Spread It Good, Spread It Fast: Identification of Influential Nodes in Social Networks

Maria-Evgenia G. Rossi[1], Fragkiskos D. Malliaros[1], Michalis Vazirgiannis[1,2]
[1]École Polytechnique, France, [2]Athens University of Economics and Business, Greece
{rossi, fmalliaros, mvazirg}@lix.polytechnique.fr

ABSTRACT

Understanding and controlling spreading dynamics in networks presupposes the identification of those influential nodes that will trigger an efficient information diffusion. It has been shown that the best spreaders are the ones located in the core of the network – as produced by the k-core decomposition. In this paper we further refine the set of the most influential nodes, showing that the nodes belonging to the best K-truss subgraph, as identified by the K-truss decomposition of the network, perform even better leading to faster and wider epidemic spreading.

Categories and Subject Descriptors

H.2.8 [**Information Systems**]: Database applications—*Data mining*

Keywords

Influential nodes; Epidemic spreading; Social network analysis

1. INTRODUCTION

The problem of identifying influential spreaders in networks has been attracting a significant part of the research community. It can reveal new insights in application domains such as viral marketing, epidemic control and more generally in information diffusion. The problem can be sub-categorized in two subtopics: identification (i) of individual influential spreaders and (ii) of a group of spreaders that render the influence more efficient. Focusing on identifying single spreaders, widely used criteria include the degree, betweenness, closeness, eigenvector, PageRank centralities and the k-core index [4]. It was recently shown that when applying SIR and SIS modeling, that describe disease spreading [2], best spreaders correspond to those identified by the k-core decomposition and not to those being highly connected or having a bigger centrality [3].

In this work we show that the K-truss decomposition [5] can serve as an even better criterion to identify privileged spreaders. The nodes belonging to the maximal K-truss of the network, show better spreading behavior compared to previously used importance criteria. Our analysis on real datasets shows that, not only more nodes get infected during the outbreak of the epidemic, but also the total number of nodes infected at the epidemic's fadeout is greater.

WWW'15 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742736.

Preliminaries. Let $G = (V, E)$ be an undirected 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. We denote as C the set of nodes with the maximum core number $k_{\max}$ (i.e., the nodes of the k-core subgraph of G that corresponds to the maximum value of k). The K-truss decomposition extends the notion of k-core using triangles, i.e., cycle subgraphs of length 3. 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. Respectively, an edge $e \in E$ has truss number $t_e = K$ if it belongs to T_K but not to T_{K+1}. Since the definition of K-truss is per edge, 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, with the latter being a subgraph of the former; in fact, K-truss represents the *core* of a k-core that filters out less important information. Here, we are only interested for sets C and T, that can be computed efficiently [5].

2. METHODOLOGY AND EVALUATION

In this paper we aim to identify those single spreaders in a network that will achieve an efficient spreading of information. We argue that those nodes are located in the previously defined set T of the graph, produced by the K-truss method. To simulate the spreading process, we use the Susceptible-Infected-Recovered (SIR) model where the nodes can be in the respective state that the names suggest [2]. Initially, we set one node to be infected (our single spreader, as chosen by different methods described later) and the rest of the nodes at the susceptible state. At each time step, the infected nodes that were infected at a previous time step, can infect their neighbors with probability β (i.e., infection rate). At every time step, a node that has been previously infected can recover from the disease with a probability γ (i.e., recovery rate). The process is repeated until there are no infected nodes in the network. Here we set the parameter β close to the epidemic threshold $\tau = 1/\lambda_1$, where λ_1 is the largest eigenvalue of the adjacency matrix of the network [1]. We also set parameter $\gamma = 0.8$, as used in Ref. [3].

We are comparing the average spreading behavior of the nodes belonging to the set T (**truss** method), to those belonging to the set $C - T$ (**core** method) (i.e., the nodes belonging to the k-core excluding those that belong to the K-truss of the graph – since T is subset of C) and those belonging to the set D that contains the highest degree nodes in the graph (**top degree** method); we

choose $|C| - |T|$ high degree nodes to achieve fair comparison between the different methods. We have performed experiments with the following real-world networks: EMAIL-ENRON, EPINIONS and WIKI-VOTE (`snap.stanford.edu`). All graphs are considered undirected and unweighted (see Table 1).

Table 1: Network datasets used in this study.

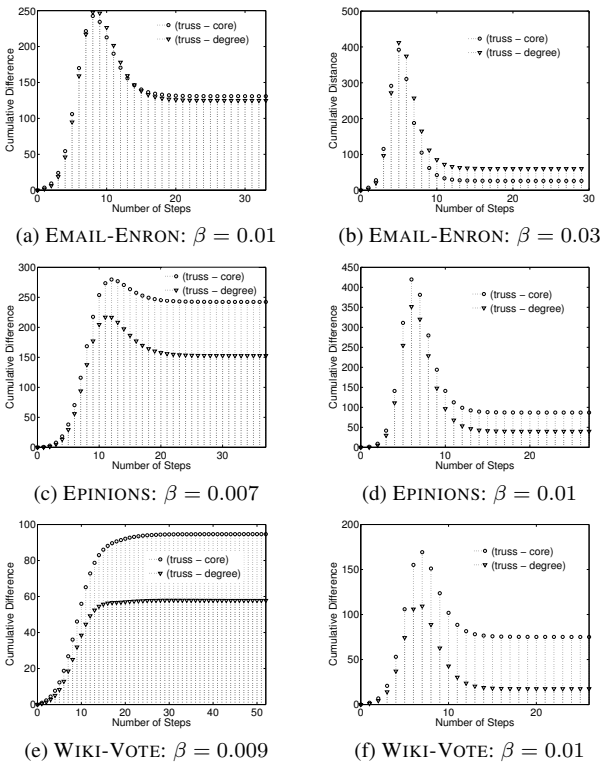
Network Name	Nodes	Edges	$\|C\| - \|T\|$	$\|T\|$
EMAIL-ENRON	33,696	180,811	231	44
EPINIONS	75,877	405,739	425	61
WIKI-VOTE	7,066	100,736	286	50

