Semester Projet
Detection of fake news in sites and online networks

Maelle LE CLAINCHE

Fall Semester 2018

1 Introduction

This last years rumors, fake-news and other wrongful accusations have been spreading over medias, and especially web medias. According to academic studies at Princeton, Dartmouth and the University of Exeter, "about 25 percent of Americans visited a fake news website in a six-week period around the time of the 2016 US election" [1] : this number shows how important this phenomenon is. Towards the rise of misinformation and political manipulation, checking the veracity of news is becoming a serious issue for our society.

During this study we will examine the news website environment and try to detect fake information using data science tools such as graph visualization and label propagation.

2 Dataset

The dataset used for this study is GDELT 2.0, available on the GDELT project blog [2]. This dataset is a "real time translation of world's news in 65 languages". It provides every 15 minutes a set of themes and events in a given time span. For this study, we would only focus on the "Mentions Table" which is a list of mentioned events in website articles. This table can be presented as a data frame (Type *DataFrame* in Python's pandas library) with columns as shown in figure 1.

	GlobalEventID	EventTimeDate	MentionTimeDate	MentionType	MentionSourceName	MentionIdentifier	SentenceID	Actor1CharOffset	Actor2CharOffset	ActionCharOffset	InRawText	Confidence	MentionDocLen	MentionDocTone	MentionDocTranslationInfo	Extras
0	667330433.0	2.017062e+13	2.018062e+13	1.0	9news.com.au	https://www.9news.com.au/national/2018/06/22/1...	4.0	-1.0	2220.0	2249.0	1.0	80.0	3233.0	4.063604	NaN	NaN
1	667332152.0	2.017062e+13	2.018062e+13	1.0	lasalle.edu	https://www.lasalle.edu/blog/2018/06/22/la-sal...	5.0	-1.0	462.0	477.0	1.0	100.0	1273.0	5.309735	NaN	NaN
2	766498492.0	2.018062e+13	2.018062e+13	1.0	irishexaminer.com	https://www.irishexaminer.com/breakingnews/vie...	25.0	-1.0	7867.0	7859.0	0.0	20.0	15858.0	-5.619266	NaN	NaN
3	766498493.0	2.018062e+13	2.018062e+13	1.0	centralmaine.com	https://www.centralmaine.com/2018/06/22/usda-t...	5.0	-1.0	2209.0	2193.0	0.0	40.0	4916.0	0.000000	NaN	NaN
4	766498494.0	2.018062e+13	2.018062e+13	1.0	chronicle.co.zw	http://www.chronicle.co.zw/ed-recounts-poisoni...	5.0	1078.0	-1.0	1101.0	0.0	20.0	3716.0	-0.768049	NaN	NaN

FIGURE 1: Example of DataFrame view from Mention table of GDELT 2.0

In this DataFrame we will exclusively use :
— GlobalEventID is the ID of an unique event
— EventTimeDate is the date of the unique event
— MentionTimeDate is the date where article has been published
— MentionType is the type of news : it will always be 1 (website)
— MentionSourceName is the news website URL
— MentionIdentifier is the URL of the article

1. https ://www.bbc.com/news/blogs-trending-42724320
2. https ://blog.gdeltproject.org/gdelt-2-0-our-global-world-in-realtime/

The other columns are linked to the content of the article : the sentence where the event is mentioned (SentenceID), the location of actors in the article (Actor1CharOffset and Actor2CharOffset), the location of the core action description (ActionCharOffset), and other indicators linked to language translation or tone. We did not used those indicators because the way they are obtained is not given so this is a black box. Furthermore, the way they link articles together (by unique GlobalEventID) is also not given and will be put into question during this study.

3 A very first graph

First of all, to understand the dataset, we used the "best partition" tool, that will highlight clusters in our graph. The graph used for this, is a very simple one : each node represents a website, and there is a link between two nodes with a weight w if the two websites share w events. Then, to visualize the results we used is *force one layout* that provides a graph where two distant points are loosely linked. The results of this partition for 3 different days are shown figures 4, 2 and 3.

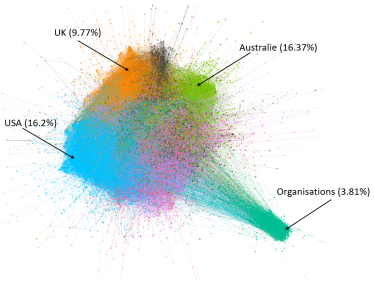

FIGURE 2: Result of "best partition" for 2018-08-30

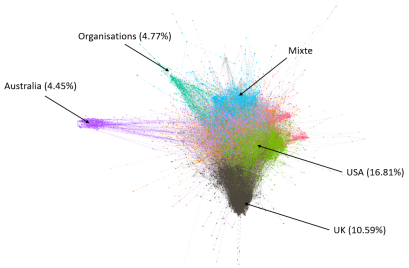

FIGURE 3: Result of "best partition" for 2017-09-30

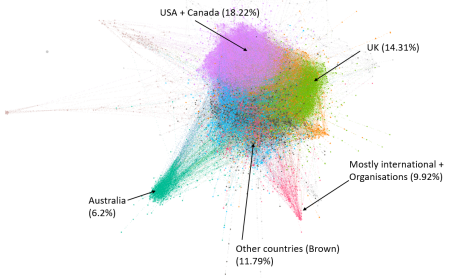

FIGURE 4: Result of "best partition" for 2017-08-30

By checking some nodes in the clusters we can assume that the "best partition" explains mostly geographical features. However some clusters are very difficult to identify, maybe the type of website (political, sports etc) has also an importance on this partition. It shows recurrent patterns along the days : some countries such as Australia or the "organizations" (NGO) are always apart. It can be explained by the fact that Australia is a country very isolated geographically. NGO probably shared issues that differed from classical medias also. Moreover we see that USA and UK are always closed to each over : it is logical in the sense where they shared a lot of common events. We can also notice that the percentage is not constant so the information is varying over time, which is rational because events do not always take place on the same geographical area.

This graph can then also tell us how global or local is an event. The figure 5 compares different types of news in term of scale. We can see that a purely local information will not spread around all the graph but only in the cluster corresponding to the area when a more global information will be more spread across graph. A good example is the event 790928951, which is a commonwealth news : we see that this event is spread across, not all the world, but the commonwealth only.

