SPREADVIZ: Analytics and Visualization of Spreading Processes in Social Networks

Konstantinos Skianis, Maria Evgenia G. Rossi, Fragkiskos D. Malliaros and Michalis Vazirgiannis
Computer Science Laboratory
École Polytechnique, France
Email: {kskianis, rossi, fmalliaros, mvazirg}@lix.polytechnique.fr

***Abstract*—In this paper, we propose SPREADVIZ, a web tool for exploration and visualization of spreading properties in social networks. SPREADVIZ consists of three main modules, namely graph exploration and analytics, detection of influential nodes, and interactive visualization. More precisely, SPREADVIZ offers the following functionalities: (i) It computes and visualizes various centrality criteria towards understanding how the position of a node in the network affects its spreading properties; (ii) It offers a wide range of criteria for the detection of single and multiple influential nodes and comparison among them; (iii) It effectively visualizes the spread of influence in the network as well as the performance of each method. In our demonstration, we invite the audience to interact with SPREADVIZ, exploring, analyzing, and visualizing the spreading processes over various real-world social networks.**

***Keywords*—Social Network Analysis; Graph Mining; Influential Spreaders; Influence Maximization**

I. INTRODUCTION

Spreading processes in social and interaction networks have gained great interest in the research community due to the plethora of applications that they occur. Characteristic examples include the spread of news, ideas and rumors in social networks, influence propagation as well as disease spreading. Being able to model and analyze the underlying mechanisms that occur in such processes is a crucial task with direct applications in a wide range of interdisciplinary fields, including social network analysis, epidemiology, viral marketing and computational social science.

In the core of all those application domains lie the identification of *influential nodes*, that are able to spread information to a large portion of the network. For example, in the domain of viral marketing, we are interested to promote a product in order to be adopted by a large fraction of individuals in the network. The basic idea behind viral marketing is the *word-of-mouth* effect, where individuals that have already adopted the product, recommend it to their own friends forming a cascade of recommendations. The fundamental question behind viral marketing is how to efficiently locate a few initial individuals with good spreading properties, that will lead to an effective product promotion campaign by maximizing the spread of influence in the network.

The task of identifying influential nodes in networks can be sub-categorized in two subtopics: (i) identification of single influential spreaders and (ii) identification of a group of spreaders that maximize the total spread of influence in the network. For example, in disease spreading, the process is typically triggered by a single individual node in the network. To this direction, several node centrality criteria have been proposed, including degree, betweenness and PageRank centralities [12], as well as criteria based on the concept of graph decomposition, such as the ones of k-core and K-truss decomposition [9]. On the other hand, in the case of viral marketing, the goal is to convince a small subset of individuals to adopt a new product, in such a way that, at the end of the process, a large number of individuals will be influenced. The latter problem is known as *influence maximization* and approximation algorithms have been proposed [6].

In this paper, we propose SPREADVIZ, a web-based tool for analytics and visualization of spreading processes in complex networks. SPREADVIZ consists of three main modules and offers the following functionalities:

- *Graph exploration and analytics*: analysis of the structural characteristics of the graph (e.g., degree distribution, distribution of PageRank scores), which are used for both graph exploration as well as for the detection of influential nodes.
- *Detection of influential nodes*: SPREADVIZ offers a wide range of criteria for the detection of both single influential nodes [9], [12] as well as for multiple influential nodes. In the *system-guided* spreading mode, SPREADVIZ detects the best spreaders of a network based on list of node importance criteria. Furthermore, combining the results produced by the graph exploration module, SPREADVIZ offers *user-guided* spreading, where the user selects the nodes of interest and then, the system computes the corresponding influence.
- *Interactive visualization*: SPREADVIZ combines data produced by the previous two components, to offer an interactive visualization to the end-user. For a given graph, SPREADVIZ displays plots of the structural characteristics of the graph, that can give further insights about the spreading properties of the nodes (e.g., distribution of the k-core numbers of the nodes). The end-user can examine how the epidemic process, that starts from a specific node or a set of nodes, spreads over the network step

Corresponding author: Fragkiskos D. Malliaros (fmalliaros@lix.polytechnique.fr)

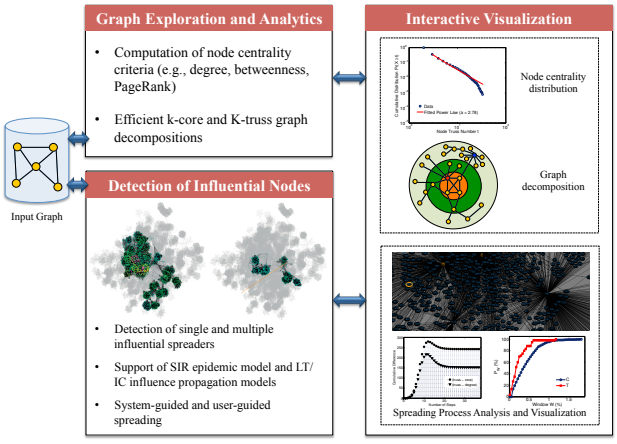

Fig. 1. Overview of SPREADVIZ.

by step. Furthermore, SPREADVIZ plots characteristics of the spreading properties that can be useful to the data analyst (e.g., number of infected nodes per step of the spreading process).

