Identifying Influential Spreaders in Complex Networks with Probabilistic Links

Pavlos Basaras and Dimitrios Katsaros

Abstract Dynamic complex networks illustrate how "agents" interact by exchanging information in a constantly changing network. Typical examples of such networks are online social networks or human contacts. This article contemplates the common distribution of time that user-nodes spend on their activities, and describes a method for identifying real-time influential spreaders. We model the reciprocal activities of actor-nodes with probabilistic links and propose a technique for identifying influential spreaders in complex networks with probabilistic edges. The proposed measure, namely, *ranged Probabilistic Communication Area* ($rPCA$), is evaluated under the susceptible-infectious-removed (SIR) model, where the results illustrate that $rPCA$ can detect very effective spreaders in a networked environment with probabilistic edges.

1 Introduction

Real-world entities often interconnect with each other through explicit or implicit relationships, by transient and continuous ways to form a complex network. Social Networks (SNs) illustrate such complex interactions between individuals, and show how information, political views, frauds, advertisements, or rumors (*data*) flow through networked populations. Consider the most popular SNs (Facebook or Twitter), where users gain access to the Internet and their social activities through diverse wireless devices (smartphones, laptops, ipads) and become embedded to the Internet infrastructure swiftly for various and different time spans of their everyday lives, to interact, exchange opinions and ideas, or simply act like tuners for advertisements. Facebook self-reported statistics note that smartphone users check online 14 times a day, while an average user spends daily 40 min on the site. Now meditate on the vast amount of data traversing through such networks and how this

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

T. Özyer et al. (eds.), *Social Networks and Surveillance for Society*, Lecture Notes in Social Networks, https://doi.org/10.1007/978-3-319-78256-0_4

magnitude of information has evolved through time. As reported in [1] in 2007 we had an average of 5000 tweets per day whereas in 2013 we were at 500 million tweets on a daily basis [2], representing a five orders of magnitude increase. From the above considerations one could argue on what share of these vast data is actually being 'seen' by its corresponding *audience*, that is, friends, followers or broadly speaking from the connected society, and on how this is further affected by the different time spans that individuals spend on their social activities.

It is evident that users cannot follow such immense traffic of data, but what of time-limited messages or alerts? As an example let's reminisce the *Twitter, Faster Than Earthquakes* event. On August 23, 2011, it took 30 s for an earthquake to travel from Washington DC to New York, but tweets were fast enough to reach NY quicker than half a minute. To account for many such cases for example of natural disasters, Twitter has launched the *Twitter Alerts: Critical information when you need it most* program in September 2013 for its users to receive reliable information during these times. In this study we emphasize on such *Real-Time Data, RTDs*, that need to be 'made known' to the largest possible portion of a social network at a short time interval (i.e., within a few minutes or hours) and on the fact that this particular info will serve no further purpose in larger time spans (e.g., days or weeks). Consider an enterprise announcing a discount of a certain 'hot product' but only for a limited stock or a limited time offer, aiming to attract large masses of consumers. A preeminent question arises; which users should be the targets for incentive that will initiate a cascade of informed-interested people and increase as much as possible the number of potential buyers?

Although we presented the problem in terms of activities over technological social networks, the issue of the effect of concurrent 'activity' is present in other types of complex networks as well, such as human contact networks and their relationship to infectious disease transmission. Theoretically, in such networks a short interaction between a susceptible and an infectious person could lead to a comparable amount of ingested infectious material as that of a long interaction assuming that the short interaction is more intensive than the long one. However, prolonged contacts tend to be more intensive than short contacts [3].

Motivation and Contributions

The issue of identifying influential spreaders in complex networks is a well-studied topic that received increased attention in recent years [4–6]. However, for this particular framework of data that we are addressing in the present study, the different patterns in the concurrent activities of 'connected' users will constitute the most essential ingredient for detecting the *Real-Time Influential Spreaders, RTISs*, rather than simply focusing in a static image of a social network and traditional approaches. At this point, we should note that both RTDs and RTISs are connotations to characterize data with relative short lifetimes and influential spreaders for such cases, respectively.

Empirical observations [7–9] note that users in SNs are not active around the clock, and they show a complex behavior and distribution over the time they spend on their social activities. A probabilistic framework that follows such complex behavior could portray the possibility of a link-connection to exist, that is, when connected users are active, and the dissemination process is in progress. A relative approach is that reported in [8] where the authors illustrate a probabilistic model that accounts for a node-user to be active or not (and thus his connections to be present or not) at the time for example of a disease outbreak or broadly speaking a diffusion process. It is thus an important feature that we need to consider in order to quantify the strength of the corresponding propagation.

In a similar approach to [8], we model the existence or absence of connections, rather than users, by annotating weights on links that correspond to the *mutual time* that connected users spent on their separate social activities. Intuitively if we could locate those nodes that are the starting points for paths of users which *share at a great degree common time in their online social activities*, it could provide valuable insights into better approximate the spreading capability of users and thus more efficiently 'control' the spreading process of RTDs. By conducting simulations and experiments in different Social Networks, we will see how the proposed identification technique, namely, *ranged Probabilistic Communication Area* ($rPCA$) effectively combined the activity schedules of connected users, identified the most influential spreaders and outperformed the competing techniques in various scenarios.

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

- Investigates the issue of detecting real-time influential spreaders by considering the mutual time connected users spend on their online social activities.
- Proposes an adjustable centrality measure, the range Probabilistic Communication Area ($rPCA$) that accounts for such characteristic and real- time data.
- Thoroughly evaluates this centrality measure under diverse competitive techniques in different real networks.

The rest of this article is organized as follows: An overview of relevant important works for the identification of influentials is presented in Sect. 2. Section 3 presents the proposed algorithm. In Sect. 4, we describe our experimental environment, competing techniques, and evaluation criteria. In Sect. 5 we evaluate the performance of the adversaries and finally in Sect. 6 the conclusions.

2 Related Work

The literature on the problems of maximizing the spread of influence and of identifying influential spreaders in complex networks is quite rich during the last decade. In this section, we only mention but a few among many important studies. We should also categorize networks depending on the pattern of their connectivity, that is, directed or undirected networks in order to discuss the direction of the propagation and finally emphasize on directed networks. The first problem was posed in [10] and later investigated further providing more efficient algorithms, for example, in [11–13]. Newer approaches to the design of centralities include concepts such as κ-path centrality [14] and distributed algorithms for identifying influentials based on random walks [15]. Other graph-theoretic methods include the k-shell decomposition of a network [6], where the authors contend that a node's location may be the determining factor that defines the influence potential of that node. Other approaches based on several shortcomings of k-shell are presented in [5, 16, 17], whereas local techniques that combine effectiveness and efficiency are proposed in [4, 18].

All these works concern single-layer complex networks. However, the last few years, we are witnessing an initiative in the analysis of new kinds of complex networks, where the interacting entities are assumed to belong to more than one network, called layers. Online social networks, financial systems, transportation networks are such networks to name a few; more detailed examples can be found in [19]. The study of spreading processes in multilayer networks has started to attract significant interest [20]. The works most closely related to the current article, that is, to influentials detection, are those reported in [21–24]. The blending of all layers into a single one and then application of traditional options for influentials detection are proposed in [21]. A generalization of the k-core is proposed in [22] but it results in a vector of values that cannot be used in a straightforward manner for detecting effective influential spreaders. In [23], the authors proposed an called KS, which follows the intuition of [25], that is, aggregates the shell indexes of its neighbors, and moreover combines the intra- and interlayer spreading rates. However, to our understanding, incorporating the unknown spreading rates, of (and between) the layers, is not realistic. In [24], very elegant methods based on tensor analysis are proposed.

Considering a directed social network, a user i is called a follower of j if there is a directed link from i to j ($i \rightarrow j$), namely, i can receive information from j. Thus for these network cases, the diffusion takes place through the incoming connections of a node-user. To detect the most influential spreaders in directed social networks, researchers often apply the $PageRank$ algorithm [26] where a node i is considered as influential if it is pointed by many other and important nodes. It is a random walk algorithm that was first used for ranking relative contents of web pages. A variation of PageRank, namely, $LeaderRank$, was proposed in [27] by introducing a ground node to the initial network, connected to all other nodes through a bidirectional link. LeaderRank identifies nodes which lead to quick and extensive spreading. On the other hand, LeaderRank is tolerant of spurious and missing links, which benefits applications with noisy data. In summary, LeaderRank outperformed PageRank not only by better identifying the influence potential of nodes, but also by converging faster to the final scores, and by being more robust to noise and spammers.

Via assigning degree-dependent weights onto links associated with the ground node, $weightedLeaderRank$ was presented in [28]. For this approach, the authors allow nodes with higher in-degree to get more scores from the ground node. Since the in-degree of a node directly indicates its influence, it is natural to weigh nodes according to their influence. Weighted LeaderRank outperformed its immediate predecessor by identifying more influential spreaders, by having higher tolerance to noisy data, and by having higher robustness to intentional attacks.

Finally, TwitterRank [29], also a variation of PageRank, was developed for identifying influential spreaders in Twitter. The fundamental difference of the two algorithms is that TwitterRank develops a topic-sensitive random walk, that is, the transition probability between users in Twitter is topic-dependent; in a way this generates a topic-sensitive network structure. Despite being better than PageRank, the design of TwitterRank takes into account a number of tweets a twitterer publishes, which makes it susceptible to manipulations if a twitterer deliberately publishes a large number of tweets.

