Analyzing Polemics Evolution from Twitter Streams Using Author-Based Social Networks

Arnaud Quirin[1], Rocío Abascal-Mena[2], Florence Sèdes[3]

[1] Institut de Recherche en Informatique (IRIT),
Université Paul Sabatier de Toulouse,
France

[2] Universidad Autónoma Metropolitana - Cuajimalpa,
Mexico

[3] Institut de Recherche en Informatique (IRIT),
Université Paul Sabatier de Toulouse,
France

aquirin@gmail.com, mabascal@correo.cua.uam.mx, sedes@irit.fr

Abstract. The construction of social network graphs from online networks data has become nowadays a common track to analyze these data. Typical research questions in this domain are related to profile building, interest's recommendation, and trending topics prediction. However, few work has been devoted to the analysis of the evolution of very short and unpredictable events, called *polemics*. Also, experts do not use tools coming from social network graphs analysis and classical graph theory for this analysis. In this way, this article shows that such analysis lead to a colossal amount of data collected from public social sources like Twitter. The main problem is collecting enough evidences about a non-predictable event, which requires capturing a complete history before and during the course of this event, and processing them. To cope with this problem, while waiting for an event, we captured social data without filtering it, which required more than a TB of disk space. Then, we conduct a time-related social network analysis. The first one is dedicated to the study of the evolution of the actor interactions, using time-series built from a total of 33 graph theory metrics. A Big Data pipeline allows us to validate these techniques on a complex dataset of 284 millions of tweets, analyzing 56 days of the Volkswagen scandal [12].

Keywords. Author-based social networks, social network analysis, topic evolution, Twitter microblogging website.

1 Introduction

Nowadays, online social networks are a quite popular resource for connected people to express what is happening in the world and share their opinions about it. Defending specific opinions or positions often generates never-ending debates, sometimes attacks between disputants, especially for controversial topics. So, *polemics* arise naturally on large online media such as forums, social networks or microblogging websites. In our case, polemics is defined as the exchange occurring when different users speak about a specific and controversial topic in a short period of time. There are characterized by an aggressive attack on or refutation of the opinions or principles of another. There are also viewed as the art or practice of disputation or controversy [20]. During a polemic, the main theme of the discussion will not be reused over time. This is for us a key difference in comparison to a bursting or a trending topic, for which the same topic will attract new comments later, such as a famous actor or a large concert event.

Analyzing such social data becomes a crucial aspect. Social networks create new opportunities for companies to interact with their customers through online campaigns and mining these

Computación y Sistemas, Vol. 22, No. 1, 2018, pp. 35–45
ISSN 1405-5546
doi: 10.13053/CyS-22-1-2762

data is increasingly common to support digital marketing initiatives as well as a variety of business intelligence applications [18]. Social data can provide a sharp view into trends in user interests and behaviors, thus to guide governments and businesses to make "better" decisions.

Currently, social networks analysis techniques are viewed as a typical tool to get insights about social data. However, up to our knowledge, few attention has been dedicated to the applicability of this method to short-timed events, such as the polemics.

In the current contribution, we propose a technique based on classical Social Network Analysis (SNA), to analyze author interactions during the timeline of a polemic. Our objective is to answer the following research question: how can we detect and predict common behaviors between different users taking part into a polemic, even if they have a different vocabulary? In this paper, we will be particularly interested by the question of whether or not a predefined set of users or keywords should be followed to detect a polemic in the tweets.

For this purpose, we collected data from two real time Twitter streams, before, during and after a global and impacting event, the Volkswagen scandal occurred at the end of September 2015 [12]. We mine them to understand the events and then drivers which propagate the polemic, from its early starting point to its maximum peak and to its exhaustion. Due to the considerable amount of data, up to 1.5 TB, we relied on a Big Data pipeline to process them.

The structure of the current paper is as follows: in the next section, we review the current SNA techniques to mine a large amount of data, as well as the techniques used to study social behavior in Twitter. The third section details our main contribution, with the necessary steps to download, process and clean such amount of data. In the fourth section, we check which stream of data is most relevant to collect polemical tweets. The fifth section presents the Volkswagen case study. Finally, some concluding remarks and future perspectives are pointed out in the last section.

