Social Networks Analysis: A Case Study on the Twitter Network

Cassio Melo
Universidade Federal de Pernambuco
Recife, Brazil
cam2@cin.ufpe.br

Yves Lechevallier
INRIA
Rocquencourt, France
Yves.Lechevallier@inria.fr

Marie-Aude Aufare
École Centrale de Paris
Paris, France
Marie-Aude.Aufaure@ecp.fr

ABSTRACT

Most of research in the domain of Social Network Analysis is conduced using a static version of the network. Evolutionary social networks, however, exhibits a number of properties that only possible to determine by looking to several snapshots network. These properties include node interactions through time, community formation, information dissemination, among others.

In this work we present a exploratory study on the Twitter network with emphasis on its dynamic aspects. We have defined operators to be applied in the evolving networks in order to get insights on questions like: How are the patterns of the information being published? Is there a correspondence between the community structure and the subject of the published topics? How does information propagates through the network?

Categories and Subject Descriptors

D.2.8 [**Database Management**]: Database Applications—*Data mining*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information Filtering*; J.4 [**Computer Applications**]: Social and Behavioral Sciences—*Economics*

General Terms

Algorithms, Experimentation

Keywords

Community, Evolution, Information Spreading

1. INTRODUCTION

Social Network Analysis (SNA) is a multidisciplinary research field that provide methods for understanding the structure and behavior of interconnected systems where entities are represented as nodes and their relationships represented as ties or edges. Many nature phenomena features a Social Network structure which explains why SNA has applications in several domains such as HIV-prevention strategies [1]; analyzing protein interaction networks [14], evaluating the effectiveness of organizational roles [11], patterns of learning in collaborative systems [13].

Most real networks typically contain parts in which the nodes (units) are more highly connected to each other than to the rest of the network. The sets of such nodes are usually called clusters, communities, cohesive groups or modules. Those networks are often dynamic[1] in the sense that nodes joins or leave communities, new interactions are created, information flows through connected peers, among many others events. Such behavior has been receiving some attention recently [4] [15] [6] [8].

In this work we present a exploratory study on the Twitter network with emphasis on its dynamic aspects. We would like to get insights about questions like: How are the patterns of the information being published? Is there a correspondence between the community structure and the subject of the published topics? How does information propagates through the network? To answer such questions we defined SNA operators and techniques described throughout this study.

The work is organized as follows: Section 2 briefly overviews the related work on the analysis of evolving communities and introduces the indexes for evolving communities used in the experiment; Section 3 presents a case study on the Twitter network; and Section 4 discusses the key findings of the experiment followed by the conclusions and future work.

2. ANALYSIS OF EVOLVING COMMUNITIES

Most of research in the domain of Social Network Analysis is conduced using a static version of the network. Evolutionary social networks, however, exhibits a number of properties that only possible to determine by looking to several snapshots network. The basic events that may occur in a network throughout the time are a node joins the network; a node links to someone; a node can join in a community; communities can form or dissipate, a community can grow or contract; groups may merge or split. All the events related to the behavior of the community dynamically over time characterizes what we call "Evolving Community". A

[1] We may refer to a changing network as "evolving", "dynamic" network.

clustering algorithm works by producing a sequence of clusterings, one for each snapshot and measuring for comparing those clusters [4]. There are basically two approaches of clustering: clustering of evolving communities and incremental clustering. The last is incrementally updated as new data arrive. In evolutionary clustering, however, the focus is upon optimizing a new quality measure which incorporates deviation from history.

Researchers are beginning to uncover the potential of the analysis of time-evolving networks recently. [4] proposes an evolutionary clustering framework using modifications of classic clustering algorithms to incorporate time features. In [15] authors define a generative model for network evolution which captures the power law distribution, network densification and shrinking diameter providing insights into the evolution of networks with both social and affiliation links.

In our experiment we used data fetched systematically during two weeks from the Twitter social network. In particular, we wanted to investigate the following questions:

- How is the network changing?
- Is there a correspondence between structural and functional communities?
- How is the information spreading through the network?
- Who are the most influential people of the network?

2.1 Indexes for Evolving Communities

We next define two indexes which provide us insights about the above mentioned questions. They are applied taking into account the sequence of snapshots of the network.

2.1.1 Influence Index

We define influence in this case as being the ability of a person propagate their information to other peers. It works works by relating the people who has published a tweet about something with the probability of their followers doing the same at subsequent time intervals (including or not the "ReTweets").