The current version of SPREADVIZ can be accessed at: www.lix.polytechnique.fr/dascim/demos/SpreadViz. The rest of the paper is organized as follows: Sec. II gives a detailed overview of SPREADVIZ, Sec. III provides the demonstration plan and finally we conclude in Sec. IV.

II. SYSTEM OVERVIEW

In this section, we provide an overview of SPREADVIZ system. In particular, after introducing the necessary background, we describe the three main components of SPREADVIZ ((i) graph exploration and analytics; (ii) detection of influential nodes; (iii) interactive visualization), towards exploratory analysis of the spreading processes in networks. As the interactive visualization component of SPREADVIZ is used by both the other modules, we describe its main functionalities alongside each one of them. Figure 1 depicts an overview of the SPREADVIZ system.

A. Preliminaries

Let $G = (V, E)$ be an undirected graph. Below, we provide the definitions of the node centralities that are most commonly used towards locating influential spreaders in networks.

Degree **(d_v).** Each node $v \in V$ has a *degree* $d_v = d$ if it is connected with d nodes in the graph.

Core number **(c_v).** 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.

Truss number **(t_v).** 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.

Betweenness centrality **(b_v).** Let $\sigma_{uw} = \sigma_{wu}$ denote the number of the shortest paths from u to w, where $\sigma_{uu} = 1$ by convention, and let $\sigma_{uw}(v)$ denote the number of shortest paths from u to w that some $v \in V$ lies on. Then *betweenness* centrality is defined as $b_v = \sum_{u \neq w \neq v \in V} \frac{\sigma_{uw}(v)}{\sigma_{uw}}$.

PageRank score **(p_v).** Let $A = (a_{ij})$ be the adjacency matrix of a directed graph. The *PageRank* centrality of node v is given by: $p_v = \alpha \sum_u \frac{a_{uv}}{d_u} p_u + \beta$ where α and β are constants and d_u is the out-degree of node u if such degree is positive, or $d_u = 1$ if the out-degree of u is null.

B. Graph Exploration and Analytics

The first module of SPREADVIZ provides an exploratory analysis and basic information for the given network by computing, analyzing and visualizing (i) the distribution of the node centralities and (ii) the graph decomposition using the k-core and K-truss decompositions.

1) Node centrality distribution: The first functionality of the module computes the distributions of the different node centralities of the graph (i.e., *degree*, *core number*, *truss number*, *betweenness* and *PageRank*) that serve as influence metrics in the second module of the system (detection of influential nodes), and visualizes the results. This first exploratory step is of great importance in order for the user to be able to observe the structural characteristics of the graph and the behavior of these node centralities of the network that are frequently used to locate influential spreaders.

2) Graph decomposition: As we will present in Sec. II-C, the k-core number of a node plays a significant role towards its influential power. For that reason, the second functionality was chosen to provide a schematic representation of the network under the k-core decomposition [1]. After computing the decomposition of the graph using the efficient algorithm of [1], SPREADVIZ visualizes the graph coloring each node according to its k-core number.

C. Identification of Influential Nodes

One of the most common tasks in the spreading process analysis is the identification of those nodes that will maximize information diffusion throughout the network, which constitutes the second component of SPREADVIZ. As we have already discussed, the problem is further split into the identification of individual influential nodes and the identification of a set of nodes that can maximize the total spread of influence, commonly known as the *Influence Maximization* problem. Both tasks are supported by SPREADVIZ.

Towards the first direction, several approaches have been proposed. The majority of them are considering node centralities to rank a node's effectiveness as a spreading predictor. A straightforward approach has been to consider the one of *degree centrality* [11]. Nevertheless, there exist cases where a node can have arbitrarily high degree, while its neighbors are not well-connected. Based on this fact, global node centrality criteria have been proposed for the problem of influential node detection. Those include using *closeness*, *betweenness* [4] centralities as ranking methods as well as other heuristic algorithms. Random-walk based methods such as well-known *PageRank* [4] and *LeaderRank* [8] have also received great attention. Of particular importance is the work by Kitsak et al. [7], showing that less connected but strategically placed nodes in the core of the network are able to disseminate information to a larger part of the population. To quantify the core-periphery structure of networks, the graph-theoretic notion of *k-core decomposition* [1] is applied. That way, the nodes that belong to the maximal k-core subgraph (i.e., maximum value k of the decomposition) are able to infect a larger portion of the network, compared to other well-know centrality criteria, such as node degree or betweenness centrality in a more efficient way.