2 Related Works

Mining polemics in Twitter, sometimes also called *trending*, *hot* [11], or *bursting* [3], topics in the literature, is an increasing research question since few years. Typically, most propositions focus on the detection of events. Atypical behaviors such as natural disasters have been detected in Twitter using probabilistic models [28, 29, 33]. Di Eugenio et al. [7], employed Natural Language Processing (NLP), and classifiers to detect life events such as marriage, graduation, or birth in Twitter. Local events can also have been detected in real time by a clustering algorithm [3]. Guo et al., [11], proposed a frequent pattern recognition method to track trending topics in Twitter streams. Musto et al., [21], build *hate maps* based on a semantic analysis of Twitter streams to identify risk zones in Italy. In the later case, Big Data techniques have been used to filter and process large amounts of data.

However, we found few papers going through the analysis and the interpretation of Twitter streams covering short events, which still seems to be a fresh research topic. We could cite Lipizzi et al., [18], who analyzed the structure of conversations in Twitter following the launch of two commercial products. They only analyzed a three-day period but they show how *concept maps*, together with a time-slicing technique, help to study structural differences between the conversations. Wu et al., [36] proposed a propagation model to track popular news in Twitter. This model can be used to predict the final number of retweets of a piece of information. In conclusion, we found no paper analyzing short events with SNA techniques, which is an established tool when dealing with large amount of data. In this case, SNA techniques have emerged since a long time as tools in computational sociology to model and analyze an increasing amount of social phenomena. One of their main features is their ability to scale to very large and complex electronic datasets [5]. SNA applied to large networks can help to represent the data, measure local and global properties of the network and have effectiveness visualization techniques in order to analyze data. However, it seems that the interpretation of the evolution

of large-scale events through time is often easier with new visualization techniques. For instance, Dörk et al., [6] proposed Topic Streams, a kind of stacked area chart displaying the evolution of topics in Twitter. We propose, also, a valuable contribution in this domain with our social network based analysis.

3 Methodology

In this section, we describe our complete methodology from the data collection of the Twitter streams and their preprocessing using a Big Data pipeline to the construction of time-related social networks.

3.1 Background

Twitter is a microblogging website allowing users to share short messages, up to 140 characters, called tweets. The main characteristic of Twitter is this short length, forcing users to summarize their opinions in a quick and essential way. Tweets can wrap specific elements, namely the hashtags, which are words preceded by the symbol '#' (called a pad). With them, it is possible to link conversations to a common topic, and search and filter them. Social media engagement statistics for 2017 show a staggering of 319 millions of monthly active Twitter users worldwide from 1st quarter 2010 to 4th quarter 2016 [27]. Social media has played an increasingly important role in social participation by encouraging message exchange and converting Twitter into a large space of debate.

3.2 Data Collection

Data have been collected using two real-time, public and free streams of tweets provided by Twitter. The first one, called *statuses/sample* [32], is a never-ending stream collecting 1% of all tweets published globally. The second one, called *statuses/filter* [31], is a never-ending stream collecting only the tweets containing a given set of keywords specified by the user. Twitter will deliver all the tweets matching the criteria, providing that their amount never exceeds the 1% of the tweets published globally.

The advantage of the *filter* stream is that we can collect freely all the tweets relevant to a set of keywords: however, as this set has to be defined before connecting to the stream, it is not possible to predict which set of keywords is relevant before a polemic actually occurs.

All the data, for our research, was collected continuously starting from September 11, 2015. From this, 1563 GB of data have been collected (in this case, the sample stream represents a 73 % of the total). To ease the collection, storage and processing of this huge amount of data, an Apache Hadoop [1][26], pipeline has been used, based on Flume [2] to collect the tweets and the Hadoop Distributed File System (HDFS), to store them. This pipeline is run on two nodes, each equipped with 10 bi-core 1.6GHz AMD Opteron and 16 GB of memory. Sequential computation is performed on a cluster with 10 x 64-core 1.6GHz AMD Opteron CPUs with 512 GB of memory. Apache Hadoop and Flume are both free technologies from the Apache Software Foundation.

3.3 Selected Fields

Table 1 shows the relevant fields we have selected to perform our current study among the ones available from Twitter. Note that an additional field, namely `lang`, can be used to filter the tweets according to their language, but is computed automatically by Twitter, thus we did not use it here.

Table 1. List of relevant tweet fields

Field	Description
`text`	Text of the tweet
`created_at`	Publication date
`screen_name`	Author
`entitites.user_mentions`	List of user mentions

[1] The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. Available at: http://hadoop.apache.org/

[2] A distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of streaming event data. Available at: https://flume.apache.org/

3.4 Preprocessing

After the download of the tweets from their streams and the extraction of their relevant fields, we keep only the tweets containing a given keyword. For the sake of space, we study in this paper only the tweets containing volkswagen (independently of the case), for the sample stream. In the sec. 4.1 we show that taking the other stream would be statistically equivalent. We have filtered the tweets by following these steps:

1. Converting stopwords and punctuation to white spaces,
2. Eliminating any URLs (http, https, ftp, etc.),
3. Removing the mentions and the RT keyword,
4. Removing non-ASCII characters,
5. Converting the text in lowercase,
6. Tokenizing it, this means to convert the string to a list of tokens based on white spaces.