REMARK : We can already have a doubt about the way GDELT classify the events. On figure 5, we can see that two different GlobalEventID (791072761 and 791974306) are relying on the same event. We should be careful with that for the next parts.

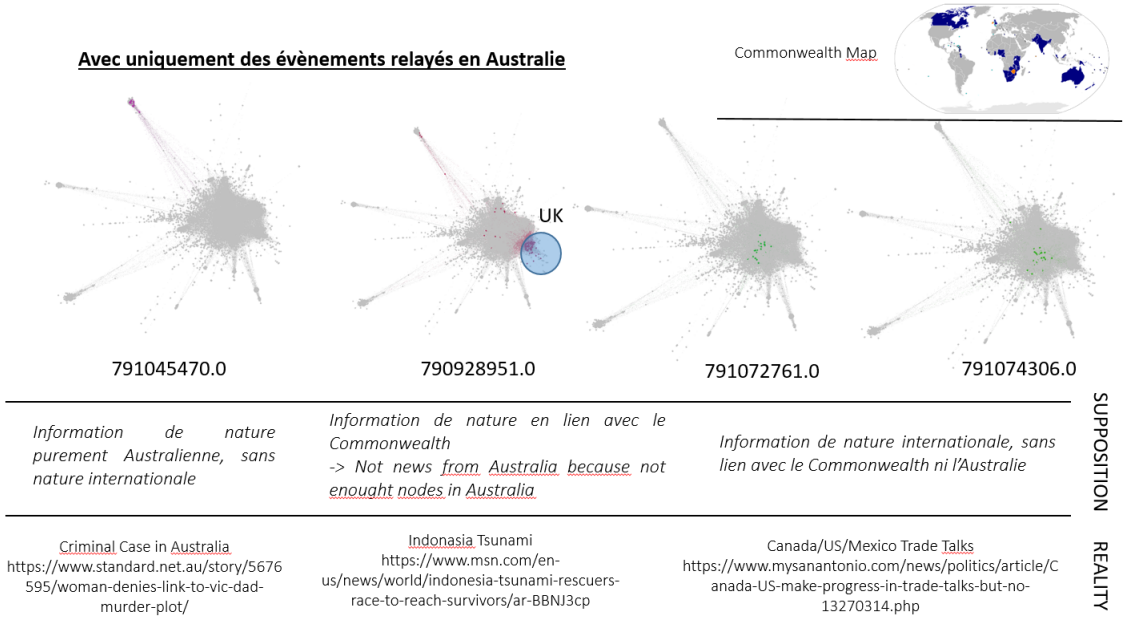

FIGURE 5: Comparison of different scaled news, relayed in Australia. The websites that mentioned the event are coloured

It is then clear that countries close geographically (USA and Canada for example) or by history and culturally (Commonwealth) are more likely to be on clusters closed to each over (and sometimes in the same cluster).

The first problematic will be then to verify that our "best partition" is mostly lead by geography.

4 Geographical Label Propagation

To verify that the "best partition" is mostly lead by geography we need to classify our graphs' nodes (website) by country or region. Obviously founding a reference dataset linking every websites of our graphs with a country or geographical area is very difficult. That is why we need to found some references and extend them across our graphs : this is a method called "label propagation".

4.1 References

We have chosen to create references from two different sources :

— Website Domain : it is, of course, a good indicator of the country. In our case we identify Australia (.au), UK (.uk), India (.in), New Zealand (.nz), USA (.us), Zambia (.zm), Canada (.ca), South Africa (.za), but also the organizations (.org) and the international sources (.international). The creation of this reference is very fast because we only need to look at the website URL. Moreover this is a very good reference because the confidence is very high. The web domain clearly indicates the country except for the .com domain.

— External Source : this source is needed because a good percentage of websites has .com as domain and then do not allow us to classify thanks to it. To solve this issue, we used some references from ABYZ News Links [3], which is a web page that references a lot of news website thanks to their country. We had noticed that, with website domain reference, only a few US websites where detected so we used this external reference to create mostly references for US, but also India, News Zealand and International.

Of course, this method of "label propagation" would require much more information to classify perfectly each different country. But the idea here is to verify the assumption that our "best partition" is mostly linked to geographical feature, so we only need to make appearing the biggest clusters.

4.2 Algorithm

The algorithm used has several steps that are enumerated just after :

1. Prepare the graph : we need to add the url labels to the graph where we want to propagate the geographical labels.
2. Create a reference DataFrame. If a website has a conflict between the URL domain and the CSV file from the external source, we use the URL domain. This choice has been done considering that URL domain is more reliable than the CSV file list, where we do not know really what it is based on. The number of countries with different label for URL and CSV is mostly due to the "international" classification of the CSV file.
3. Propagate the labels over the graph. This is the heart of the label propagation. This is done by the *label_propagation_pays* function. This function uses 3 parameters : the graph with the url labels, the list of nodes that we want to classify and the reference DataFrame. Each loop of the algorithm we :
   — Choose a random node in the list to approximate the fact we want to start with nodes where we have a high number of known neighbors nodes.
   — Go over the neighbors and count the country labels that appears.
   — We attribute the most appearing country label to the node.
   — If there is no labelled neighbors the chosen node, we increase the repeat counter and restart the all loop.
   — If the algorithm fails to classify 100 times successively, it stops. It is important to avoid infinite loop for not connected nodes.

4.3 Results

The idea is to use this label propagation in a lot of graphs, to gather the results in a DataFrame. The next step is then to take the most appearing country for each website over the days. This study shew that each node is not always present in each graph. By observing 15 different

3. http ://www.abyznewslinks.com/

days (graphs), we found that the mean of the number of appearance is close to 4 (3.715). This is why the number of appearance is not really showing how good the label propagation is. A better choice is to check the confidence : a ratio between the number of appearance of the "chosen country" over the number of graphs where this node is present. The mean of this indicator over all the nodes is close to 0.75 : it shows that **the label propagation is quite stable over the graphs**.

