Exploring Network Centralities in Spreading Processes

Maria Evgenia G. Rossi, Michalis Vazirgiannis

Exploring Network Centralities in Spreading Processes

Maria-Evgenia G. Rossi
École Polytechnique, France
rossi@lix.polytechnique.fr

Michalis Vazirgiannis
École Polytechnique, France
mvazirg@lix.polytechnique.fr

Abstract

Information and influence spread have been attracting a lot of attention due to the important role they play to numerous applications. Various previous studies have been focused on understanding the patterns during a spreading process. Most studies have been focusing on individual-based diffusion data and on inferring the diffusion network. In this work, we investigate the topological characteristics of individuals that are influenced and that participate in a diffusion process and present the patterns that are detected. We furthermore compare the individuals' characteristics between a simulated and a real world spreading process and show the need for a more comprehensive model in this area.

I. Introduction

Spreading of ideas is an important social phenomenon which in recent years has been influencing areas such as viral marketing, adoption of innovations and more generally spread of behavior and social norms. Many previous studies have been motivated by the analogy between an information diffusion process and that of the spreading of an epidemic. Such models imply that a specific idea or piece of information is diffused through the links connecting entities in a network and eventually impact a part of the network.

These epidemic approaches, led to important results concerning the identification of those entities that will trigger an efficient information diffusion. It was specifically shown that best spreaders correspond to those identified by the k-core and K-Truss decomposition [1, 2, 3] and not to those being highly connected or having a bigger node centrality (e.g., degree centrality). There exist cases where a node can have arbitrarily high degree, while its neighbors are not well-connected, making degree a not very accurate predictor of the spreading properties. As the influential spreaders, identified by the aforementioned graph degeneracy algorithms, are the ones responsible for the greatest part of the spreading activity, we have decided to study the topological characteristics of the individuals taking part in a spreading process that was triggered by such entities.

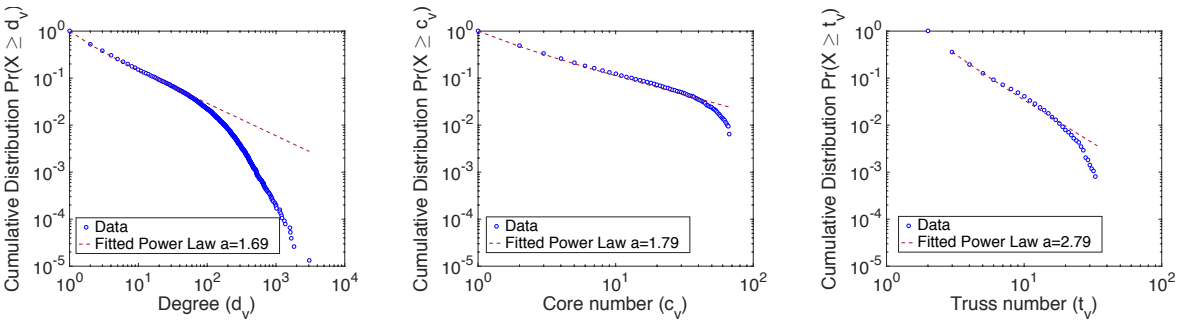
In this work we explore the centralities of the entities that are involved in a spreading process which is triggered by different groups of influential spreaders of a network. We analyze the patterns that occur by simulating the spreading process with the SIR and SIS epidemic models [4]. We finally compare the simulated diffusion process with real influence, in terms of the evolution of the centralities of the infected nodes and conclude that there is need for a diffusion model that fits the real world process.