As an example, the preprocessing step converts the following tweet: “RT: Great! This is a Beautiful Day @harry! http://webpage.com/” to the following list of three tokens: great, beautiful, and day.

4 Determining the Right Stream of Tweets to Use

In this section, we analyze which stream of tweets is better to collect a relevant set of polemical tweets. In the first study, we compared the *sample* and the *filter* streams to determine if predefining a keyword before the data collection is relevant. In the second study, we demonstrate that working with a whole stream of data is more useful than with individual users. This is crucial to answer our research question, as anticipating which users will publish during a polemic regardless of the theme.

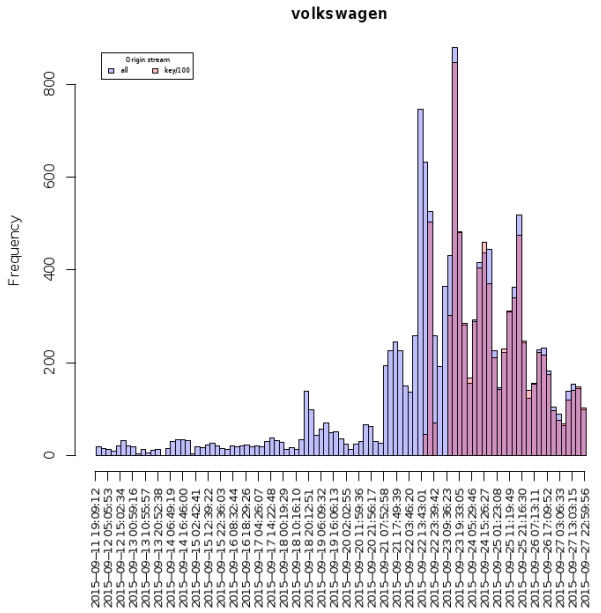

Fig. 1. Count of the number of collected tweets for the sample (in blue) and for the filtered (in red) streams, for different keywords: mecca, mecque (in French), refugiados (in Spanish), refugee (in English), and volkswagen. The count for the filtered stream has been divided by 100 to reflect the fact that only 1% of the sample stream is available

4.1 Comparison of Both Streams

In Fig. 1, we compare the raw number of tweets collected by the sample stream containing a given keyword, with the raw number of tweets collected by the filtered stream containing the same keyword. We selected four keywords in different language: mecca (in English), mecque (in French), refugiados (refugee, in English), and volkswagen. To reflect the fact that Twitter gives us only 1% of the sample stream publicly, we divided the number of filtered tweets by 100. On these datasets, it is clear to see that both streams of data are perfectly aligned, even though there are some little deviations due to the discretization.

This is an important assessment for us, because we can hypothesize both streams are consistent. If it would not be the case, probably the filtered stream would be a better source of data because it is not limited. However, due to the fact we can

only collect a filtered stream *after* deciding which keywords are relevant during the configuration of the Twitter API, it would be very hard to use this source to monitor the messages sent before or just after the origin of the polemic. Having consistent streams means that a passive collector of tweets can not only provide a reliable source of data to compute relevant statistical metrics (at least based on the frequency), but we can plug it in much before the origin of the polemic, avoiding any loss of data. From now on, we consider that any frequency-based statistics conducted on the filtered stream, corresponding to all the tweets containing a given keyword, is still valid (by a scalar factor), for the sample stream.

We can observe two quick surges in the case of Mecca. It corresponds to the horrified reaction of people after a crane collapsed killing 111 people [2], the 11th of September 2015, and the stampede incident, in the 2015 Hajj, killing at least 1470 people [10], the 24th of September 2015. In the case of the refugees, we spotted a more continuous stream of tweets, as the news stories were uninterrupted about this topic during all the month of September 2015.

In Twitter, from the Volkswagen scandal case, we found numerous surges corresponding to the never-ending bounces that this international event generated. They correspond to the initial announcement by the US Environmental Protection Agency (EPA), to recall a large amount of cars (18/9/2015, see the bottom right of Fig. 1), the announcement by Volkswagen that 11 millions diesel cars worldwide have the same ”defeat device” (22/9/2015), the resignation of Volkswagen CEO Martin Winterkorn (23/9/2015), and the nomination of the new Volkswagen CEO Matthias Muller (25/9/2015), [19].

