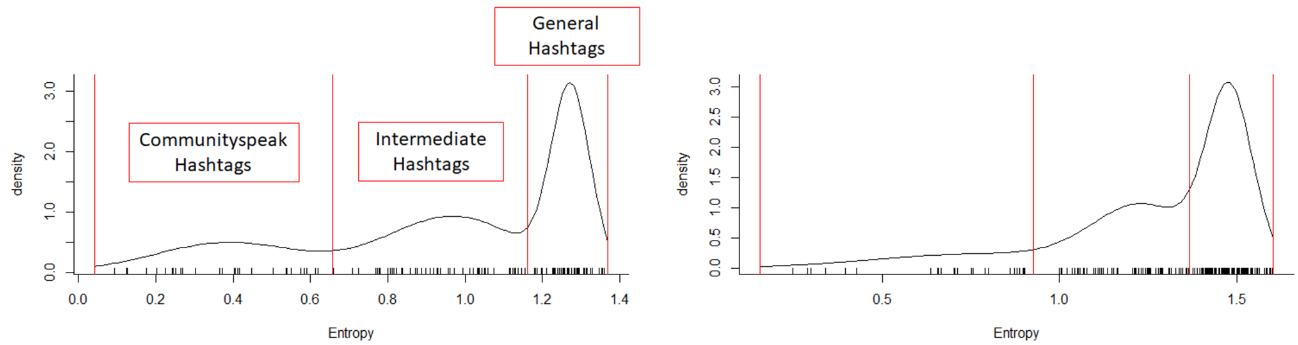



**Fig. 9** Gaussian-fitted clusters of entropy for USA (left) and France (right). Red lines are clusters' limits. We observe three well-defined clusters: the first one contains hashtags with low entropy, specific to one community (`#ImWithHer`, `#MAGA`). General hashtags are defined by a high entropy, i.e., they are used by many communities (color figure online)

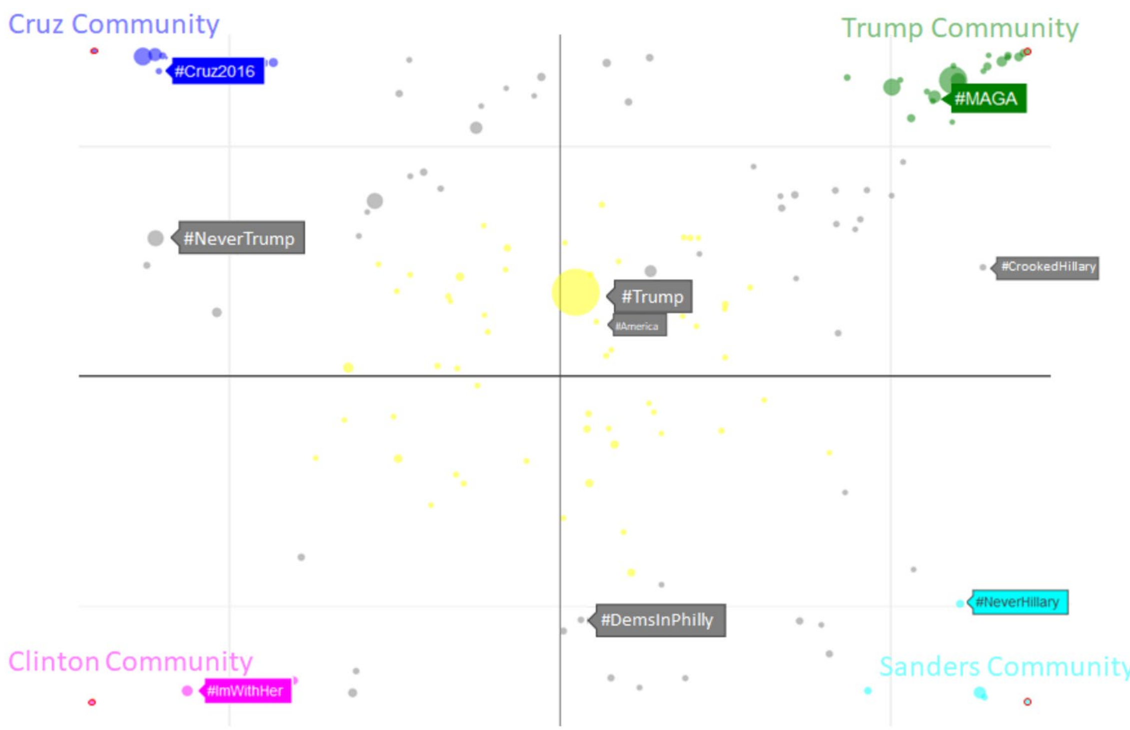



**Fig. 10** Visualization of hashtags usage by communities during the US election. The size of the dot represents the hashtag popularity (`#Trump` is used most often). Yellow hashtags are general, gray intermediate and colors other than yellow are community-speak hashtags. The community-speak hashtag color corresponds to the community they are associated with. For example, the community-speak hashtag `#ImWithHer` is mostly used by the Clinton community (color figure online)

section, we extract the entropy of hashtags distribution over communities *but for each time bin*. For a specific hashtag, we have a signal of entropy, similar to what is presented in Fig. 12 but now for a time bin which in our case is a month.

As we want to observe general patterns of entropy, we are not interested in the timescale information but only on the *evolution* of the entropy. We therefore compute a dynamic time wrapping (DTW) distance for each hashtags pairs[7] and cluster times series using a spectral approach (Ng et al. 2002). We use a classical multidimensional scaling (Torgerson 1958) to embed the obtained kernel. We still need to choose a dimension for the dimension reduction. Since the

[7] using the DTW package (Giorgino and Giorgino 2018).

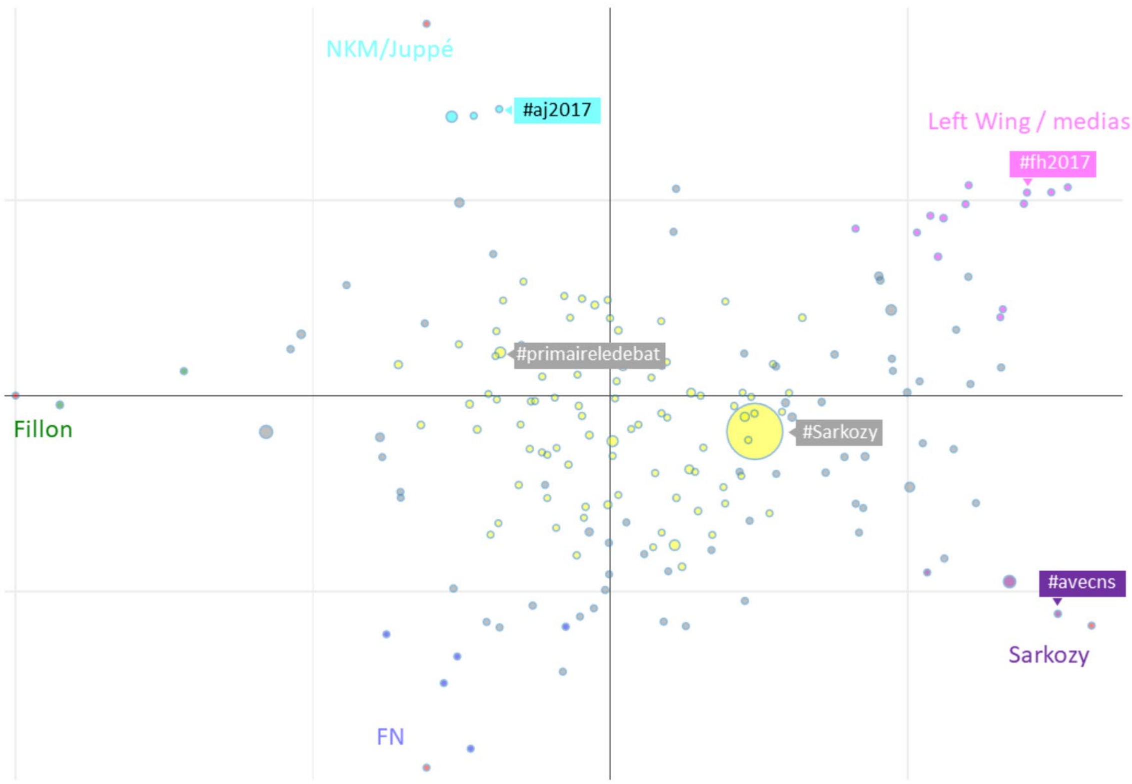



**Fig. 11** Visualization of hashtags usage by communities during the French election. The size of the dot represents the hashtag popularity (`#Sarkozy` is used most often). Yellow hashtags are general, gray intermediate and colors other than yellow are community-speak hashtags. The community-speak hashtag color corresponds to the community they are associated with. For example, the community-speak hashtag `#aj2017` is mostly used by the NKM-Juppé community (color figure online)

clustering is unsupervised, we want the obtained results to be consistent with our previous static profiling. Therefore, we use grid search to find the dimensionality which allows to better separating the community-speak hashtags and the others. To evaluate this, we perform linear SVM on the embedded hashtags with dimension from 1 to $h-1$. The minimum is obtained when taking the first three dimensions for both datasets, with a 4% error rate in tenfold cross validation for the US dataset and 5,4% for the French dataset.

The two first axes of the embedding are provided in Fig. 13. The red dots are the community-speak hashtags. We can see that the DTW is a good distance to use, as it separates very well the community-speak hashtags and the other hashtags in a dynamic context. We then use Mclust with an automatic parameter selection and obtain four clusters for the US dataset and five for the French dataset. The mean entropy “signals” of the clusters are provided in Figs. 14 and 15 for both datasets.

The dark line in front of a figure with k shapes is the entropy value of a hashtag used only by k communities. The mean signals are very easy to interpret, providing explainable profiles of hashtags.

For the US dataset, mean signals are easy to interpret. The blue curve is for hashtags used by several communities at the beginning of the observed period, then by mainly one. The second is for hashtags used by everyone, the third one for hashtags that started in one community and propagated to the others, and the last is for community-speak hashtags used by only one community.