As we mentioned earlier, users gain access to their networked environment through diverse wireless devices for arbitrary lengths of time and different frequencies. Such interacting behavior in social platforms resembles that of temporal networks. Quite often temporal networks are separated into two categories based on time sequences and time intervals for the interactions between connected individuals in communication networks. In our study, however, we are searching for connected individuals who have common online activity, that is, they do not necessarily exchange messages at arbitrary times but rather they are concurrently active at regular times. This can be considered as another simplification of temporal networks where we discuss the probability of existence of interacting paths based on such observations. For more details on temporal network analysis, readers are referred to [30] and references therein.

3 Proposed Technique

In this section, we present our proposal, the *range Probabilistic Communication Area* ($rPCA$).

Complex Networks with Probabilistic Links

A complex network $G(V, E, w)$ is a directed graph where V is the set of vertices (nodes), and E is the set of pairs of vertices (edges). Every edge is described by a weight $w \in [0, 1]$ and a direction. Each vertex involves in- and out-neighbors. As usual, the number of head endpoints adjacent to a user-node is called its *inDegree* (k_{in}), and the number of tail endpoints defines the node's *outDegree (k_{out})*. The weight values associated with every edge define a network structure which describes the probability for any two connected nodes to be both active, for example, during a diffusion process. As we will see later in our experimentation the mining and efficient use of such information will prove a valuable asset for the spreading of RTDs.

r-Hop User Communication Paths (UCPs)

A user communication path (UCP) on a directed complex network is a directed path consisting of *n* individuals and *n-1* connections among them such as no user appears more than once, for example, $a \rightarrow b \rightarrow e \rightarrow j$ in Fig. 1. For simplicity, the example network is a Directed Acyclic Graph (DAG). To complete our definition, we also need to define the range for such interacting paths as the number of connections that form it or the hop distance from the initial node, for example, a to j. For our technique, the communication paths emanating from each individual node will define its significance in the network. The weight values on the connections will be used to investigate on the quality of paths through which a user i "sees" the rest of the network in range or in other words *to search for users which share common time to their social activities*. Ideally we would like to identify user paths such as $a \rightarrow x \rightarrow z$, however, in a realistic networked environment, we cannot expect users to have identical online activity schedules. Furthermore since those node paths are probabilistic in nature, we need to also quantify the strength of those paths. Hence we apply the following formula to measure the strength of an r-hop interacting path ($SUCP^r$):

$$SUCP^r = \sum_{j=1}^{r-1} w_j \cdot w_{j+1} \tag{1}$$

where r defines the range of a particular UCP and w_j is the weight value at j hop distance from the originator, that is, the weight of the corresponding connection. Intuitively if we could rate nodes on the basis of their UCPs, we could set the right paths for the spreading of real-time data.

This methodology can sustain the case when we want to model topic similarity among nodes. In this case, each link connecting nodes i and j can have one more weight factor, namely, $\sigma_{i,j} \in [0,1]$ that will describe the topic coherence among neighboring nodes [31]. Therefore, the total link strength will be evaluated as $w_{i,j} \times \sigma_{i,j}$. In this article, we assume $\sigma_{i,j} = 1 \ \forall i, j$.

Up to this point, we presented our proposal for quantifying the strength of a UCP. However, how to effectively combine the weight values associated with the corresponding connections in a communication path and define its significance is still an open issue. Another formula could be to simply acquire the product of its weights, however, such consideration will provide no distinction for paths with relatively equal-weight probabilities. For example in Fig. 1, for the interacting path $a \rightarrow b \rightarrow e \rightarrow i$, we would obtain a value of 0.063. The same value, however, would be attained if we sorted the weights in any possible way, for example, by reversing the probabilities of $b \rightarrow e$ and $e \rightarrow i$ or by placing the weakest interaction first and thus decreasing the probability of existence for the path. Another policy could be to assign a measure of importance for a specific weight depending on its hop distance from the originator, that is, weights closer to the initial node in a UCP are perceived as more vital. However, except for the fact that a tunable parameter

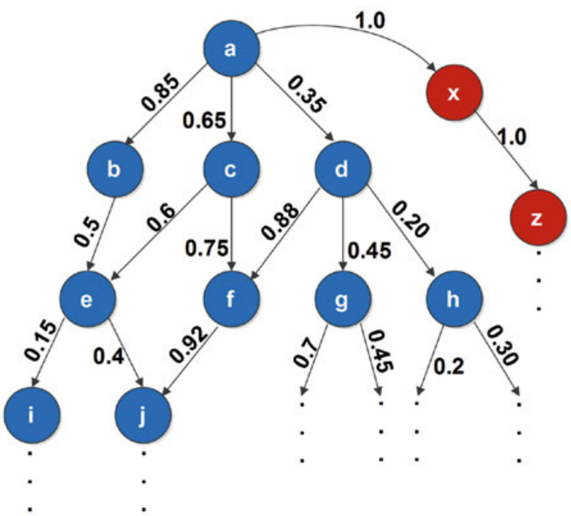

Fig. 1 $rPCP$ identifies nodes which possess the characteristic that from these nodes emanates "strong" paths. For 2 hops distance: $2PCA(a) = 17.283$ and $2PCA(b) = 1.1$ assuming that both i and j have 2 outgoing neighbors and x, z are hypothetic nodes, that is, not included

would have to be added, the significance of an interacting path like $a \to d \to f \to j$ which starts with a relatively weak weight and henceforth is composed of strongly connected users would be belittled with such consideration.

Range Probabilistic Communication Area (rPCA)

Following on these requirements, we built our proposal for defining centrality measures over graphs with probabilistic edges for range-limited neighborhoods. The $rPCA$ value of a node i within a specified range r is computed as the sum of $SUCP^r$'s emanating from i as follows:

$$rPCA(i) = \sum_{j=1}^{n} SUCP^r(j). \tag{2}$$

Note that nodes quite often share similar vicinities, that is, they may have a large number of common friends, and thus a certain path may be traversed by more than one way, for example, $a \to b \to e \to i$ and $a \to c \to e \to i$. For paths of interaction with hop distance greater than 2, the appearance of cycles, for example, $i \to j \to k \to j$ is a frequent phenomenon, especially when studying social networks often characterized as community networks, that is, dense connections within neighbors in the same community. However, considering "cycles of interaction" and thus returning to previous paths (or revisited node regions) are very likely to degrade an algorithm's performance, and thus these occasions are omitted by definition from our algorithm. In Fig. 1, we illustrate a toy example for the $rPCA$ method.

The proposed centrality measure can be *defined for both, the entire network ($*PCA$), and for neighborhoods around each node*. It is within our scope to maintain locality in order to provide an effective and efficient algorithm that can be applied in large-scale networks and real-time applications, and thus the range of UCPs is limited at low values, that is, 2 and 3. Generally, we could search to any number of hops; however, we understand that increasing the range of UCPs beyond the 90-percentile-diameter of a network (cf. Sect. 4) will provide little additional information to our approach since only 10% of the total size of the network is yet uncovered.

Although we have presented our method via the out-links of a node, when information flows through the in-neighbors of the network nodes as in our evaluation of Sect. 4, the implementation of UCPs is straightforward by following the in-links.

4 Performance Evaluation

For the evaluation purposes, we had to select appropriate competing methods, use networks with probabilistic edges, and also propagation models. In this section we describe our simulation environment and data sources.

Competing Techniques

In this subsection, we briefly describe the competing algorithms used in our simulation to evaluate the proposed method. Note that since information will flow through the in-neighbors of the network nodes the competitors are computed accordingly.

A diverse list of competitors were chosen regarding geodesics, the position of a node in the network, local techniques, or random walk-based approaches. A plethora of studies so far use the local degree centrality of a node to provide a baseline method for measuring the influence of nodes in complex networks.

(1) Likewise in our experimentation, we apply the weighted version of the approach. The *weighted degree centrality* ($wDeg$) of a node-user i or equivalently the strength of i is defined as the sum of the weights from the connections incident on i:

$$wDeg(i) = \sum_j w_{ji} \tag{3}$$

where j depicts the neighbors of i, that is, those nodes that i can exert influence, and w_{ij} stands for their associated weights.

(2) The farness of a user-node i is defined as the sum of its shortest distances to all other nodes of a network and the inverse of farness is noted as the closeness centrality of i. For its weighted implementation ($wClo$), the weights will describe how close or how far connected individuals are to each other as given by the formula:

$$wClo(i) = \sum_{j} \frac{1}{d_{ji}^{w}} \tag{4}$$

for all different j nodes of a network. In our framework, $wClo$ aggregates the weights on the shortest path and thus likewise our approach combines the weight values to provide an alternate technique that measures the strength and probability of existence for those paths.

(3) Shortest-path betweenness centrality describes the number of shortest paths that use a node i in order to reach other nodes of a network. Previous studies [5, 6, 18] found its performance insufficient to measure the influence potential of nodes in complex networks. Here we evaluate its performance in a relatively different experimental environment of weighted interactions and find similar conclusions ($wBet$):

$$wBet(i) = \sum_{s \neq i \neq t} \frac{\sigma_{st}^{w}(i)}{\sigma_{st}^{w}} \tag{5}$$

where σ_{st} is the total number of shortest paths from s to t and $\sigma_{st}(i)$ depicts the number of those paths that pass through i.