4.2 Is it Necessary to Follow Single Users?

Currently, most studies based on Twitter consider single users as their units of analysis [35, 17, 15, 22, 13, 37, 16]. This allows to perform precise profiling on them, but this method might not be suitable to analyze unpredictable events. For instance, this method requires that a specific set of users to follow is predetermined before the event occurs, which is impossible to do in practice. Perhaps collecting a whole stream of data *anticipating* the event is better in this case. However, moving from a single-user analysis to a stream-analysis would mean shifting to a new paradigm of analysis. We study in this section its pertinence. To characterize the polemics and extract interesting features from them, we have studied if it is relevant to follow and access the tweets of single users, which would be possible using another part of the Twitter API. Or, instead, by taking the filtered stream to discriminate the tweets only by their keywords. Note that the download of the tweets from single users is a lengthy process, and Twitter limits its API to only the last 3200 tweets for each user [30].

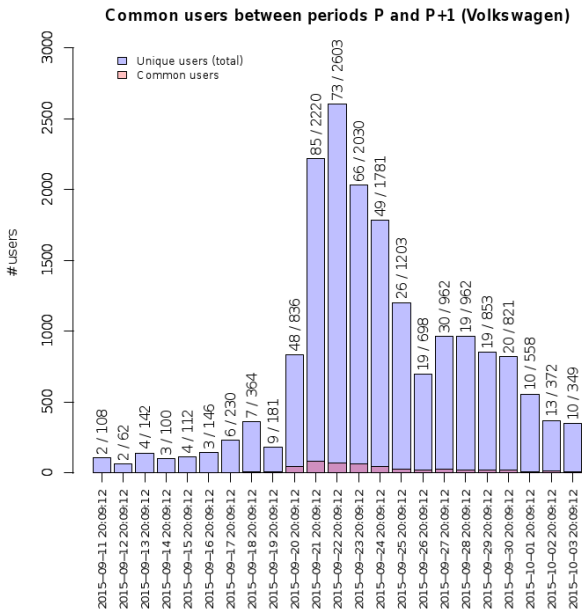

Fig. 2. For a given period P and for the volkswagen-filtered stream, number of unique users authoring at least one tweet inside P (in blue), and number of common users authoring at least one tweet both in period P and in the next one (in red). Exact values of both counts are shown in the top of each bar

So, we took a stream of tweets containing only a given keyword (volkswagen), and sliced it in several periods of 24 hours covering the duration of the polemic (see Fig. 2), so that no time-zone effect could have biased the results. For each of these

periods P, we computed the number of unique users who had published at least one tweet inside P (in blue) and the number of common users who had published at least one tweet both in period P and in the next one (in red). We observed that the number of common users are marginal compared to the total number of users involved in a reaction for this particular polemic (the maximum is at a 3.8%), so consequently it seems that very few users participate in two consecutive days about a particular event. This could be explained by the fact users, once their opinions are given (in Twitter at least, to specifically react to a new story, to post a caricature, etc), do not continuously feed the debate. We conclude that it is not useful to follow predefined set of users and collect their tweets over time.

We suggest, however, that it is still useful to follow the evolution of the polemics over time by observing short timed reactions of given users in the flow of the collected data, but that such information cannot be gathered easily by looking inside the whole content of these user timelines. In fact, it seems easier to detect these users from what they are publishing in the complete stream of tweets or by observing their behavior in the filtered stream. These *sentinels*, small group of users generating the first tweets just before the polemic is inflating on the online network, are then more easily spotted. This is why we decided to configure our stream collector architecture, presented in the previous sections of this article, to retrieve the keywords and not the tweets of the users.

5 Case Study: Evolution of Author-Based Social Networks in the Volkswagen Polemic

From this preprocessed data, we generate author-based social networks, in which the author of the tweet is represented for each node of the network. This is inspired from a previous work of Abascal-Mena et al. [1].

As we have a dataset for each day, we can generate as many author-based social networks and study the evolution of a polemic from two different points of view. Both constructions use their own way of selecting the nodes and the edges to be included in the networks, and their own visualization techniques, because of the type of information we want to emphasize.