For the French dataset, results are much different. community-speak hashtags are concentrated in clusters 2 and 5, meaning that they were, at least at one time bin, used by two communities. Cluster 3 has a particular dynamic. It

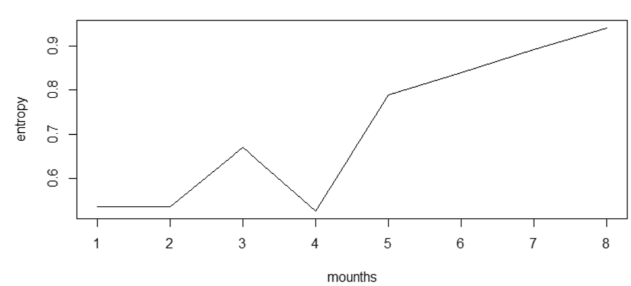



**Fig. 12** Signal of entropy for the hashtag `#NeverTrump`. Its entropy grows, meaning that its usage becomes more and more global as time goes by

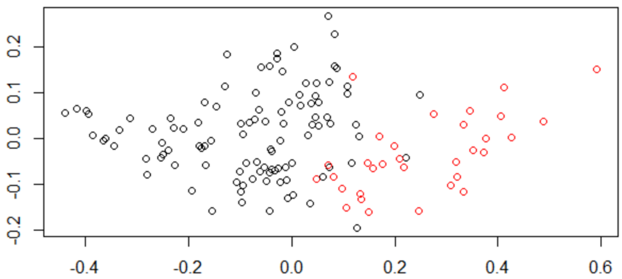



**Fig. 13** First axes of the US DTW kernel embedding. Red dots are **hashtags** previously labeled as community-speak. Using the DTW metric allows to easily separate community-speak and other types of hashtags (color figure online)

presents some kinds of oscillation of the number of communities; it is still ascending (but non-monotonous). Clusters 4 and 1 are clusters for general or intermediate hashtags, used by a group of communities. It is not surprising to find two clusters of that kind as there is one more community in the French dataset.

### 3.3 Comparisons and insights from mapping to the full taxonomy

It is interesting to compare the hashtags' distributions in both datasets for the full taxonomy proposed in Fig. 7. As explained earlier, we simplified the dynamics as "constant," "ascending" and "descending." The frequency of hashtags is displayed in this figure and reveals insights on the difference of usage of Twitter between the two countries.

It seems that the usage of community-speak hashtags was different in the French election: 24% of hashtags are community-speak for the US dataset, whereas they are only 15% for the French one. A quick exploration of Tables 15 and 16 shows that there is less propaganda hashtags (e.g., `#nkm`) in the French community-speak hashtag set than in the USA. Furthermore, the dominant category of hashtags is mainly community-speak and constant in the US dataset, while they are general and constant in the French dataset. Both the intermediate and general hashtags are in the majority descending (they become specific to a community) and constant in the US dataset. On the contrary, French intermediate and general hashtags are mainly constant and ascending. (They are used first by a few number of communities then become general.) Hashtags seem to be picked up and used exclusively by a specific community as time goes by in the US election, whereas the tendency in the French election seems to be the spreading of usage.

Hence, we can reach a conclusion that, based on our experiments, Twitter was used mainly for debate in the French election, whereas it was intensively used for propaganda in the US election. In this section, we demonstrated how we could classify hashtags in both static and dynamic contexts. We will focus now on describing communities with specific hashtags.

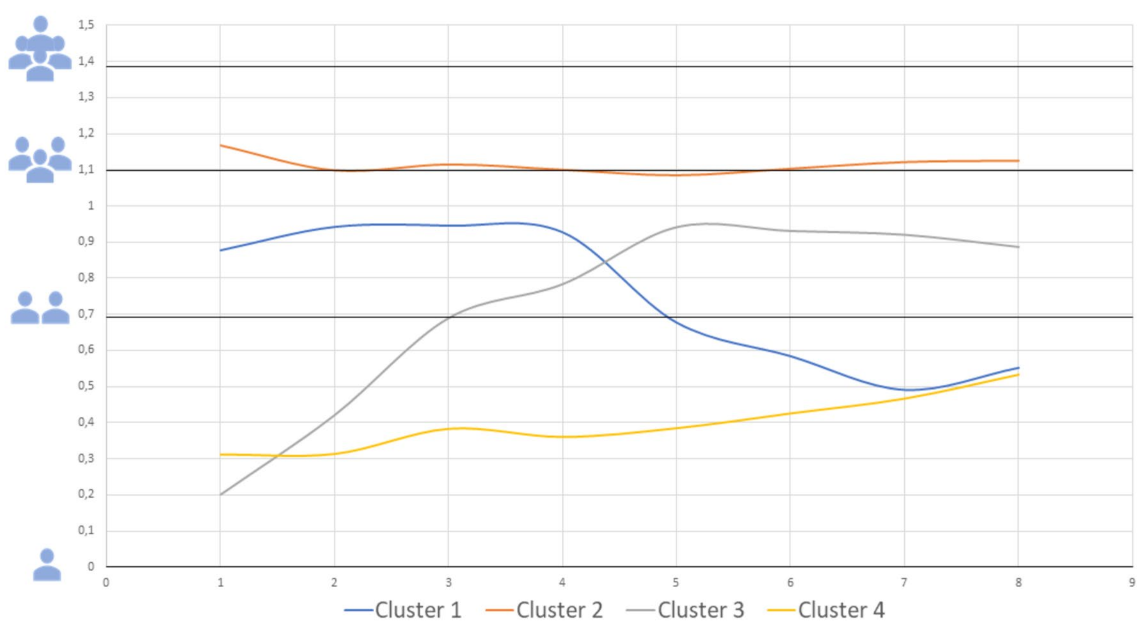



**Fig. 14** Mean entropy signals for the four obtained clusters on **hashtags** on the US dataset over time. We provide baselines for the interpretation of the values. The line in front of a figure with k shapes is the entropy value of an hashtag used only by k communities (from 1 to 4, here). For example, the blue cluster contains hashtags used by few communities first and used by many communities seven months later

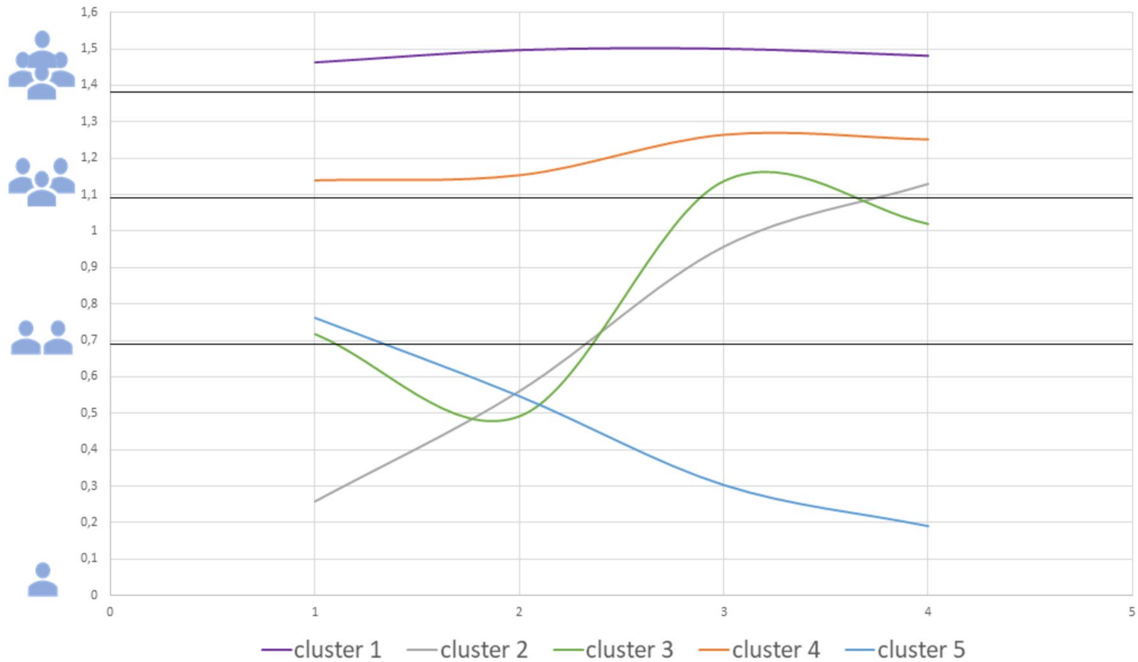



**Fig. 15** Mean entropy signals for the five obtained clusters on the French dataset. We provide baseline for the interpretation of the values. The line in front of a figure with k shapes is the entropy value of an hashtag used only by k communities. For example, the gray cluster contains hashtags used by few communities first and used by many communities four months later

## 4 Toward an explanation of community structure using behavior

Our aim here is to take a set of communities found by another algorithm (in this case the Louvain method as described earlier) using the structure/topology of the graph (retweet behavior) and explain it using the vertex labels (hashtag usage). This is naturally an optimization problem as we wish to search for explanations that are (i) compact/succinct/simple, (ii) wide in coverage in that they explain most or all users in the community. Finding communities and explanations simultaneously is a challenging area, especially for large networks. We explored this research direction in IJCAI 18 (Kuo et al. 2018), which was to *simultaneously* find a community and its explanation. That research direction has several challenges. Firstly, it requires a Pareto optimization setting since there is no need to believe a compact explanation corresponding to a compact clustering; hence, those methods will not scale easily. Secondly (and most importantly), it requires the creation of new algorithms when there are already many existing entrenched high-quality methods that have been extensively used such as the Louvain method.

### 4.1 Community explanation formulation

Here, we describe a variation of our recently published (Davidson et al. 2018) work on cluster description adapted to the community explanation.

**Notation** We explore the idea of taking an *existing* set of communities ($\pi = \{C_1, C_2, \dots, C_k\}$) defined over the instance $S = \{s_1, s_2, \dots, s_n\}$ found using dataset $X$ and explaining it using another dataset $Y$. For example, $X$ would be (as they are in this paper) the $n \times n$ adjacency matrix of a graph showing the *retweet* relation between individuals. For each individual/account, we have a set $y_i \subseteq H$ of tags, $1 \leq i \leq h$ which together define $Y$ an $n \times h$ *behavioral* information matrix showing how often each individual posted on each of $h$ different hashtags. Importantly, only $X$ and *not* $Y$ was used to find the communities; hence, this is not a semi-supervised setting.

**Problem definition** The goal of community explanation is to find a subset $H_j \subseteq H$ of tags for each community $C_j$ $(1 \leq j \leq k)$ such that all the following conditions are satisfied.

(a) For each community $C_j$ and each instance/account $s_i \in C_j$, $y_i$ (the tags for the instance) has at least one of the tags in $H_j$; formally, $|H_j \cap y_i| \geq 1$, for each $s_i \in C_j$ and $1 \leq j \leq k$.
(b) The sets $H_1, H_2, \cdots, H_k$ are pairwise disjoint.

