Chapter 32
Detecting Influential Nodes in Complex Networks with Range Probabilistic Control Centrality

Dimitrios Katsaros and Pavlos Basaras

32.1 Introduction and Motivation

Real-world entities often interconnect with each other through explicit or implicit relationships, by transient and continuous ways to form a complex network. Such networks are studied in many fields of science like bioinformatics, statistical mechanics, sociology, and computer science [1]. Complex networks have provided a wealth of evidence for their ability to disseminate information rapidly among other node users [2]. Rumors and fashion, but also social unrest or the spreading of infectious diseases among those people networks, highlight the need for identifying those entities to either boost or hinder spreading. A great deal of research into the structure of complex systems has focused on trying to identify such entities in an attempt to efficiently control complex systems. For the identification of such entities, traditional centrality measures have been proposed such as the shortest-path betweenness centrality or spectral centrality measures, e.g., PageRank [3]. More sophisticated methods for the detection of influentials are reported in [4, 5]

The common characteristic of the aforementioned efforts is that they all deal with static complex networks, i.e., they apply for a specific instance of the network's lifetime; at that specific instance, a link between a specific pair of nodes either exists or not. However, many of the real-world complex networks are continuously evolving, and their links rapidly appear and disappear. Examples of such complex systems are vehicular ad hoc networks [6] whose links live for only a few seconds and are usually characterized by a link quality parameter, ranging from a zero value indicating an absent link, to a value of one, indicating a perfect communication link. Moreover,

D. Katsaros (✉) · P. Basaras
Electrical and Computer Engineering Department,
University of Thessaly, Volos, Greece
e-mail: dkatsar@inf.uth.gr

P. Basaras
e-mail: pabasara@uth.gr

J.H. van Schuppen and T. Villa (eds.), *Coordination Control of Distributed Systems*,
Lecture Notes in Control and Information Sciences 456,
DOI 10.1007/978-3-319-10407-2_32

many evolving complex networks are examined from an 'aggregated' perspective, associating to each link the percentage of time that this link existed.

The study of influentials in complex networks with probabilistic links is a challenging, new task, because apart from the 'neighborhoods' that an influential can exert influence, we should also take into account the 'strength' of the links. We could resort to the old ideas finding stochastic shortest paths and computing analogous betweenness centralities, but these centralities have already been shown that they do not perform well for static networks either [5].

In this article, we develop a semi-local centrality measure for dynamic, complex networks with probabilistic links, the *range probabilistic control centrality* (RPCC), which considers both the 'strength' of links emanating from each node, and it additionally estimates the influence region of the node based on ideas from the literature of control theory. In the absence of relevant methods, the proposed centrality measure is compared against a baseline method, namely the localized weighted-degree centrality [7], for a couple of networks with various distributions for the probabilities of the links.

32.2 Utility Examples

Consider the vehicular ad hoc network (VANET) where the existence and quality of connections between vehicles (e.g., time of active connection, signal strength) is a factor of several parameters such as the vehicle's direction, acceleration–deceleration of vehicles, the underlying road network topology, possible obstacles or interference, and so on. The aggregate effect of these factors results in having a temporal network. A vital operation in a VANET is that of locating the nodes that can disseminate a safety message to as many vehicles as possible within the whole network or focused parts of it, e.g., safety geocasting messages.

Apart from these ad hoc communication networks, a wide variety of complex systems in nature, society, and technology can be represented as graphs with entities linked by probabilistic edges. A couple of other examples include a transportation or airline network [8], where schedules of transportations vary or change, and examples of phone, email, or social networks, depicting contacts as entities and the amount of time of their interaction as their links strength [9], and we need to determine the entities that can exert the maximum influence over the network. Earlier works such as betweenness centralities based on stochastic shortest paths suffer from the inability to detect influential spreaders [5]. Recent efforts using positive and negative links [10] are not rich enough to address the present problem.

The present article investigates the issue of detecting influential nodes in temporal networks with probabilistic links and makes the following contributions:

- Investigates the issue of influential spreaders in complex networks with probabilistic links.

- Extends the concept of control centrality [11] and proposes an adjustable centrality measure, the range probabilistic control centrality (RPCC), based on control theory, to help identify such nodes.
- Evaluates this centrality measure across a range of complex networks and distributions of probabilities over the links, and compares it with a baseline method, namely weighted-degree centrality [7].

