Identifying Influential Spreaders by Graph Sampling

Nikos Salamanos, Elli Voudigari and Emmanuel J. Yannakoudakis

Abstract The complex nature of real world networks is a central subject in several disciplines, from Physics to computer science. The complex network dynamics of peers communication and information exchange are specified to a large degree by the most efficient spreaders - the entities that play a central role in various ways such as the viruses propagation, the diffusion of information, the viral marketing and network vulnerability to external attacks. In this paper, we deal with the problem of identifying the influential spreaders of a complex network when either the network is very large or else we have limited computational capabilities to compute global centrality measures. Our approach is based on graph sampling and specifically on *Rank Degree*, a newly published graph exploration sampling method. We conduct extensive experiments in five real world networks using four centrality metrics for the nodes spreading efficiency. We present strong evidence that our method is highly effective. By sampling 30% of the network and using at least two out of four centrality measures, we can identify more than 80% of the influential spreaders, while at the same time, preserving the original ranking to a large extent.

The original version of this chapter was revised. An erratum to this chapter can be found at 10.1007/978-3-319-50901-3_66

Nikos Salamanos
Athens University of Economics and Business, 76 Patission Str. GR10434 Athens Greece, e-mail: salaman@aueb.gr

Elli Voudigari
Athens University of Economics and Business, 76 Patission Str. GR10434 Athens Greece, e-mail: elliv@aueb.gr

Emmanuel J. Yannakoudakis
Athens University of Economics and Business, 76 Patission Str. GR10434 Athens Greece, e-mail: eyan@aueb.gr

H. Cherifi et al. (eds.), *Complex Networks & Their Applications V*,
Studies in Computational Intelligence 693,
DOI 10.1007/978-3-319-50901-3_9

1 Introduction

Understanding spreading process in real world complex networks is of high importance due to the variety of applications that they occur, such as the acceleration of information diffusion, the control of the spread of a disease and the improvement of the resilience of networks to external attacks.

Key role to spreading dynamics plays the heterogeneity of nodes in terms of *spreading efficiency*. High spreading efficient nodes are called *influential spreaders*, representing the nodes that are more likely to spread information or a virus in a large part of the network. Therefore, thorough research has been realized in order to connect the topological properties of network nodes with their spreading efficiency.

In this paper, we deal with the problem of identifying the influential spreaders of a complex network when we are not able to analyze directly the whole network, either because of its large size or of our limited computational resources which are necessary for estimating global centrality measures or other advanced nodes properties.

Our approach is based on graph sampling, the problem of selecting a small subgraph which will preserve the topological properties of the original graph. In our case, the central question is whether the top-k spreaders in the samples correspond to the top-k spreaders in the original graph. Thus, a sampling method could be served effectively as an *influential spreaders identifier* if and only if: (a) the fraction of top-k common nodes in the samples and in the graph is on average sufficiently large and (b) the rankings of these nodes in the samples are close to the original ranking in the graph.

We address this question using *Rank Degree* [18], a graph exploration sampling method which as proven outperforms other well known methods such as *Forest Fire* and *Frontier sampling* [11, 10, 14].

We conduct extensive experiments in five real world networks using four centrality metrics in order to rank the nodes, with respect to spreading efficiency. In order to emphasize the efficiency of Rank Degree, we compare our method with that of Forest Fire. The results show that Forest Fire is inadequate in identifying the best spreaders, while our method is highly effective. Studying the samples of Rank Degree, we are able to identify in every network, at least 80% of the influential spreaders by sampling 30% of the network, using at least two out of four centrality measures.

Finally, and more importantly, in four out of five networks, the rank correlation between the top-k nodes in the samples and the top-k nodes in the original graph is very large.

The rest of the paper is organized as follows. Sect. 2 describes the related work. Sect. 3 presents our method. Sect. 4 describes the experimental analysis and provides information on the methods and datasets used and Sect. 5 concludes the paper.

2 Related Work

The problem of identifying the influential spreaders in a network is a central subject in complex networks analysis and therefore, several approaches have been proposed in the literature.

Kitsak *et al.* [9] proposed the *k*-shell decomposition method [15, 16] as an *influential spreaders identifier*, showing that the *k*-core values constitute a more reliable measure than *degree centrality* and *betweenness centrality*. One of the core results is that the placement of a node (node global property) is more important than its degree (node local property). Two nodes with the same degree but different placement, where the one is connected with the periphery of the network and the other with the innermost core will not have equal spreading efficiency. Thus, highly connected nodes are not always the best spreaders, while less connected nodes but well connected with the core of the network may strongly affect the spreading process. In addition, Zeng *et al.* [19] investigated the limitations of the *k*-shell method and they proposed a mixed degree decomposition procedure which performs more accurately than the *k*-shell approach.

Chen *et al.* [2] proposed the *local centrality*, a semi-local centrality measure as a tradeoff between the degree centrality (local measure) and the computationally complex betweenness and closeness (the global measures). They showed that local centrality is more effective to identify influential nodes than the degree centrality.

LeaderRank [13] is a ranking algorithm for identifying influential nodes in directed social networks. LeaderRank is a parameter-free random walk algorithm analogous to PageRank [1]. Moreover, Li *et al.* [12] proposed a weighted variation of Leader Rank which outperforms LeaderRank. Furthermore, in [3] the authors introduced *ClusterRank* a local ranking algorithm for directed graphs that takes into account the nodes *clustering coefficient* and they proved that ClusterRank outperforms other approaches such as LeaderRank.

