INFLUENTIAL NODES PREDICTION USING LINKS INFORMATION IN SOCIAL NETWORKS

Klotilda Nikaj
University of Tirana, Albania, klotildanikaj@hotmail.com
Margarita Ifti
University of Tirana, Albania, margarita.iftit@rambler.ru

Abstract: Identifying influential nodes and measure the influence of nodes in social networks, has been inspired by analogies between social behavior and statistical mechanics. Social interactions among humans create complex networks, and despite an increase of online communication, the interaction between physical proximity remains a fundamental way for people to connect. Here we can initiate a research on the foundations of ranking nodes, a fundamental ingredient of analyzing social systems. In order to understand the essence and the exact rationale of node ranking algorithms we suggest the axiomatic approach of agent based model taken in the formal theory of social choice. Based on essential factors of influence propagation (such as the location and neighborhood of source node, propagation rate) and network invulnerability, we propose a novel strategy to search the influential nodes in terms of outgoing and ingoing links to the node. The aim of this work is to identify the influential nodes as they affect the hierarchical structures of the network. By analyzing the data and describing how these nodes affect the network structure, we aim to obtain new tools and methodology which will help us to describe how networks grow and fall apart in smaller structures, which have similar features with the large network, but different dynamics. In order to characterize this phenomenon and explore the correlation between collective behaviors and locally interacting elements, we use statistical methods and visualization software as a combined approach to understand the behavior of the network for a given behavior of the influential nodes that we use to recreate our network. The results of our research on real-world networks' dataset show that the proposed method outperforms state-of-the-art influence algorithms.
Keywords: social systems, influential nodes, statistical mechanics

1. INTRODUCTION

Nowadays, more than ever, the world seems to be a complex social system[1], whose behavior is primarily the result of the behavior of social agents. Detecting influential agents [2,3,2]depends heavily on one basic premise about the structure of the network: Nodes that are more useful to people will also be more popular, and will accordingly have more links pointing to them from other nodes. A very simple approach to figure out which nodes are most useful/important would be to count the number of incoming links to node, and use that as a ranking score. However, this would be assuming that every link counts equally, which is quite wrong. A single link from an important node should count for much more than a link from some little-known nodes that presumably no one seems interested in. Thus, a network is important if many (and/or important) nodes link to it. This appears to be a rather circular definition of importance, and begs the question: how can we tell which nodes are important to begin with? Node Rank detection[1,2,3] handles this problem by initially ranking all nodes as equally important, but then it repeatedly performs a process on the rank scores of the nodes that will cause the importance rankings to change. This model presents a different way of calculating rank that would eventually converge to the exact same rankings being assigned to each network, if we could let the algorithm run forever.
Based on essential factors of influence propagation (such as the location and neighborhood of source node, propagation rate) and network invulnerability, we propose a novel strategy to search the influential nodes in terms of the local and global topology[3,4]. Two important indicators are node diffusion degree and node cohesion degree, which are used to increase the probability of influence diffusion and reduce the feasibility of network collapse.

2. MATERIALS AND METHODS

The proposed method, represents agents, arranged in a network, provided with an individual behavior, that change rank in function of the outgoing and ingoing links. Ranking nodes[4,5] is a technique for ranking the relevancy of nodes on the network, through analysis of the link structure that links nodes together. This model demonstrates one agent-based method for calculating the Rank of interconnected nodes. The use of an agent-based perspective attempts to provide a deeper understanding of this algorithm and the mathematics behind it. However, Ranking Nodes it is technically a ranking algorithm, which provides importance weights for each node in a network. These rankings turn out to be very useful when performing an internet search, because they can be used to help determine the order in which search results are displayed to the user. Yet, many of these important nodes are similar, meaning

that they can be transformed into one another through continues topological deformations. At the same time, a network can also have multiple non similar embeddings, each defining a distinct topological deformation class. To determine whether two network embeddings are non-similar, we start from the linking number, that measures the number of times two closed cycles wind around each other, capturing the number of tangles. The graph linking number, which for a network with embedding represents the sum of the linking numbers of all pairs of nodes in the graph

$$G(\varepsilon) = \sum_{c,c' \epsilon \{C\}} l(\varepsilon, c, c') \quad (1)$$

Where $\{C\}$ is the set of the cycles in the network, determined only by the adjacency matrix, and $l(\varepsilon, c, c')$ is the linking number between cycle c and c'.

