LaSNE: Visualising retweet networks through time. A case study of the Ukrainian crisis.

Christina Bogdan, Varun D N and Alexandre Sablayrolles

***Abstract*— The emergence and predominance of social media has made it a valuable tool for political analysis of major events, such as the 2013 Ukrainian crisis. Standard statistical and visualisation tools allow for an overall analysis of social media activity during these events, but there is currently no satisfactory tool for visualising influence networks on Twitter and their evolution through time. Our project, LaSNe, develops a new model to summarize retweet networks and visualise them. We show results on retweet networks collected by the SMaPP[1] lab during the Ukrainian crisis.**

I. INTRODUCTION

On November 21 2013, Ukraine's president Viktor Yanukovych suspended preparations for a trade deal with the European Union. This resulted in mass protests by its proponents, known as the *Euromaidan*. The Euromaidan is the first truly successful social media uprising, as earlier movements have stayed either out of social media (such as the Arab Spring), or have been confined to it and failed to have major political impact (such as Occupy Wall Street), according to a study lead by the Social Media and Political Participation (SMaPP) lab.

After months of such protests, Yanukovych was ousted by the protesters on 22 February 2014, when he fled the capital city of Ukraine, Kiev.

The SMaPP lab was created to analyze how social media impacts political participation. During major political events, they collect data on the main platforms (Twitter, Facebook, etc.) to track the evolution of the political movements. Given the scale of these political movements, data collection results in large-scale datasets. There are up to 10 million Tweets related to the Euromaidan movement collected from November 2013 to February 2014, and most of them occur on peak days of the crisis.

This large scale makes it impossible to visualise twitter communities and influence circles, not only because standard algorithms are not designed for this purpose but also because the computational problem requires hours of runtime, making it impossible to have a quick feedback loop. Gephi[1] is a very popular tool for visualising graph with a force-directed layout. Running it on a retweet network of 1 day takes several hours, and results in a cluttered graph (see figure 1).

In this project, we develop a tool to both reduce the number of nodes to the leaders of the movement, and visualise them. Our tool allows for fast preprocessing (a couple of minutes for the all 96 graphs) and instant rendering in the browser using d3.js [2].

[1]https://wp.nyu.edu/smapp/

[2]https://d3js.org/

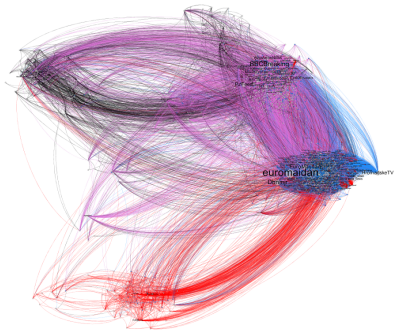


Fig. 1: Gephi visualisation of a retweet network.

II. DATA

A. Retweet networks

The data collection and acquisition was conducted by the SMaPP lab. The euromaidan hashtag was monitored as soon as the protests started, and millions of tweets were collected. Retweet networks are created from these tweets, where a node represents a Twitter user and a directed edge represents one user retweeting a message from another user. Retweet networks are a standard way to create graphs from a collection of tweets. As many real-world graphs, they have the power law property: a small portion of nodes are massively connected, whereas the bulk of the nodes have only very few connections. This can be seen in figure 2: on two different days, the number of tweets or retweets follow the power law, because their representation on a log-log plot is a line.

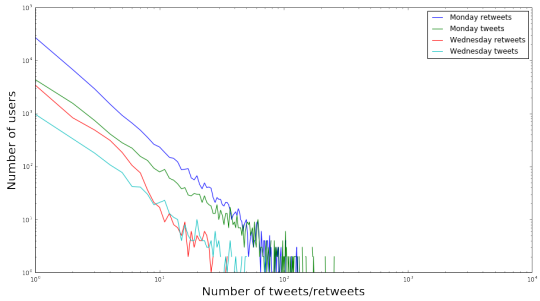


Fig. 2: Log-log plot of nodes connectivity. The horizontal axis is the number of tweets (resp. retweets) and the vertical axis is the number of nodes who have this number of tweets (resp. retweets)

B. The Ukrainian crisis

The Ukrainian crisis data covers 96 days, from November 25, 2013 to February 28, 2014. The activity gained momentum after February 18, 2014 beacuse that day saw violence and deaths. The peak day is February 20, 2014 with 47,743 users and 125,046 edges. This was the day when the protesters occupied central Kiev from the police. Figure 3 shows the evolution of the number of nodes and edges during these 96 days: there is a clear peak at the end, February 28, 2014, when Ukraine's president fled the country.