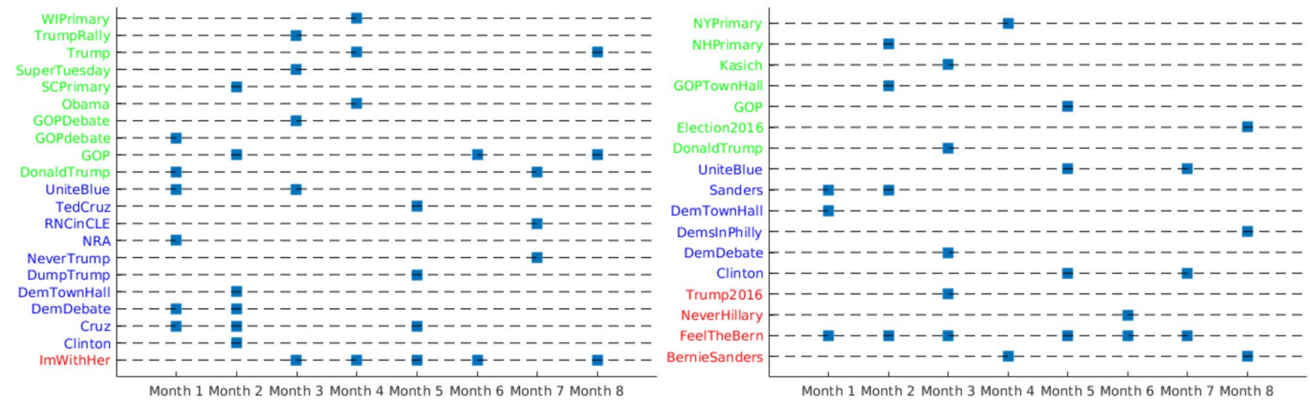



**Fig. 16** US election. The evolution of community explanation (using hashtag usage) by month (x-axis) for democratic candidates. Left figure Clinton, right Sanders. c.f. Table 3

For $1 \le j \le k$, the set $H_j$ will be referred to as the **descriptor** for community $C_j$. Later in Sect. 4.1, we will use this definition to formulate a minimization problem which adds the requirement that the size of the description is minimized, that is:

(c) $\sum_j |H_j|$ is minimized.

**Interpretation of explanation** The interpretation of $H_j$ as an explanation is as follows. Each instance in $C_j$ has at least one of the tags in $H_j$. That is if $H_j = \{Maga, MakeAmericanGreatAgain, Trump\}$ then all instances in community $j$ use one or more of these hashtags.[8] Furthermore, since we require these explanations to be pairwise disjoint, common tags can only be used to explain one community.

**An ILP formulation** As before, we are given a community structure $C_1, C_2 \dots C_k$ of $n$ instances with each instance described by a subset of the $h = |H|$ tags. These tags are in the $n \times h$ matrix $Y$. We solve for the $k \times h$ binary matrix $X$ where $X_{i,j} = 1$ iff community $i$ is described by tag $j$. One objective function then is simply to find the most concise overall community description:

$$argmin_X \sum_{i,j} X_{i,j} \tag{2}$$

Hence, the number of variables in this version of the formulation is $k \times h$ where $k$ is the number of communities and $h$ is the number of tags.

Our first basic constraint includes the set coverage requirement for each different community/universe. Here we must define the matrices $S^1, \dots, S^k$, where $S^a_{i,j} = 1$ iff the $i^{th}$ instance is actually in community $a$ and has tag $j$. Note that $S^i, 1 \le i \le k$, can be pre-computed. Since each instance must be explained/covered, there will be $n$ constraints of this type.

$$s.t. \ \sum_j X_{k,j} S^k_{i,j} \ \ge \ 1 \ \ \forall \, i \in C_k, \ \forall \, k \tag{3}$$

Our next basic constraint requires that the tags chosen to represent each community do not overlap that is they must be disjoint ($w_j = 1$) or minimally overlap ($w_j > 1$), where $w_j$ is the maximum number of times tag $j$ can be used in all descriptors. In all our experiments $w_j = 1, \ \forall j$. This is simply an OR constraint and can be encoded as:

$$s.t. \ \sum_i X_{i,j} \ \le \ w_j \ \ \forall j \tag{4}$$

There will be $t$ constraints of this type where $t$ is the number of tags. So overall the number of variables to solve for is $O(hk)$, and the total number of constraints is $O(n + h)$.

## 5 Discussion on behavioral differences

Community discovery algorithms only find: (1) interacting people but they do not find the second requirement of a community and (2) which is the common behavior members in a community share. Here we attempt to address that second requirement experimentally with the following questions. We outline several common questions to experimentally address for each Twitter dataset separately. Those questions are:

1. What behavior can be used to describe all communities? (USA: Table 2, France: Table 7)
2. What behavior explains each community? (USA: Table 3, France: Table 8).

[8] When there is no ambiguity, we remove the sharp symbol, that means Trump stands for #Trump.

**Table 2** US election. What tags explain (cover) all main four communities. Note these are heavily skewed toward Republican hashtags and belonging to our general usage taxonomy category (see Table 16) (color figure online)

| Groups | Description |
|---|---|
| All Four Main Communities | GOPdebate ∨ Trump ∨ ∨ GOP ∨ CruzCrew |

**Table 3** US election. The four main communities found by the Louvain method (the first is an amalgamation of the smaller communities) on the retweet graph and their description using hashtags. Red = Community-speak, blue = intermediate and green = general per our taxonomy in Table 16 (color figure online)

| Community | Description |
|---|---|
| Other (C1) | GOP ∨ BernieSanders |
| Pro-Clinton (C2) | ImWithHer ∨ DemDebate ∨ Sanders |
| Pro-Sanders (C3) | DemsInPhilly ∨ IowaCaucus ∨ FeelTheBern ∨ DonaldTrump |
| Pro-Trump (C4) | Trump∨ SuperTuesday ∨ MakeAmericaGreatAgain |
| Pro-Cruz (C5) | GOPDebate ∨ Cruz ∨ Clinton ∨ Breaking |

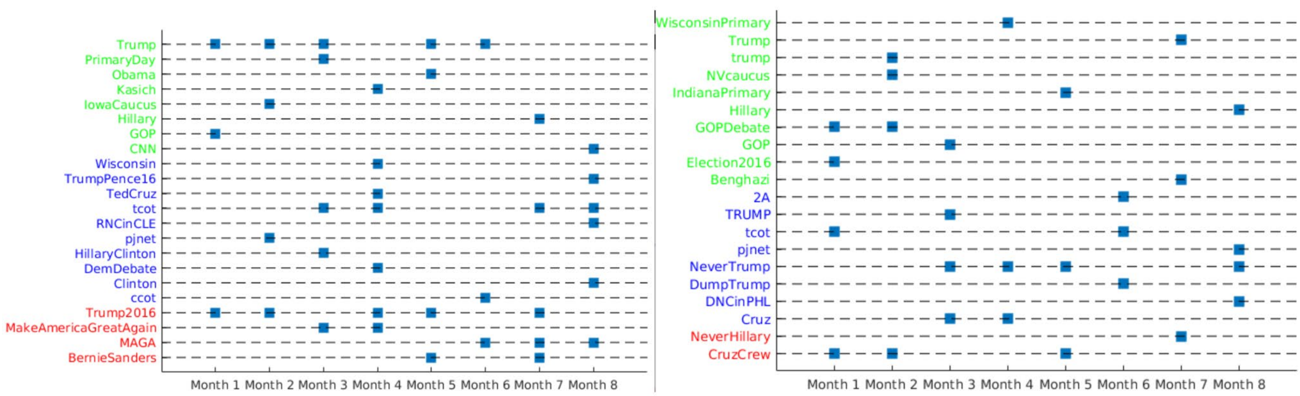



**Fig. 17** US election. The evolution of community explanation (using hashtag usage) by month (x-axis) for Republican candidates. Left figure Trump, right Cruz. c.f. Table 3

3. What behavior explains *merged* communities? (USA: Tables 4, 5 and 6 , France: Table 9)

The explanations we find are color-coded to match some aspects of the taxonomy created in Sect. 3 and detailed in the appendices so we can easily see the differences between the communities.

## 5.1 Results: US election

**Q1: What behavior explains all communities** Here we apply the community explanation method described except rather than have *k* communities we have just one community (all four communities amalgamated into one). Table 2 shows interestingly that the description for these four communities is heavily skewed to using Republican-related hashtags with there being only one Democratic hashtag (`#DemDebate`). This indicates that the democratic leaning Twitter accounts of the pro-Clinton and pro-Sanders communities were using not only hashtags about their own candidate but other candidates as well. As expected, all hashtags (except `#DemDebate`) belong to the general category in our taxonomy (see Table 16).

**Q2: What behavior is specific to each community** Table 3 shows the behavior that explains each community separately. We see that the majority of hashtags used to describe each community is an example of *community-speak*, that is hashtags that are used almost exclusively by only one community (see Table 16). Figures 16, 17 show how the four main communities' behavior evolves. We see that all communities have a few tags that are used over many months: pro-Clinton (`#ImWithHer`), pro-Sanders (`#FeelTheBern`), pro-Trump (`#Trump`, `#MAGA`, `#Trump2016`), pro-Cruz (`#NeverTrump`, `#CruzCrew`).

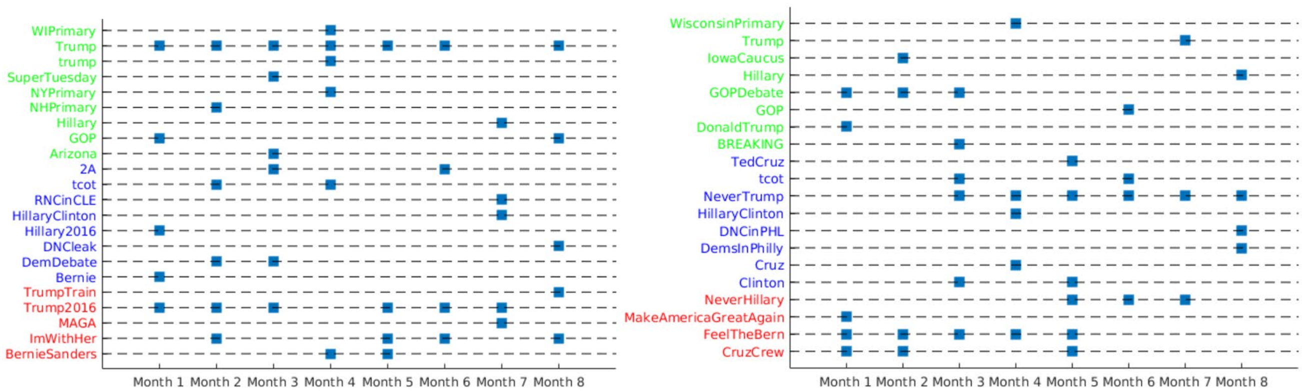



**Fig. 18** US election. The evolution of joint-community explanation (using hashtag usage) by month (x-axis). Left figure Clinton/Trump, right Sanders/Cruz. c.f. Table 4

**Table 4** US elections. What tags explain the Clinton/Trump communities (C2, C4) and the Sanders/Cruz communities(C3, C5)

| Groups | Description |
|---|---|
| C2, C4 | GOPdebate ∨ DemDebate ∨ IowaCaucus ∨ tcot ∨ trump ∨ TrumpTrain ∨ TRUMP |
| C3, C5 | Trump ∨ SCPrimary ∨ FeelTheBern ∨ iacaucus ∨ Cruz2016 ∨ Breaking |

**Table 5** US elections. What tags explain the Clinton/Sanders communities (C2, C3) and the Trump/Cruz communities(C4, C5). note tcot = top conservatives on Twitter

| Groups | Description |
|---|---|
| C2, C3 | FeelTheBern ∨ ImWithHer ∨ Ohio ∨ UniteBlue ∨ Trump |
| C4, C5 | GOPdebate ∨ Clinton ∨ tcot ∨ MakeAmericaGreatAgain ∨ TrumpTrain ∨ TedCruz |

**Table 6** US elections. What tags explain the Clinton/Cruz communities (C2, C5) and the Sanders/Trump communities(C3, C4). note tcot = top conservatives on Twitter

| Groups | Description |
|---|---|
| C2, C5 | ImWithHer ∨ iacaucus ∨ Kasich ∨ Bernie ∨ Trump ∨ DumpTrump ∨ Cruz2016 |
| C3, C4 | GOPdebate ∨ GOP ∨ FeelTheBern ∨ tcot ∨ Veterans ∨ MakeAmericaGreatAgain |

Interestingly, we see that pro-Cruz and pro-Sanders community have the most simplest explanations (indicated by the sparsest figure).