2.1.2 Geographical Connectivity Index

The geographical connectivity index is given by sum of all edges of the minimum spanning tree where nodes are geographically located points. This index reveals how connected is the network, with higher values of it meaning a sparse network, dense network otherwise.

3. ANALYSIS OF THE TWITTER NETWORK

Twitter is a social networking site focused on the publication of short messages that enables its users to send and read each others' updates, known as tweets. Tweets are text-based posts of up to 140 characters, displayed on the author's profile page and delivered to other users - known as followers - who have subscribed to them. The Twitter network has actually over 75 million users and 21 million daily visitors. We have chosen it as the sample for our experiment due to its intense activity.

3.1 Fetching and Preparing Data

The Twitter Application Programming Interface (API) was used in order to request trending topics, tweets data, and user information. The data collection was authorized by Twitter and we are able to make up to 20,000 requests per hour. We have crawled information using an ego-centric multi-focal approach - each node which was added to the sample was connected to at least other node. This avoids having single nodes, i.e., nodes not connected with anyone and, because its focal nodes were assigned randomly, we have a better picture of the whole network.

Concerning location fields, Twitter allows users to optionally tell their location from a simple text field. That leads to a very unstructured format containing sometimes invalid data (e.g. "Planet Earth"); location at different levels of granularity ("latitude and longitude" to "539 Bryant Street, San Francisco"); and multilingual text (e.g. "Brazil", "Brasil", "Brésil"). We used the Google Maps API for batch geocoding the location due to its capacity of translating text into coordinates. Users without an valid location data are automatically discarded by the API. We retrieved the geographical coordinates and as such, location is represented as a point on the map. In the future we plan to use Gmap generator and ArcGis in order to get a fashioned view of the geographical data, for example, using the shapefiles[2] of territories.

3.1.1 Data format

User Attributes

Each user entry is formatted as follows:

```
#ID LOCATION DATE_REGISTERED FOLLOWERS_COUNT
STATUSES_COUNT CURRENT_STATUS_TEXT
```

Followers Network

The network explicits the relationship between users in this case who "follows" whom, in other words, who is subscribed to receive status updates of whom. The network data is formatted as follows:

`FOLLOW_ID FOLLOWED_ID` (e.g. 637073 18685460)

3.2 Analysis of Twitter Followers Network

In the analysis of the followers network we sought to investigate the relation between the structural groups identified by the clustering algorithm and the social functional groups, i.e., people who shared similar interests. Our hypothesis is that if the nature of the content published by a group of people is similar among them and differs from others, this should reflect in a similar structure from the followers network point of view. That means that connected people are more susceptible to share the same interests.

While there are popular techniques for addressing the similarity of a given text, for example, Content Analysis [12], we opted to conduce a heuristic evaluation on a tag-like visualization of the content published for each group. The proce-

[2] Shapefiles spatially describe geometries for geographic information systems

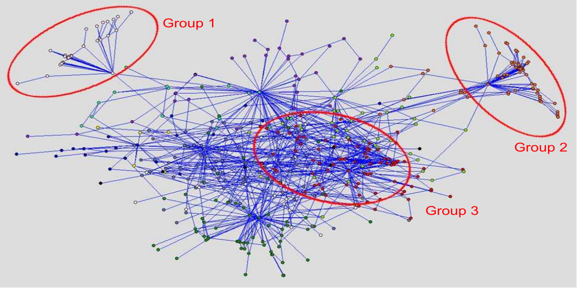


Figure 1: Three groups were clearly identified by the algorithm

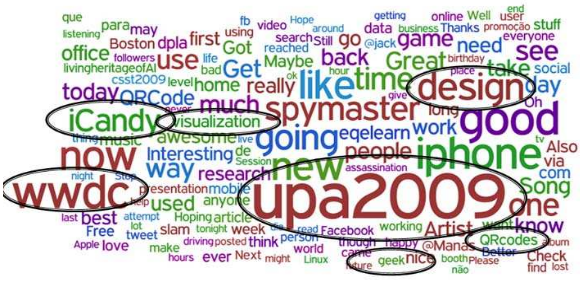


Figure 2: Group 1 tweets

dure of this experiment is divided in three steps: the community identification, the collection of community tweets and the analysis of their relation.