32.3 Range Probabilistic Control Centrality

The concept of *control centrality* was introduced in [11] based on the work of [12]. Their motivation was to detect the nodes of the network that can control a directed network, i.e., to drive, based on specific inputs, the 'controlled' nodes to the state required by the control goal. They described the notion of a *stem*, which is a directed path starting from an initial node, such that no nodes appear more than once along the path, e.g., $j \rightarrow k \rightarrow l \rightarrow m$. A *stem-cycle disjoint subgraph* of G is the subgraph of G consisting of stems and cycles with no common nodes. The control centrality of a node is defined as the largest number of edges among all possible stem-cycle disjoint subgraphs.

Whereas their definition of centrality is very interesting from a control-theoretic perspective, our needs for addressing the requirements of all the aforementioned application fields demand two major reconsiderations. The first one concerns the fact that our *links are probabilistic*, and this must be incorporated in the definition of a control-theoretic type of centrality. Additionally, it does not make sense, for a VANET for instance, to demand from a single vehicle to be able to 'control' the whole ad hoc network; we need to redefine the centrality measure in a way that it can be *defined for both the entire network, and for neighborhoods around each node*.

Following on these requirements, we define the generic concept of *stem significance* (ssf), as the product of two scalar terms:

$$: \mathrm{sf} = \mathrm{sizeOfStem} \;\; x \;\; \mathrm{weight\,OfStem} \tag{32.1}$$

where sizeOfStem is the number of edges of the stem and weightOfStem is the product of its weights.

Based on this, we build two approaches for defining centrality measures over probabilistic graphs for range-limited neighborhoods. In the first approach, we adjust appropriately the ideas of [11], but in the second approach, we depart significantly from them and rely on the graph-theoretic concept of the influence range of a node, which is defined as the set of nodes reachable from a specific node.

32.3.1 RPCC with Cycle Extraction (RPCC$_{CE}$)

In our fist attempt to identify the most influential users following the idea of stem-cycle disjoint subgraphs, we denote the *cycle significance, csf,* in a similar way as *ssf*:

$$\text{csf} = \text{cyclePointer} * \text{weightOfCycle} * (\text{sizeOf Cycle} + 1) \tag{32.2}$$

where cyclePointer is the weight of the edge through which we visit a node of the cycle, weightOfCycle is the product of the weights of the edges that form it and sizeOfCycle is the number of its edges.

We compute the k-RPCC of a node i as the sum of the significances of the disjoint stems and cycles within the k-specified range. The pseudocode for the algorithm is as follows:

Step 1: Remove all incoming links of node i.
Step 2: Perform the Cycle Extraction procedure.
Step 3: Calculate and sum: cycle significances.
Step 4: Calculate and sum: stem significances of the remaining graph.
Step 5: Sum results of Steps 4 and 5.

For $n = 1$, this method is identical to the weighted-degree centrality. For $n = 2$, we exclude Steps 2 and 3, since there can be no cycles within such range. For $n \geq 3$, the computation is as given. The procedure *Cycle Extraction* is described in the next paragraphs.

32.3.1.1 Cycle Extraction

In each step, one cycle is removed from the graph. The first identified cycle becomes a candidate for extraction. Cycle Weight is the average sum of the weights of the cycle and Cycle Size *i* the number of edges that form it. When multiple cycles exist, the criteria to change candidates are as follows:

1. CycleWeight > CandCycleWeight and CycleSize/CandCycleSize > 0.7
2. CycleSize > CandCycleSize and CycleWeight/CandCycleWeight > 0.7.

CandCycleWeight and CandCycleSize denote the characteristics of the previous cycle candidate, and CycleWeight and CycleSize are the characteristics of the newly found one. The first criterion is to prevent small-sized cycles with high weights to be chosen over larger ones with high-quality links, due to their small number of edges. We use the second criterion to account for cases where the significance of a newly found cycle might be lower than that of the candidates, but if its number of edges is greater, and their significances are not far off (e.g., are more than 70 % equal), then the new cycle may be a better choice. If at least one of the above criteria is true, then the newly found cycle becomes the candidate. Finally, the candidate is removed and the process is repeated until there are no cycles in the graph. The choice 0.7