**Q3: What behavior explains pairs of communities** Tables 4–6 show community explanations for pairs of communities. First main candidates (Clinton/Trump) versus secondary candidates (Sanders/Cruz) in Table 4 and then democratic versus republic candidates in Table 5. We see in the first table that the Republican leaning hashtags are almost exclusively used, whilst in the second table we find the expected results. Finally, for completeness we show the last combination of main candidates paired with secondary candidates from the other party in Table 6, and again, we find that the behavior of a Democrat/Republican community is best explained using mainly Republican-focused hashtags.

Figures 18, 19 and 20 show the evolution of community description. We find (Fig. 18) for the Trump/Clinton, pairing that the hashtag `#Trump` is almost always used to describe this community as is `#Trump2016` and to a lesser extent `#ImWithHer`. No such dominating tag is found for the Cruz/Sanders pairing. In Fig. 19, an interesting observation is found. The democratic pairing of candidates creates explanations that to large extent cover the candidates in the **other** party, but this is not true for the Republican pairing of candidates.

## 5.2 Results: French election

We will now address the three question presented above for the French results.

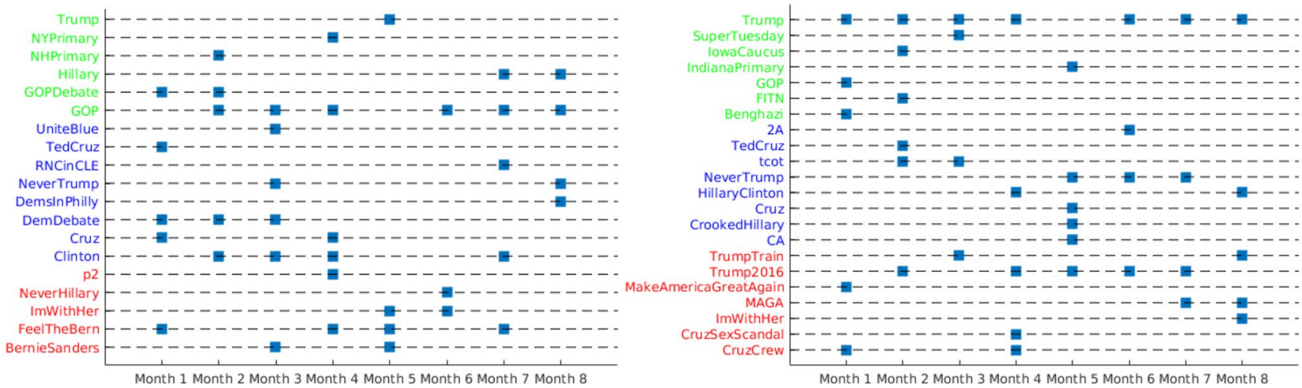



**Fig. 19** US election. The evolution of joint-community explanation (using hashtag usage) by month (x-axis). Left figure Clinton/Sanders, right Trump/Cruz. c.f. Table 5

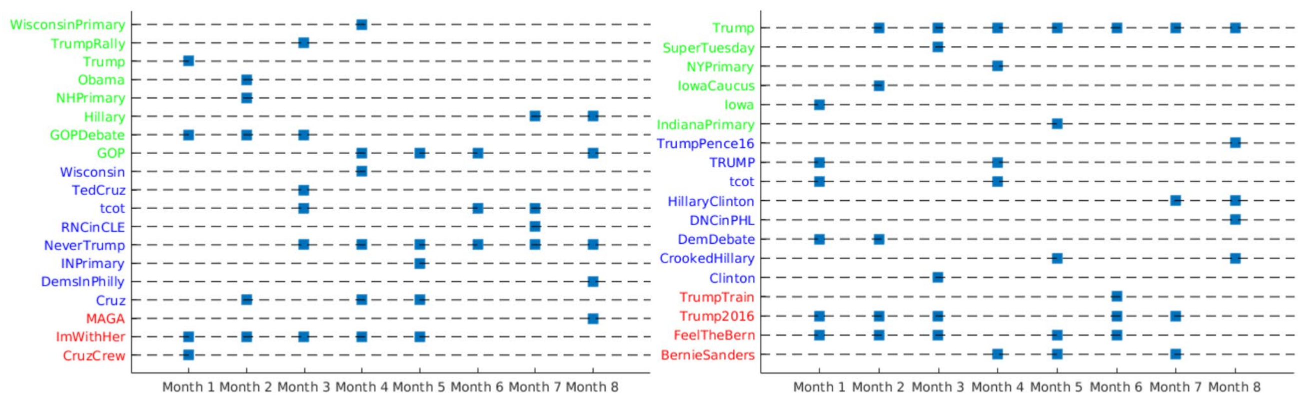



**Fig. 20** US election. The evolution of joint-community explanation (using hashtag usage) by month (x-axis). Left figure Clinton/Cruz, right Trump/Sanders. c.f. Table 6

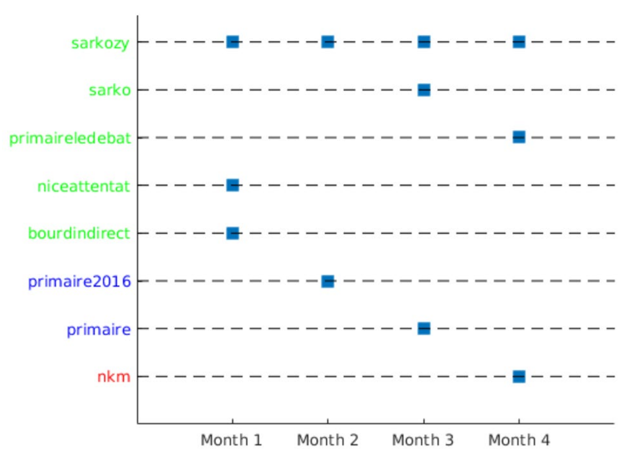



**Fig. 21** French election. The evolution of community explanation for all the main communities. c.f. Table 7. It is interesting to note that the only ""community-speak" hashtag #nkm (in red) occurs in October 2016 (Month 4) when the candidate Nathalie Kosciusko-Morizet released her program for the country, named "Nouvelle société, nouvelle France" (c.f. https://fr.wikipedia.org/wiki/Nathalie_Kosciusko-Morizet#Campagne) (color figure online)

**Q1: What behavior is common to all communities** Table 7 presents the hashtag description for the five main communities altogether. The `#bygmalion` case was about illegal funding of the `#sarkozy` campaign during the previous election. It seems that it was central in Twitter. `#primaireledebat` is a general event and as expected most users tweeted about it. `#quotidien` is a very famous news TV showing that gained popularity during the elections. When looking at the evolution of the hashtag description in Fig. 21, it is clear that we can separate a background topic (`#sarkozy`) and hashtags related to temporary topics such as events. It is not surprising to find `#niceattentat`, which relates to the Nice terrorist attack that happened during the first observed month. Nevertheless, it is striking to observe the lack of interest in the following months.

**Q2: What behavior is specific to each community** Table 8 shows the behavior that explains each community. Surprisingly, there is not that much propaganda hashtags, in contrast to the US dataset. The FN community talks a lot about the candidates, which is consistent with the

**Table 7** The results for the main communities together (C2–C6) on the whole period

| Groups | Description |
| --- | --- |
| C2, C3, C4, C5, C6 | sarkozy ∨ primaireledebat ∨ bygmalion ∨ quotidien |

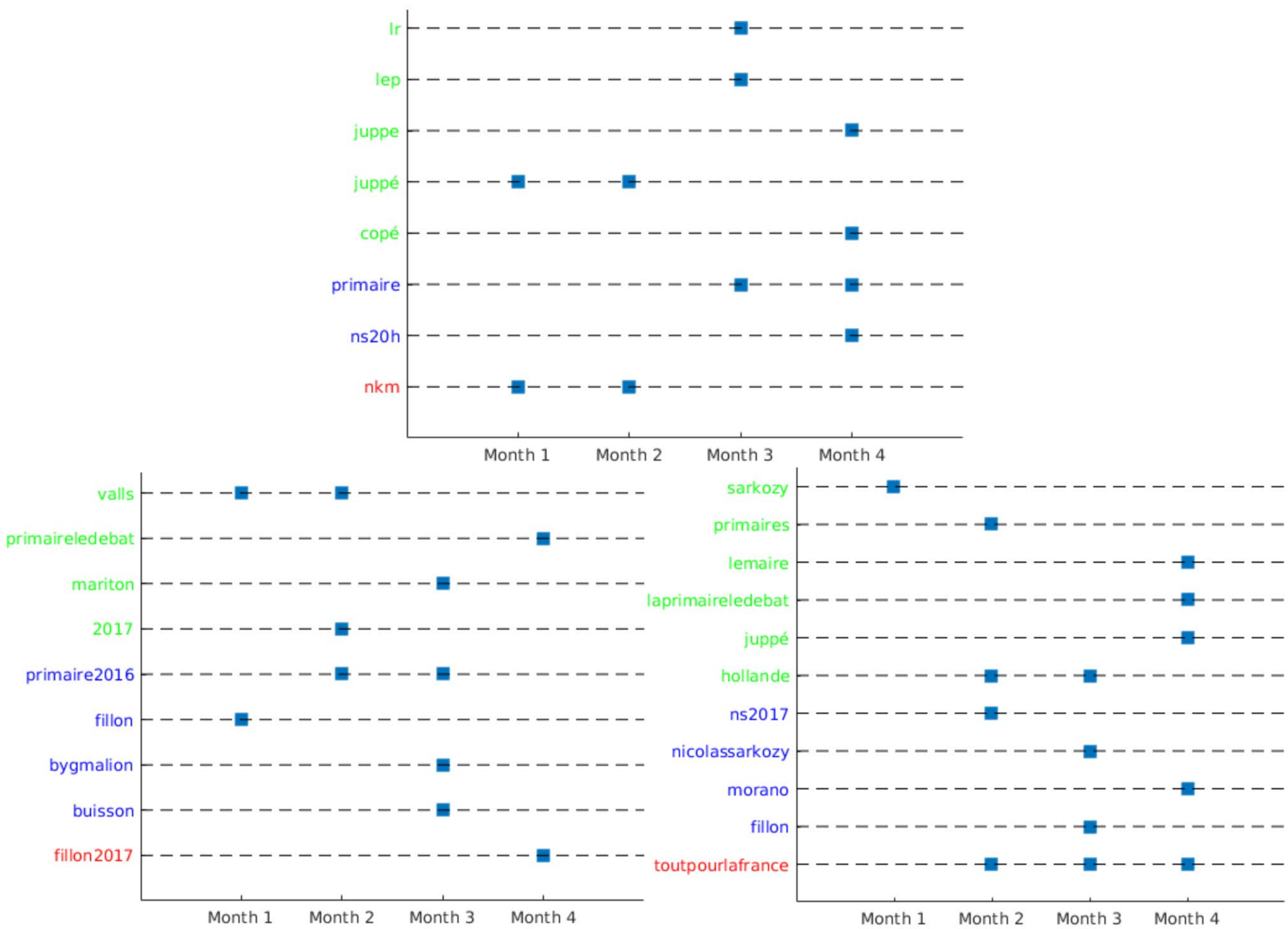



**Fig. 22** French election. The evolution of community explanation (using hashtag usage) by month (x-axis) for NKM-Juppé, Fillon and Sarkozy . c.f. Table 8