On the GDELT blog a master file is available with the country of each website. It is a great opportunity to test the result of our algorithm. We take it from the study before (on 15 different days). The result of the verification is shown in figure 6.

```
Number of well classified : 4915
Number of bad classified : 8851
----------------------------
Nb of found countries : 11
Nb of real countries : 178
----------------------------
Just for the 11 countries we know :
Number of well classified : 4915
Number of bad classified : 5209
```

FIGURE 6: Result of the GDELT master file checking

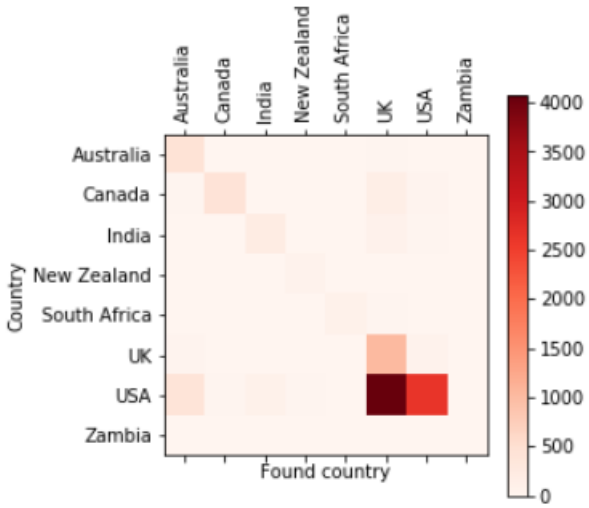

FIGURE 7: Confusion Matrix. The graduation indicates the magnitude on the confusion matrix.

It shows that we have two times more bad classify than well classify websites. It can be easily explained by the fact that we have a reference basis of 11 countries when there is in reality (according to GDELT master file) 178 countries. The good indicator to look is the number of well classified for the 11 countries we have as references. We see that we have the same order of magnitude for well classified and bad classified websites. The confusion matrix (figure 7) shows where the errors are and where the classification is good. We can conclude that

- The diagonal presents higher levels than other possibilities. It shows that the label propagation provides a good level of true positives.
- There is a lot more of true positive for the USA than for other countries. It can be explained by the fact that our references have a lot more 'USA' labels (via the external source) than other countries.
- There is a lot of errors between USA and UK : quite a lot of USA are classified as UK. It can be explained by the fact that UK and USA are very connected (same subjects).
- One more thing not to forget is that our reference is not containing a lot of countries.

We have seen that our propagation is quite stable over the graphs (=days). We can then test our algorithm in each day separately (without counting the most appearance country over the days). We can, at the end, visualized the result in percentage of good and bad classified websites. We conclude that there is **always less than 2% error on the label propagation**. We must stay careful with this good result. On the complete graphs there could be more error

because in our test the nodes. This is because the nodes where we verify the labelisation may be more connected to "known" sources than in the real complete graph.

5 Type of site Label Propagation

It is very difficult to found some reference fake-news to study because there are quickly removed from websites. This is why the choice has been done to study the differences between controversial and confidence website rather than fake-news and real news. We found a dataset of references, created by Melissa Zimdars who is an assistant professor of communication at Merrimarck College. This list, called "False, Misleading, Clickbait-y, and/or Satirical "News" Sources" [4] has been relayed by traditional medias and was a guideline for students : it should then be a confident source of our work.

This dataset classify 1001 website into several types of fake-news and confidence sources.

5.1 Neighbor Criteria

As the label propagation based on neighborhood has shown quite a good result on geographical labels, we have tried to see how well/bad it can work with type labels. If this label propagation based on neighborhood is well working, it will show that fake-news website are sharing more common events together than with confidence websites, and conversely.

The algorithm is very similar to the geographical label propagation seen before : the only variation is the introduction of a criteria for the classification and the train/test ratio which is fixed to 2/3 (which is a good ratio because on average, each graph contains 51 labeled websites). The criteria is that if the confidence weight of all the neighbors is bigger than the criteria multiply the controversial weight of the neighbor, the node will be classified as confident one. On the over side, if the controversial weight of all the neighbors is bigger than the criteria multiply by the confidence weight of the neighbor, the node will be classified as controversial one.

We have decided to used normalized edges in our graph to reduce the order of magnitude of our criteria. Each edges will then have his weight divided by the square of the degrees of his two nodes. The normalized graph and the non-normalized graph give the same result, considering a larger scale for the criteria on non-normalized graph. One example of result for the same day but for a graph normalize or not is shown in figures 8 and 9. We will then **use the normalized graph** for the rest of this label propagation.

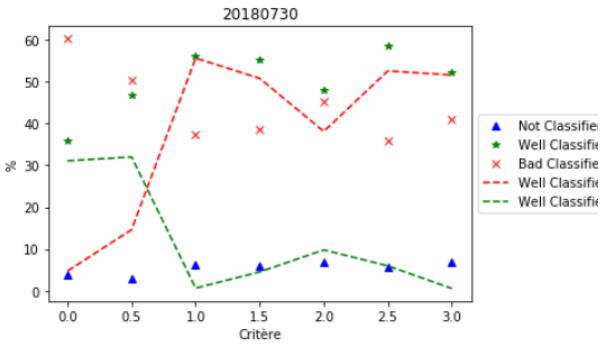

FIGURE 8: Result of type label propagation on 2018-07-30 normalized graph

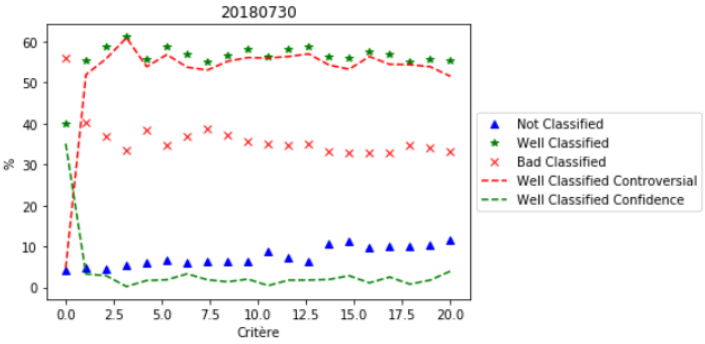

FIGURE 9: Result of type label propagation on 2018-07-30 not normalized graph

4. https ://docs.google.com/document/u/1/d/10eA5-mCZLSS4MQY5QGb5ewC3VAL6pLkT53V_81ZyitM/mobilebasic

The results of the type label propagation is quite fluctuating over the tests (it is big graphs with a lot of nodes) : that is why we have chosen to use f-fold cross-validation. The compromise to do is between time consuming and number of fold that ensure the stability of the result : after several tests, the chosen number of fold is f= 40. Moreover the number of retry for the algorithm (number of successively failed to classify a random node) is set a 50 because putting higher number will increase the time running of the algorithm and will not increase significantly the results (when it failed 50 times successively, knowing that nodes are picked randomly, shows that the algorithm is close to have finish classifying all nodes it can).