Usually larger networks have more cycles, hence we expect more potential tangles between them, which would lead to higher values G , that's why is better to use the normalized graph linking number

$$G_n(\varepsilon) = \frac{G(\varepsilon)}{N_p} \quad (2)$$

We have to limit the above equations to find the minimal loop set, which is a computationally expensive problem that prompts us to use the method of spanning trees to sample the minimal loops. In physical networks, the links do not have arbitrary lengths. We therefore measured the total elastic energy of the layout, representing the sum of the elastic energies of all links l as defined below:

$$V_{el}[\{\gamma_l\}] = \sum_l \int_{\gamma_l} \frac{dx_l}{ds}^2 ds \quad (3)$$

Where the integral is over the path γ_l for link l. $s\epsilon[0,1]$ parameterizes the length of the link and $x_l(s)$ is the location of the segment. To avoid the crossing of links and nodes , we add a short range node-node repulsion $V_{nn} \approx A_n \sum_{ij} \exp\{-[(X_i - X_j)/2r_n]^p\}$ and the link-link repulsion $V_{ll} \approx A_{nl} \sum_{lm} \iint ds_l\, ds_m \exp\{-[(X_i - X_j)/2r_n]^p\}$ with $p \geq 2$, where the A_n, A_l are the amplitudes for the potentials X_i, X_j are the location for node i, j, x_l, x_m are directed segments on link l, m, r_N, r_l are parameters for node and link interaction ranges, and the exponent p determines how hard or soft the potentials are. As the total elastic energy increases monotonically with the total link length. Indeed, all similar layouts can be continuously transformed into one another, implying that they belong to the same energy well[7,8].

In the diffusion approach, the nodes themselves are the central agents we are concerned with. Each node starts with some rank value, which is a measure of how important it is in the network. Initially, every node gets the same rank value as every other node, and the sum of all the nodes rank values is 1. Then, in each time step, every node distributes its rank value (influence) to those nodes that it has outgoing links to[11,12,13]. Each node's new rank value will thus be based on how much rank it receives from each of the nodes that link to it, combined in a weighted average with a baseline amount of rank value which each website gets each time step regardless of its neighbors. Over time, this process causes the rank values of each node to converge to the actual values for each node. In more formal mathematical terminology, this method is similar to using the "power method" for finding the principal eigenvector associated with a modified adjacency matrix of the directed link graph.

Data source. To test the effectiveness of our research, we conduct experiments on real world networks. The real world network dataset that we use include one online social network [6] that is composed by 4039 nodes and 88234 edges.

3. RESULTS

By combining statistical mechanics analysis and the centrality metrics[9,10] for the network, we can identify and highlight the most influential nodes, based on the number of ingoing and outgoing links of the nodes. At the same time, by deleting from the network the most influential nodes, we can change or even destroy the network topology of the system (Fig.3)

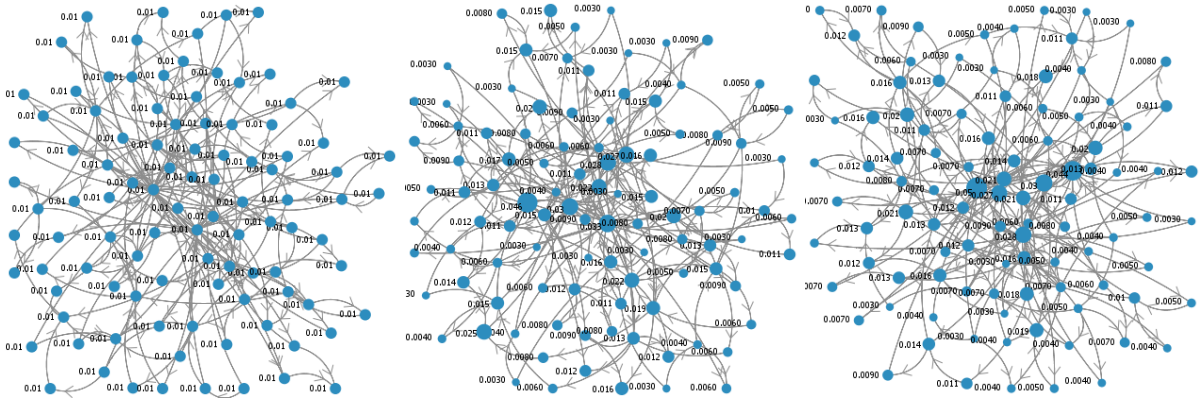

Figure 1 a) Initially ranking all nodes as equally important, where the diffusion probability is uniform for all the links b) Repeatedly performing the process on the rank scores of the nodes that will cause the importance rankings to change. c) Here we have performed the same steps as before, but we have changed the damping factor that expresses the probability at each step that the surfer will not continue with a link but will jump to a random node.

