Community structure and information diffusion in social networks

Antoine Gourru[1,2], Erwan Giry-Fouquet[1,2], Ian Davidson[2,3], Julien Velcin[2]

Université Lumière Lyon 2, ICOM
Master 2 Data Mining

Université de Lyon, Lyon 2
ERIC EA3083

Computer Science Department University of California Davis
Collegium de Lyon - France , Institute of Advanced Studies Fellow 2017-2018

Abstract

Social networks are now the dominant source of information in society. However, little is understood on how information diffuses in these networks beyond simple node level models. This work explores the topic of understanding if community structure influences information spreading in social networks. Our work demonstrate that individuals behave most similarly to other individuals within their own community. This supports the notion of structural homophily implying behavioral homophily. Based on this conclusion, we use NMF to explore the activities and their similarities between communities.

Higgs Boson Dataset

The Higgs Twitter Dataset [2] provides the activity on twitter during the week of the Higgs Boson discovery announcement. Two networks are provided :

- a structural network (follower/followee), with 456k nodes and 15M edges.
- an interaction timestamped network, with 304k nodes and 563k edges (retweets, mentions and replies)

Community detection

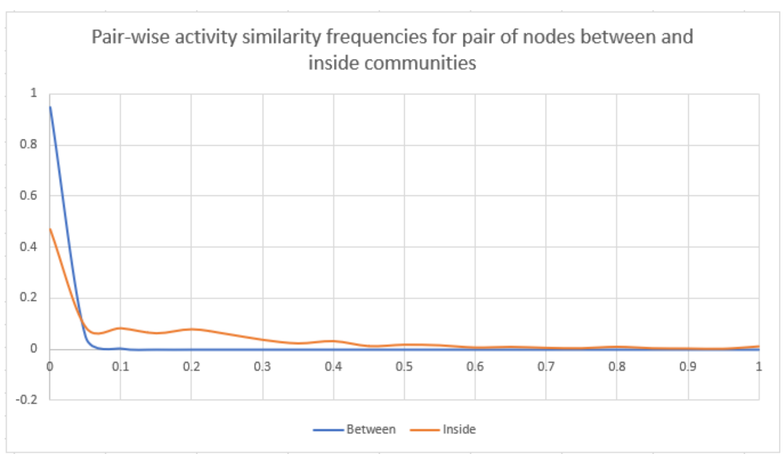
We use the igraph[1] package in order to test community detection algorithms on the structural network.

#nodes(~)	#edges	Louvain	Fast Greedy	Edge betweeness	Label Propagation	Info Map	Walk Trap
800	0.00001%	0.02	0.01	5	0	0.59	0.21
3000	0.0001%	0.05	0.27	469	0.01	7	1.04
5 500	0.001%	0.1	0.86	3290	0.03	12	3.13
20 000	0.01%	0.39	14.14	NC	0.15	131	33
33 000	0.1%	0.79	38.49	NC	0.99	337.42	105
94 000	1%	4.8	381.53	NC	1.17	NC	Inf(Memory issues)
135 000	10%	5.39	977.73	NC	1.73	NC	Inf(Memory issues)
450 000	100%	94.06	NC	NC	53.77	NC	Inf(Memory issues)

Only Louvain and Label Propagation methods scale well. Louvain is more robust than Label Propagation when adding and removing random edges. Thus, we used Louvain method to find communities.

Activity similarity inside communities

We follow [3] method, by measuring the activation of nodes, by interval of 4 hours (i.e. how many times it interacts). Then, we compute cosine similarity between every node activity vector.



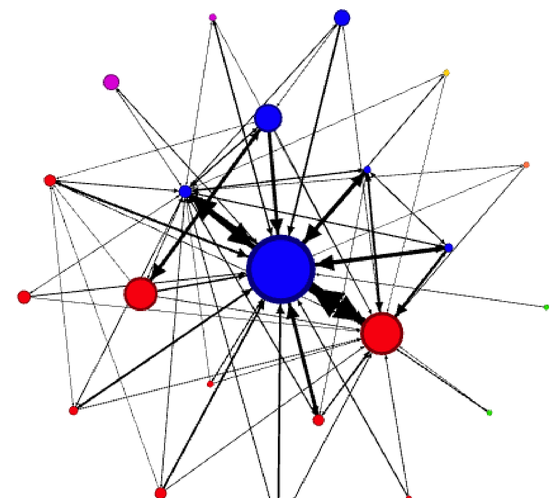
We can see on this chart that nodes in the same community usually share similar cosine distances than the nodes in different communities : they are active at the same time.

Clustering activities

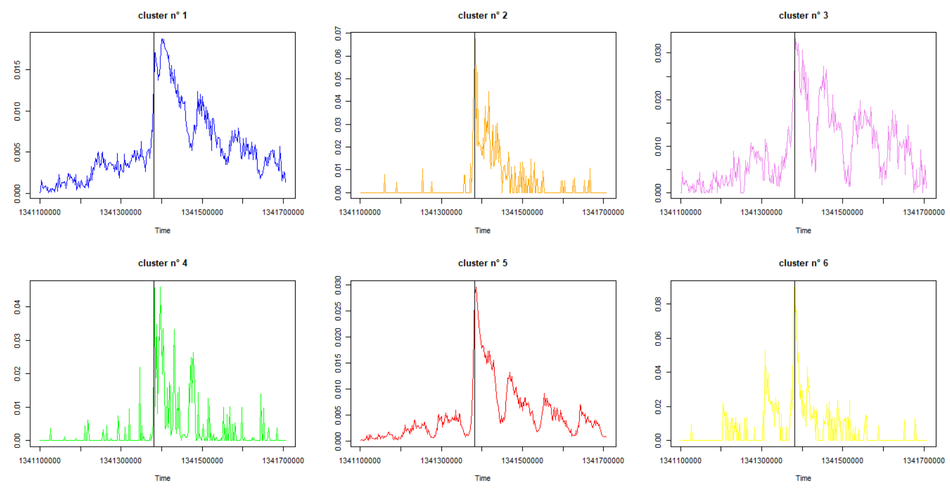
We now consider the activity by communities as the number of covered edges, in a time window of 30 minutes. We then get, for the i-th community, a vector A_i containing activity by time period. A is the row concatenation of A_i for each i. We use Non Negative Matrix Factorization, as positivity assure the interpretability of the dictionary entries $G_{i,:}$. The optimization problem is :

$$\begin{aligned} &\arg\min_{H,G} ||A - HG||_F \\ &\text{s.t} : H \geq 0, G \geq 0 \end{aligned} \tag{1}$$

We choose $k = 6$ as the factorization rank. We then use a simple k-means on H to get hard clusters for visualization :



The following charts are the mean activity signals for each cluster.



This clustering shows two majors clusters (1 & 5), with different mean signals, and four small clusters, with unusual activities.

Conclusion

We show that users from the same structural communities tweet at the same time. With a NMF approach, we cluster communities based on the similarity of their activity, highlighting "outsider" signals and frequent activity patterns.

References

[1] G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695(5):1–9, 2006.

[2] M. De Domenico, A. Lima, P. Mougel, and M. Musolesi. The anatomy of a scientific rumor. *Scientific reports*, 3:2980, 2013.

[3] S. Lin, Q. Hu, G. Wang, and S. Y. Philip. Understanding community effects on information diffusion. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 82–95. Springer, 2015.