However, it is quite usual for a large number of nodes to belong to the maximal k-core subgraph even if they differ with respect to their spreading capabilities. To deal with this issue, a triangle-based extension of the k-core decomposition has been used, namely the *K-truss decomposition* [3], which has be shown to detect nodes that show better spreading behavior compared to the previously described criteria, leading to faster and wider epidemic spreading [9].

Concerning the case of multiple influential spreaders, Kempe et al. [6] are the first to formulate influence maximization as a combinatorial optimization problem. They considered probabilistic cascade models from the sociology and marketing literature – which are presenting in a following paragraph – in order to simulate a spreading process. While they introduce a *Greedy* approach which provides (1-1/e-ϵ)-approximate solutions, their algorithm is computationally expensive thus inefficient even for networks of a few thousands of nodes and edges. The research community has since been focused on introducing algorithms that reduce the computation overhead of influence maximization [2], [5].

In SPREADVIZ, the user can specify which of the above-mentioned criteria will be used for the detection of single influential nodes. Then, after the nodes of interest have been extracted, SPREADVIZ simulates the spreading process over the network, in order to determine the spreading effect of the selected nodes using epidemic models, such as the SIR model (see Sec. II-D). For the case of multiple influential nodes, the SPREADVIZ automatically determines the set of k nodes that maximize the spread of influence, using the *Greedy* algorithm and its variants [5].

D. Simulating spreading processes

1) Epidemic models: The most common epidemic models that are used to simulate a spreading process are the Susceptible-Infected-Recovered (SIR) and Susceptible-Infected-Susceptible (SIS) models [10], where the nodes can be in one of the states that the names suggest. In the current version of SPREADVIZ, we have mainly focused on the SIR mode, as it is widely used in the related literature [12]. In this model, initially a single or a set of nodes are set to be in the infected state 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 some probability γ. In SPREADVIZ, the user can specify both the infection and recovery probabilities. As a rule of thumb, it is suggested to set the parameter β close to the epidemic threshold of the graph [4] and parameter γ close to $\gamma = 0.8$, in order to amplify the influential properties of the nodes [7].

2) Linear Threshold model (LT): In this model, a node v is influenced by each neighbor u according to a weight b_{vu} [6]. The value of this weight is such that the sum of all the weights towards all neighbors of v is less or equal to 1. Each node v chooses a threshold θ_v uniformly at random from the interval $[0, 1]$ which represents the weighted fraction of v's neighbors that must become active in order for v to become active. Given a random choice of thresholds and an initial set of active nodes (with all other nodes being inactive), the diffusion process unfolds deterministically in discrete steps. A node v can be activated when the total weight of its active neighbors is at least θ_v.

3) Independent Cascade model (IC): In this model, when a node v first becomes active in timestep t, it is given a single chance to activate each neighbor u – which is currently inactive – and succeeds with a probability $p_{v,u}$. If v succeeds, then u will become active in the next timestep. If v does not succeed it cannot further attempt to activate u in future timesteps. The process runs while node activations are possible.

E. Spreading Process Analysis and Visualization

SPREADVIZ computes the individual influence power of every node in the network. While there exist many models that simulate a spreading process, the user is capable of choosing the model of their preference – among those frequently used in most studies: *SIR/SIS*, *LT*, *IC* – in order to calculate the influence power of all the nodes of the network. The results are given as a plot of the graph where each node is colored accordingly to its influential power. It also provides further functionalities about visualization and analytics of spreading processes in the given network. The spreading process that is simulated can be triggered by a single or multiple nodes. The model that is used in most of the functionalities of the module to simulate the spreading process is the *SIR* model. Nevertheless, as mentioned above, for the case of multiple spreaders there exist methods that provide the optimal influential nodes by assuming the spreading process evolves as

the *LT* and *IC* models suggest. For that reason, when multiple nodes are chosen to start a spreading process, there is the possibility to use the aforementioned models in order to have a fair comparison with the respective methods.

1) Single Spreaders: The module also provides a comparison of the most common methods used to identify single influential spreaders that can trigger a fast and efficient spreading process in the network. The comparative results are visualized as the per-step or cumulative influence – in number of nodes – triggered by spreaders chosen as the different methods suggest. As the *SIR* model that is used to simulate the spreading is a probabilistic model, a comparison to the average performance after 100 or 1000 times (depending on the size of the dataset) is provided. That way, the best influential node as identified by the different criteria can be found (*system-guided* spreading). There is also a choice for the user to choose a node of his/her preference or choose a node with specific centrality characteristics of his/her choice and observe the information flow on the network for every step of the process (*user-guided* spreading). Furthermore, feeding those results to the visualization module, the end-user is able to observe the influence network and its evolution that is caused by the propagation of information initiated by a single spreader (i.e., the nodes of the graph that have been influenced by the propagation are colored accordingly).