Preliminaries Let $G = (V, E)$ be an undirected graph. Then, each node $v \in V$ has a degree $d_v = d$ if it is connected with d nodes in the graph. Let set D denote the set of nodes with the highest degree in the graph. C_k is defined to be the k-core subgraph of G if it is a maximal connected subgraph in which all nodes have degree at least k. Then, each node $v \in V$ has a core number $c_v = k$, if it belongs to a k-core but not to a $(k+1)$-core. Let C denote the set of nodes with the maximum core number. The K-truss decomposition extends the notion of k-core using triangles. The K-truss subgraph of G, denoted by T_K, $K \geq 2$, is defined as the largest subgraph where all edges belong to $K - 2$ triangles. An edge $e \in E$ has truss number $t_e = K$ if it belongs to T_K but not to T_{K+1}. We define the node's truss number $t_v, v \in V$ as the maximum t_e of its adjacent edges. Then, T denotes the set of nodes with the maximum node truss number. It has been shown that the maximal k-core and K-truss subgraphs (i.e.,maximum values for k;K) overlap [5], with the latter being a subgraph of the former. For that reason we finally denote as C' the set of nodes belonging to the k-core excluding those that belong to the K-truss of the graph.

II. Methodology and Evaluation

To simulate the spreading process, we use the Susceptible-Infected-Recovered (SIR) and Susceptible-Infected-Susceptible (SIS) models where the nodes can be in one of the states that the names suggest. Initially, we set a single node to be infected (as chosen from each of the three groups to be compared and that are described later) and the rest of the nodes at the susceptible state. At each time step, the infected nodes can infect their neighbors with probability β which corresponds to the infection rate and can recover from the disease or return to the susceptible state with a probability γ for the case of the SIR and SIS models respectively. Here we set the parameter β close to the epidemic threshold [6] and the parameter $\gamma = 0.8$, as used by Kitsak et al. [1].

We have performed experiments with the following real-world networks: EmailEnron, Epinions and Higgs Twitter (`snap.stanford.edu`). All graphs are considered undirected and unweighted (see Table 1). We have examined the distribution of the node degree (d_v), core number (c_v) and truss number (t_v) of these networks and the results for the Epinions dataset are depicted in Figure 1. The plot shows the complementary cumulative distribution function of the nodes' aforementioned centralities in log-log scale. We observe that all three distributions are skewed, indicating that few nodes have high centralities and the majority of them have "low" degree and participate in "low" k-core and



(a) DEGREE d_v (b) CORE NUMBER c_v (c) TRUSS NUMBER t_v

Figure 1: *Complementary cumulative distribution function of nodes' (a)* DEGREE d_v*, (b)* CORE NUMBER c_v *and (c)* TRUSS NUMBER t_v *of the* EPINIONS *dataset in log-log scale. The red line corresponds to the fitted power-law distribution.*

Network	Nodes	Edges	d_{max}	k_{max}	K_{max}
EMAILENRON	33,696	180,811	1383	43	22
EPINIONS	75,877	405,739	3044	67	33
HIGGS	456,626	14,855,842	51386	125	72

Table 1: *Network datasets used in this study.*

K-truss subgraphs.

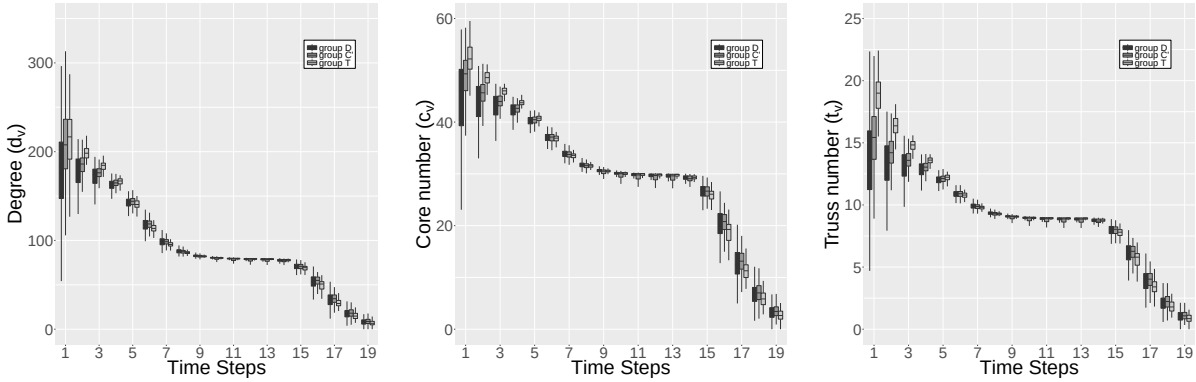
In our experiment, we compare three node centralities: (i) degree (d_v), (ii) core number (c_v) and (iii) truss number (t_v). Those are the centralities of the nodes that are being infected at every time step of the process while the epidemic was triggered from three different groups of nodes: (a) group D, (b) group C' and (c) group T. For every node of the group, we simulate the process 100 times and get the average behavior of the node. In order to get the average behavior of all the nodes of each group, we repeat the above for all respective nodes. The results from the experiments are depicted in Figure 2. We are also showing results from our previous studies [2, 3], in Figure 3 where the average number of nodes being infected at each time step during the evolution of the SIR model is depicted. The latter is for the reader to have in mind how many nodes have the centralities depicted in Figure 2.