3.2.1 Community Identification

A community detection algorithm was applied in the Twitter network dataset in order to identify densely connected groups. In our experiment, we have applied a centrality algorithm based on edge betweeness [9].It measures how many times a node n occurs in a shortest path between any other 2 nodes in the graph and removes edges with lower value. This algorithm yields good results in a relatively small amount of time. Figure 1 show the groups identified in a subset of the network containing 546 nodes. We have highlighted tree of the groups identified for the tweets collection explained in the next section.

3.2.2 Community Tweets

Community tweets are the set of the last 20 tweets published by a person in a previously identified community. We collected those tweets, eliminated the stop-words from them and created a tag-like visualization using the Wordle[3] service. This visualization displays words from those group tweets relatively to their occurrence (figures 2-4).

3.2.3 Analysis between Published Content and the Followers Network

While there are many words not so “representative” of a group (it may be because people use more twitter for telling about the day-to-day things in their lives, for example “I need to rest” or “I feel free”); Yet, a couple of words are really discriminant of what sort of group it refers to. For example, group 1 should be a functional design group because the frequency of tweets related to “iCandy”, “design”, “upa2009”, “wwdc” (both popular design conferences)(figure 2). Group 2 has “marketing”, “business”, “credit” as its most discriminant words (figure 3). Surprisingly, group 3 is a german-speakers group and almost all the words are discriminant (figure 4). As one can see, the clusters identified by the algorithm have a correspondence between the functional groups of people, who share the same interests. We are not stating however, that those groups offers a clear definition of people’s interests. People activities and interests are diverse and possible overlaps a single group definition.

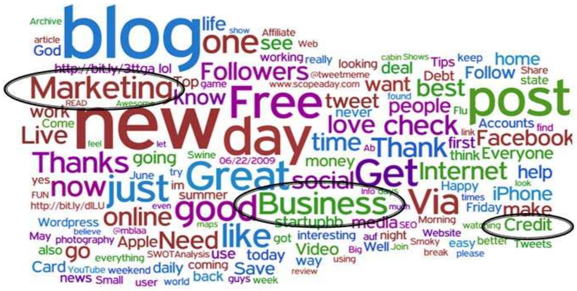


Figure 3: Group 2 tweets

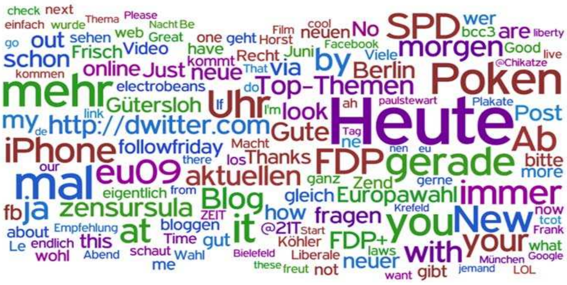


Figure 4: Group 3 tweets

A drawback with this approach is that an individual may be publishing many tweets about the same subject and therefore biasing the visualization and analysis. A possible solution for this is the size of an word in the visualization not be relative to word occurrence, but to the number of people who have published tweets containing that word. The `tf/idf` measure [10] can also be used to discriminate groups, considering groups as documents.

3.3 Analysis of Twitter Trending Topics

He have systematically collected all tweets being published about four important concurrent events at the time, being: “Air france” (the Air France plane crash in June 2009), “Élections Européennes” , “Grippe Porcine” (swine flu) and “Roland Garros”. The data was mined at regular intervals of thirty minutes over two weeks starting June 1st, 2009. The analysis of trending topics on Twitter investigated 1) the evolution of those topics related to real-time events and 2) the geographical dissemination of those topics.

3.3.1 Topics Publishing Over Time

[3] www.wordle.com

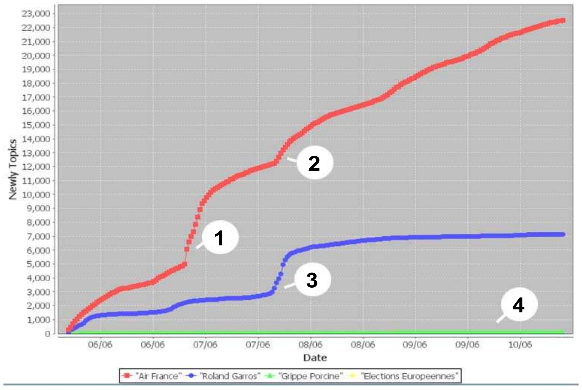


Figure 5: “Air france”, “Élections Européennes” , “Grippe Porcine” and “Roland Garros” published topics over a week.