(4) A weighted version of the PageRank algorithm where the weights are proportional to the probabilities that a random walker will select a particular edge when choosing an outgoing connection from the current user-node [32]. Therefore, edges with larger weights are assumed to be traversed more frequently and are thus more important:

$$wPR_i(t+1) = (1-d) + d \cdot \sum_{j=1}^{N} \frac{w_{ji}}{\sum_{l=1}^{N} w_{jl}} wPR_j(t) \tag{6}$$

where w_{ji} is the probability of visiting node i from j if j is an in-neighbor to i otherwise $w_{ji} = 0$, d is the damping factor accounting for random jumps (in our experimentation, we assume no such occasions) and N stands for the total number of nodes in the network.

(5) The next and final competing algorithm weightedLeaderRank (wLR) was found to be more effective for the identification of influentials than PageRank and LeaderRank in directed networks [28]. Furthermore, it was proven more tolerant to noisy data by adding or removing links from the original network. We understand, however, that the traditional wLR algorithm does not use any

information regarding the weights of the links. It is used in our framework as a baseline method to measure the loss impact for not taking into consideration information through the weighted interaction. Nonetheless, our experimentation showed interesting results. As mentioned, it is a variant of LeaderRank, which introduces a ground node connected to all nodes of a network and recursively assigns scores to nodes depending on their k_{in}:

$$wLR_i(t+1) = \sum_{j=1}^{N+1} \frac{w_{ji}}{\sum_{l=1}^{N+1} w_{jl}} wLR_j(t) \tag{7}$$

where w_{ji} is equal to 1 if there is a directed link from j to i and 0 otherwise. If the destination node is the ground node then $w_{jg} = k_{in}^{\alpha}$ of j, where α is a free parameter set to 1 in our experimentation.

For the directed and weighted implementation for most of the above algorithms—excluding wLR and $wDeg$—we use the "igraph" R package.[1] igraph considers the weights assigned to each link as costs, that is, the largest the value the weaker the path. However, in our experimentation, weights indicate the strength of a link and thus we invert the original weight values for $wBet$ and $wClo$. A very popular method for the identification of influentials is the k-shell decomposition analysis [6] and its weighted versions, for example, [33]. However, to the best of our knowledge, there is no formal definition of the algorithm for directed and weighted networks. Could we have used measures such as μ–pci ? To such methods which are based on link counting and coreness, it is not clear how to quantize a "fractional degree" to its integer counterpart. Besides, such a conversion would loose significant part of the information carried by the probabilistic link.

Simulation Settings

Datasets

Nowadays there is a wealth of real datasets which concern complex networks; however, it is hard to find many input networks with probabilistic links with varying size and topology and varying distribution in the links' probabilities. Thus, in this article, we follow a dual methodology: We work with a real complex network to prove the applicability of our method in a real setting, and also use four real (initially unweighted) complex networks, which we annotate their links with probabilities drawn from various distributions, so as to test the scalability, effectiveness, and efficiency of the proposed method across a range of network sizes and link weight distributions.

[1] http://igraph.org/r/.

Table 1 Networks base attributes

Network	Nodes (V)	Links (E)	Diameter	90-EPD	E/V	Type
ego-Twitter	81,306	1,768,149	7	4,5	21.74	Social
soc-Slashdot0922	82,168	948,464	11	4,7	11.54	Social
soc-Epinions1	75,879	508,837	14	5	6.7	Social
wiki-Vote	7115	103,689	7	3,8	14.57	Social

The real probabilistic network is a contact network measured by the SocioPatterns collaboration[2] using wearable proximity sensors in a primary school, and covers 2 days of school activity. The sensors detect the face-to-face proximity relations (contacts) of 242 children [34]. The weight of a link is the aggregated contact duration of a pair of children. We normalize the links into the $[0, 1]$ interval by dividing each weight with the maximum weight found in the network. The experimental results which concern this real network are presented in Sect. 5.

The procedure for annotating the network links with weights is described in the following lines. We obtained our experimentation networks from the Stanford Network Analysis Platform [35]. For our evaluation purposes, the experimented networks were selected based on their connectivity, that is, three networks with a relatively equal number of nodes and decreasing in the number of their respective connections and finally a significantly smaller network. Specifically, we used the *ego-Twitter* network crawled from public sources, where followers receive information from their followees; *Soc-Epinions1*, a who-trust-whom social network of a general consumer review site, where users choose whether or not to trust reviews on products; *soc-Slashdot0922*, a technology-related news website, which allows users to tag each other as friends or foes; and finally *Wiki-Vote*, where nodes represent Wikipedia users and a directed edge from node i to node j represents that user i voted on user j. The base attributes of the aforesaid networks are listed in Table 1. The 90-effective-percentile-diameter (90-EPD) denotes the number of edges needed on average to reach 90% of all other nodes.

Generation of Probabilistic Links

For our simulation, the probabilities for the edge weights are assigned based on the Zipfian distribution for a range of skew values when the parameter $s \in [0.1, 0.9]$. The Zipfian distribution depicts the frequency of occurrence, for example, of a word randomly chosen from a text or the population rank of a city randomly chosen from a country. In our framework, it will depict the frequency of strong interactions. As s increases, we increase in the skewness for the distribution of weights and thus the strong weights will become more rare. In this study, we assume than any two connected nodes would share some common time of networked social activity, but

[2] http://www.sociopatterns.org.

also there are no identical schedules, that is, $w \in [0.1, 1)$. The resultant weight values will stand for the mutual time spent by nodes on their online social activities and thus depict the probability of an edge to be present or not at the time of the diffusion process. Links with values close to 1 are mostly active in our inspection time, whereas values near 0.1 are considered mainly inactive. According to these probabilities, we take ten 'snapshots' of the input graph resulting in ten abstract network images. Similar to [28] to obtain statistically unbiased results, we repeated the computation 100 times for each vertex in each network image, that is, averages over 1000 spreading processes.

Propagation Model and Influence

As far as the diffusion model is concerned, we employ the widely used susceptible-infectious-removed (SIR) model. SIR is commonly used for studying the spreading of epidemics in complex networks, where the infected nodes will either get immunity or die [36] and thus is suitable for our experimentation. We assume that an interested user propagates "data" only once, that is, users will not repeatedly send the same information to their respective vicinities. The Susceptible-Infectious-Susceptible (SIS) model is another popular method also used for the spreading of epidemics. SIS, however, has no immunity (like flu), and thus nodes get reinfected and further contribute in the diffusion. However, such consideration in our framework would include the provision of incentives to users in order to motivate them for propagating a certain datum a number of times.

In this study, we model the penetration of RTDs in a networked environment, with fixed transmissibility (infection rate) λ, for all user-nodes. SIR models three possible states:

- The susceptible state S, in which the S nodes are vulnerable to infection.
- The infected state I, in which the I nodes try to infect their susceptible neighbors and succeed with probability λ.
- The removed state R, in which nodes have recovered from infection and cannot be reinfected.

The diffusion proceeds as follows: In the initial phase, all nodes are in the *S* state except one node in *I*. An infected node is given a single chance to infect its susceptible neighbors and succeeds with probability λ. Immediately after and without loss of generality [28], the node enters the *R* state. The process continues until there are no nodes left in the infected state. Similar to [5] given a directed network, the influence of a node i, denoted by (IF_i), is defined as the average number of removed user-nodes at the end of the spreading process if i was the initially infected node. Conventional techniques for measuring the *epidemic thresholds* [37] in the evaluated networks cannot be employed in our case study, due to the probabilistic nature of in- and out-neighbors, and thus we confined our work to a range of λ values between 1 and 10%.

Evaluation Criteria

Kendall's Correlation (τ)

To evaluate the ranking abilities of each competing method with respect to the actual spreading potential of each node, we use the Kendall's Tau 'b' rank correlation coefficient (τ) [38]. It is a statistic used to measure the association between two measured quantities, for example, (*2PCA*, *IF*). When $\tau = 1$, we have a perfect correlation, indicating that when node i is ranked before j by some method, that is, with greater *2PCA*, then its spreading capability is also higher. For $\tau = 0$, the measured entities are considered neutral, whereas $\tau = -1$ implies opposite correlation. Generally, the closer we get to 1, the better the correlation of the evaluated approach.