We have tested the algorithm with the cross-validation on several days : the results are shown in figure 10. Different conclusions can be extracted from this figure 10 :

- The results are different over the days : it seems good because each graph is different so each edges and each nodes are not the same, so the neighborhood is different.
- The global results (well/bad classified) are always closed to 50%, with some fluctuation, where we can go up to 60%/40%.
- The more higher criteria, the more not classified proportion increase : logical according to our used of the criteria in the algorithm.
- The criteria influences also the percentage of controversial and confidence website inside the well classified websites. A lower criteria will better classify confidence website rather than controversial website (which is logical because the first condition is on confidence website) when a higher criteria will benefit for controversial website good classification.

Finally, the conclusion we can have on those results is that there is no 'global' solution in term of result : even if for some graph the result is better than random classification (for example for the 2017-12-30, with criteria=3), the result on different days is really fluctuating. Moreover it is not possible to chose an optimal criteria that would be valid over all day. So the idea to chose a 'global' criteria and look over all graph to take the most appearance type for each website will not result in a proper output.

In order to improve the algorithm result, one idea is to try to give more information for the label propagation. We then tried to do the type label propagation on a two day graph. The results are shown in figure 11.

We see that the results of the 2-day graph (down of figure 11) does not provide a significantly better result than the 1-day graph corresponding (upper of figure 11). Those results may validate the hypothesis that our label propagation is independent of the number of day on the graph.

To validate this hypothesis we wanted to try our algorithm on a one-week-graph. The issue is that the graph is huge and take too many time to be constructed. Our hypothesis was that the events very popular are not significant to use because almost every website will publish about it : it will then create a lot of edges that are probably not useful for our propagation. To choose on how many publication we will "cut" the event, we rely on the figure 12. We decided to "cut" at 73 (the vertical line in the figure) in order to remove enough edges to facilitate the computation and to keep only the event that are less published on the week.

By doing the type label propagation on this "light" graph, we obtain the curves showed in figure 13. We can observe that the result is quite similar in term of percent comparing with previous results. We can then conclude that **our type label propagation is independent of the number of day of the graph**. We can also see that the reduction of the size of the graph does not seem to influence the result. To be sure of that we apply the same reduction of the graph on the 2017-09-30 : the figure 14 which is the result on the light graph shows quite similar results as the 2017-09-30 full graph of figure 11. We can then conclude that **reducing**

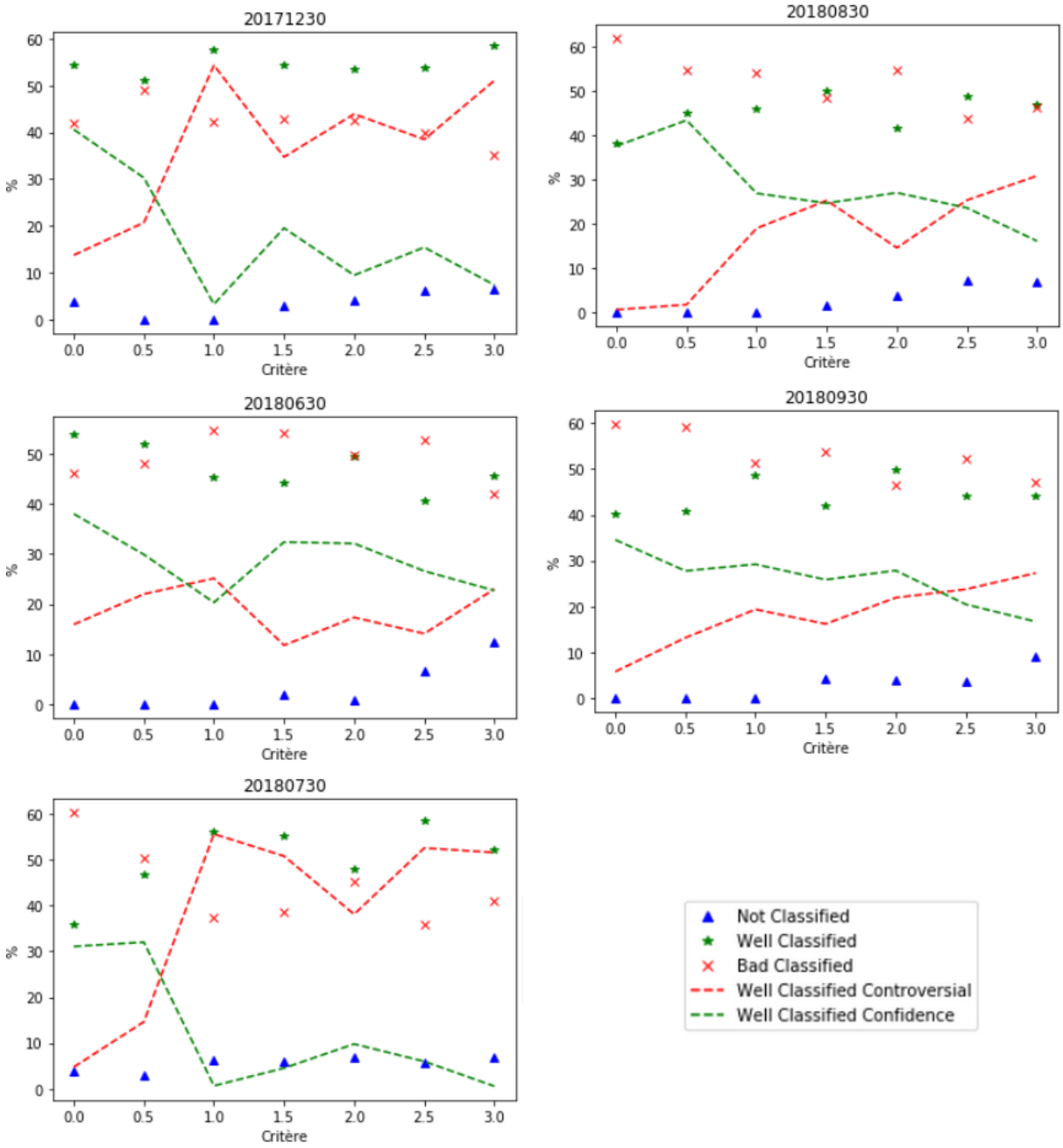

FIGURE 10: Results of label propagation on different days (normalized graphs).