**Table 8** The table for the hashtag description of the main French communities on the whole time period (static)

| Community | Description |
| --- | --- |
| LeftWingMedias (C2) | hollande ∨ bygmalion ∨ lr ∨ lep ∨ gaulois ∨ copé |
| NKM-Juppé (C3) | primaireledebat ∨ primaire ∨ politique |
| Fillon (C4) | fillon ∨ fn |
| FN (C5) | nkm ∨ bourdindirect ∨ brexit ∨ attentatnice ∨ juppé ∨ sarko ∨ morano ∨ lepen |
| Sarkozy (C6) | sarkozy |

communication strategy of this party: they spent a lot of time criticizing the other parties and politicians. The Sarkozy community only uses the `#sarkozy` hashtag. The media community is the only one to extensively use `#bygmalion` but also `#gaulois`. This hashtag was created after the controversy raised by the former president who said that "every French citizen has Gaulish ancestors."

Figures 22 and 23 show how the five main communities' behavior evolves. In Fig. 22, we clearly see topics over time in the NKM-Juppé usage of hashtags. For instance, we can clearly see the temporary interest on the invitation

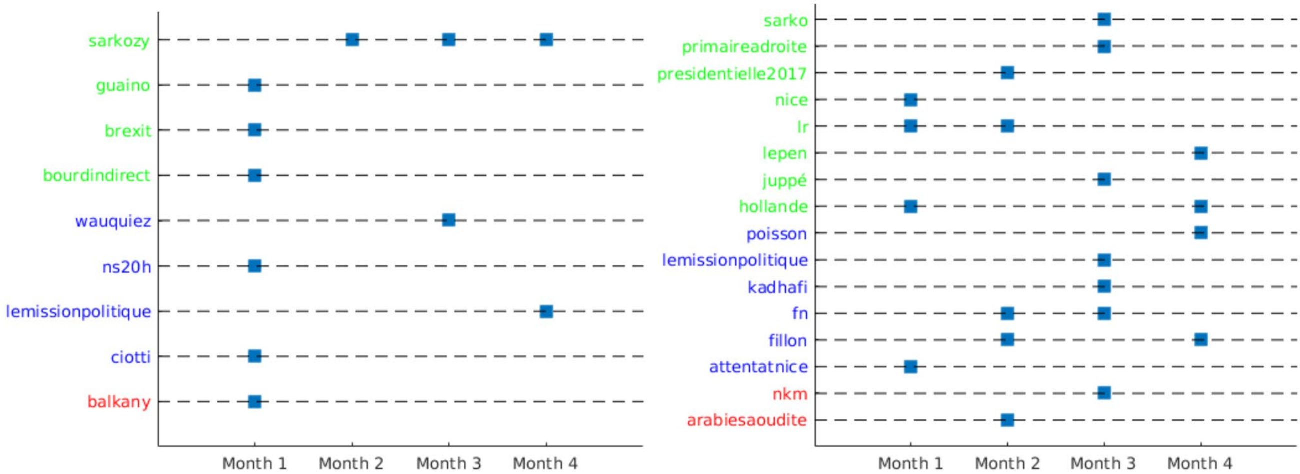



**Fig. 23** French election. The evolution of community explanation (using hashtag usage) by month (x-axis) for left wing/Medias (left) and FN (right). c.f. Table 8

of N. Sarkozy on French television (`#ns20h`) on the 13th of November (Month 4). The left-wing community usage of hashtag overtime (Fig. 23) is also interesting. They talk about general subjects at the beginning of the campaign and then focus on `#sarkozy`.

**Q3: What behavior explains merged communities** Table 9 shows community possible explanations for merged moderate right-wing communities (C3, C4, C6). It is interesting to note that the description is really compact. The only consensual topics are `#sarkozy` and the event-related hashtag `#primaireledebat` (the primary debate that occurred in October, i.e., Month 4). More interestingly, `#sarkozy` disappears when considering the description by month (see Fig. 24). The topics addressed tend to be more general even if the politician names remain.

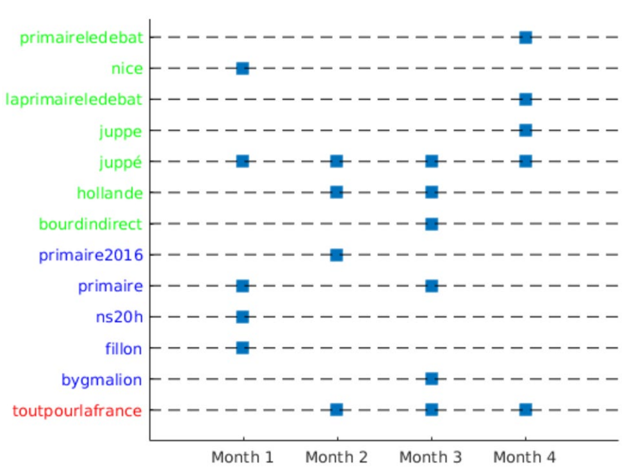



**Fig. 24** French election. The evolution of joint-community explanation (using hashtag usage) by month (x-axis) for communities C3, C4 and C6 (right wing, moderate). c.f. Table 9

## 5.3 Discussion

As shown in previous studies (Hong et al. 2011), strong behavior differences can be observed on users of different languages. In particular on Twitter user, origin matters in the way she uses and disseminates hashtags on the social network. Our work is maybe the first that highlights these differences in language usage between France and the USA. Besides the differences are observed at the level of hashtags' taxonomy, which means that the types we set up in Sect. 3 (e.g., community-speak) can be used in further studies to better understand the effect of language on information diffusion through communities, similarly to what has been done on blogs (Herring et al. 2007). It is clearly what can be seen when comparing Table 8 for the French campaign (majority of "general"—green—followed by "intermediate"—blue—and just one "community-speak"—red—) and Table 3 for the US campaign (the three types look balanced). For the USA, we can also hypothesize that this apparent balance between hashtag types is highly dependent on the current context. It is visible when looking at the evolutions in Figs. 18, 19 and 21. All those observations have been made possible by using our explanation mechanism that lets the most important hashtags emerge.

**Table 9** the results when merging community C3, C4 and C6

| Groups | Description |
|---|---|
| C3, C4, C6 | sarkozy $\vee$ primaireledebat |

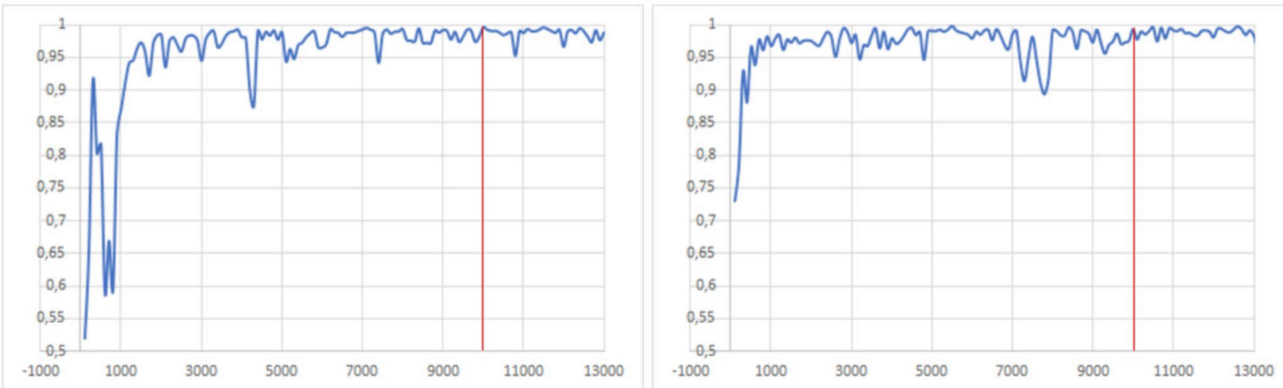



**Fig. 25** Perturbation of community structure when adding new vertices. The measure is the adjusted Rand index.(left : USA, right: France). Note the ARI is measured between adjacent datasets, that is the between a dataset of $x$ accounts and $x + 100$

## 6 Conclusion

We have modeled election data from the Twitter network as a complex graph that has an edge structure defined by structural connectivity (retweet) and with vertex labels derived from behavioral data. We have explored using well-known community detection methods applied to the structural connectivity to find communities and then trying to understand and explain these communities using the behavioral data on the labels of the vertices. To compare the elections from both countries, we created a taxonomy of hashtag usage based on how exclusive the hashtag is used by the communities and the evolution/stability of that measure (see Fig. 7). We believe this is the first time election results from two different countries are directly compared in such a way. We find that in the US Twitter data, hashtags which are examples of community-speak (spoken almost exclusively by one community) are the dominant hashtag used, whilst for the French Twitter data, hashtags that are used by many communities are common. We explore explanations of communities and also their evolution over time (month by month). Interestingly, we saw that the explanation for the entire community period where quite different to the month-by-month explanations for both the USA and French elections, indicating that the hashtag usage was quite dynamic.

## Appendix 1: community extraction, a deeper analysis

In this section, we propose a deeper analysis of the community extraction in our dataset. First, we demonstrate that filtering the users does not affect the general community structure. Second, we attempt to label the communities using well-known accounts. We next explore the properties of the subgraph for each community. For example, do some communities have more denser connections than others that is, is the Trump community subgraph more denser than the Clinton subgraph? (Yes it is, even though their clustering coefficient is similar).

### High-impact users

Despite the Louvain method's scalability, we will focus on high-impact users during this study for clarity of interpretation. We, therefore, must ensure a community structure similar to the full network found in the network of high-impact users we select. To evaluate this requirement, we analyze whether adding users with lower in-degree (less impact) will modify the community structure found using only high-impact users as follows.