Fraction of Ranked Nodes: False Index

As depicted in Fig. 2 for the lower spreading rates, there is a large number of users with zero influence, for example, over 70% for the soc-Slashdot0922 network when $\lambda = 2$. Applying Kendall's correlation to such unfiltered values will provide harsh results. In our experimentation, we take a closer look for each λ value to provide a more complete assessment and thus the ranked sample used for the ranking process will be composed of user-nodes with $IF > 0$, namely, p users. To complete the evaluation of the results and conclude on which technique better identifies the influence power of nodes, we also need to provide an assessment for the rest of the

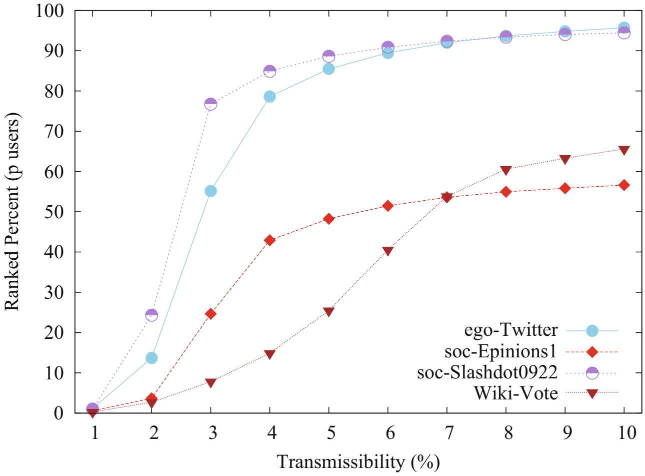

Fig. 2 Ranked percent with respect to the total number of nodes of each network case for all evaluated λ values, that is, nodes with $IF > 0$

$1-p$ non-ranked users. The *False Index* depicted in Figs. 3(upper right) to 4(bottom right) fills this void. To obtain the False Index, we calculate for each node in $1-p$ the number of nodes in p whose index is lower from that particular node's. In other words, we measure the average number of nodes which, although did not succeed in propagating, were ranked with higher index by some users in p, for example, with greater $2PCA$. Reasonably, a small False Index indicates better results.

5 Results

Impact of Infection Probability

In this section, we evaluate the efficiency of each competing method in ranking nodes according to their actual spreading potential, when varying in the strength of the propagation in four different Social Networks. For the distribution of links in Figs. 2, 3, 4, 5, 6, 7, s is set at 0.7. In almost all the evaluated networks, we observe that the most abrupt changes in the curves of correlation for all methods occur at

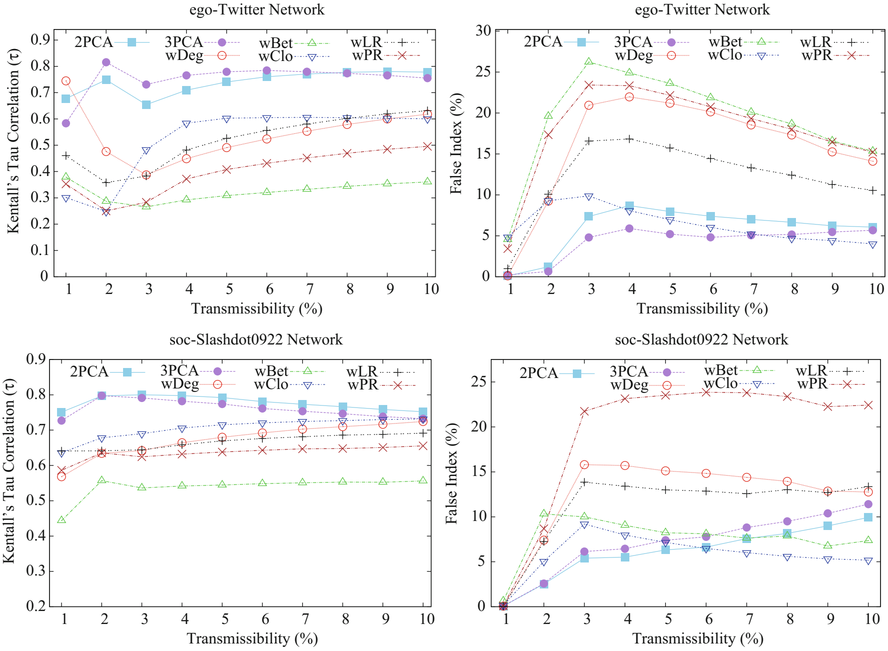

Fig. 3 (Top: Twitter plots). In almost all different spreading rates for the ego-Twitter network, the proposed technique significantly outperforms its competitors. (Bottom: Slashdot plots). For the soc-Slashdot0922 network, we observe that our approach coincides with the rest of the competing algorithms only for the higher spreading rates

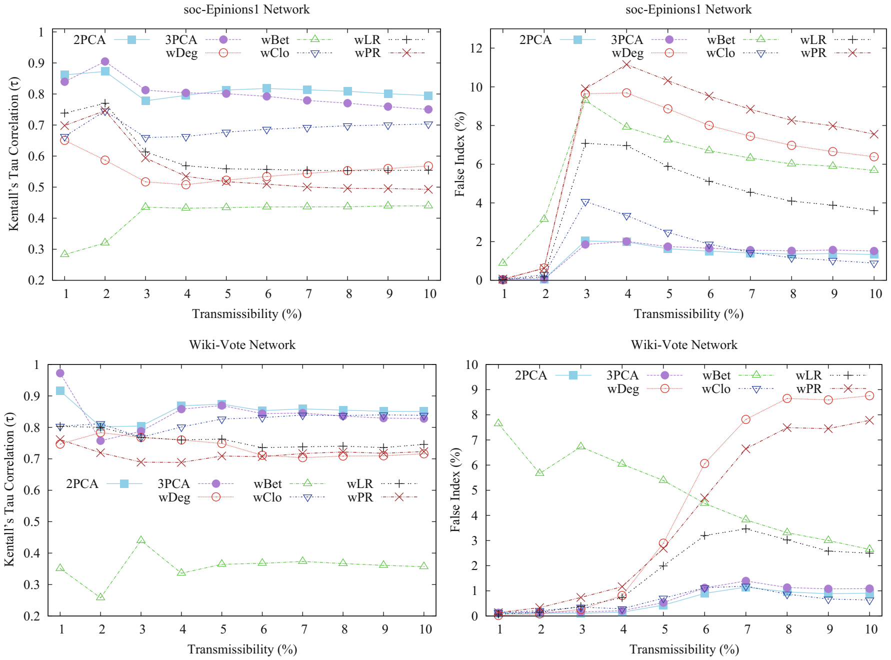

Fig. 4 (Top: Epinions plots). As the spreading rate increases, our two-fold approach maintains its superior performance as compared to the rest of the competing techniques. (Bottom: Wiki plots). For the final network case, an oscillation for the most accurate ranking is observed at the lower spreading rates. Nonetheless, the proposed technique is found within the higher τ values

the lower λ values. This is partly because the largest leaps in the percent of the ranked p users occur within the fist few increments of the spreading rate, that is, when $\lambda < 4$ for most of the evaluated networks (about 6 for the Wiki-Vote), where we observe that the fraction of ranked nodes drastically changes. For instance as illustrated in Fig. 2, for the Twitter network when $\lambda = 2$ the p nodes constitute about 15% of the total size of the network, whereas when we move to $\lambda = 3$ this percent is close to 58%. The changes in the curves of τ however are not only due to the increasing number of the p users used in the ranking process. As the spreading rate increases, the influence of nodes from previous λ values also changes and the same may happen to the ranking between those nodes in subsequent spreading rates.

Considering the results in Fig. 3(upper left), Kendall's coefficient for 2–3PCA when $\lambda = 2$ is above 0.75, whereas the rest of the competing techniques are found below 0.5. Similar observations can be made for the soc-Slashdot0922 network, that is, the largest differences in τ are found at the low spreading rates. For Fig. 4(top left) and (bottom left), however, the above observation does not hold. For these cases, we observe a more sedate behavior of the curves as we increase in λ. Apparently, the probabilistic property of the networks affects the dynamics of a cascade and thus in Sect. 5 we investigate on the quality of the probabilistic links.