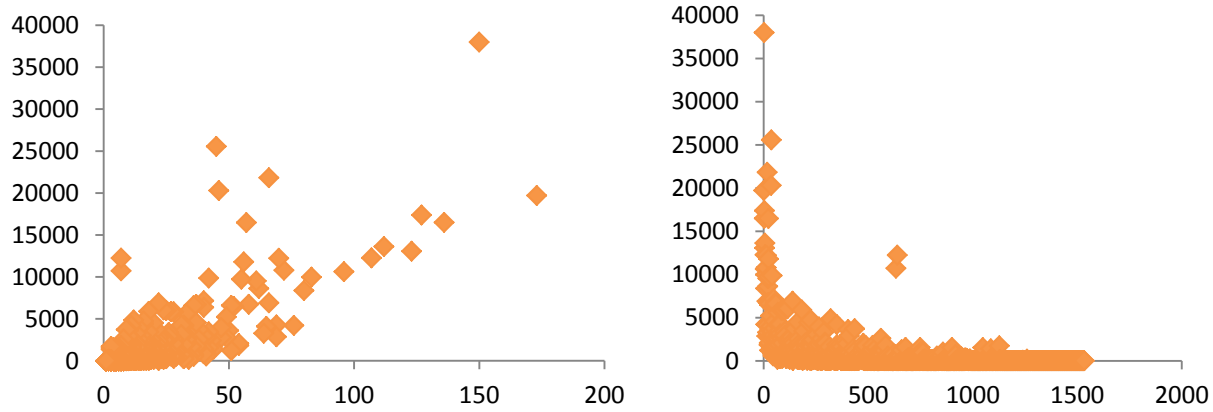

Figure 2. a)the nodes influence; b) The average distance of a node to all other nodes.

In the fig.2 a) we have computed the influence of the nodes in the network, by assuming that a node is central if it lies between many other nodes, and the linking number is given by equation (2). Fig 2. b) shows the average distance of a node to all other nodes which can affect the communicating information among the nodes in the graph. For every pair of nodes in a connected network, there exists at least one shortest path between the nodes such that either the number of links that the path passes through (or the sum of the weights of the edges (for weighted graphs) is minimized. We have compute the shortest-path between influential nodes, representing the degree of which nodes stand between each other.

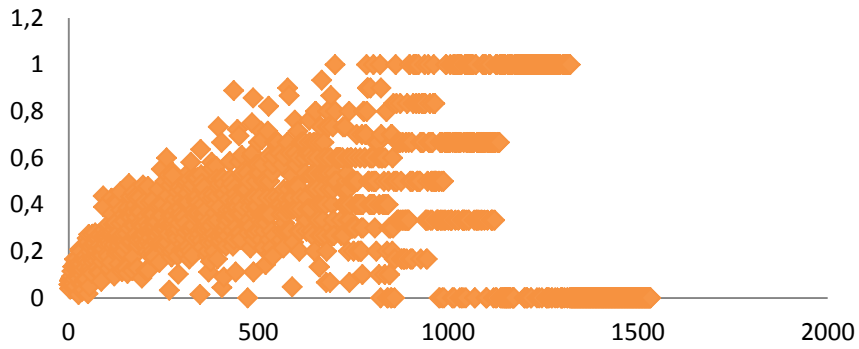

Figure 3 c) The importance of a node taking into consideration weighted links.

Here we have computed the eigenvector centrality that measures the importance of a node by assuming links from more central nodes contribute more to its ranking than less central nodes. Links are calculated again based on the equation (2).

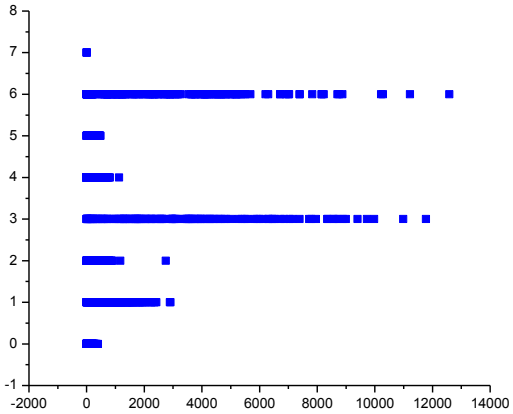

Figure 4. The number of influential nodes, and the number of links for each influential node.

The detected influential nodes are the ones that have great influence and can resist certain damage and disturbance of the networks. To reinforce this result, by deleting the most influential nodes of the network we have changed the topology of the network.