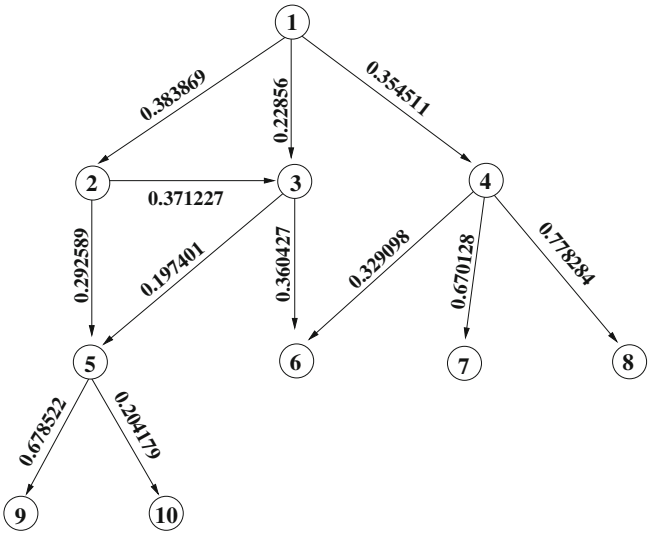

Fig. 32.1 A small complex network with probabilistic links

of the relative importance might seem arbitrary, but it does not have a significant algorithmic impact, as long as is larger than 0.4.

Figure 32.1 illustrates a small graph with probabilities on edges. The weighted degree of node 4 indicates that this node is the most influential one, however as illustrated, through 4 only three nodes are potentially accessible. From our point of view, node 1 becomes the better choice. Its RPCC value is equal to 2.02493, whereas node 1's RPCC value equals to 1.77751.

32.3.2 RPCC Without Cycle Extraction (RPCC$_C$)

In our second approach to calculate the importance of a node, we use only the sum of the significances of the stems within a specified range, leaving the cycles of the graph intact. The calculation of RPCC for each node is as previously, but now without Steps 2 and 3.

With this approach, we test the quality of paths through which a node i sees the rest of the network within the specified range. Since a path may be accessed by more than one nodes (e.g., $i \to j \to k \to l \to m$ and $i \to t \to k \to l \to m$), this approach also takes into account with how many ways a certain portion of the network can be controlled by i.

This approach targets the elimination of the burden of cycle calculation that can become significant in large networks and when k becomes relatively large. In principle, it does not differentiate significantly the performance of the method with respect to the previous method.

32.4 Simulation Parameters and Experimentation

For evaluation purposes, we had to select appropriate competitor methods, use networks with probabilistic edges, and also propagation models. As already mentioned, in the absence of competitors designed specifically for our problem, we used the weighted degree [7]. It is a straightforward generalization of the traditional unweighted degree as used in [5] for the evaluation of the spreading capabilities of a node in complex networks. Also, despite the wealth of real datasets that concern complex networks, it is hard to find appropriate input networks with probabilistic links. Therefore, we had to resort to the solution of using real complex networks and annotate their links with probabilities drawn from various distributions (uniform, zipfian, exponential, gaussian). Specifically, in the present article, we present results from a social network, namely *Wiki-Vote* which is part of the Stanford Network Analysis Platform [13]. As far as the propagation model is concerned, there is a wealth of such models in the literature, and it is worth examining the performance of the methods for each one of them. In this article, we confined ourselves to the SIR model with the characteristic that an infection originates from a single spreader, which is quite popular and has been used in similar studies [5]. We use relatively small values of infection probability to highlight the importance of influential spreaders.

The proposed centrality methods k-RPCC can be calculated for regions around the node of interest, and the whole network, as well. We experimented for values of $k = 2$, $k = 3$, and $k = 5$, where k is the distance in hops. Similar to [5, 14], we used the average size of the network's infected area as the performance measure.

For the experiment presented here due to lack of space, the probabilities of the edges are assigned based on uniform distribution and $k = 2$ for the (RPCC_C) approach. The probabilities range from 0.1 to 1. As said, these values depict the probability of an edge to be active on the graph. Links with values close to 1 are mostly active in our inspection time, whereas values near 0.1 are mostly inactive. According to these probabilities, we take 10 snapshots of the input graph resulting in 10 temporal graphs. To obtain statistically unbiased results, we repeated the computation 1000 times for each vertex in every temporal graph, i.e., 10,000 spreading processes.