We can observe that in case of the Epinions dataset, nodes originating from group T, achieve to influence on average nodes with higher degree, core and truss centralities during the outburst of the epidemic – specifically during the first four steps. In case of the Email-Enron dataset, group T and C' seem to influence nodes with similar centralities during the first timesteps. In both datasets though, the superiority of the latter groups compared to group D is easily recognizable during the first period of the spreading process. After the outburst of the epidemic (after the 6th time step), we observe in both datasets that nodes being infected are characterized by similar centralities for all the three compared behaviors. It should be noted that the centralities of the nodes infected during this "plateau" period are quite high considering the fact that most of the nodes of the network are characterized by low centralities. We realize that most of the nodes infected in all cases during such an epidemic are characterized by the centralities observed during the "plateau" period. For example for the EPINIONS dataset, from the 8th until the 14th timestep, we observe that nodes being infected have a degree ranging from 77 to 87, a core number ranging from 29 to 31 and a truss number ranging from 8 to 9. Finally, during the fadeout (during the 5 last time steps), the centralities of the nodes infected are severely decreased in all cases. Note that the process stops when no more nodes get infected (for the EPINIONS dataset this happens at time step 19).

The respective results while using the SIS model do not much differ than those depicted for the SIR model. The only difference is that, due to the SIS model's nature, the epidemic does not stop and the nodes that continue to get infected are characterized by similar centralities for all three behaviors.

Comparison to a real spreading process. In order to explore the information spreading in a real world setting we have used the Higgs Twitter dataset [7]. The dataset is built by studying the diffusion (in means of tweets) of the announcement of the Higgs boson-like particle at CERN on the Twitter social network between the 1st and 7th of July 2012. The interactions that were considered were retweets, mentions and replies. Some characteristics of the user network that was involved in the respective interactions is shown in Table 1.

In order to fairly compare the specific spreading process with the simulation models that were previously discussed, specific assumptions have to be made[8, 9]. The spreading activity that is recorded, involves around 562,556 asynchronous timestamps during which at least one spreading interaction is recorded. We have decided to study the influence that is triggered from nodes belonging to group C (i.e., the totality of the nodes participating in the maximum k-core subgraph of the network) as they have been proven to represent a great percentage of the spreading activity in a network. The timestamp where each of the respective nodes is firstly influenced by a user of its network is considered as the first timestep of the specific node's spreading activity. The following timestep is considered after 5000 consecutively recorded timestamps. We are considering the nodes being influenced during every such period by all the nodes that were influenced during the preceding periods. We have considered for our experiments totally ten such periods which we will be refering to as timesteps. We are specifically interested in the three centralities of those nodes



(a) DEGREE d_v (b) CORE NUMBER c_v (c) TRUSS NUMBER t_v

Figure 2: *Evolution of the infected nodes' average (a)* DEGREE d_v, *(b)* CORE NUMBER c_v *and (c)* TRUSS NUMBER t_v *during a simulated spreading process using the SIR model for the* EPINIONS *dataset having triggered the epidemic from nodes of sets D, C' and T.*