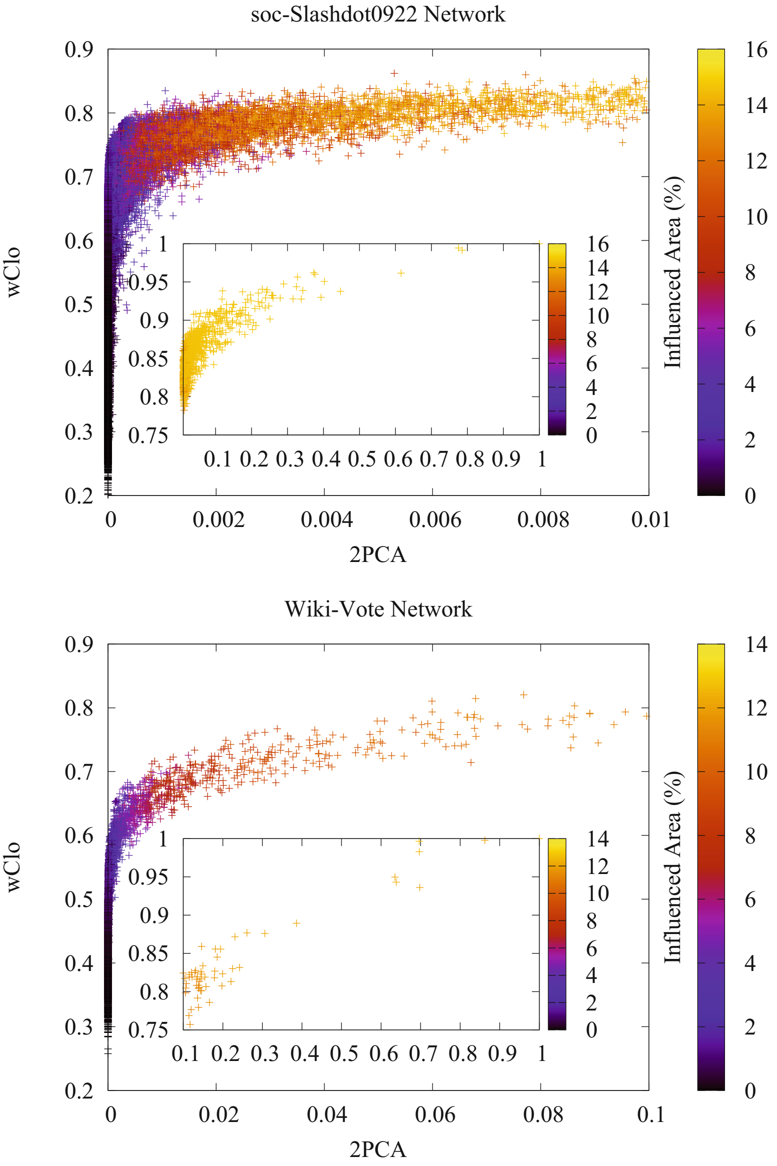

Fig. 5 $wClo$ was found to coincide with the proposed technique in a few configurations. The presented heat plots illustrate that influence is closely related to 2PCA. On the contrary, for $wClo$, we observe that the medium values depict an amplitude of influence values