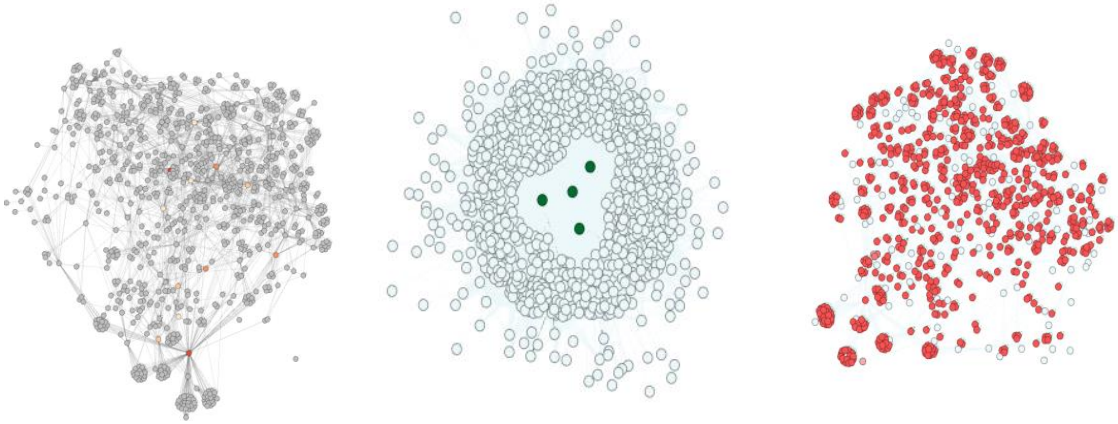

Figure 5: The network topology in three different states: a) we have some initial influential nodes on the network. b) the whole nodes of the network are connected with these influential nodes. c) By deleting the most influential nodes of the network, the network topology is transformed.

4. DISCUSSIONS

Our results show that nodes identified as influential by our method are substantially different from those by each of the conventional methods. We believe that statistical mechanics analysis effectively supports the analytical process of influential nodes detection. It is important to estimate the probability as accurately as possible in finding the influential nodes, since the probability affects the ranking. A set of interfaces provide a mechanism for guiding the end user to focus their attention on those decisions that can make those most difference.

We can see that introducing statistical mechanics in the problem of influential nodes may be a good option in order to improve the diversity of links to be predicted, which opens interesting leads for future works on the topic. Anyway, the current prototype analysis does have limitations. The most evident one is that it is most effective for smaller networks as response time of each interaction goes up with network size. In particular, finding a way to compute automatically relevant activity thresholds would be a significant improvement to the current version of the influential nodes analysis using statistical mechanics.

5. CONCLUSIONS

In this paper, we have presented and analyzed a simple agent based model combined with statistical mechanics for detecting influential nodes in social systems. We have applied these in real networks (see fig.4 and fig. 5) in the simplest setting where the diffusion probability is uniform for all the links. Further showed that the proposed method can predict the high ranked influential nodes accurately. An interesting direction for future work is to investigate which are the most general influence models for which provable approximation guarantees can be achieved.

NOTE

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s). Of course we welcome any kind of suggestion to improve this material.

REFERENCES

Arnoux, T., Tabourier, L., and M. Latapy. (2018). *Predicting interactions between individuals with structural and dynamical information*. CoRR

Arnoux, T., Tabourier, L. and M. Latapy. (2017). *Combining structural and dynamic information to predict activity in link streams*. In ASONAM

Albert, R. & Baraba´si, A.-L. (2002). Statistical mechanics of complex networks. Rev. Mod. Phys. 74, 47–97

Backstrom, L., & Leskovec, J. (2011). "Supervised Random Walks: Predicting and Recommending Links in Social Networks", ACM Internation Conference on Web Search and Data Mining (WSDM)

Barrat, A., Barthelemy, M. & Vespignani, A. (2008). Dynamical Processes on Complex Networks (Cambridge University Press, 2008).

Brandes, U. (2001). *A Faster Algorithm for Betweenness Centrality*. Journal of Mathematical Sociology 25(2):163-177, https://doi.org/10.1080/0022250X.2001.9990249

Brandes, U. (2008). *On Variants of Shortest-Path Betweenness Centrality and their Generic Computation. Social Networks* 30(2):136-145,.https://doi.org/10.1016/j.socnet.2007.11.001

Baytas, I. M., Xiao, C., Zhang, X., Wang, F., Jain, A. K., & Zhou. J. (2017). Patient subtyping via time-aware lstm networks. In *KDD*,

Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2017). Anyone can become a troll: Causes of trolling behavior in online discussions. In CSCW, [10] J. Cheng, C

Goyal, A., F. Bonchi, L.V.S. (2010). Lakshmanan. *Learning influence probabilities in social networks*. In Proc. WSDM

Leskovec, J., & Krevl, A. (n.d.). http://snap.stanford.edu.data

Liben-Nowell D., & J. Kleinberg. (2003). *The link prediction problem for social networks*. In IKM ’03, pages 556–559

Newman, M. E. J. (2010). Networks: An Introduction (Oxford University Press)