We graphically represented the accumulative number of topics on a plotted graph as shown in figure 5. Each color represents a topic: red - “Air france”; blue - “Roland Garros”; green - “Grippe Porcine” and yellow - “Élections Européennes” (both at the bottom).

One can notice sudden increases on the number of topics followed by a stabilization over time. Those increases reveal facts within the events that attracted the peoples’ attention and thereafter became popular. The related major facts concerning “Air france” resulting in a increasing of tweets were: (06/06 20:00 CEST) The first bodies of the Air France plane crash were found in Atlantic (figure 5-1); (07/06 16:00 CEST) More bodies were found, debris photos published (figure 5-2). (07/06) Roland Garros finals (figure 5-3). It is also noteworthy that those tweets are related to announcements and reveal the underlying herd behavior of the crowds.

3.3.2 Geographical Analysis of Trending Topics

We have analyzed the spreading of tweets geographically using the trending topics collected previously (figures 6-9). We used the same operator described in section 2.1.2 - which is the sum of the MST in the first snapshot over the MST in the last one - to illustrate how connected is a trending topic. In other words, if a trending topic becomes widely spread geographically it has a lower value of the geographical connectivity and vice-versa.

The “Air France” trending topic occurred most all over the world while the “Roland Garros” trending topic was mostly concentrated in France, as the tennis event is hosted there. What is important to notice here is the spreading of the trending topics. As expected, in general a trending topic starts in large cities and then progresses to less demographically dense areas in both cases.

3.3.3 Most Influential People on Twitter

We have applied the operator described in section 2.1.1 to rank the top 10 influential people on Twitter. The operator works by relating the people who has published a tweet about something with the probability of their followers doing the same at subsequent time intervals. Using the “Air

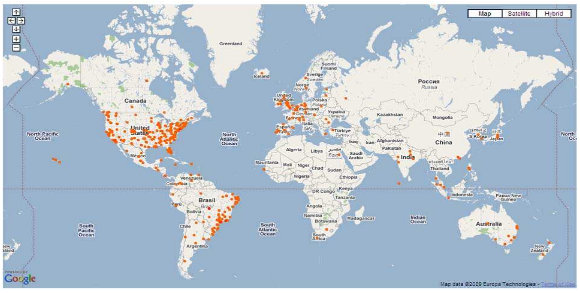


Figure 6: A world map displaying the published information in June 1st about *“Air france”*.

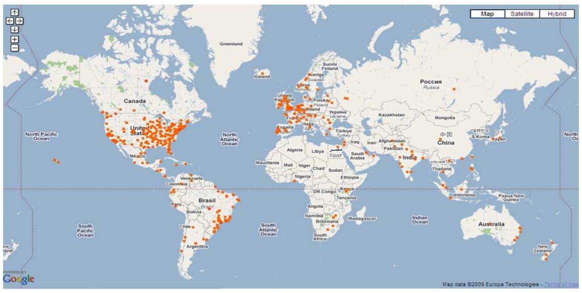


Figure 7: A world map displaying the published information about *“Air france”* a week later.

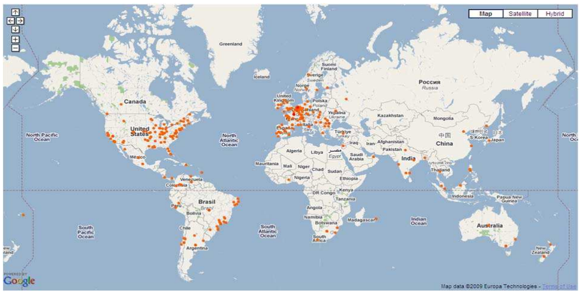


Figure 8: A world map displaying the published information in June 1st about *“Roland Garros”*.

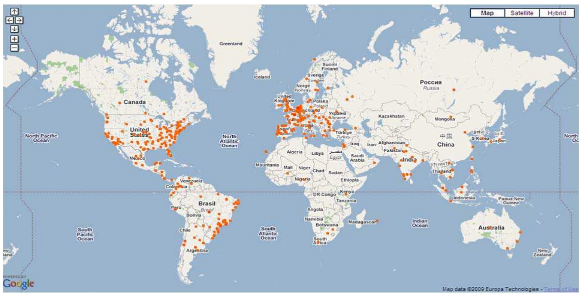


Figure 9: A world map displaying the published information about *“Roland Garros”* a week later.