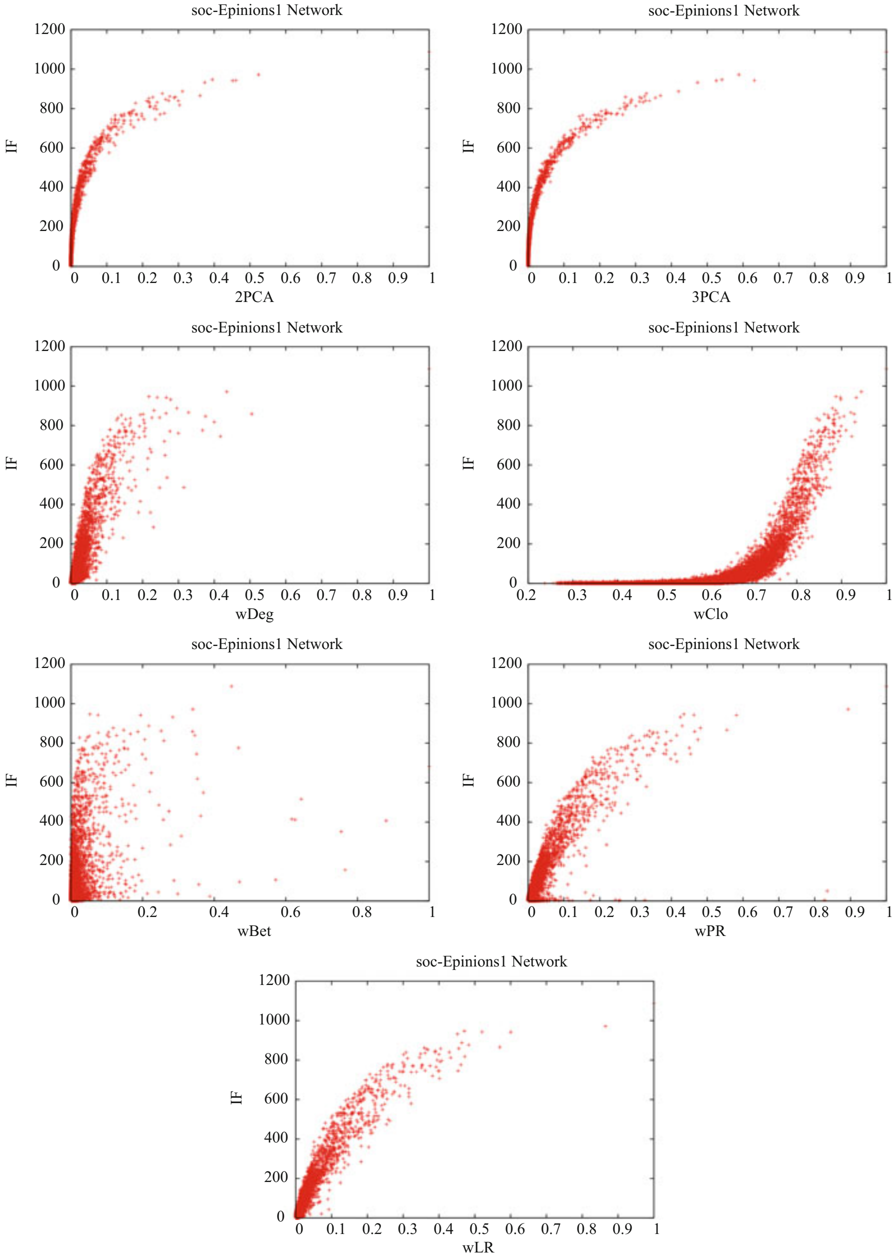

Fig. 6 Spreading rate is set at 3%

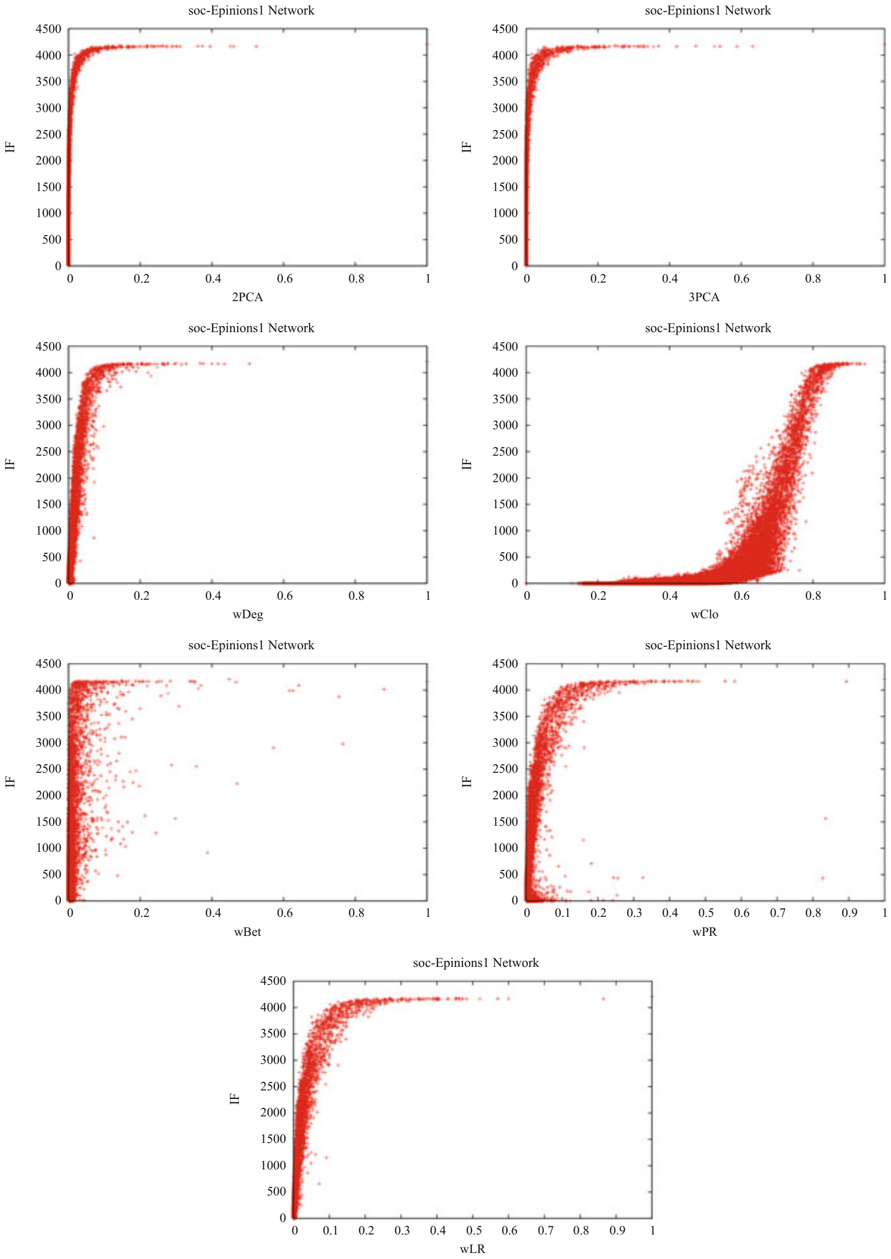

Fig. 7 Spreading rate is set at 6%

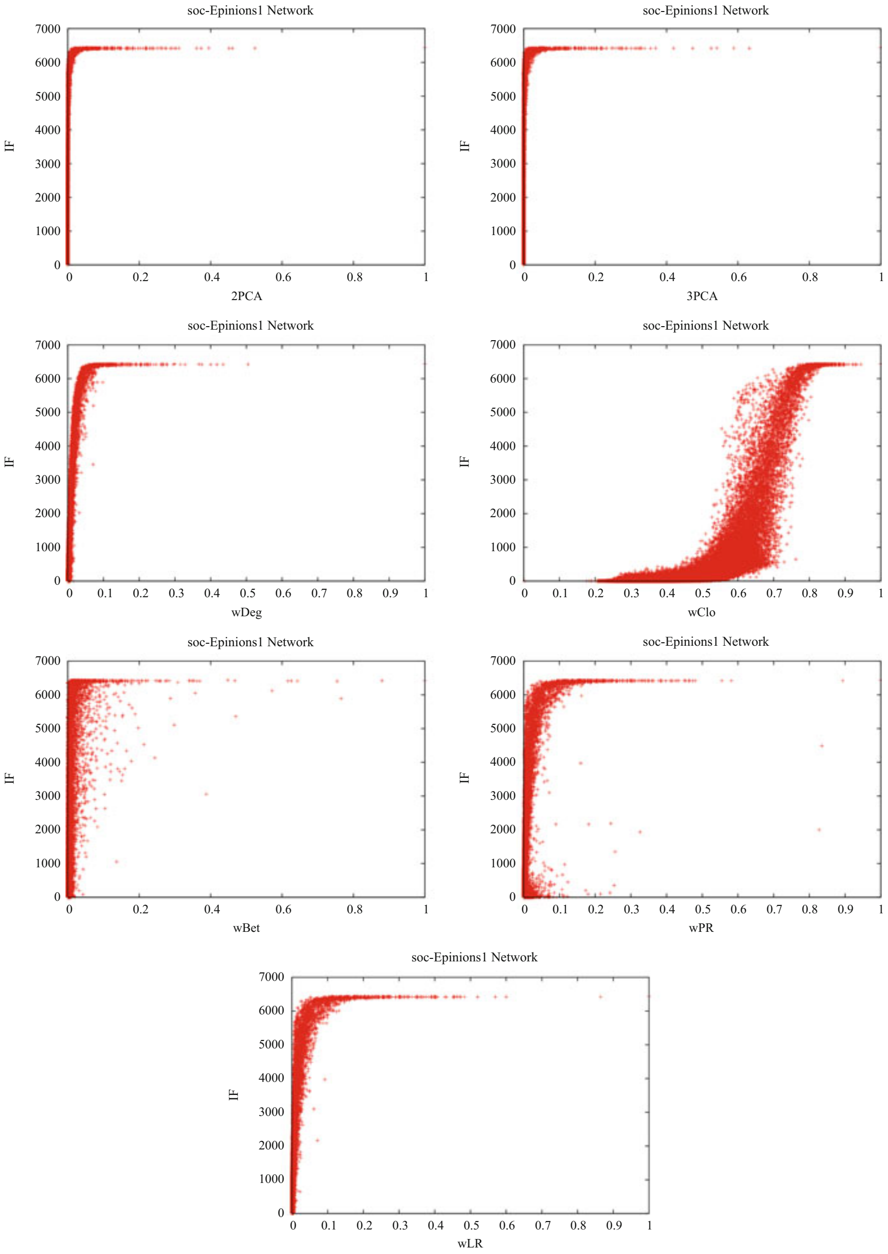

Fig. 8 Spreading rate is set at 9%

In Fig. 3(bottom plots), when λ is around 9%, $wDeg$ and $wClo$ coincide with our approach. It should be emphasized that for very large values of λ, the τ values of correlation for the competitors are bound to cross over and oscillate. This is due to the fact that on such occasions an epidemic will occur regardless of the characteristics of the originator. For the higher spreading rates, the true influential nodes are very likely to get infected at some point as the diffusion progresses, and thus result in an epidemic outbreak, even though the originator is not truly an influential. Besides by using large λ values, the role of individual nodes in the diffusion process will no longer bear significance [5, 6, 16, 17]. When considering the different ranges of our approach, we can see that for the low spreading rates there is an oscillation for the most accurate ranking between the two methods. However, as we increase in λ for all network cases $2PCA$ always obtain higher τ values. Such consideration indicates that local information of a node's surroundings (communication paths) is more favorable as we increase in the spreading rate.

For an overview on the False Index, 2–$3PCA$ is found at the lower percentages. $wClo$ illustrates similar behavior; however, the rest of the competing techniques illustrate significantly higher values. Note that the False Index does not provide any information about how accurate the ranking for the p nodes is but rather acts as a further criterion for each respective technique. Ideally, we would obtain a zero False Index indicating that none of the $1 - p$ users has higher index than any node in p. Generally, a low False Index coupled with a high τ will promote the most efficient algorithm for the addressed issue. Clearly, the proposed technique supports the desired outcome. Only at the higher spreading rates in Fig. 3(top left) and (bottom left) 2–$3PCA$ illustrates higher False Index. However, these cases are trivial since as depicted in Fig. 2, almost all nodes are within the p node set.

Focusing separately on each competitor, $wDeg$ is used as a baseline method to illustrate how complete locality serves in quantifying the spreading power of a node. For its overall evaluation, it is indicated by our results as a moderate approach for our ranking purposes and real-time data. When considering its False Index, we can see that $wDeg$ is rated among the three worst-performing methods in all evaluated network cases. This observation indicates that simply considering the total strength of a node's local connections is not a good indicator to quantify its spreading influence. For example, a high $wDeg$ index may be accumulated by many but otherwise weak interactions which in our framework is interpreted as regularly absent connections. To our perception, such occasions will result in insignificant influence results, and may be the reason for $wDeg$'s high False Index values. As another contributing factor to its medium performance, we can say that $wDeg$ does not 'carry' any information about the position of the node in a network. Therefore, although a node might me connected to its immediate vicinity with strong links, if it is positioned in the periphery of a network [6], reasonably we expect that its influence will be rather diminished.

Another interesting point seen through our simulation is the performance of wLR. Reminisce that for this particular method no information about the activity schedules is used, and thus we expected a relatively low correlation in our framework of weighted interactions. Nonetheless, it was proved rather compensatory as a competing technique which indicates that wLR may indeed be a good indicator for the spreading potential of nodes in unweighted networks. Although the τ values of correlation for wLR are significantly lower from our approach, it was found to be comparable and even better on many occasions when considering the rest of the evaluated techniques, for example, as illustrated in Fig. 3(bottom left) or 4(top left). Generally, its performance can be considered relatively similar to $wDeg$'s; however, we can conclude that wLR provides a more accurate ranking if we consider the False Index of the two aforementioned techniques, that is, $Wdeg$'s False Index is always higher.

In contrast to wLR, wPR accommodates information from the weighted interactions in the sense that links with higher weights are traversed more often. Both techniques were found to follow approximately the same trend in all evaluated networks as the spreading rate increases, that is, their illustrated curves either both ascend or descend. However, our experimentation showed that wLR obtained higher correlation with influence and also significantly lower False Index values. Nodes with no outgoing links, the sink nodes, which are indeed present in the evaluated networks are not well handled by PageRank, since they decrease the PageRank overall [39]. To our understanding, such inefficacy overestimates the spreading power of a node and may be the reason for wPR's low correlation and the highest False Index values. Generally, through such methods users pointed by many other and important users are elected as strong influencers; however, as also noted in [26, 40] quite often the k_{in} of a node is not sufficient to characterize its influence capacity.

Next we investigate on $wBet$ and find that this particular method has the worst performance in all evaluated networks, while other studies [5, 6, 18] also note its inability to capture a node's influence capacity. Its low efficiency can be explained if we consider that through $wBet$, node-users who are unique intermediates for some other nodes (or medians leading to different communities) are elected as important entities. However, in such cases their capability for influence and propagation may well be overestimated if these nodes lead to regions with sparsely connected nodes or small sized communities. In our simulation where the problem of identifying influential spreaders is further enhanced by considering the time distribution of nodes social activities, $wBet$ will be at a further disadvantage if those links correspond to nodes with highly uncommon time spans. As a final observation for $wBet$ in our experimentation, we found that among the p user-nodes, there was a significant amount of nodes with zero betweenness scores, which also explains the high False Index values of the competitor. This observation indicates that nodes which do not reside in any shortest path may be more influential from nodes with higher betweenness scores, and further confirms that the influence cannot be measured through the shortest paths that pass through a node.

Finally, $wClo$ utilizes useful data through the weighted interactions, in the sense that nodes connected through weak links are considered to be relatively far to each other. However, as shown in our simulation in most of the illustrated results, simply aggregating the strength of the connections to obtain the average distance of a node to the remaining nodes of the network lacks when compared to our approach.

We attribute its lower performance to the following: first, although the effective diameter for all network cases is relatively small, for example, between 4.5 and 5, there are still more than 8000 nodes for Twitter and Slashdot0922 networks, and more than 7500 for soc-Epinions within a diameter of 7,11, and 14 hops, respectively. However, considering long interacting paths would include a mixed set of connections, that is, a relatively long path may be composed of both strong and weak links. To this end, we expect that techniques that utilize global information of a network's connections to define the significance of a node in the network will furnish varying results. Figures 3 to 4 confirm our statement. Lastly, unlike our approach, $wClo$ considers a single communication path to all other nodes from the focal node, and in particular the shortest (strongest) paths to those nodes. Nonetheless, rather than a single strong path, it may be more favorable to take into account a number of interacting paths that reach a single user-node, that is, multiple paths, in our framework of complex networks with probabilistic links.