2) Multiple Spreaders: This functionality visualizes the information flow triggered by multiple nodes of the user's choice. The results can also be provided in the form of a plot where the per-step and cumulative influence is provided, as in the single spreader case. While for the above only the *SIR* model is used to simulate the spreading, as an extra functionality our system can provide a comparison of the influence induced by the nodes chosen by the methods using the *LT* and *IC* models and the set of nodes chosen by the user. Lastly, the *user-guided* spreading is also an option for the multiple spreaders case.

F. Implementation Details

Our tool is essentially a browser-based application, without any need for the user to install anything. It uses D3.js for the visualization part, which is a JavaScript library for producing dynamic, interactive data visualizations in web browsers. It can be viewed in Google Chrome and Mozilla Firefox.

III. Demonstration of SpreadViz

A. Present State of Demo

In the current state of SpreadViz, the system is demonstrable, as some of the basic features have been already implemented. The user can analyze the spreading properties of a pre-loaded graph. With respect to the graph exploration and analytics module, the degree, PageRank and k-core metrics have been implemented along with the SIR epidemic model. The pathway of information spreading is also visualized in the network. In the remaining time upon acceptance of the demo, we will implement the remaining features of SpreadViz, namely the IC and LT models for influence maximization and the visualization of the plots (number of infected nodes per step of the epidemic process).

B. Demonstration Plan

The audience will be invited to load and process a medium size network, which will be imported as an edge-list format. Then, SpreadViz will perform exploration and visualization of the structural properties of the underlying graph. Based on this step, the audience can get further insights about the spreading performance of the nodes, as spreading capability of the nodes is highly related to the notion of centrality. Then, the audience will be able to (i) locate the most influential spreader or the set of k nodes that maximize the spread of influence in the network, using various algorithmic approaches (e.g., nodes of the maximal k-core subgraph combined with the SIR model), (ii) examine how influence spreads over the network (based on the visualization tool) and compare the performance of the difference criteria, (iii) or even to determine the spreading properties of a set of knows that are chosen manually.

1) Equipment: For the demonstration, we will bring our laptop into which SpreadViz will run. As additional equipment, we will need either a monitor or a projector.

IV. Conclusion

In this demo paper, we have proposed SpreadViz, a web tool for exploring, analyzing and visualizing spreading processes in social networks. In particular, SpreadViz explores centrality criteria of the nodes of the graph as they are related to spreading properties, it offers a wide range of criteria for the detection of single and multiple influential spreaders, it supports both system-guided and user-guided spreading, and finally it visualizes the results in an interactive manner.

References

[1] Vladimir Batagelj and Matjaz Zaversnik. An O (m) algorithm for cores decomposition of networks. arXiv, (2003).

[2] Wei Chen, Yifei Yuan, and Li Zhang. Scalable influence maximization in social networks under the linear threshold model. In *ICDM*, 2010.

[3] Jonathan Cohen. Trusses: Cohesive subgraphs for social network analysis. National Security Agency Technical Report (2008).

[4] Easley, David, and Jon Kleinberg. Networks, crowds, and markets: Reasoning about a highly connected world. Cambridge University Press, 2010.

[5] Amit Goyal, Wei Lu, and Laks VS Lakshmanan. Celf++: optimizing the greedy algorithm for influence maximization in social networks. In *WWW*, 2011.

[6] David Kempe, Jon Kleinberg, and va Tardos. Maximizing the spread of influence through a social network. In *KDD*, 2003.

[7] Maksim Kitsak, Lazaros K. Gallos, Shlomo Havlin, Fredrik Liljeros, Lev Muchnik, H. Eugene Stanley, and Hernn A. Makse. Identification of influential spreaders in complex networks. Nature Phys, 6.11 (2010): 888-893.

[8] Linyuan L, Yi-Cheng Zhang, Chi Ho Yeung, and Tao Zhou. Leaders in social networks, the delicious case. PloS One, 6.6 (2011): e21202.

[9] Fragkiskos D. Malliaros, Maria-Evgenia G. Rossi, and Michalis Vazirgiannis. Locating influential nodes in complex networks. Sci. Rep. 6 (2016).

[10] Mark EJ Newman. Spread of epidemic disease on networks. Phys. Rev. E, 66.1 (2002): 016128.

[11] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. Phys. Rev. Let., 86.14 (2001): 3200.

[12] Sen Pei and Hernn A. Makse. Spreading dynamics in complex networks. Journal of Statistical Mechanics: Theory and Experiment 2013.12 (2013).