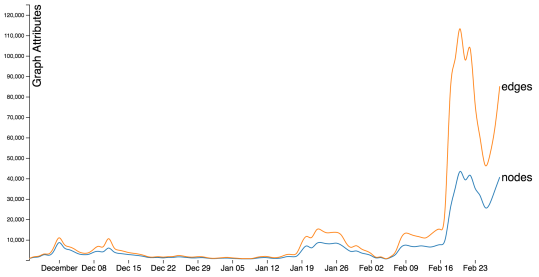


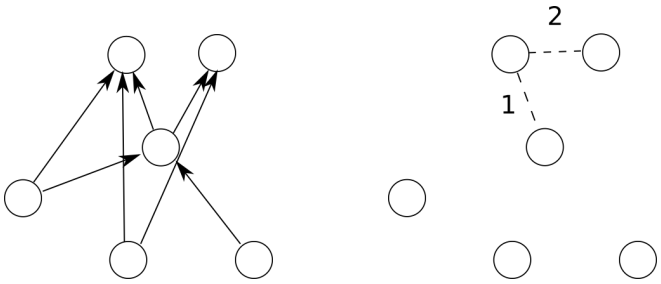
Fig. 3: Number of nodes and edges of RT networks during the Ukrainian crisis.

III. MODEL

A. CoRetweet networks

The first step of our model is to summarize the information of the graph into a more compact and meaningful way. As the graph is summarized, minimal information is lost but the result is more easily interpretable, and gives insights for further exploration.

Our process for summarization consists in generating co-retweet (coRT) graphs: if two nodes A and B were retweeted by nodes $C_1, \ldots, C_n$, we remove the (directed) edges from C_i to A and B, and add an (undirected) edge between A and B, of weight n. This process is shown in figure 4. As a lot of nodes are just retweeting, and not being retweeted, this process disconnects a lot of nodes and results in a graph with an order of magnitude less nodes.



(a) Retweet graph (b) Co-retweet graph

Fig. 4: Generating a co-retweet graph.

Our analysis is motivated by the numbers shown in table I: 50% of the nodes retweet one user, so they are only "attached" to this user and not connected to anyone else

Is retweeted \Retweets	0	1	2	3+
0	0	25634	5988	6975
1	2727	617	287	696
2	844	235	134	340
3+	1372	437	294	1163

TABLE I: Number of nodes by tweets/retweets count for February 20, 2013.

and thus they can be safely removed for the force-directed layout. Moreover, 80% of the nodes only retweet from other users, but are never retweeted.

B. Preprocessing pipeline

Our processing pipeline consists in three stages:

- Extract *leaders* of the protest
- For this leaders, compute the coRT graph
- Sparsify this graph by keeping only meaningful edges

The first step allows to limit the visualisation to the most prominent figures. To determine *leaders*, we look for the subset of users that have the highest number of retweets. There are two different strategies for that: either aggregating the retweet count over the whole period, and then extracting the top users, or extracting the top users for each day. We found in our analysis that the second strategy was better, because there is a lot of variance of retweets from one day to the next. After the second step, we sparsify the graph to keep only the meaningful edges. For this step also, there are multiple strategies: keeping edges with the largest weight, keeping edges with the largest weight relative to the retweet count of the nodes, etc. We find a robust strategy was to keep the edges that were the top 5 largest for either of the nodes. This step allows to remove the less important edges, so as not to clutter the graph.

Properties of the graph need to be kept in the CoRT graph as well. For the Ukraine data, edges have a *language* attribute. This attribute is conveyed into the coRT by counting the number of retweeters for which both edges had the same language. Edges thus have a weight, and a weight relative to each language.

Another property is kept in the CoRT graph: the raw retweet count of nodes. Because of the way our coRT graph is created, the number of times a node is retweeted is not the sum of incident edge's weights. An example of this can be seen in figure 4: the center node is being retweeted twice in total, but the sum of incident edges'weights is 1. As we want to minimize information loss, we add the attribute "tweet count" to the nodes in the summarized graph.

This surfaces different behaviors in the graph: some nodes have a big retweet count but are not connected to the rest, whereas other are very central on the contrary.