Table 1: 10 most influential people on “Air France” on Twitter.

User	Followers count	Tweets count	Location
metheoro	792	14,647	SAO PAULO, BR
AHN Breaking News	529	32,482	??
Eduardo Santos	444	226	??, BR
Telegraph World News	371	3,909	LONDON, UK
Examiner Weather	61	152	DENVER, USA
Jorge Salgado	2,106	1,061	LISBON, PT
notivagos	1,584	106,936	PELOTAS, BR
Devesh Agarwal	219	1,317	MUMBAI, IN
GMTV news editor	323	836	LONDON, UK
El Nuevo Diario	227	7,289	MANAGUA, NIC

France” topic on Twitter we have got the results as shown in Table 1.

One important thing to notice is that the operator is robust enough to deal with variations of either followers and tweets count. It is expected that people with a large number of followers to have a greater influence index. As shown in Table 1., it’s not the case since the “Examine Weather” for example, has only 61 followers and 150 tweets published so far and it’s placed as the fifth most influential entity on Twitter according to our operator. It roughly means that when this entity publishes a new Tweet their followers, though few, tend to publish something related. Of course, for making such assumption it’s necessary a longitudinal approach considering many key-words to be assessed by the operator, which is left as an exercise for the reader.

By looking at the name of the entities on this list we have found that many of them are representative of media publishers (in bold). It makes sense, since those entities are often associated to reliable sources of information and therefore people trust and comment. This reinforces the accuracy of the influence operator.

4. CONCLUSIONS AND FUTURE WORK

In this exploratory study we have conduced some experiments on the Twitter network taking into account its evolutionary properties. The experiment has demonstrated aspects of how people receives and propagates information. The use of operators can reveal implicit relations on the evolving social networks, for example, showing us that there are some people that are more influential than others and most of them are trustworthy media representatives. We deduced that in general there’s a correspondence between the structural groups on Twitter and the content of their posts and people publish more tweets according to events occurring in real-time.

In spite of the insights resulted from this experiment, we are aware of its limitations. The size of the sample and the results must be validated in a formal way. Apart the “Air France” and “Roland Garros” which are internationally known names, the trending topics analyzed were in French language, thus it is expected to have more tweets coming from francophone countries.

We plan to surpass those limitations in the next study and extend the analysis of groups evolution over time. We intend to see patterns of interaction intra and extra groups as well as the movement of an individual in a network, like for example, how a newcomer is being integrated with a group in time and how do people plays different roles in the network at different times. Another interesting research subject is studying how groups propagates information. Some groups can spread information quickly and broadly when this information is of a common interest.

5. ACKNOWLEDGMENTS

The paper is based on work funded by Égide - file number 658322J, AxIS, and the Institut National de Recherche en Informatique et Automatique (INRIA).

6. REFERENCES

[1] Amirkhanian, Kelly, Kabakchieva, McAuliffe, and Vassileva. Evaluation of a social network hiv prevention intervention program for young men who have sex with men in russia and bulgaria. *AIDS Education and Prevention*, 15(3):205–220, November 2003.

[2] C. Butts. Carter’s archive of s routines for the r statistical computing environment. http://erzuli.ss.uci.edu/R.stuff/, Apr. 2010.

[3] C. T. Butts. Social network analysis with sna. *Journal of Statistical Software*, 24(6):1–51, May 2008.

[4] D. Chakrabarti, R. Kumar, and A. Tomkins. Evolutionary clustering. In *KDD ’06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 554–560, New York, NY, USA, 2006. ACM.

[5] W. de Nooy, A. Mrvar, and V. Batagelj. *Exploratory Social Network Analysis with Pajek (Structural Analysis in the Social Sciences)*. Cambridge University Press, January 2005.

[6] P. Holme, C. R. Edling, and F. Liljeros. Structure and time-evolution of an internet dating community. *Social Networks*, 26(2):155–174, May 2004.

[7] M. Huisman and M. A. van Duijn. Software for social network analysis. In P. J. Carrington, J. Scott, and S. Wasserman, editors, *Models and Methods in Social Network Analysis*, pages 270–316. Cambridge University Press, 2005.