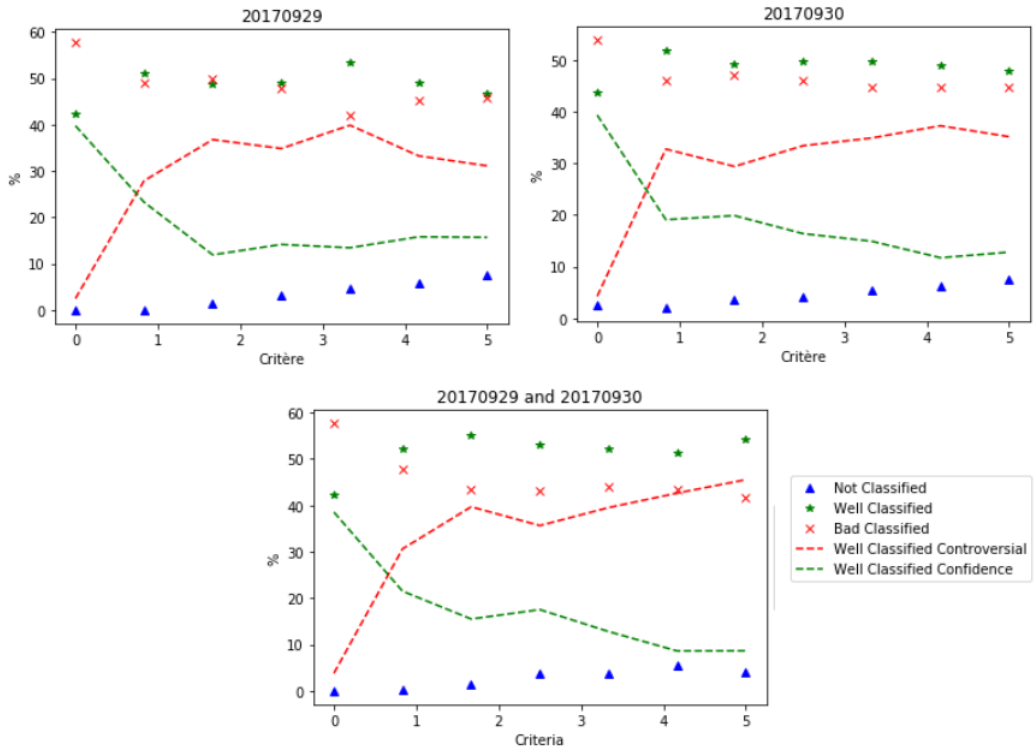

FIGURE 11: Results of type label propagation for 2-days-graph and one day graph corresponding.

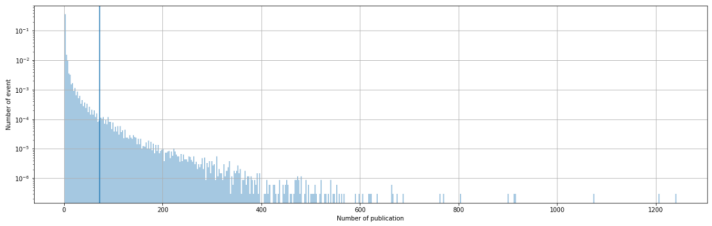

FIGURE 12: Histogramme of the number of publication for the 1 week graph.

the number of events considering only the less published events does not influence the results and reduce the computing time of the algorithm.

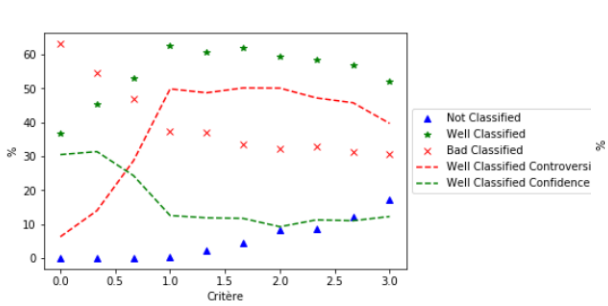

FIGURE 13: Result of the type label proapagation on the light graph of 1 week

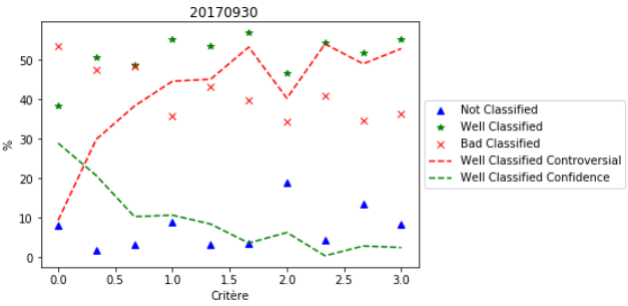

FIGURE 14: Result of type label propagation on 2017-09-30 light graph

To see how much we can reduce the original graph and still getting good results, we tried to

reduce the 1 week graph and see how the graph is impacted. By testing with different graphs, we found that the graph with events that are published more than 1 time and less than 10 times shows similar results as before (figure 15), without considering the issue of not classified website (that can be explained by the fact that the size of the graph have really decrease a lot). The figure 16 shows how the information is spread on the several "layer" of event classified on the number of publication. We see that the information allowing the classification of the website is well spread over this "layer" of number of publication. Of course, the less we have layers, the worse the classification is. However, we can denote that the graph constructed with the events that are published between 3 and 10 times shows a very good performance, even better than the graph between 1 and 10. We can then conclude that the information contained between 3 and 10 is more relevant (for this particular graph).