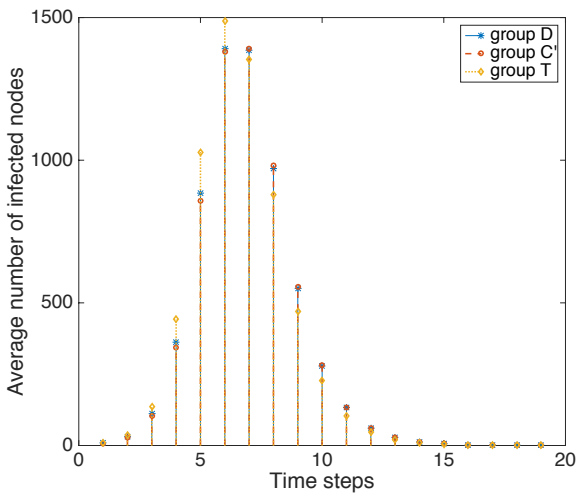


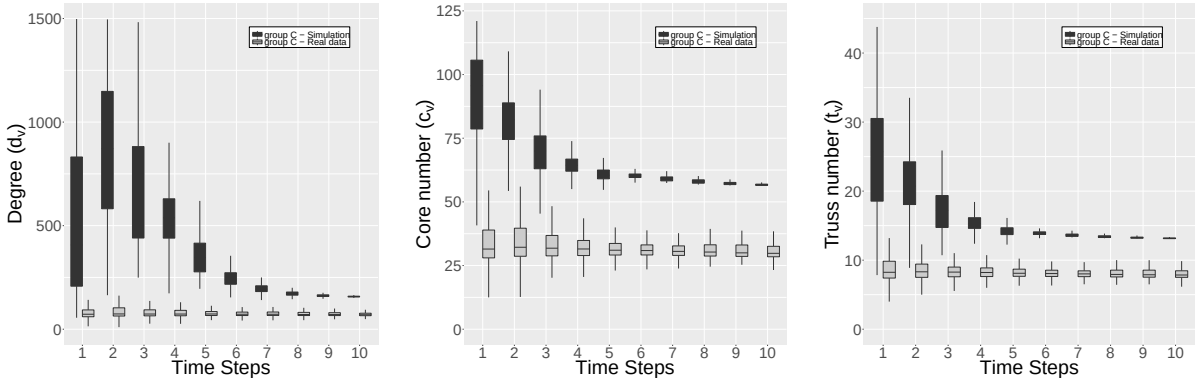
Figure 3: *Average number of infected nodes per step of the SIR model having triggered an epidemic from nodes of sets D, C' and T for the* EPINIONS *dataset. We are using as value β, a value close to the epidemic threshold (here β=0.01) and γ= 0.8.*

that are being infected during these timesteps which we have compared with the respective centralities after running the SIR model for ten timesteps starting from the same C nodes. As in our previous experiments, the process is simulated 10 times for every node of the group (due the dataset's size) and the average behavior of the node is calculated. The above is repeated for all the nodes of the C set. Results from the experiments are shown in Figure 4.

We observe that there are great differences between the two settings. While the simulation shows that during the first steps, nodes with high centralities are influenced, real data show that the nodes that are influenced do not differ much in terms of centralities during these 10 time steps that we study. Similar results are observed after running the SIS model and comparing it with real data. It has indeed been proven that epidemic models fail to reproduce the realistic viral spreading pattern [9] in terms of i) number of nodes being infected and ii) of the characteristics of the diffusion trees created during the process. We prove that the model also fails to indicate the centrality characteristics of the nodes being infected during the process. This can be explained by the definition of the models. First and foremost the probability of an entity influencing a neighboring entity shouldn't be the same for all entity relations. Moreover, considering the SIR model, an entity does not get "recovered" while in a spreading process such as an information diffusion in a Twitter network. User behavior contains more complex patterns concerning the way information is disseminated. Users may stop diffusing information for some period of time but start "spreading the word" again in a later period for indefinite reasons. This resembles the SIS model where infected nodes can return to the susceptible state and with a probability can start again infecting their neighbors. But unfortunately, neither this model can be compared with the real influence data of our study. While the latter may be extremely hard to model, we believe that there exist "who-influences-whom" patterns in influence data that can help towards a better definition of the probabilities of an entity influencing a fellow neighbor. Those patterns can be found while exploring the aforementioned centralities of entities influencing their peers between steps of the spreading process.