[8] P. C. H. Ma, K. C. C. Chan, X. Yao, and D. K. Y. Chiu. An evolutionary clustering algorithm for gene expression microarray data analysis. *IEEE Transactions on Evolutionary Computation*,

10:296–314, 2006.

[9] M. E. J. Newman. Detecting community structure in networks. *The European Physical Journal B-Condensed Matter*, Jan. 2004.

[10] S. Robertson. Understanding inverse document frequency: On theoretical arguments for idf. *Journal of Documentation*, 2004.

[11] A. V. Shipilov. Firm scope experience, historic multimarket contact with partners, centrality, and the relationship between structural holes and performance. *Organization Science*, 20(1):85–106, 2009.

[12] P. J. Stone, D. C. Dunphy, M. S. Smith, and D. M. Ogilvie. *The General Inquirer: A Computer Approach to Content Analysis.* MIT Press, Cambridge, MA, 1966.

[13] Y. Wang and K. Li. An application of social network analysis in evaluation of cscl. In *Proceeding of the 2006 conference on Learning by Effective Utilization of Technologies: Facilitating Intercultural Understanding*, pages 353–356, Amsterdam, The Netherlands, The Netherlands, 2006. IOS Press.

[14] Q. Yang and S. Lonardi. A parallel edge betweenness clustering tool for protein interaction networks. *Int. J. Data Min. Bioinformatics*, 1(3):241–247, 2007.

[15] E. Zheleva, H. Sharara, and L. Getoor. Co-evolution of social and affiliation networks. In *15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, June 2009.

APPENDIX

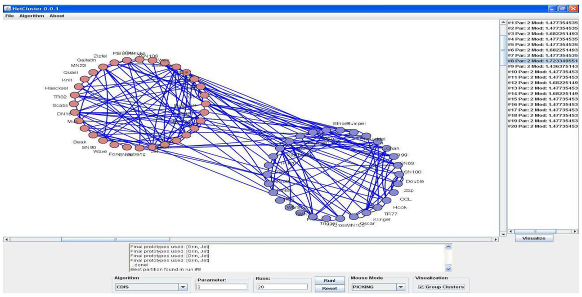
A. SOCIAL NETWORKS ANALYSIS SOFTWARE

There is a number of SNA tools and programming packages, from generalists such as Matlab to specialized SNA applications like UCINET. There are two basic types of software for SNA: software application (e.g. Pajek) and programming language with SNA package (e.g. R and SNA package). In the following we briefly introduces the most popular SNA tools reviewed during our experiment. For a systematic review of software packages for social network analysis see [7].

Pajek. Pajek[5] is a popular SNA software that provides analysis tools for networks and visualization capabilities. It provides many of the usual centrality measures, clustering algorithms and visualizations. It can output the results to R to calculate additional statistics.

UCINET. UCINET offers a comprehensive package of SNA methods including centrality measures, subgroup identification, role analysis, elementary graph theory, and permutation-based statistical analysis.

R and SNA package. The R is a general-purpose statistical computing environment. There is a collection of R routines for social network analysis called SNA package: utilities included range from hierarchical Bayesian modeling of informant accuracy to logistic network regression. Quite a few low-level utilities for plotting and transforming networks are available as well, along with many of the usual centrality and distance measures. There is a gentle introduction by Butts [3]. For a comprehensive Social Network Analysis routines for R, see [2].



#1 Par: 2 Mod: 1.477354535
#2 Par: 2 Mod: 1.477354535
#3 Par: 2 Mod: 1.682251493
#4 Par: 2 Mod: 1.477354535
#5 Par: 2 Mod: 1.477354535
#6 Par: 2 Mod: 1.682251493
#7 Par: 2 Mod: 1.477354535
#8 Par: 2 Mod: 1.723349551
#9 Par: 2 Mod: 1.436375143
#10 Par: 2 Mod: 1.47735453
#11 Par: 2 Mod: 1.47735453
#12 Par: 2 Mod: 1.68225149
#13 Par: 2 Mod: 1.47735453
#14 Par: 2 Mod: 1.68225149
#15 Par: 2 Mod: 1.47735453
#16 Par: 2 Mod: 1.47735453
#17 Par: 2 Mod: 1.47735453
#18 Par: 2 Mod: 1.47735453
#19 Par: 2 Mod: 1.47735453
#20 Par: 2 Mod: 1.47735453

```
Final prototypes used: [Grin, Jet]
Final prototypes used: [Grin, Jet]
Final prototypes used: [Grin, Jet]
Final prototypes used: [Grin, Jet]
...done!
Best partition found in run #8
```

Algorithm: CDIS
Parameter: 2
Runs: 20
Mouse Mode: PICKING
Visualization: ☑ Group Clusters

Figure 10: NetCluster interface.