32.5 Evaluation and Overview of Research Contributions

Figure 32.2 illustrates the results of the comparison of RPCC_C versus weighted degree for $k = 2$. The y-axis corresponds to the portion of the temporal network that got infected in percentage, and the x-axis depicts the values of the respective centrality measure. An ideal performance curve would be a very 'slim' one; in this curve, a very small number of infection percentages (values at y-axis) correspond to the same centrality value (value at x-axis). This would mean that the centrality measure would be able to divide the nodes on non-overlapping classes based on the

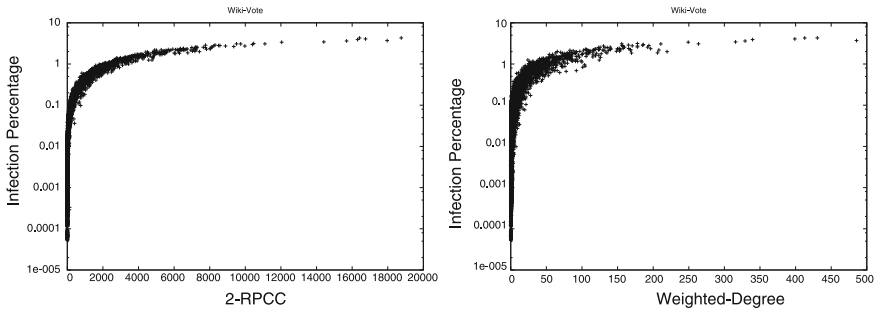

Fig. 32.2 2-RPCC$_C$ versus weighted degree

network percentage that can infect. For our competitor, there is no such observation since nodes with value of 100 are equal spreaders to those of 200, as depicted in (b). For a fixed k-RPCC value, there is a small deviation in the spreading capabilities, converging to a thinner curve, whereas for the weighted-degree, the deviation is much wider. Overall, the performance curve of the proposed method is much closer to the ideal one, than the competitor's curve.

In general, we expect that the network topologies and link probability distributions will affect the algorithms' performance, but for any influential spreader detection algorithm in order to be characterized as an efficient one, it is important that the algorithm has a steep ascending curve, which is 'thin,' especially as we move to larger values along the x-axis.

The study of complex, dynamic networks with probabilistic links arises naturally in some application fields, such as vehicular ad hoc networks, and aggregated descriptions of evolving complex networks. The identification of influential nodes in such networks is a new and interesting topic of investigation. This article takes a first step toward exploring this area and develops a measure of significance for the nodes of such complex network that quantifies whether each node is the starting point of 'strong' (i.e., almost permanent) paths that subsequently can 'control' the rest of the nodes. For the future, it is interesting to investigate the RPCC from a control theory perspective, instead of a pure engineering aspect, as it was done in the present article.

References

1. Newman MEJ, (2010) Networks: aN iNtroductioN. Oxford UNiversity Press, Cambridge
2. Doerr B, Fouz M, Friedrich T, (2011) Social networks spread rumors in sublogarithmic time, In Proceedings of ACM STOC, 2011, pp 21–30
3. Langville A, Meyer C (2006) Google's PageRank and beyond: the science of search engine rankings. Princeton University Press, New Jersey
4. Basaras P, Katsaros D, Tassiulas L (2013) Detecting influential spreaders in complex, dynamic networks. IEEE Comput Mag 46(4):26–31

5. Kitsak M, Gallos LK, Havlin S, Liljeros F, Muchnik L, Stanley HE, Makse HA (2010) Identification of influential spreaders in complex networks. Nat Phys 6:888–893
6. Pallis G, Katsaros D, Dikaiakos MD, Loulloudes N, Tassiulas L (2009) On the structure and evolution of vehicular networks. In: Proceedings of IEEE/ACM MASCOTS, pp 502–511
7. Fountalis I, Bracco A, Dovrolis C (2013) Spatio-temporal network analysis for studying climate patterns. Climate Dynamics, 2013, (to appear)
8. Barrat A, Barthelemy M, Pastor-Satorras R, Vespignani A (2004) The architecture of complex weighted networks. Proc Nat Acad Sci 101(11):3747–3752
9. Lambiotte R, Blondel VD, de Kerchove C, Huens E, Prieur C, Smoreda Z, Van Dooren P (2008) Geographical dispersal of mobile communication networks. Physica, 1(1)
10. Li Y, Chen W, Wang Y, Zhang Z.-L (2013) Influence diffusion dynamics and influence maximization in social networks with friend and foe relationships. In Proceedings of ACM WSDM, 2013
11. Y.-Y. Liu, J.-J. Slotine, and A.-L. Barabasi (2012) Control centrality and hierarchical structure in complex networks. PLOS One, 7(9)
12. Hosoe S (1980) Determination of generic dimensions of controllable subspaces and its applications. IEEE Trans Autom Control 25(6):1192–1196
13. The Stanford Network Analysis Project, Available at http://snap.stanford.edu/data
14. Pei S, Muchnik L, Andrade Jr J, Zheng JSZ, Makse HA (2014) Searching for superspreaders of information in real-world social media, 2014, Available at http://arxiv.org/abs/1405.1790