In Figs. 3(bottom plots) and 4(bottom plots), we found that $wClo$ coincides with our approach significantly and thus we advance to thoroughly understand the relation of the two methods, that is, $2PCA$ and $wClo$ with influence in Fig. 5. The spreading rate is set at 10% for both networks where the aforementioned techniques are closer. The heat values depict the IF in percent, for each user-node with pair values ($2PCA$,$wClo$). For nodes with the same pair values, the average IF is used. Note that each axis is normalized to its largest corresponding index. Moreover the outer plots are ranged up to a certain value of $2PCA$ which is then resumed in the embedded charts of each corresponding network to illustrate more precise results. From these figures, we can further argue that $2PCA$ is the better indicator for the spreading influence of nodes in complex networks with probabilistic links and further strengthen the superiority of the proposed technique. From the embedded charts, we can understand that the highest index values for both methods indeed correspond to the most influential spreaders. However, from the outer plots, for example in soc-Slashdot09022, we can see that for a range of values in 0.7 to 0.8 for $wClo$ there is a wide variety of influence scores, that is, approximately between 4 and 14%. Such observation indicates that the medium values of the competitor cannot distinguish the influence potential of nodes in contrast to $2PCA$ which provides a more accurate ranking. We found similar conclusions when comparing $wClo$ to $3PCA$.

Overall, our experimentation showed that for our technique, paths limited in the near neighborhood of the focal node, that is, two-hop UCPs, are usually sufficient to characterize its role in an epidemic. In our framework, the probabilistic property of the networks affects the diffusion dynamics and thus we urge for a technique that effectively handles the different probabilities for connected nodes. Our ranged approach was found quite effective and efficient that better identified influential spreaders in most of the observed network scenarios.

Spreadability

In this section, we illustrate the results in Figs. 6, 7, 8 where the x-axis represents the values of each competing technique and the y-axis the corresponding influence (IF). For a better overview on the competitors, we illustrate the results for a range of λ values, that is, 3,6 and 9%, respectively. We found similar qualitative results for the rest of the evaluated networks. In this section, we measure the different influences of user-nodes with approximately equal index scores, that is, for relatively equal $2PCA$s what is the range of IF values. Apparently the smaller the amplitude of IF the better the technique. It is evident that among all competing methods $2PCA$ illustrates the best correlation with influence which is indicated by its thin ascending curve which increases in IF as $2PCA$ increases. Similar conclusions can be made for $3PCA$; however, as we increase in the spreading rate $2PCA$ illustrates better performance. These observations are consistent with the results in Fig. 4(top left). For wLR and $wDeg$ as previously noted, we found similar behavior. For the lower spreading rate ($\lambda = 3$), we observe that there is a wider range of IF values for $wDeg$ and moreover the slope of the curve for wLR is greater. These observations explain the higher τ for the latter in Fig. 4(top left). Nonetheless, as the spreading rate increases, their difference diminishes and when $\lambda = 9$, $wDeg$ has a thinner curve and thus better correlation.

When $\lambda = 3$, for wPR we notice that there are a few nodes with very large values, for example, around 0.8, which have insignificant or zero influence. Moreover in 0.1 to 0.3 the amplitude of IF ranges from almost zero influence up to the largest value. These observations persist as we increase in the spreading rate; however, such inefficacies will significantly affect the correlation of the technique with influence. A similar amplitude of IF values is illustrated for $wBet$, although for significantly more user-nodes, which further distances the algorithm for use in the identification of influentials. Finally, following on the performance of $wClo$, we understand that the lower values of the technique correspond to users with low or insignificant IF, whereas the highest index scores indicate the most influential nodes. The problem of the competitor lies to its medium values, that is, for a range around 0.7 (which increases as λ increases) where we observe a wide variety of influence scores, quite similar to Fig. 5(upper) for soc-Slashdot0922 network. Such observation afflicts the competence of the algorithm for the addressed issue and further strengthens the superiority of the proposed technique.

For any influential spreader detection algorithm in order to be characterized as an efficient one, it is important to have a steep ascending curve, which is ‘thin’, especially as we move to larger values of a technique along the x-axis. 2-$3PCA$ was found to adopt such behavior, taking the lead on its competitors in both steep upward slope and smaller deviation in IF in all network cases.

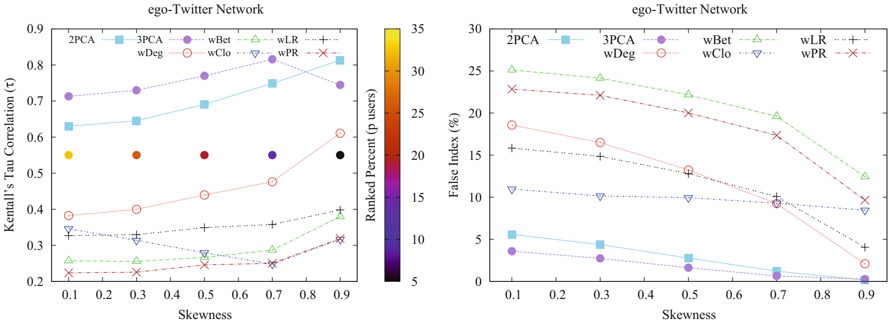

Fig. 9 Ranging in skewness for the distribution of links (spreading rate is set at 2%)

Impact of Zipfian Skewness

In this set of experiments, we investigate on the skewness of the Zipfian distribution as s increases. Due to similar results, we present only those for the ego-Twitter network in Fig. 9. The spreading rate is set at 2%. The percentage of nodes that succeeded in propagating (p users) is illustrated with the colored cycles mapped to the corresponding heat values in the palette. As a first observation, we note that as we increase in s for the distribution of links, the number of users that are able to propagate in their respective vicinities decreases. This phenomenon is anticipated as we distance our experiments from uniform distribution and gradually force the weights toward the lower possible values. In our framework, such configuration results into frequently absent connections resembling a realistic social environment where we cannot expect node-users to have largely common time spans for their social activities.

As shown in Fig. 9 most of the competing techniques illustrate similar behavior in both evaluation criteria, that is, decreasing and increasing trend for the False Index and τ, respectively. For the lower s values, we observe only small increases in τ. However, as we further increase in s, the changes in τ become more evident. This is due to the fact that for the larger skews, the now fewer strong links and interacting paths become more clear for the competitors. This remark is most visible when $s > 0.7$ where we observe the most significant changes for all methods. wLR, however, shows minor changes in τ, an observation somehow coherent with [28] where the authors explain the robustness of the technique in "noisy" networks, that is, missing links.

When we have a fairly good distribution for the weights (low skewness), we observe that $3PCA$ obtains the highest correlation followed by $2PCA$, whereas the rest of the competing techniques obtain significantly lower values in τ. This observation indicates that when we have many strong interactions, that is, nodes with highly common activities, accumulating information from relatively long UCPs indeed results in better correlation. In an opposite scenario where node-users have significantly different schedules (large skewness), the strong weights become more rare. Using long paths composed of weak interactions will degrade our algorithms performance which explains the steep fall of $3PCA$ for the higher s values. Conversely thinking, we can understand the illustrated behavior of $2PCA$ which uses short-ranged communication paths and takes the edge on our ranged approach in the aforesaid cases. The significant difference in the False Index values between the competitors and 2–$3PCA$ further strengthens the superiority of our method. For instance, 2–$3PCA$'s "misjudgment" near 0.9 becomes almost zero, whereas in most of the evaluated scenarios (different skews) it is found below 5%. Finally we conclude that in a framework with probabilistic links that portray the property of active nodes as described in our work, considering multiple paths and moreover multiple alternative paths (unlike $wClo$) is a first step for devising an appropriate method for the identification of real-time influential nodes.

Evaluation with a Real Complex Network

After the detailed performance evaluation of the methods across a range of network sizes and link weight distributions, we use a real weighted complex network in order to confirm the practicality of the problem examined and also to further support the superiority of the proposed method. Recall from Sect. 4 that this is a contact network measured by the SocioPatterns collaboration[3] in a primary school. The sensors detect the face-to-face proximity relations (contacts) of 242 children [34]. The resulting network has 242 nodes and 4024 links, after removing the nodes terms as "Teachers" and their interactions, because the network offers no possibility to differentiate between different teachers. Figure 10 depicts the number of interactions

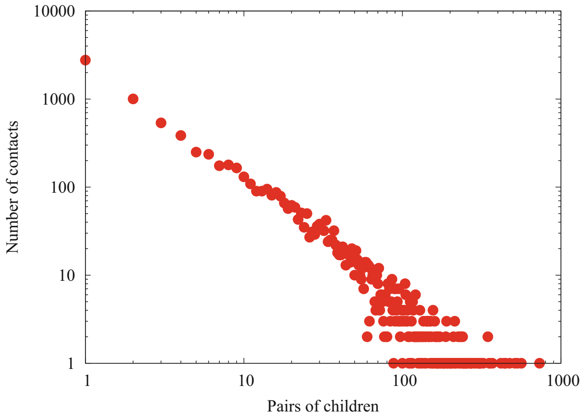

Fig. 10 Distributions of the link weight (i.e., aggregated contact duration) of the real weighted network

[3] http://www.sociopatterns.org/2015/01/a-high-resolution-social-network-measured-in-a-primary-school/.