Figure 5 shows an example of a graph where the most retweeted node is completely separate from the core community. These cases are rare in our data.

IV. VISUALISATION

Our visualisation tool was developed with the goal of conveying the maximum amount of information. Our tool

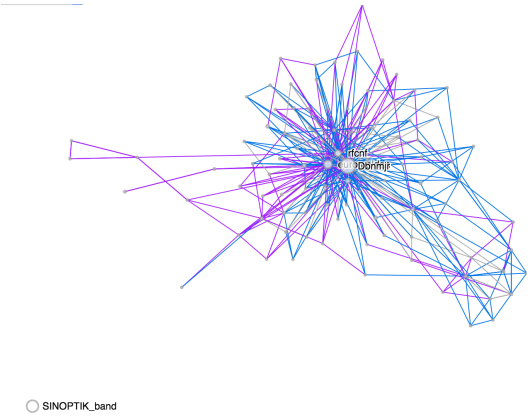


Fig. 5: Retweet network of January 30. The most retweeted node (sinoptik_band) is not connected to all other nodes.

allows to alternate between a view of a force-directed layout and a connectedness graph. Outside of the examples outlined here, our tool is also available online for use[3] .

A. *Force-directed layout*

The force-directed layout renders the coRT graph of the leaders on a particular day. This algorithm puts nodes who are densely connected close together, and nodes losely connected to them further away. This layout clusters communities. Edges are colored by the language property, by using the following rule: if one language takes up more than 90% of the weight, then the edge is colored with that language's color (blue for Ukrainian, purple for Russian), and the default color is gray. Figure 6 shows an example of a force-directed layout.

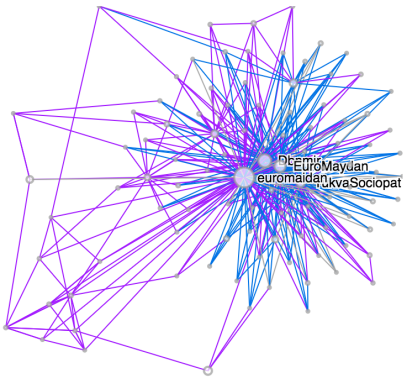


Fig. 6: Force-directed layout.

B. *ConnectedNess*

As the force-directed layout connects users very densely, an alternate view of the graph gives more perspective on which users are connected. This view is the "ConnectedNess" view, and it represents users according to their degree of separation to an initial user. A ConnectedNess view of the graph is shown on figure 7. At the left of the figure is a node chosen as the root of the view. The first column of nodes represents the root's direct relationships (i.e. users who are one edge away from the root). The second column represents nodes who are two edges away from the root, and so on. This allows to have a more precise idea of how users connect.

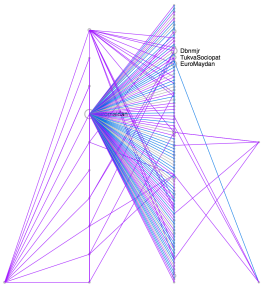


Fig. 7: ConnectedNess.

C. *Displaying more metrics*

To display more information about the graph, we show internal information on hovering. These include:

- Names of users on hovering
- Weights of edges on hovering
- Highlighting edges when hovering a node

Overall, the features of our visualisation allows someone to explore the data while having a precise idea of what metrics drive the visualisation. This leads to good insight and important results on the data, as we will see in the next section.

V. RESULTS

A. *Communities and leaders*

One of the main results of the SMaPP analysis of the Ukraine crisis was the pro-EU / pro-Russia divide could not be reduced to the language of the tweets (Ukrainian or Russian). This conclusion is surfaced by our network visualisation: in figure 8, we can see two clusters: the top one is centered around Euromaidan, and the bottom one is more spread-out and Russian-speaking. However, the top cluster has a significant number of edges in Russian.

On February 20, 2014, protesters regain control of Kiev from the police, with many dead. A strong division of language shows up in figure 9.

At the height of the crisis in February 2014, we see more and more predominant figures emerging in the pro-russian cluster, such "ARTEM_KLYUSHIN" and "ruredaktor", right after Yanukovych's flight of the country, and before Russia's annexion of Crimea.