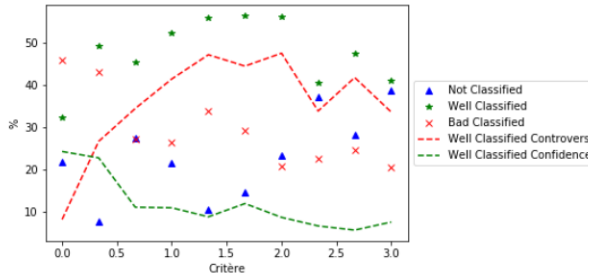

FIGURE 15: Result of the type label propaga- tion on the light graph of 1 week (with events on one week light graph with more than 1 publication and less than 10)

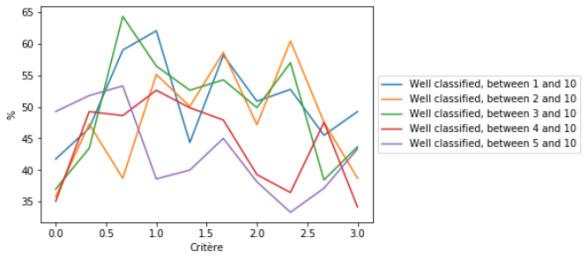

FIGURE 16: Result of type label propagation with several size

So we have tried to improve the results by adding more days on our graph, but we concluded that this is not useful because our algorithm is independent of the number of day given. We have also see that taking only a few event with some properties (number of publication) can help to improve a little bit the result. However, those results are still close to 50%. To understand why the algorithm is not that efficient, we take a look to the inside of the graphs. The figure17 shows the sub-graphs of labeled websites on the 2017-09-29 and 2017-09-30 graph. We see that there is **as many edges between the intra-classes as between the inter-classes** (the degree is two times higher for all the labeled website than for only one type of websites). This is why our algorithm have difficulties to classify the type of website : there is 50% of good classification (intra-class edges) and 50% bad classification (inter-class edges).

According to those results, we can conclude that neighbor criteria can present some results for particular days, but it is not enough to globally well classify the type of website : we need to introduce one or more criteria to reach our goal.

5.2 Republication Criteria

An other criterion that can be used for classifying the type of website is the republication criteria. The republication percentage is defined, for each website, as shown in equation 1.

$$\%republication = \frac{numberofeventpublicated - numberofuniqueeventpublicated}{numberofeventpublicated} \quad (1)$$

For this part of the study, we will work on the DataFrame containing all the data from the 2018-06-23 to the 2018-06-30 (1 week of data), because it will be more significant than working

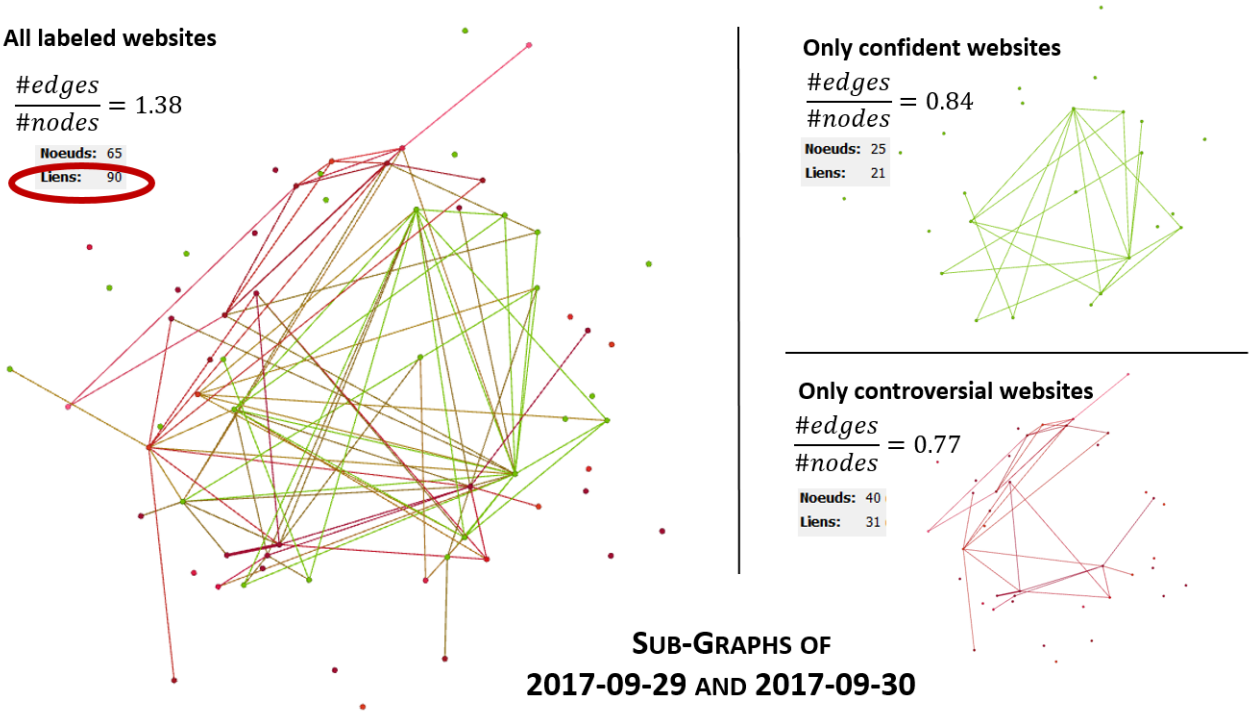

FIGURE 17: A look into labeled websites on the 2017-09-29 and 2017-09-30 graph

type	mean	median	max	min	quartiles	% of no republication
Controversial	0.034090	0.586585	0.586585	0.0	0.25 0.000000 0.50 0.000000 0.75 0.01...	0.684211
Confidence	0.006632	0.061856	0.061856	0.0	0.25 0.000000 0.50 0.001272 0.75 0.00...	0.500000

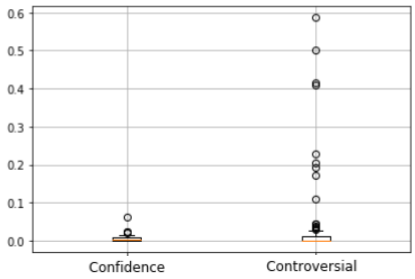

FIGURE 18: Statistics of the republication criteria for the week 2018-06-23 to 2018-06-30.