We first build the retweet network of the 100 users with the higher in-degrees and then perform a Louvain community detection (Blondel et al. 2008). We then compute the ARI between this clustering output and the clustering of the 200 users Retweet network and then the ARI between the 200-node clustering and a 300-node clustering and repeat this process until we reach 13,000 users. This produces a measure of the perturbation of the community structure when adding nodes with lower and lower in-degree. The results are shown in Fig. 25. Starting from 7000, perturbation seems to be minimal for US retweet network. For the French networks, convergence seems to happen around 10,000 users. For ease of comparison, we used the 10,000 most influential accounts for both datasets. Hence from now on all experimental results will be on the 10,000 users with the highest in-degree.

### Validation of communities

For further analysis, we would like to label the community obtained in the previous section. A straightforward approach for getting a better understanding of a group of users lies in the identification of the most central users. To this end, we extract the subgraph of each community and we compute the Page Rank score for every user. Results are shown in Tables 10 (US politics, 6 clusters) and 11 (French politics, 7 clusters). It is a simple way to check whether there is a clear political leaning for an observer of US or French politics. There is a clear political polarization of the communities we found. For instance, if we look at the nodes most central to the cluster 4 (US dataset) we can see many accounts that have a clear political orientation in favor of D. Trump including the candidate, his children and the news outlets that favor him (e.g., @realDonaldTrump, @DonaldTrumpJr, @FoxNews, @foxandfriends). We choose to simply label this cluster as "Trump." Similarly, cluster 3 has the political candidate Bernie Sanders and the news outlets that are supportive of him (e.g., @CNNPolitics, @ABC, @GMA); hence, we label it as "Sanders." For the French dataset, most of the discovered clusters are identifiable. For instance, the cluster 2 is related to political personalities and media that can be associated to the left wing of French politics (e.g., @najatvb of Najat Belkacem, radio account @lelab_e1, @ellensalvi of the Mediapart journalist Ellen Salvi). The cluster 5 is clearly related to the Front National party (e.g., @f_philippot of F. Filippot, @tprincedelamour and @avec_marine that are a clear support of the candidate M. Le Pen). Therefore, it is rather straightforward to give them a first label to help in understanding the analysis that is carried out in the next sections. There is no surprise we cannot assign a good label to the Lambda cluster since it is composed of a myriad of small communities. In both datasets, we can also see one cluster (number 6 for the US and 7 for France) that has no clear political orientation and we choose to label these two clusters as "Others."

Table 12 shows a set of classic graph measures that have been calculated on the same set of communities (US dataset, top table, and French dataset, bottom table). It is not surprising that the Lambda community's diameter is small due to its low size. Furthermore, this pseudo-community is characterized by a very low mean in-degree (at least ten times less than the others are) and eigen centrality (at least five times less). This is also the case for the Others community as it is a very sparsely connected community. For these reasons, we choose not to focus on these two loosely defined communities for the rest of our analysis. For the US dataset, the higher clustering coefficient can be explained: Locally, the neighborhood of a node has higher chances to get connected together because of the components size (small diameter). However, we do not observe the same for the French dataset.

The remaining communities (4 for the USA and 5 for France) share that they are much denser and with a higher in-degree. However, it is interesting to note some differences. For instance, the Trump community is much denser than all the other communities if we look at the mean degree, so is the Sarkozy community. Their respective diameter values seem to be lower (8 for Trump and 7 for Sarkozy). It is also interesting to note that the clustering coefficient is uniformly smaller for the French communities than the USA communities indicating a more spoke to hub arrangement for the later.

## Appendix 2: The allocation of hashtags to our taxonomy

See Tables 10, 11, 12, 13, 14, 15 and 16.

Here we list the allocation of hashtags to the taxonomy created in Fig. 7.

**Table 10** Most central users based on page rank and their in-degree for the US dataset

| Community 1 (Lambda) | | Community 2 | | Community 3 | | Community 4 | | Community 5 | | Community 6 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Highest PR | In-degree | Highest PR | In-degree | Highest PR | In-degree | Highest PR | In-degree | Highest PR | In-degree | Highest PR | In-degree |
| kokilamodizee | 566 | maggieNYT | 10390 | CNNPolitics | 19358 | r**realDonaldTrump** | 230068 | Mediaite | 12827 | ShopGVNG | 174 |
| InfobaeAmerica | 311 | SopanDeb | 20578 | davidsirota | 28161 | FoxNews | 27621 | teddyschleifer | 23284 | ClassicPict | 175 |
| gopibahuzee2 | 552 | PhilipRucker | 6886 | CNN | 13212 | foxandfriends | 12225 | PatrickSvitek | 26606 | Jay_IDK | 46 |
| nishabharti0001 | 529 | mmurraypolitics | 9213 | ABC | 9909 | mitchellvii | 293453 | allahpundit | 9144 | classifiedfact | 36 |
| RocanrolNa | 13 | BuzzFeedAndrew | 17441 | GMA | 1405 | AnnCoulter | 78124 | ajjaffe | 7038 | Uber_Pix | 74 |
| InfobaeEEUU | 190 | aseitzwald | 9518 | **BernieSanders** | 15677 | FoxNewsInsider | 2225 | reidepstein | 3411 | SadHappyAmazing | 59 |
| 20m | 3 | ABCPolitics | 15979 | ZaidJilani | 12460 | DanScavino | 70756 | NoahCRothman | 11841 | HeymanHustle | 20 |
| diomedes66 | 4019 | daveweigel | 11148 | ggreenwald | 7478 | LindaSuhler | 109965 | baseballcrank | 12269 | LAClippers | 13 |
| supriyar054 | 397 | alexburnsNYT | 5367 | thehill | 96226 | DRUDGE_REPORT | 21263 | EWErickson | 15965 | RowanKavner | 53 |
| Pasion_Basket | 36 | jmartNYT | 4883 | the_intercept | 2269 | kausmickey | 7129 | guypbenson | 16447 | Eating | 201 |
| Fan_2_Football | 2 | danmericaCNN | 22617 | CatchEmAll | 10 | CLewandowski_ | 19894 | rickklein | 3017 | DeadlineDayLive | 68 |
| sakchipriya0002 | 334 | BenjySarlin | 7595 | ajplus | 2100 | wpjenna | 18098 | NumbersMuncher | 16131 | BroHumors | 25 |
| FuckTimeOnly | 1757 | **HillaryClinton** | 26325 | andrewperezdc | 914 | DonaldJTrumpJr | 11544 | **tedcruz** | 24542 | RUCKIN | 217 |
| rajnishraj00051 | 329 | alivitali | 6095 | Dena | 337 | Ricky_Vaughn99 | 57926 | BillKristol | 11612 | DooleyFunnyAf | 124 |
| farukijoya951 | 320 | sahilkapur | 12696 | Dory | 75 | JaredWyand | 47635 | benshapiro | 27694 | TalkSportJosh | 1265 |

**Table 11** Most central users based on page rank and their in-degree for French dataset

| Community 1 (Lambda) | | Community 2 | | Community 3 | | Community 4 | | Community 5 | | Community 6 | | Community 7 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Highest PR | In-degree | Highest PR | In-degree | Highest PR | In-degree | Highest PR | In-degree | Highest PR | In-degree | Highest PR | In-degree | Highest PR | In-degree |
| talk2meradi-ouk | 3 | najatvb | 258 | visactu | 25 | jeudybruno | 1343 | tprincedelam-our | 2081 | lecio13002 | 7301 | thabestvines | 4 |
| kartalyavuz43 | 2 | lelab_e1 | 4979 | equipenkm | 1647 | dr_davidson | 6 | avec_marine | 464 | nousles-jeunes95 | 3469 | tam1j | 6 |
| flvccooo | 5 | ellensalvi | 1700 | elabe_fr | 110 | florencedesr-uol | 1896 | f_philippot | 1010 | claudarnaud | 1928 | mouyoumed_ | 3 |
| alijeancorakci | 1 | matthieusuc | 108 | bluteaujm | 6 | spiceemedia | 11 | marion_m_le_pen | 634 | bleusudlor-raine | 9 | breflesgars | 2 |
| abdullah-ciftcib | 1 | alni92 | 5797 | lafrancequiose | 3 | fillon_78 | 1414 | fn_officiel | 629 | nicolassarkozy | 2599 | so_mkb | 1 |
| facts444 | 1 | le_gorafi | 902 | jnpaquet | 22 | votons2017 | 524 | aloisnavarro | 319 | audrey_crespo | 53 | nordeleouf | 11 |
| neeroziaa | 2 | fabricearfi | 1316 | gillesboyer | 223 | info_et_savoir | 28 | jeunedissident | 152 | marcelinof54 | 4711 | riafeee | 6 |
| lookmavideo | 7 | ornikkar | 1522 | j_vitbr | 1981 | monsieur-scotch | 5 | olivierlu-cas_35 | 134 | marcdhere | 2549 | thisiskiyemis | 72 |
| mustafarma-gan | 1 | marian-nelemag | 2043 | flefebvre_rf | 129 | ssoumier | 16 | bfmtv | 2449 | sarkoziste | 3252 | afpfr | 357 |
| trk_peants_ | 20 | mompontet | 723 | nk_m | 313 | fillon2017_fr | 449 | herstalle | 436 | verlainedjeni | 3078 | keblo92i | 7 |
| arnaud-baumela | 4 | lemonde_pol | 741 | bardessaoli | 59 | mano_seanis | 43 | gregoryroose | 213 | cd_land | 3596 | samygerie | 19 |
| que_rule | 3 | qofficiel | 678 | peddersophie | 5 | laurencepari-sot | 47 | betty_debeketch | 66 | ntwolfmother | 2502 | m_brantt95 | 9 |
| 9687_____ | 1 | nrenard75 | 1625 | larrain_fr | 156 | bettonkevin | 450 | lgatto4 | 13 | marsupi_l_ami | 2913 | un_idiot_dit | 4 |
| northafricans | 1 | hafidam57 | 1452 | lexpress | 507 | aurelien_veron | 10 | mlp_officiel | 412 | europe1 | 1402 | cyarnoz | 3 |
| caapulcukiz | 1 | nardinemo-rano | 1709 | ajpourlafrance | 54 | lordmaham-mer | 141 | nicolasbayfn | 472 | c8c57 | 3177 | zaksnow_ | 15 |