III. CONCLUSIONS AND DISCUSSION.

In this work, we explored the centralities of the entities that are involved in a spreading process which is triggered by different groups of influential spreaders of a network. We obtain interesting results by simulating the spreading process with the SIR and SIS epidemic models that let us conclude that i) degeneracy algorithms help us detect groups of nodes that will influence nodes with high centralities during the outburst of the epidemic and ii) there exists a "plateau" period during the spreading process where a significant part of the nodes are influenced and iii) the nodes influenced in this "plateau" period have relatively high degree, core and truss centralities considering the respective centrality distributions of the net-



(a) DEGREE d_v (b) CORE NUMBER c_v (c) TRUSS NUMBER t_v

Figure 4: *Comparison of the evolution of the infected nodes' average (a)* DEGREE d_v*, (b)* CORE NUMBER c_v *and (c)* TRUSS NUMBER t_v*, between a simulated spreading process using the SIR model and real influence data for the* HIGGS-TWITTER *dataset having triggered the epidemic from nodes of set* C*.*

work. Finally, by comparing the simulated diffusion process with real influence, we observe that epidemic models cannot reproduce the real diffusion in terms of the evolution of the centralities of the infected nodes. Thus we conclude that a further research direction could be the search for a diffusion model fitting the real world process.

Acknowledgments. Maria-Evgenia G. Rossi is funded by a DigiCosme Ph.D. Fellowship.

REFERENCES

[1] Maksim Kitsak, Lazaros Gallos, Shlomo Havlin, Fredrik Liljeros, Lev Muchnik, H. Eugene Stanley, and Hermán Makse. Identification of influential spreaders in complex networks. *Nature Phyics 6*, 888 - 893, 2010.

[2] Fragkiskos D. Malliaros, Maria-Evgenia G. Rossi and Michalis Vazirgiannis. Locating influential nodes in complex networks. *Scientific reports 6*, 19307, 2016.

[3] Maria-Evgenia G. Rossi, , Fragkiskos D. Malliaros and Michalis Vazirgiannis. Spread it good, spread it fast: Identification of influential nodes in social networks. Proceedings of the 24th International Conference on World Wide Web Companion, pages 101-102, International World Wide Web Conferences Steering Committee, 2015.

[4] Mark E.J. Newman. Spread of epidemic disease on networks. Physical review E 66, 016128, 2002.

[5] Jia Wang and James Cheng. Truss decomposition in massive networks, Proceedings of the VLDB Endowment 5, 812-823, 2012.

[6] Deepayan Chakrabarti, Yang Wang, Chenxi Wang, Jurij Leskovec and Christos Faloutsos. Epidemic thresholds in real networks. ACM Transactions on Information and System Security (TISSEC), 10 (4), 2008.

[7] Manlio De Domenico, Antonio Lima, Paul Mougel and Mirco Musolesi. The anatomy of a scientific rumor. Scientific reports 3, 2980, 2013.

[8] Sen Pei, Lev Muchnik, Jr, José S Andrade, Zhiming Zheng and Hernán A Makse. Searching for superspreaders of information in real-world social media. *Scientific reports 4*, 5547, 2014.

[9] Sen Pei, Lev Muchnik, Shaoting Tang, Zhiming Zheng, and Hernán A. Makse. Exploring the complex pattern of information spreading in online blog communities. *PloS one 10*, no. 5, e0126894, 2015.