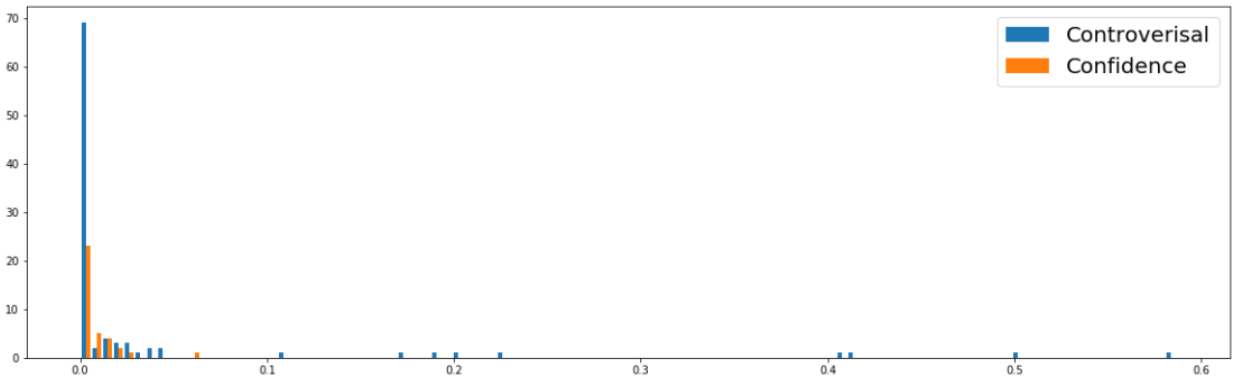

FIGURE 19: Republication criteria distribution for the week 2018-06-23 to 2018-06-30.

only on one day. The results are presented on figure 18 and 19.

We can observe on figure 18 that confidence and controversial websites have republication means very close to 0 : it can be explained by the fact that both have a lot of website that don't do republication at all ($\%republication = 0$). However, the controversial website republication is more spread when the confidence website one is more compact. The figure 19 show the spreading of this republication criteria for both types. We observe that there is a threshold (in this case 0.1) above which there is only controversial website that appears. **We can then conclude that some controversial website publish a lot more than other websites. We can also conclude that a majority of controversial website have a behavior close to confidence website in terms of republication.** Theses properties can be explained with two different points :

- Some controversial website are more likely to re-publish a lot more than other websites because they may have considerably less reporters. It follows from this that controversial websites produces probably less articles so it is easier for them to republish to attract readers.
- For the majority of controversial website that try not to re-publish too much, they try to imitate the stylish, in term of new/old article, of the confidence website. It may be a strategy to confuse reader and increase the confidence readers can have in this controversial website.

By checking the republication on different graph, it appears that the threshold seems to fluctuate slightly. T**o be able to chose one 'global' threshold in order to apply this new criteria for classification, we need to merge a lot of data into one DataFrame (maybe one or two months) and choose a threshold that will indicate, if the republication percentage is over it, that it is, for sure, a controversial website.**

Remark : The dataset is not clear about what is a republication, ie when an event is going to re-appear for the same website. Some republication leads to the same articles even with different links. Sometimes, there is troubles because it is not even the same event in reality. We should be very careful with that when choosing the threshold.

6 Time Line of Fake News

The two criteria developed before will probably not be enough to classify properly confidential and controversial website. An other clue to extract some properties may be in the time line of news spreading around controversial and confidence website : how are they interacting in time ?

To try identifying some new properties, we decided to look over several event over a full day. We used a new type of graph where each node is a unique website, and each edge is directed and represent the time line of the event over the day. The visualization on Gephi is such that the beginning of the day is on the upper left part of the circle, and the time spread out in the trigonometric direction, and end on the upper right part of the circle. For our work, we selected some events from the 2017-08-30, they are detailed in table 1. We see that even if we take different events, they are very linked to Hurricane Harvey. It is very difficult to see if there is a different tone or point of view (for the helping or Trump reaction for example), but we decided to keep those events like this for the study, doing the hypothesis that the 'cut' of event done by GDELT dataset should be having a sense.

GlobalEventID	Description	Number of publication
685158928	Hurricane Harvey (Texas)	90
685157951	Hurricane Harvey	1950
685157677	Hurricane Harvey	1118
685157902	Hurricane Harvey	450
685157780	Several subject link to Trump	400
685157906	Hurricane Harvey	1729

TABLE 1: Studied events of the 2017-08-30.

According to the figure 20 it is quite difficult to extract a global property, because each event seems to have differences. As an example we can see that for the event 685157677, there is only one website that repost a lot : it may be a local journal that give a lot of detail about a "local" news. However, for the over events, we can see that **the republication is higher for the website that post "early" about the event** : it may be logical because the event may have update so the website are more able to republish to inform they readers that the situation have changed. It is then difficult to extract over features :

- For some events, the controversial websites post quickly (event 68515790, 68515892), but it is not always the case (event 685157906).
- We can see that the controversial website does not repost about the same event (event 685157902,685157780, 685157906). It is logical to observe this, according to the republication criteria we have construct before, because we only observe one event during one day, and globally only few controversial website repost a lot.
- For some event, the confidence website have a tendency to repost about the same new several times (event 685157906, 685158928). It may be not visible in our "republication" criteria of before, because we only count the percentage of total republication, not for each event (a website that repost one time 4 different events will have the same republication percentage that a website that repost 4 times on only one event).

One great way to continue this study in order to extract feature on time line of news would be to check an event from the very first time it has been published to the end of it lifetime, not only looking on the timeline on one day.

7 Conclusion

This study on controversial and confidence websites, thanks to the GDELT 2.0 dataset, has put several finding into highlights.

Dataset The dataset GDELT 2.0 is very promising but has shown some limitations during this study, and they definitely should not be forgotten.