B. *Ranking leaders*

LaSNe also provides a feature to rank leaders every day based on the number of retweets for that day. Ranks of some interesting leaders are presented in figure 11. `@Dbnmjr`

[3] `http://cims.nyu.edu/~ceb545/LaSNe/viz/graph_line.html`

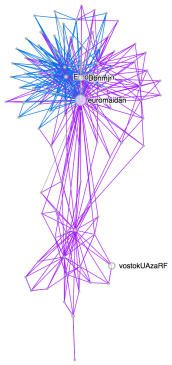


Fig. 8: Communities and language: the top community seems to be pro-EU and the bottom one pro-Russia. However, the top community also tweets in Russian.

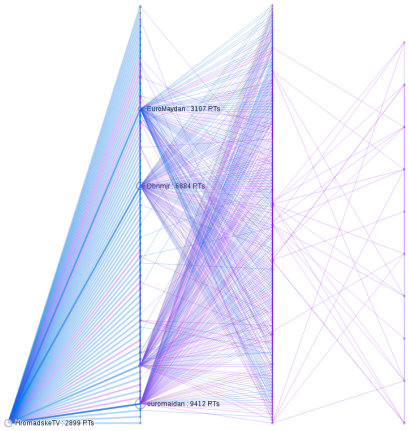


Fig. 9: Protesters regain control of Kiev.

and `@euromaidan` consistently stay on the top. This can be explained by the fact that `@Dbnmjr` was the official account of a Ukrainian newspaper and `@euromaidan` was the official account of the movement. `@EuroMaydan` was another account which was part of the same movement but it did not gain much popularity during the protest period. The popularity of the account `@euromaiden` fell over time. It was ranking high during the early stages of the protest, but dropped significantly towards the end. `@aronets` was the account of a Ukrainian public figure, Alexander Romanovich Aronets and his ranking was not consistent throughout the protest period although there were significant spikes in his rank during the key days of the movement. Similar plots of different leaders of interest can be obtained with our tool.

C. Limitations

One of the main limitation of our model applied to the Ukraine data is that the centrality of the Euromaidan user: because of the way the data is collected (following euromaidan tag), the Euromaidan user (the official twitter account of the protests) is predominant and connected to all the other users. The fact that our method surfaced this connection means that our method is representative of the data. However, it makes this central community very densely connected.

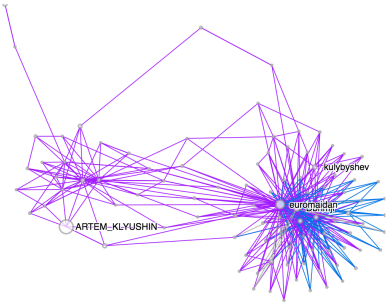


Fig. 10: Artem Klyushin becomes a major figure in the debates on Twitter.

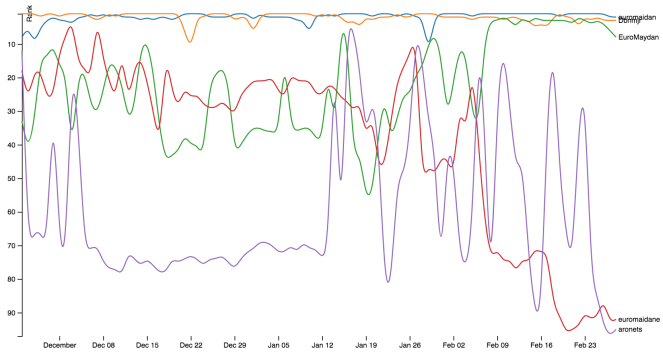


Fig. 11: Ranking some interesting leaders over time based on number of retweets.

VI. Conclusion

Our newly developed tool, LaSNe, is thus a very efficient tool for retweet network summarization and exploration. It allowed us to inspect how the protagonists of the Ukraine crisis interacted, and how their relationship evolved over time. As future work, we would like to extend this tool to other retweet networks collected by the SMaPP lab. An extension of our work could incorporate text analysis using Natural Language Processing tools on the raw tweet messages. Also, incorporating clustering techniques like community detection, backbone decomposition and k-core decomposition into the retweet network will provide valuable information about different groups with a network. Finally, we think that working upstream in the pipeline with the data collection and acquisition could help us make the most out of our tool, by collecting more data, with more various connections (avoiding the euromaidan-centric problem).

Acknowledgements

We would like to thank **Lauro Lins** for his guidance, as well as **Duncan Penfold-Brown** and **NYU SMaPP Lab** for providing us the data and supporting us throughout our project.

References

[1] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: An open source software for exploring and manipulating networks, 2009.