**Table 12** Network statistics of each community subgraph for US and French networks

| | Lambda | Clinton | Sanders | Trump | Cruz_Republicans | Others |
|---|---|---|---|---|---|---|
| mean degree | 243.7 | 641.2 | 357.8 | 2266.5 | 1492.7 | 24.3 |
| eigen_centrality | 11765.2 | 6941.9 | 6547.7 | 24471.2 | 9959.5 | 1260.0 |
| clustering coefficient | 0.8 | 0.4 | 0.5 | 0.5 | 0.5 | 0.3 |
| diameter | 6.0 | 9.0 | 11.0 | 8.0 | 8.0 | 18.0 |
| size | 717 | 2870 | 1820 | 2559 | 1529 | 505 |

| | Lambda | LeftWingMedias | NKM-Juppé | Fillon | FN | Sarkozy | Others |
|---|---|---|---|---|---|---|---|
| mean degree | 0.4 | 45.4 | 21.4 | 21.0 | 22.3 | 92.0 | 3.1 |
| eigen_centrality | 140.1 | 1946.6 | 1345.2 | 521.2 | 311.7 | 1646.5 | 25.0 |
| clustering coefficient | 0.1 | 0.1 | 0.1 | 0.2 | 0.2 | 0.2 | 0.1 |
| diameter | 4.0 | 11.0 | 12.0 | 11.0 | 12.0 | 7.0 | 8.0 |
| size | 1400 | 3916 | 478 | 717 | 1398 | 1717 | 374 |

**Table 13** All the hashtags for the US dataset. Color corresponds to the life cycles taxonomy

| Descending | Constant(many) | Constant(few) | Ascending |
|---|---|---|---|
| #GOPDebate | #Trump | #NeverTrump | #RNCinCLE |
| #DemDebate | #GOP | #Trump2016 | #SuperTuesday |
| #IowaCaucus | #1 | #FeelTheBern | #DemsInPhilly |
| #SCPrimary | #DNCinPHL | #TrumpTrain | #NYPrimary |
| #NHPrimary | #Cruz | #ImWithHer | #PrimaryDay |
| #GOPdebate | #Clinton | #MakeAmericaGreatAgain | #TrumpRally |
| #Rubio | #tcot | #MAGA | #IndianaPrimary |
| #iacaucus | #DonaldTrump | #DemTownHall | #Florida |
| #NewYorkValues | #trump | #NeverHillary | #INPrimary |
| #TedCruz | #Hillary | #CruzCrew | #Indiana |
| #WisconsinPrimary | #Election2016 | #BernieSanders | #CAPrimary |
| #NVcaucus | #BlackLivesMatter | #AlwaysTrump | #MDPrimary |
| #Sanders | #HillaryClinton | #TCOT | |
| #Bernie | #Benghazi | #AmericaFirst | |
| #IACaucus | #BREAKING | #p2 | |
| #CruzSexScandal | #Iowa | #trump2016 | |
| #Wisconsin | #GOPTownHall | #VoteTrump | |
| #FITN | #WIPrimary | #NeverCruz | |
| #TrumpPence16 | #TRUMP | #UniteBlue | |
| #2A | #CNN | #PJNET | |
| #Hillary2016 | #Obama | #TeamTrump | |
| #CrookedHillary | #Kasich | #ImWithYou | |
| #WakeUpAmerica | #DNCleak | #trumptrain | |
| #SouthCarolina | #ISIS | #1A | |
| #Hannity | #Ohio | #TRUMP2016 | |
| #NY | #USA | #LyinTed | |
| #NewYork | #ccot | #CCOT | |
| #OhioPrimary | #gop | #Cruz2016 | |
| #FloridaPrimary | #NRA | #VoteTrump2016 | |
| #FLPrimary | #MSM | #SECPrimary | |
| #AZPrimary | #NYC | #BernieOrBust | |
| #PAprimary | #America | #DumpRyan | |
| #NewYorkPrimary | #Arizona | #BuildTheWall | |
| #pjnet | #CA | #IN | |
| #PAPrimary | #Republican | #YUGE | |
| #DumpTrump | #Texas | #LyingTed | |
| #MarcoRubio | #Breaking | | |
| #PA | #Veterans | | |
| #OHPrimary | | | |
| #WI | | | |
| #Utah | | | |
| #maga | | | |
| #AZ | | | |
| #CRUZ | | | |
| #NYValues | | | |
| #UtahCaucus | | | |
| #California | | | |
| #CT | | | |
| #MD | | | |
| #ArizonaPrimary | | | |

**Table 14** All the hashtags for the French dataset. Color corresponds to the life cycles taxonomy

| Constant(almost all) | Ascending(monotonous) | Ascending (non-monotonous) | Constant(many) | | Descending |
|---|---|---|---|---|---|
| #sarkozy | #nkm | #toutpourlafrance | #fillon | #laprimaireledébat | #balkany |
| #juppé | #bygmalion | #lemissionpolitique | #ns2017 | #rtlmatin | #fillon2017 |
| #primaireledebat | #envoyespecial | #buisson | #primaire | #bourdin | #avecns |
| #hollande | #nszenith | #primairedroite | #primaire2016 | #attal | #sarkozycourttoujours |
| #lr | #poisson | #labaule2016 | #wauquiez | #justice | #saintetiennedurouvray |
| #lep | #jdd | #cahuzac | #morano | #poutine | #attentatnice |
| #sarko | #aj2017 | #migrants | #fn | #droite | #cnlr |
| #copé | #tapie | #nice06 | #lesrepublicains | #baroin | #fh2017 |
| #laprimaireledebat | #traduisonsles | #honte | #politique | #salamé | #ripoublicains |
| #juppe | #pujadas | #nicolassarkozy | #ns20h | #primairelr | #niceattentat |
| #bourdindirect | #ambitionintime | #cope | #valls | #ggrmc | #ue |
| #bayrou | #udi | #bismuth | #bfmtv | #itele | #doublepeine |
| #france | #croisonsles | #sarkozy2017 | #guaino | #chateaurenard | #europe |
| #gaulois | #primaireadroite | #e1sarkozy | #ciotti | #lesrépublicains | #euro2016 |
| #macron | #pécresse | #painauchocolat | #cdanslair | #médias | #arabiesaoudite |
| #lemaire | #punchline | #19hruthelkrief | #estrosi | #primaires2016 | #gueant |
| #primaires | #touquet | #bfmpolitique | #mariton | #taubira | |
| #chirac | #montebourg | #sa | #ploucs | #zemmour | |
| #rediff | #20h2017 | #présidentielle | #lepen | #2017 | |
| #primaireledébat | #lagarde | #woerth | #ns | #presidentielles2017 | |
| #afp | #pecresse | #gauche | #ps | #s | |
| #8h30aphatie | #alstom | #l | #nice | #nekkaz | |
| | #grandesvoix | #primaireslr | #brexit | #p | |
| | #le | #lmpt | #calais | #marinelepen | |
| | #jouyet | #policiers | #france2 | #rtl | |
| | #bfm | #rocard | #burkini | #lyon | |
| | #qatar | #referendum | #tf1 | #loitravail | |
| | #field | #guéant | #onpc | #franceinfo | |
| | #zenith | | #presidentielle2017 | #immigration | |
| | #labaule | | #kadhafi | #carlabruni | |
| | | | #trump | #fh | |
| | | | #sarkosy | #lci | |
| | | | #e1matin | #sar | |
| | | | #libye | #squarcini | |
| | | | #mlp | #franais | |
| | | | #blm | #sark | |
| | | | #quotidien | #paris | |
| | | | #cazeneuve | #mariagepourtous | |
| | | | #ump | #bfmstory | |
| | | | #terrorisme | #islam | |
| | | | #sondage | #syrie | |

**Table 15** All the hashtags for the French dataset classified according to our static taxonomy. The number after the hashtag is its ranked popularity

| community-speak | | Intermediate | | General | | | |
|---|---|---|---|---|---|---|---|
| #aj2017 | 78 | #alstom | 139 | #19hruthelkrief | 126 | #policiers | 161 |
| #arabiesaoudite | 174 | #attal | 91 | #2017 | 128 | #politique | 25 |
| #avecns | 48 | #attentatnice | 98 | #20h2017 | 121 | #poutine | 99 |
| #balkany | 32 | #bfmstory | 173 | #8h30aphatie | 113 | #présidentielle | 132 |
| #bismuth | 94 | #bourdin | 89 | #afp | 73 | #presidentielle2017 | 58 |
| #croisonsles | 101 | #buisson | 23 | #ambitionintime | 92 | #primaireadroite | 103 |
| #doublepeine | 151 | #bygmalion | 8 | #baroin | 105 | #primairedroite | 27 |
| #envoyespecial | 36 | #cahuzac | 65 | #bayrou | 28 | #primaireledebat | 6 |
| #fh2017 | 102 | #carlabruni | 153 | #bfm | 158 | #primairelr | 111 |
| #field | 164 | #cazeneuve | 75 | #bfmpolitique | 127 | #primaires | 60 |
| #fillon2017 | 35 | #ciotti | 38 | #bfmtv | 33 | #primaireslr | 149 |
| #guéant | 178 | #e1sarkozy | 115 | #blm | 71 | #ps | 49 |
| #gueant | 176 | #estrosi | 40 | #bourdindirect | 24 | #rediff | 63 |
| #jouyet | 146 | #euro2016 | 163 | #brexit | 51 | #rocard | 165 |
| #laprimaireledébat | 19 | #europe | 156 | #burkini | 54 | #rtl | 142 |
| #nkm | 4 | #fh | 154 | #calais | 52 | #s | 134 |
| #qatar | 159 | #fillon | 2 | #cdanslair | 39 | #sa | 130 |
| #ripoublicains | 109 | #fn | 21 | #chateaurenard | 117 | #saintetiennedurouvray | 90 |
| #sarkosy | 66 | #gauche | 137 | #chirac | 61 | #sark | 167 |
| #sarkozycourttoujours | 77 | #grandesvoix | 143 | #cnlr | 100 | #sarko | 13 |
| #squarcini | 162 | #honte | 84 | #cope | 93 | #sarkozy | 1 |
| #tapie | 81 | #immigration | 152 | #copé | 14 | #sondage | 83 |
| #toutpourlafrance | 3 | #islam | 175 | #droite | 104 | #taubira | 124 |
| #traduisonsles | 82 | #kadhafi | 62 | #e1matin | 68 | #terrorisme | 80 |
| #udi | 97 | #lagarde | 129 | #franais | 166 | #tf1 | 56 |
| #ue | 148 | #lemissionpolitique | 15 | #france | 29 | #valls | 30 |
| | | #libye | 69 | #france2 | 53 | | |
| | | #lmpt | 157 | #franceinfo | 150 | | |
| | | #loitravail | 147 | #gaulois | 31 | | |
| | | #mariagepourtous | 171 | #ggrmc | 112 | | |
| | | #marinelepen | 140 | #guaino | 34 | | |
| | | #migrants | 72 | #hollande | 7 | | |
| | | #mlp | 70 | #itele | 116 | | |
| | | #morano | 18 | #jdd | 59 | | |
| | | #nekkaz | 136 | #juppé | 20 | | |
| | | #nice06 | 76 | #juppe | 5 | | |
| | | #nicolassarkozy | 85 | #justice | 96 | | |
| | | #ns | 46 | #l | 141 | | |
| | | #ns2017 | 9 | #labaule | 170 | | |
| | | #ns20h | 26 | #labaule2016 | 43 | | |
| | | #nszenith | 37 | #laprimaireledebat | 86 | | |
| | | #poisson | 47 | #lci | 155 | | |
| | | #presidentielles2017 | 133 | #le | 144 | | |
| | | #primaire | 11 | #lemaire | 55 | | |
| | | #primaire2016 | 16 | #lep | 12 | | |
| | | #primaireledébat | 67 | #lepen | 45 | | |
| | | #primaires2016 | 123 | #lesrepublicains | 22 | | |
| | | #pujadas | 87 | #lesrépublicains | 120 | | |
| | | #punchline | 108 | #lr | 10 | | |
| | | #quotidien | 74 | #lyon | 145 | | |
| | | #referendum | 172 | #macron | 44 | | |
| | | #rtlmatin | 88 | #mariton | 41 | | |
| | | #salamé | 107 | #médias | 122 | | |
| | | #sar | 160 | #montebourg | 118 | | |
| | | #sarkozy2017 | 95 | #nice | 50 | | |
| | | #syrie | 177 | #niceattentat | 110 | | |
| | | #touquet | 114 | #onpc | 57 | | |
| | | #trump | 64 | #p | 138 | | |
| | | #ump | 79 | #painauchocolat | 119 | | |
| | | #wauquiez | 17 | #paris | 168 | | |
| | | #woerth | 135 | #pecresse | 131 | | |
| | | #zemmour | 125 | #pécresse | 106 | | |
| | | #zenith | 169 | #ploucs | 42 | | |