- A lot of event are doubled with, as the only change on the DataFrame, an exchange between actor 1 and actor 2.
- The way GlobalEventID are attributed still very opaque and we noticed that some events are put into the same GlobalEventID whereas they have nothing in common except the footer of the page (with a survey common to all articles of the website for the day for example). The other situation has also been seen : two articles that are talking of the

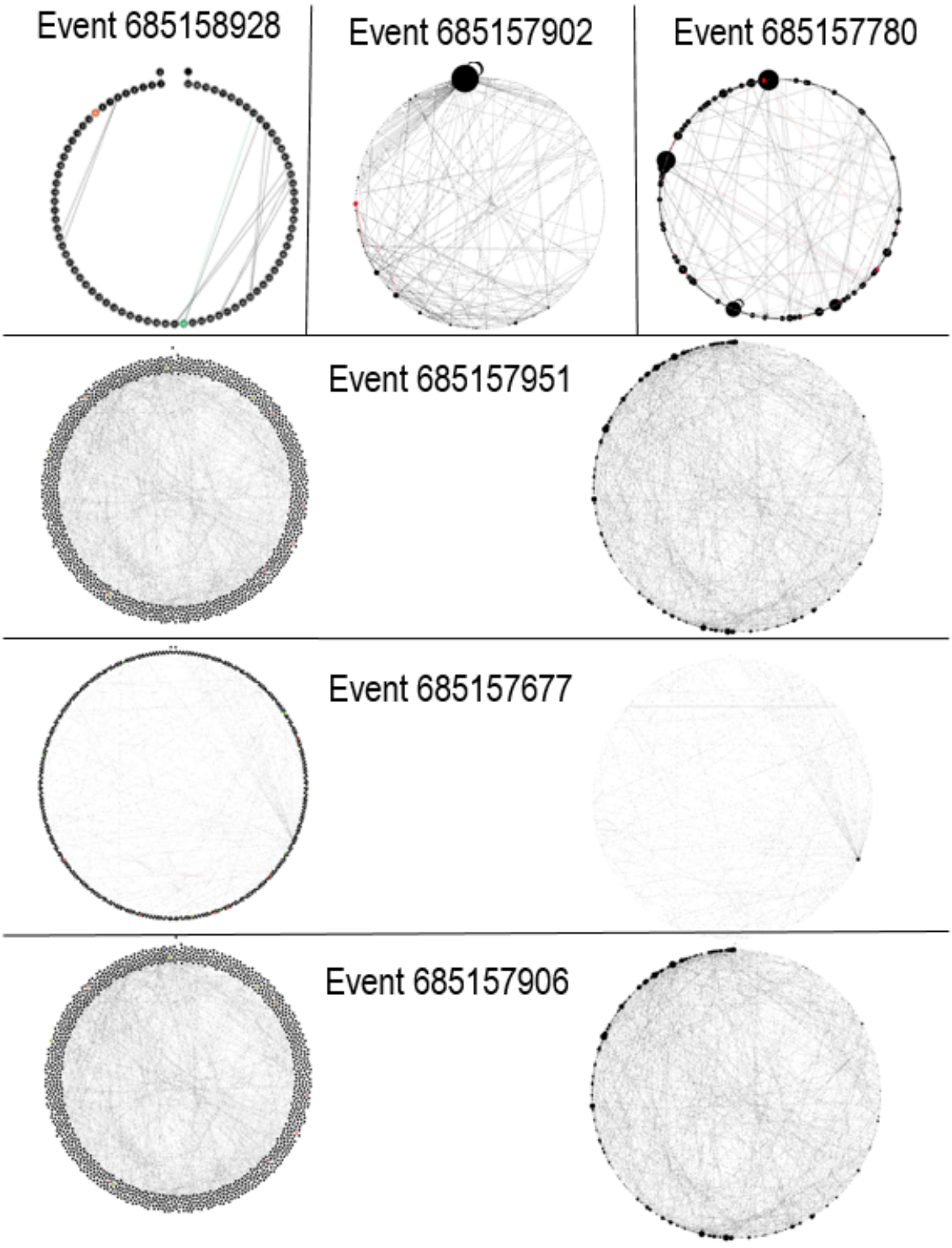

FIGURE 20: Graph of studied events. When the size of node vary it depends on the degree of the node (ie the number of republication of the event by the website.) The red nodes correspond to controversial website and the green ones correspond to the confidence website.

same event are classified as different GlobalEventID. **This issue of GlobalEventID is the most annoying for the dataset.**
— Some websites are also listed as different but are in reality the same (for example : abcd.it and abcd.com). This provides some issues when construction the graph : two websites will seems extremely connected compared to all the others, but it will not be relevant because it is the same website in reality.
— The point detected in the part of the republication criteria is also that we do not really know when an article is taking into account or not : we do not know if an update of an existing article will be counted.

Classification of controversial and confidence website The classification of the websites' type has been used based on the neighborhood. This method provides results that stay close to a random classification, but is able to tell when the classification can not be done. Moreover, the variation over the graph is quite important : some graph will provide a very good result and some a bad one. The choice of the criteria for the classification influences the percentage of confidence/controversial/not classified websites. This classification has been shown to be independent of the number of day given in term of information. We discovered that the most important events to classify are the events that are not republished a lot (less than 10 in our case). However, the results are not good enough to classify only on this criteria.
An other criteria that we found is based on the republication : we clearly found that over a threshold, there is only controversial websites. It could clearly allows us to classify with insurance that they are controversial websites.
However, it would be very interesting to found new criteria to increase definitely the results of our classification : the time line may be a good path to found it, but it will imply to look at a lot of data to found a global criteria.

Conclusion on controversial website behavior Thanks to this different criteria and according to the several studies done, we can extract some behaviour of controversial websites. The majority has a behaviour very close to confidence website. They are as linked with confidence website as with controversial one. In general, they republish the same percentage of information as confidence websites. This may be a choice in order to increase their credibility in the eyes of readers to raise their hearing whether for money or to broadcast fake information.
We have also seen that some controversial websites have a very different behaviour in term of republication, comparing to confidence websites : republication is a very important part of what they share. It may be explained by the fact that those controversial websites are very little companies or associations so they have not enough people to produce a lot of content : they are then much more likely to repost a lot.

Finally, we should keep in mind that all those results can fluctuate over time. News websites are very active and may change rapidly their publication policy. A few years ago, nobody was worried about fake-news, and now they are a growing concern in our society.