5.1 Design of Author-Based Social Networks

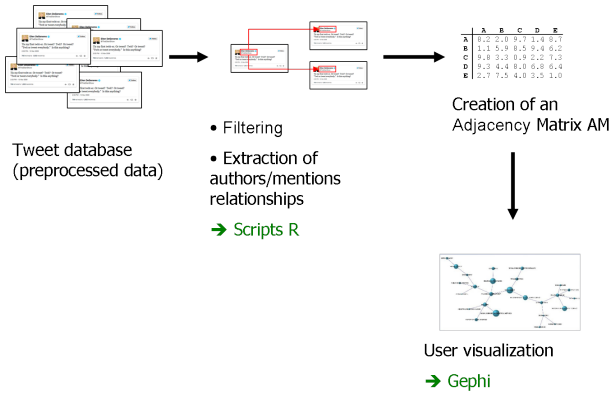

Fig. 3. Process of the design of author-based social networks

The generation of an author-based social network is based on a sequential series of steps (see Fig. 3). We selected the top-k most mentioned authors as the units of analysis and representation. Note that this would exclude *spammers* as they can follow a large number of users but are rarely mentioned. Note also that the networks do not have a constant size: some authors will inevitably appear or disappear during the course of the polemic depending on their behavior. This will allow us to uncover the variation of the growth of their relationships. As polemics and rumors are mainly propagated by individuals citing themselves, we computed an adjacency matrix AM for each pair of users i and j in which $AM(i,j) = 1$ if user i wrote a tweet mentioning user j. Note that AM is not a symmetric matrix.

The 21st of September, Volkswagen confirmed that it has ordered dealers to stop the sales of all four-cylinder diesel cars, after the surge of the polemic. In Fig. 4 and 5, we drawn two author-based social networks for this date and the day before. To build the set of nodes for each author-based social network, thus for each day,

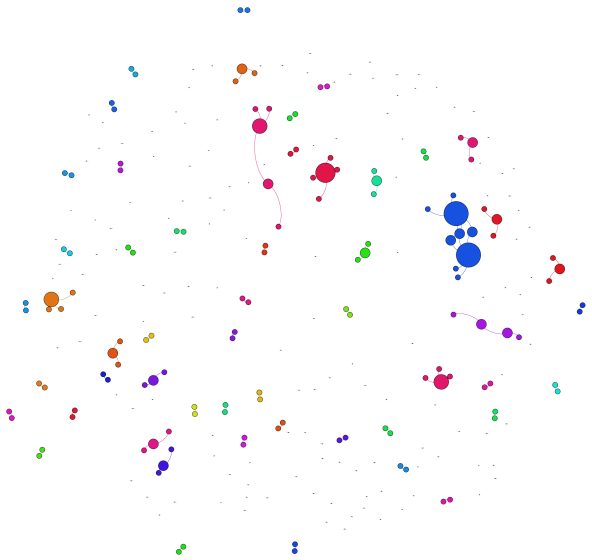

Fig. 4. Author-based social network the 20th of September

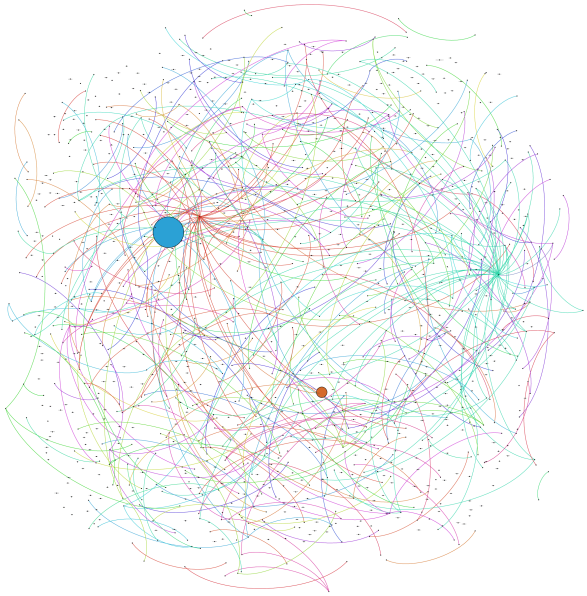

Fig. 5. Author-based social network the 21st of September

we used the 60% of the most mentioned authors. Note that the number of nodes in a network for a given day varies accordingly to the total number of mentioned authors for this day. We followed this methodology because we did not want to fix a given number of authors monitored for all the period, yet we wanted to observe more than the top-half of the mentioned authors.

Nodes are then colored according to their betweenness centrality [9], using the Gephi freeware. This software allows to quickly check and compare several layout algorithms. We chose the betweenness centrality over another metric because it exhibits some features highlighting important aspects of the polemic. Technically, it indicates if a node is located on many shortest paths between any pairs of nodes. In our case, this is translated to the fact that an author with a high betweenness value is a necessary intermediary in comparison to other authors, and thus is an interesting metric to consider. Note that no parameters are involved in this step, other than the choice of the top-k most mentioned authors, which make the generation of author-based social networks a truly automatic tool.