**Table 16** All the hashtags for the US dataset classified according to our static taxonomy. The number after the hashtag is its ranked popularity

| community-speak | | Intermediate | | General | |
|---|---|---|---|---|---|
| #AlwaysTrump | 14 | #1 | 43 | #America | 109 |
| #AmericaFirst | 65 | #1A | 33 | #Arizona | 124 |
| #BernieOrBust | 54 | #2A | 22 | #ArizonaPrimary | 134 |
| #BernieSanders | 34 | #AZ | 93 | #AZPrimary | 88 |
| #BuildTheWall | 112 | #Bernie | 46 | #Benghazi | 94 |
| #CCOT | 48 | #BlackLivesMatter | 105 | #Breaking | 136 |
| #Cruz2016 | 66 | #CA | 64 | #BREAKING | 127 |
| #CruzCrew | 3 | #ccot | 42 | #California | 102 |
| #CruzSexScandal | 26 | #Clinton | 41 | #CAPrimary | 72 |
| #DumpRyan | 97 | #CrookedHillary | 57 | #CNN | 77 |
| #FeelTheBern | 11 | #CRUZ | 108 | #DonaldTrump | 24 |
| #ImWithHer | 13 | #Cruz | 10 | #Election2016 | 110 |
| #ImWithYou | 111 | #CT | 95 | #FITN | 130 |
| #IN | 75 | #DemDebate | 47 | #Florida | 73 |
| #LyingTed | 85 | #DemsInPhilly | 56 | #FloridaPrimary | 107 |
| #LyinTed | 67 | #DemTownHall | 92 | #gop | 20 |
| #MAGA | 9 | #DNCinPHL | 101 | #GOP | 99 |
| #MakeAmericaGreatAgain | 7 | #DNCleak | 116 | #GOPdebate | 128 |
| #NeverCruz | 52 | #DumpTrump | 49 | #GOPDebate | 15 |
| #NeverHillary | 31 | #FLPrimary | 87 | #GOPTownHall | 120 |
| #p2 | 29 | #Hannity | 125 | #Hillary | 30 |
| #PJNET | 8 | #Hillary2016 | 58 | #IACaucus | 78 |
| #TCOT | 17 | #HillaryClinton | 45 | #iacaucus | 115 |
| #TeamTrump | 37 | #INPrimary | 59 | #Indiana | 82 |
| #TRUMP2016 | 71 | #maga | 114 | #IndianaPrimary | 81 |
| #trump2016 | 27 | #MarcoRubio | 119 | #Iowa | 74 |
| #Trump2016 | 2 | #MD | 96 | #IowaCaucus | 55 |
| #trumptrain | 76 | #NeverTrump | 6 | #ISIS | 135 |
| #TrumpTrain | 4 | #NRA | 84 | #Kasich | 69 |
| #VoteTrump | 18 | #NY | 53 | #MDPrimary | 104 |
| #VoteTrump2016 | 28 | #NYC | 122 | #MSM | 123 |
| #YUGE | 132 | #NYValues | 131 | #NewYork | 68 |
| | | #PA | 62 | #NewYorkPrimary | 117 |
| | | #pjnet | 40 | #NewYorkValues | 118 |
| | | #RNCinCLE | 44 | #NHPrimary | 63 |
| | | #Rubio | 38 | #NVcaucus | 91 |
| | | #Sanders | 70 | #NYPrimary | 25 |
| | | #SECPrimary | 79 | #Obama | 50 |
| | | #tcot | 5 | #Ohio | 80 |
| | | #TedCruz | 16 | #OhioPrimary | 83 |
| | | #Texas | 121 | #OHPrimary | 113 |
| | | #TRUMP | 12 | #PAPrimary | 90 |
| | | #TrumpPence16 | 51 | #PAprimary | 89 |
| | | #UniteBlue | 19 | #PrimaryDay | 100 |
| | | #USA | 61 | #Republican | 126 |
| | | #Utah | 106 | #SCPrimary | 21 |
| | | #WakeUpAmerica | 23 | #SouthCarolina | 103 |
| | | #WI | 86 | #SuperTuesday | 35 |
| | | #Wisconsin | 39 | #trump | 32 |
| | | | | #Trump | 1 |
| | | | | #TrumpRally | 98 |
| | | | | #UtahCaucus | 133 |
| | | | | #Veterans | 129 |
| | | | | #WIPrimary | 36 |
| | | | | #WisconsinPrimary | 60 |

## References

Aragón Pablo, Kappler Karolin Eva, Kaltenbrunner Andreas, Laniado David, Volkovich Yana (2013) Communication dynamics in Twitter during political campaigns: the case of the 2011 Spanish national election. Policy Internet 5(2):183–206

Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech Theory Exp 2008(10):P10008

Bryden J, Funk S, Jansen VAA (2013) Word usage mirrors community structure in the online social network Twitter. EPJ Data Sci 2(1):3

Conover M, Ratkiewicz J, Francisco MR, Gonçalves B, Menczer F, Flammini A (2011) Political polarization on Twitter. Proc Int Conf Web Soc Med (ICWSM) 133:89–96

Davidson I, Gourru A, Ravi SS (2018) The clustering description problem: complexity results, algorithms and applications. In NIPS 2018

Enli G (2017) Twitter as arena for the authentic outsider: exploring the social media campaigns of Trump and Clinton in the 2016 US presidential election. Eur J Commun 32(1):50–61

Fraisier O, Cabanac G, Pitarch Y, Besançon R, Boughanem M (2018) #lyse2017fr: The 2017 French presidential campaign on twitter. Stanford, California, pp 501–510

Fraley C, Raftery AE (2006) Mclust version 3: an r package for normal mixture modeling and model-based clustering. Technical report, Washington Univ. Seattle Dpt. of Statistics

Gaumont N, Panahi M, Chavalarias D (2018) Reconstruction of the socio-semantic dynamics of political activist Twitter networks—method and application to the 2017 French presidential election. PLoS One 13(9):e0201879

Giorgino T, Giorgino MT (2018) Package dtw

Girvan M, Newman MEJ (2002) Community structure in social and biological networks. Proc Natl Acad Sci 99(12):7821–7826

Herring SC, Paolillo JC, Ramos-Vielba I, Kouper I, Wright E, Stoerger S, Scheidt LA, Clark B (2007) Language networks on livejournal. In: 2007 40th Annual Hawaii International Conference on System Sciences (HICSS'07). IEEE, pp 79–79

Hong L, Convertino G, Chi EH (2011) Language matters in Twitter: a large scale study. In: 5th International AAAI conference on weblogs and social media

Johnson TJ, Perlmutter DD (2013) New media, campaigning and the 2008 Facebook election. Routledge, Abingdon

Kouloumpis E, Wilson T, Moore JD (2011) Twitter sentiment analysis: the good the bad and the omg! In: Proceedings of the international conference on web and the social media (ICWSM) 11(538–541):164

Kuo C-T, Ravi SS, Vrain C, Davidson I, et al (2018) Descriptive clustering: ILP and CP formulations with applications. In: IJCAI-ECAI 2018, the 27th International Joint Conference on Artificial Intelligence and the 23rd European Conference on Artificial Intelligence

Larsson AO, Moe H (2012) Studying political microblogging: Twitter users in the 2010 Swedish election campaign. New Med Soc 14(5):729–747

Ng, AY, Jordan MI, Weiss Y (2002) On spectral clustering: analysis and an algorithm. In: Advances in neural information processing systems, pp 849–856

Perrin A (2015) Social media usage. Pew Research Center, pp 52–68

Poblete B, Garcia R, Mendoza M, Jaimes A (2011) Do all birds tweet the same?: characterizing Twitter around the world. In: Proceedings of the 20th ACM international conference on Information and knowledge management. ACM, pp 1025–1030

Quraishi M, Fafalios P, Herder E (2018) Viewpoint discovery and understanding in social networks. In: Proceedings of the 10th ACM Conference on Web Science. ACM, pp 47–56

Romero DM, Meeder B, Kleinberg J (2011) Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on Twitter. In: Proceedings of the 20th international conference on World wide web. ACM, pp 695–704

Smith MA, Rainie L, Shneiderman B, Himelboim I (2014) Mapping Twitter topic networks: from polarized crowds to community clusters. Pew Research Center (https://www.pewinternet.org/2014/02/20/mapping-twitter-topic-networks-from-polarized-crowds-to-community-clusters)

Torgerson WS (1958) Theory and methods of scaling

Velcin J, Kim Y-M, Brun C, Dormagen J-Y, SanJuan E, Khouas L, Peradotto A, Bonnevay S, Roux C, Boyadjian J et al (2014) Investigating the image of entities in social media: Dataset design and first results. In: LREC, pp 818–822

Wang Y, Zheng B (2014) On macro and micro exploration of hashtag diffusion in Twitter. In: Proceedings of the IEEE/ACM international conference on advances in social networks analysis and mining. IEEE Press, pp 285–288

Williams C, Gulati G (2009) What is a social network worth? Facebook and vote share in the 2008 presidential primaries. In: Proceedings of the annual meeting of the American Political Science Association

Wong FMF, Tan CW, Sen S, Chiang M (2016) Quantifying political leaning from tweets, retweets, and retweeters. IEEE Trans Knowl Data Eng 28(8):2158–2172

Yang J, Leskovec J (2010) Modeling information diffusion in implicit networks. In: IEEE international conference on data mining. IEEE, pp 599–608

Yang Z, Algesheimer R, Tessone CJ (2016) A comparative analysis of community detection algorithms on artificial networks. Sci Rep 6:30750