Table 2: Average number of infected nodes for some steps of the SIR model, using β close to the epidemic threshold of each graph and $\gamma = 0.8$. *Fin. step* column shows the total number of infected nodes at the end of the process (with std. deviation σ).

		Time Step					*Max*
	Method	2	6	10	*Fin. step*	σ	*step*
EMAIL-ENRON	**truss**	8.44	204.08	355.84	2,596.52	136.7	33
	core	4.78	152.55	364.13	2,465.60	199.6	37
	top deg	6.89	155.48	357.08	2,471.67	354.8	36
EPINIONS	**truss**	4.17	75.04	329.08	2,567.69	227.8	37
	core	3.45	55.27	280.03	2,325.37	327.2	43
	top deg	4.22	58.84	289.49	2,414.99	331.7	47
WIKI-VOTE	**truss**	2.92	15.27	42.46	560.66	114.9	52
	core	1.92	10.65	32.40	466.01	104.5	57
	top deg	2.43	12.05	35.55	502.88	104.5	62

To evaluate the performance of the methods, we perform the SIR simulation starting from a single node each time. For each node, we repeat the simulation 100 times to get the average behavior of the node. For each of the settings, we repeat the above for all the respective nodes and calculate the average behavior for the nodes of each set. Results from the experiments are shown in Table 2. The **truss** method achieves significantly higher infection rate during the first steps of the epidemic. Also, the total number of infected nodes at the end of the process is larger (*Fin. step*), while the fade out occurs earlier (*Max step*). Lastly, the number of nodes in the truss set T is much smaller compared to the set $C - T$ (Table 1). By refining significantly the set of influential nodes in truss set T, the "weaker" spreaders of C are left in core set $C - T$ explaining the inferior behavior of the **core** method compared to the **top degree**.

We have also computed the cumulative differences of the number of infected nodes per step achieved by the methods. Let $I_t^{\textbf{truss}}$ be the number of infected nodes at step t achieved by the **truss** method (similar for **core** and **top degree**). We define the cumulative difference for the **truss** and **core** methods at step t as $\mathcal{D}_t^{\textbf{truss-core}} = \text{cumsum}_{z=1\ldots t}(I_z^{\textbf{truss}} - I_z^{\textbf{core}})$ (similar for **truss** vs. **top degree**). The results are shown at Fig. 1. We observe that the cumulative difference of the number of nodes that are being infected at every step is always bigger between **truss** and **core** than between **truss** and **top degree**. Both differences increase during the outbreak of the disease until they stabilize to the number of nodes which is actually the final difference of the number of nodes that got infected during the epidemic of the two compared methods. Clearly, as the differences are always above zero, one can conclude to the effectiveness of information diffusion when starting from nodes belonging to **truss**. Worth mentioning is that on average, the epidemic stops at an earlier time step when the spreading is triggered from the **truss** nodes.



(a) EMAIL-ENRON: $\beta = 0.01$ (b) EMAIL-ENRON: $\beta = 0.03$

(c) EPINIONS: $\beta = 0.007$ (d) EPINIONS: $\beta = 0.01$

(e) WIKI-VOTE: $\beta = 0.009$ (f) WIKI-VOTE: $\beta = 0.01$

Figure 1: Cumulative differences of infected nodes per step achieved by the truss method vs. the core (truss - core) and top degree (truss - degree) methods.

3. CONCLUSIONS

In this work, we showed that the K-truss decomposition of a network can help towards identifying single influential spreaders. K-truss, being a subset of the k-core of the network, contributes in the reduction of the set of privileged spreaders for information diffusion [3]. Using the SIR epidemic model, we show that such spreaders will influence a greater part of the network during the first steps of the process, but will also cover a larger portion of it at the end. Future research includes finding multiple influential spreaders utilizing the properties of k-core and K-truss decompositions.

Acknowledgments. Maria-Evgenia G. Rossi is partially funded by a DigiCosme Ph.D. Fellowship. Fragkiskos D. Malliaros is a recipient of the Google Europe Fellowship in Graph Mining, and this research is supported in part by this Google Fellowship.

4. REFERENCES

[1] D. Chakrabarti, Y. Wang, C. Wang, J. Leskovec, and C. Faloutsos. Epidemic thresholds in real networks. *ACM Trans. Inf. Syst. Secur.*, 10(4):1:1–1:26, 2008.

[2] E. David and K. Jon. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, New York, NY, USA, 2010.

[3] L. Gallos, S. Havlin, M. Kitsak, F. Liljeros, H. Makse, L. Muchnik, and H. Stanley. Identification of influential spreaders in complex networks. *Nature Physics*, 6(11):888–893, Aug 2010.

[4] S. B. Seidman. Network structure and minimum degree. *Social Networks*, 1983.

[5] J. Wang and J. Cheng. Truss decomposition in massive networks. *Proc. VLDB Endow.*, 5(9):812–823, 2012.