We can clearly see a set of sharp evolution between these two networks. The 20th of September, authors appear working independently, with authors citing themselves, mainly two by two but rarely exceeding five authors. The next day, a much broader scope of authors are replying to others, as can be seen by the numerous edges crossing the social network.

To obtain a better insight about this phenomenon, we computed specific network-based metrics to detect specific features, such as local maxima, and relate them to the original polemic. For instance, specific network-based metrics could incorporate a spam-detection algorithm which would combine classical spam detection by text analysis and scores computed from the friends of a given user. This kind of analysis would not have been possible without the use of graph-based representations. One of the best candidate we found that could be a good indicator to detect polemic in the real world was the number of communities, as we will see in the next section.

5.2 Evolution of Author-Based Social Networks

In order to get more insights from these authors-based social networks, one way to do it would be to convert them into time-series, enabling their analysis with more traditional tools such as signal processing, or *social signal processing* which is still an emerging topic [25, 34]. This way, further processing of the signals, to detect local maxima, duration of the maxima and the gaps, can easily be performed automatically. We are particularly interested of any metric highlighting the most important heartbeats of the polemic, which occurred the 22nd of September with the initial recognition by Volkswagen of the presence of defeat devices, and the 23rd of September with the resignation of CEO Martin Winterkorn.

We tried different network-related metrics, such as the average degree, the diameter, the edge density, as the main indicator to build the time-series. We show here the results we obtained with one of the metric best correlated with the evolution of the real events. In the figure 6, we present the results of the Walktrap algorithm [23]. We observe that this metric properly catches the two main events at the end of September, along with several replica. The most important replica, around the 28th of September, corresponds to the announcement of a refit plan from Volkswagen and the rumor that an alternative solution to not cheat would have added a cost of only 300 euros per vehicle [8]. The second replica, around the 6th of October, corresponds to the cancellation from

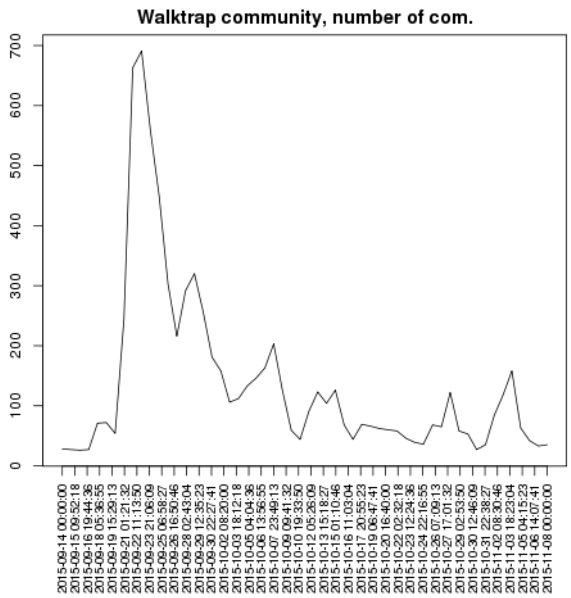

Fig. 6. Number of communities detected by the Walktrap algorithm during the Volkswagen polemic

'VW Group of America' of three Cars.com awards for TDI clean-diesel versions of VW vehicles. The 27th of October, a smallest replica corresponds to two main events: the Volkswagen CEO publicly apologizing at a Tokyo show [14] and Toyota becoming again the world's largest automaker [4]. It is worthy to mention that both words apologizes and Toyota can be found in the concept-based visual polemic maps of the 27th of October, for $k = 200$, not reproduced here.

6 Conclusion and Future Work

In this contribution, we proposed a time-related SNA technique and we show how it can be successfully applied to extract relevant information from Twitter streams. A large amount of data (284 millions of tweets, 1563 GB of data), have been collected to study the evolution of a polemic during a period of 56 days. For this purpose, we designed author-based social networks which allowed us to discover interesting features such as an explosion of the number of authors and interactions between them shortly after the apogee of the polemic. We

further used these networks as a base to build time-series using classical graph theory metrics. They clearly exhibit some recognizable patterns such as peaks when the press had something new to tell and users reacted to it. Finally, we shown that it is not useful to follow predefined set of users and collect their tweets over time.

Our methodology is scalable and did not suffer from massive load of data. It is worthy to note that the processing of the data is fully automatic and require very few parameters, which is simply the number of nodes (top-k most mentioned authors), we keep in the final representation. Many parts of the technique is generic (such as the co-citation metric we used) and can be customized for other use.