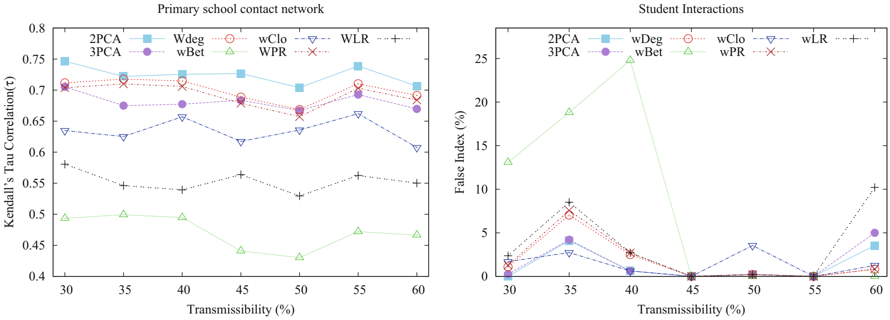

Fig. 11 Evaluation of competing algorithms over the real weighted network

per pair of children. According to the methodology of data collection (sensor beaconing), each contact lasts for 20 s. Thus, this figure shows in an equivalent way the aggregated contact duration of a pair of children, which is the link weight in our case. Evidently, this distribution follows a power-law, where the majority of the pairs of children have less than ten contacts.

The evaluation of the competing algorithms is presented in Fig. 11. The first comment concerns the transmissibility rates in order to achieve high enough infection. The generic comment is that the infrequent student interactions require higher transmissibility rates for successful transitions. Specifically, for the lower λ value, only about 2% of the network is infected, for example, from an emerging flu originating from the most influential student, whereas when $\lambda = 60$, the infected students rise up to 30%.

Regarding the performance of methods, we observe that the best strategy—consistent with our previous results—is $2PCA$, whereas $wBet$ is the worse strategy. The position of the second best-performing strategy is now occupied by $3PCA$, $wDeg$, and WPR (subject to some variation). The interesting thing is that $wClo$, which was steadily the third winner in our earlier finding, now is fifth. Based on the rankings we obtained for this real network and the conclusions by Figs. 10 and 9, we can say that the link weight distribution of this network is highly skewed for which networks we already have seen that the performance of $3PCA$ and $wClo$ degrades significantly. Finally, complementary to the False Index illustrated for the artificial networks, we observe (right plot Fig. 11) no different qualitative results, that is, the proposed technique is found at the lower false values, which further strengthens the superiority of 2–$3PCA$ for the addressed issue.

6 Conclusions

The evolution of social networks to date indicates that the amount of information flowing though user interactions is only going to increase. In this article, we argued on what portion of information remains 'unseen' from interested users due the continuous flow of data in such networks. With this consideration, we focus on 'pieces' of information with limited life spans, that is, for data that are interesting to some users but only for a limited time (RTDs). In order to push information into a network and spread RTDs to the largest possible extent, we need to account for users who share at a great degree common time in their social activities. With this demand, social networks must be remodeled to probabilistic structures. In this study, we used probabilistic links to simulate the probability of connected users with common social activity, and proposed a centrality metric, namely, $rPCA$, which accounts for probabilistic communication paths around the focal node. The proposed technique was evaluated under different spreading rates and distribution for the weight probabilities, and proved superior from its competitors in ranking nodes according to their true spreading potential. Finally, to our understanding, how each method uses-filters the lower weight values is a determinant factor to its performance, since users with low common time spans will contribute little to each other's influence. Moreover in order for RTDs to be substantially propagated, we need not only consider the strength of each individual link separately but rather as combined attributes within the interacting paths. For our future direction, we intend to apply different approaches for quantifying the strength of the UCPs and further improve our formula for the identification of influential spreaders. Also, other factors could be considered in defining the weights on edges, for example, the characteristics of the individual nodes, or the characteristics of the communities, could play a significant role in communication.

Acknowledgements This work was supported by the project "REDUCTION: Reducing Environmental Footprint based on Multi-Modal Fleet management System for Eco-Routing and Driver Behaviour Adaptation", funded by the EU.ICT program, Challenge ICT-2011.7.

References

1. K. Weil, Measuring tweets (2010). Twitter Official Blog, February 22
2. R. Krikorian, New tweets per second record, and how! (2013). Twitter Official Blog, August 16
3. T. Smieszek, Theoretical Biology and Medical Modelling **6**, 2–15 (2009)
4. P. Basaras, D. Katsaros, L. Tassiulas, IEEE Comput. Mag. **46**(4), 26 (2013)
5. B. Joonhyun, K. Sangwook, Physica A **395**(1), 549 (2014)
6. M. Kitsak, L.K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H.E. Stanley, H.A. Makse, Nat. Phys. **6**, 888 (2010)
7. A.L. Barabasi, Nature **435**(6), 207 (2005)
8. J.B. Holthoefer, S. Meloni, B. Goncalves, Y. Moreno, J. Stat. Phys. **151**, 383 (2013)
9. J.B. Holthoefer, A. Rivero, Y. Moreno, Phys. Rev. E **85**, 066123:1 (2012)
10. P. Domingos, M. Richardson, *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)* (2002), pp. 57–66
11. D. Kempe, J. Kleinberg, E. Tardos, *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)* (2003), pp. 137–146
12. J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J.M. van Briesen, N.S. Glance, *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)* (2007), pp. 420–429

13. C.T. Li, T.T. Kuo, C.T. Ho, S.C. Hong, W.S. Lin, S.D. Lin, Soc. Netw. Anal. Min. **3**(3), 341 (2013)
14. N. Kourtellis, T. Alahakoon, R. Simha, A. Iamnitchi, R. Tripathi, Soc. Netw. Anal. Min. **3**(4), 899 (2013)
15. B. Han, A. Srinivasan, *Proceedings of the ACM International Symposium on Mobile Ad Hoc Networking and Computing (MOBIHOC)* (2012), pp. 5–14
16. J.G. Liu, Z.M. Ren, Q. Guo, Physica A **392**(18), 4154 (2013)
17. A. Zeng, C.J. Zhang, Phys. Lett. A **377**(14), 1031 (2013)
18. D. Chen, L. Lu, M.S. Shang, Y.C. Zhang, T. Zhou, Physica A **391**(4), 1777 (2012)
19. S. Boccaletti, G. Bianconi, R. Criado, C.I. del Genio, J. Gomez-Gardenes, M. Romance, I. Sendina-Nadal, Z. Wang, M. Zanin, Phys. Rep. **544**, 1 (2014)
20. M. Salehi, R. Sharma, M. Marzolla, M. Magnani, P. Siyari, D. Montesi, IEEE Trans. Netw. Sci. Eng. **2**(2), 65 (2015)
21. M.A. Al-garadi, K.D. Varathan, S.D. Ravana, E. Ahmed, V. Chang, J. Intell. Fuzzy Syst. **31**(5), 2721 (2016)
22. N. Azimi-Tafreshi, J. Gomez-Gardenes, S.N. Dorogovtsev, Phys. Rev. E **90**(3), 032816 (2014)
23. Z. Dawei, L. Lixiang, L. Shudong, H. Yujia, Y. Yixian, Phys. Scr. **89**(1), 015203 (2014)
24. M. De Domenico, A. Solé-Ribalta, E. Omodei, S. Gómez, A. Arenas, Nat. Commun. **6**, 6868 (2015)
25. B. Joonhyun, K. Sangwook, Physica A **395**, 549 (2014)
26. S. Brin, L. Page, Comput. Netw. ISDN Syst. **30**(1–7), 107 (1998)
27. L. Lu, Y.C. Zhang, C.H. Yeung, T. Zhou, PLoS ONE **6**, 0021202:1 (2011)
28. Q. Li, T. Zhou, L. Lv, D. Chen, Physica A **404**, 47 (2014)
29. J. Weng, E.P. Lim, J. Jang, Q. He, *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)* (2010), pp. 261–270
30. P. Holme, Proc. IEEE **102**(12), 1922 (2014)
31. F. Menczer, Proc. Natl. Acad. Sci. **99**(22), 14014 (2002)
32. R. Baeza-Yates, E. Davis, *Proceedings of the ACM International World Wide Web Conference (WWW)* (2004), pp. 328–329
33. M. Eidsaa, E. Almaas, Phys. Rev. E **88**, 062819:1 (2013)
34. V. Gemmetto, C. Barrat, A. Cattuto, BMC Infect. Dis. **14**, 694:1 (2014)
35. J. Leskovec, A. Krevl, SNAP datasets: Stanford large network dataset collection (2014). http://snap.stanford.edu/data
36. B.A. Prakash, D. Chakrabarti, N.C. Valler, M. Faloutsos, C. Faloutos, Knowl. Inf. Syst. **33**(3), 549 (2012)
37. L. Cong, H. Wang, P.V. Mieghem, Phys. Rev. E **88**, 062802:1 (2013)
38. M. Kendall, Biometrika **30**, 81 (1938)
39. P. Devi, A. Gupta, A. Dixit, Int. J. Adv. Res. Comput. Commun. Eng. **3**(2), 5749 (2014)
40. J.K. Kleinberg, J. ACM **46**(5), 604 (1999)