Java Universal Graph Framework (JUNG). JUNG is a Java API for the modeling, analysis, and visualization of relational data. It enables customizable graph, properties, algorithms, visualization and comprises a range o social network analysis methods (e.g., clustering, decomposition, optimization, random graph generation, statistical analysis, distances, flows, and centrality measures.

B. NETCLUSTER

NetCluster is a software and library build upon JUNG for Social Network Analysis created at the Institut National de Recherche en Informatique et Automatique - INRIA. In its alpha version it provides the basic methods for fetching, preparing, clustering and analyzing data from online Social Networks. Among other utilities for social network analysis, NetCluster has a graphical interface for evaluating clustering algorithms on social networks (figure 10). It allows you for example, to see how different algorithms behaves when applied to a network, compare and analyze results.

C. ANALYSIS WORKFLOW

The workflow of the analysis conduced in this work is shown in figure 11.

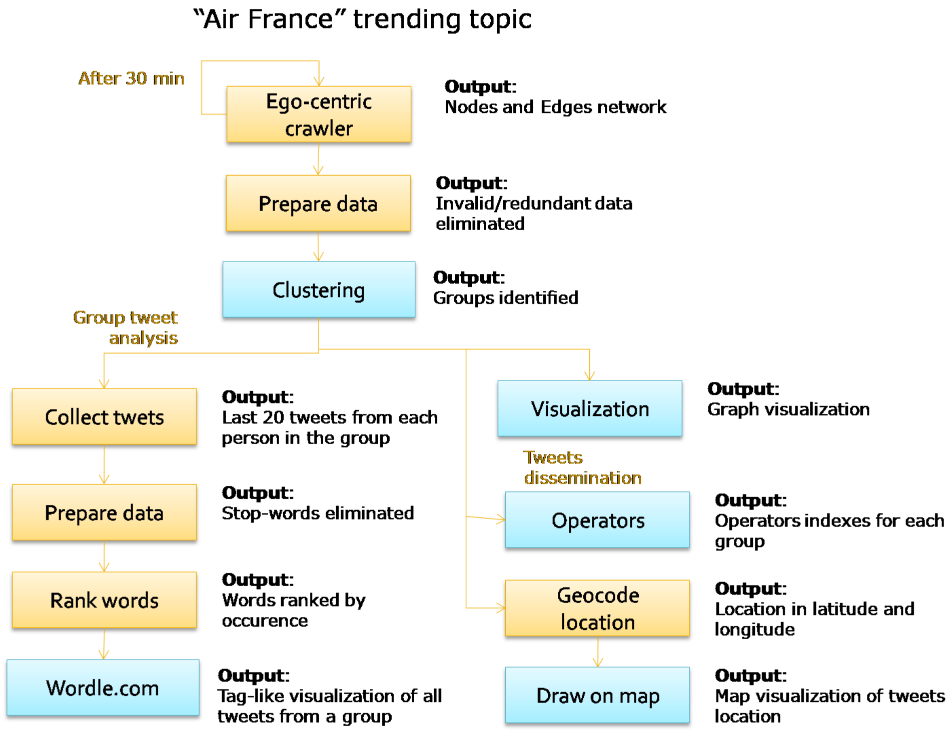

"Air France" trending topic
After 30 min
Ego-centric crawler
Output: Nodes and Edges network
Prepare data
Output: Invalid/redundant data eliminated
Clustering
Output: Groups identified
Group tweet analysis
Collect twets
Output: Last 20 tweets from each person in the group
Prepare data
Output: Stop-words eliminated
Rank words
Output: Words ranked by occurence
Wordle.com
Output: Tag-like visualization of all tweets from a group
Visualization
Output: Graph visualization
Tweets dissemination
Operators
Output: Operators indexes for each group
Geocode location
Output: Location in latitude and longitude
Draw on map
Output: Map visualization of tweets location

Figure 11: The experiment workflow