However, we are still in the edge of a world of research questions still unresolved. We plan to discover if our representation has any predictability power: is it possible to predict when a polemic will reach a peak, given the history of the time-series? This may depends on time-series of other events, as well as the weight of these events (importance). Could these values be combined with the other Twitter fields, such as the profile of the authors (number of published tweets, friends count, etc.) to improve the accuracy of this prediction? This could have a particular value for social marketing companies. How to filter or remove spammers? Probably, how they interact with their *friends* and the variance of the user profile creation dates could be a clue for this question. Monitoring in parallel other social networks is also an interesting topic. Finally, we plan to apply on the top of these networks Graph-Based Data Mining (GBDM), techniques [24], to get even more insights from this pile of data.

Acknowledgements

The author of the paper, Arnaud Quirin, is funded by the Xunta de Galicia, including funding from the operative program FSE Galicia 2007-2013, under the grant agreement IN809A 19/2015. Experiments presented in this paper were carried out using the OSIRIM platform that is administered by IRIT and supported by CNRS, the Region Midi-Pyrénées, the French Government, and ERDF (http://osirim.irit.fr/site/en). We thank also the responsibles of the Hadoop Cluster of the Blagnac IUT School to make their architecture available for this project. Finally, we thank our two reviewers for their interesting suggestions allowing us to improve the quality of this paper.

References

1. **Abascal-Mena, R., Lema, R., & Sèdes, F. (2014).** From tweet to graph: Social network analysis for semantic information extraction. *IEEE International Conference on Research Challenges in Information Science (RCIS 2014)*, Marrakesh (Morocco), pp. 1–10.
2. **Batrawy, A. (2015).** Pilgrims Traumatized, Asking How Mecca Crane Could Collapse. `https://tinyurl.com/kxbxqnv`. [Online; accessed 8/3/2017].
3. **Boettcher, A. & Lee, D. (2012).** Eventradar: A real-time local event detection scheme using twitter stream. *Green Computing and Communications (GreenCom), 2012 IEEE International Conference on*, IEEE, Besancon (France), pp. 358–367.
4. **Box, T. (2015).** Toyota once again world's largest automaker. `https://tinyurl.com/mj7rvw5`. [Online; accessed 8/3/2017].
5. **Butts, C. T. (2008).** Social network analysis: A methodological introduction. *Asian Journal of Social Psychology*, Vol. 11, No. 1, pp. 13–41.
6. **Dörk, M., Gruen, D., Williamson, C., & Carpendale, S. (2010).** A visual backchannel for large-scale events. *Visualization and Computer Graphics, IEEE Transactions on*, Vol. 16, No. 6, pp. 1129–1138.
7. **Eugenio, B. D., Green, N., & Subba, R. (2013).** Detecting life events in feeds from twitter. *2013 IEEE Seventh International Conference on Semantic Computing*, IEEE, Irvine (CA), pp. 274–277.
8. **Europe, A. N. (2015).** Bosch warned VW about illegal software use in diesel cars, report says. `https://tinyurl.com/nnylduq`. [Online; accessed 8/3/2017].
9. **Freeman, L. C. (1977).** A set of measures of centrality based on betweenness. *Sociometry*, Vol. 40, No. 1, pp. 35–41.
10. **Gambrell, J. (2015).** Saudi crush was deadliest hajj tragedy ever. `https://tinyurl.com/l43mgfu`. [Online; accessed 8/3/2017].

11. **Guo, J., Zhang, P., & Guo, L. (2012).** Mining hot topics from twitter streams. *Procedia Computer Science*, Vol. 9, pp. 2008–2011.
12. **Hotten, R. (2015).** Volkswagen: The scandal explained. http://www.bbc.com/news/business-34324772.
13. **Huang, X., Yang, Y., Hu, Y., Shen, F., & Shao, J. (2016).** Dynamic user attribute discovery on social media. *Asia-Pacific Web Conference*, Springer, pp. 256–267.
14. **Kageyama, Y. (2015).** Volkswagen chief executive apologizes for emissions scandal at Tokyo auto show. https://tinyurl.com/m4ghw4g. [Online; accessed 8/3/2017].
15. **Kim, E. D.-j., Keng, B. J.-l., & Padmanabhan, K. (2016).** Systems and methods for dynamically determining influencers in a social data network using weighted analysis. US Patent 9,262,537.
16. **Kooti, F., Moro, E., & Lerman, K. (2016).** Twitter session analytics: Profiling users' short-term behavioral changes. *International Conference on Social Informatics*, Springer, pp. 71–86.
17. **Lampos, V., Aletras, N., Geyti, J. K., Zou, B., & Cox, I. J. (2016).** Inferring the socioeconomic status of social media users based on behaviour and language. *European Conference on Information Retrieval*, Springer, pp. 689–695.
18. **Lipizzi, C., Iandoli, L., & Marquez, J. E. R. (2015).** Extracting and evaluating conversational patterns in social media: A socio-semantic analysis of customers' reactions to the launch of new products using twitter streams. *International Journal of Information Management*, Vol. 35, No. 4, pp. 490–503.
19. **Mays, K. (2015).** VW Diesel Crisis: Timeline of Events. https://tinyurl.com/lwg4ret. [Online; accessed 8/3/2017].
20. **Merriam-Webster (2017).** Merriam-Webster Dictionary. https://www.merriam-webster.com/dictionary/polemic/. [Online; accessed 15/3/2017].
21. **Musto, C., Semeraro, G., Lops, P., & de Gemmis, M. (2015).** Crowdpulse: A framework for real-time semantic analysis of social streams. *Information Systems*, Vol. 54, No. C, pp. 127–146.
22. **Piao, G. & Breslin, J. G. (2016).** Exploring dynamics and semantics of user interests for user modeling on twitter for link recommendations. *Proceedings of the 12th International Conference on Semantic Systems*, ACM, pp. 81–88.
23. **Pons, P. & Latapy, M. (2006).** Computing communities in large networks using random walks. *Journal of Graph Algorithms and Applications*, Vol. 10, No. 2, pp. 191–218.
24. **Quirin, A., Cordón, O., Vargas-Quesada, B., & de Moya-Anegón, F. (2010).** Graph-based data mining: A new tool for the analysis and comparison of scientific domains represented as scientograms. *Journal of Informetrics*, Vol. 4, No. 3, pp. 291–312.
25. **Shuman, D., Narang, S. K., Frossard, P., Ortega, A., & Vandergheynst, P. (2013).** The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *Signal Processing Magazine, IEEE*, Vol. 30, No. 3, pp. 83–98.
26. **Shvachko, K., Kuang, H., Radia, S., & Chansler, R. (2010).** The hadoop distributed file system. *Proceedings of the 26th IEEE Symposium on Mass Storage Systems and Technologies (MSST 2010)*, IEEE, Incline Village, NV, pp. 1–10.
27. **Statista (2017).** Social networks. https://www.statista.com/topics/1164/social-networks/. [Online; accessed 8/3/2017].
28. **Takahashi, B., Tandoc, E. C., & Carmichael, C. (2015).** Communicating on twitter during a disaster: An analysis of tweets during typhoon haiyan in the philippines. *Computers in Human Behavior*, Vol. 50, pp. 392–398.
29. **Takahashi, T., Tomioka, R., & Yamanishi, K. (2014).** Discovering emerging topics in social streams via link-anomaly detection. *Knowledge and Data Engineering, IEEE Transactions on*, Vol. 26, No. 1, pp. 120–130.
30. **Twitter (2015).** GET statuses/user_timeline. https://dev.twitter.com/rest/reference/get/statuses/user_timeline. [Online; accessed 8/3/2017].
31. **Twitter (2015).** Statuses/filter. https://dev.twitter.com/streaming/reference/post/statuses/filter. [Online; accessed 8/3/2017].
32. **Twitter (2015).** Statuses/sample. https://dev.twitter.com/streaming/reference/get/statuses/sample. [Online; accessed 8/3/2017].
33. **Vigliotti, M. G. & Hankin, C. (2015).** Discovery of anomalous behaviour in temporal networks. *Social Networks*, Vol. 41, pp. 18–25.

34. Vinciarelli, A., Pantic, M., & Bourlard, H. (2009). Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, Vol. 27, No. 12, pp. 1743–1759.

35. Wei, W., Joseph, K., Liu, H., & Carley, K. M. (2016). Exploring characteristics of suspended users and network stability on twitter. *Social Network Analysis and Mining*, Vol. 6, No. 1, pp. 51.

36. Wu, B. & Shen, H. (2015). Analyzing and predicting news popularity on twitter. *International Journal of Information Management*, Vol. 35, No. 6, pp. 702–711.

37. Zarrinkalam, F., Fani, H., Bagheri, E., & Kahani, M. (2016). Inferring implicit topical interests on twitter. *European Conference on Information Retrieval*, Springer, pp. 479–491.

Article received on 27/03/2017; accepted on 30/06/2017.
Corresponding author is Arnaud Quirin.