3 The Rank Degree Method

Algorithm 1 presents briefly the *Rank Degree* (*RD*) sampling method. *RD* is a graph exploration sampling algorithm which outperforms several other well known approaches. A detailed analysis of the algorithm is out of the scope of this paper and we refer to [18] where the authors studied thoroughly the properties and the efficiency of the algorithm.

The main characteristic of the method is that the graph traverse is based on a deterministic selection rule, the ranking of nodes according to their degree values (see Steps 9-10). The algorithm is specified by two parameters: (a) the number s of the initial starting nodes (seeds) and (b) the parameter ρ which defines the top-k, that is, the selected fraction of nodes from each ranking list. Hence, we use the notation $RD(\rho)$. The extreme case is for top-k with k=1, in other words when we

Algorithm 1 Rank Degree Algorithm

Set parameters: (i) s: number of initial seeds, (ii) ρ (see Step-10), (iii) target sample size x
Input: undirected graph $G(V,E)$
Output: sample of size x
Initialization: $\{Seeds\} \leftarrow s$ *nodes selected uniformly at random*
$Sample \leftarrow \emptyset$
while sample size $<$ target size x **do**
  $\{New\ Seeds\} \leftarrow \emptyset$
  for $\forall w \in \{Seeds\}$ **do**
    Rank w's friends based on their degree values
    Selection rule:
    (i) $RD(max)$: select the max degree (top-1) friend of w
    (ii) $RD(\rho)$: select the top-k friends of w, where $k = \rho \cdot (\#friends(w))$, $0 < \rho \leq 1$
    Update the current sample with the selected edges $(w,\ friend(w)\ on\ the\ top-k)$ along with the symmetric ones
    Add to $\{New\ Seeds\}$ the top-k friends of w
  end for
  Update graph G: delete from the graph all the currently selected edges
  $\{Seeds\} \leftarrow \{New\ Seeds\}$
  If $\{New\ Seeds\} = \emptyset$ then repeat Step-4 (random jump)
end while

select only one node from each ranking list - that node having the maximum degree. For simplicity, we refer to this case as $RD(max)$.

The algorithm, starting from s initial nodes, performs s parallel graph traverses. At each time step, the number of visited nodes (current seeds) varies and depends on the set of selected nodes at the previous time step.

As referred to, in [18], the algorithm generates the most representative samples for $RD(max)$ and $RD(0.1)$, i.e. when we select either the top-1 or the top-10% from the ranking lists. In this paper, we concentrate our analysis to $RD(max)$ studying its performance with respect to influential spreaders.

4 Experimental Analysis

4.1 Methods

Sampling: Apart from our method, *RD*, we study the *Forest Fire (FF)*, a well known sampling method introduced by Leskovec *et al.* [11]. FF starts from a randomly selected node (seed) w and at each step, the algorithm moves from the current set of seeds to the next one as follows: from each node w in the set of current nodes, a random number x is generated which is geometrically distributed with mean $p_f(1-p_f)$. The parameter p_f is called *forward burning probability* which is set to 0.7. Then, x outgoing edges are selected from the set of $w's$ outgoing edges. The end nodes of the selected edges constitute the next set of current nodes. At each step,

the visited nodes are considered as burned and are removed from the graph. Hence, they cannot be traversed for a second time. Finally, the process is repeated until a sample of the requested size is reached.

Spreading efficiency: In the absence of ground truth information with regard to nodes spreading efficiency, several approaches have been proposed in the literature such as the *Linear Threshold* and *Independent Cascade* models [7], as well as the basic epidemic models *Susceptible Infected Recovered (SIR)* and *Susceptible Infectious Susceptible (SIS)* [9, 2] which tend to simulate the spreading process in a graph.

In this paper, we use local and global topological properties, centrality measures, in order to estimate the nodes spreading efficiency in the original graph and in the samples: (a) *k-core decomposition*, a subgraph with nodes of degree at least k (on the subgraph). k-shell: the set of nodes that belong to the k-core but not to the $k+1$-core. For the rest of the paper, when we refer to nodes k-core values we imply the max k-shell that these nodes belong to, (b) *degree centrality*, (c) *betweenness centrality* and (d) *closeness centrality* [5].

It has been proved that most of the centrality measures are positive correlated [17] and also that some measures are less effected by sampling [4].

Sampling evaluation: We study the efficiency of the sampling methods with regard to node influences using two measures:

(a) *OSim* [6], an object similarity measure (in our case the objects are the nodes), the overlap between the elements of two ranking lists A and B (each of size k), without taking into account their ordering. It is defined as $OSim(A,B) = \frac{|A \cap B|}{k}$. In our case, the lists A and B correspond to the ranking lists $r_G(top-k)$ and $r_S(top-k)$ which are computed as follows: for a given centrality measure we calculate the nodes centrality values for both the original graph G as well as each of the collected samples S and we rank the nodes accordingly (in descending order) creating the ranking lists r_G and r_S. Then, for a given k, we create the $r_G(top-k)$ and $r_S(top-k)$ collecting the top-k nodes of the ranking lists r_G and r_S.

(b) *Kendall tau* [8], the well known rank correlation coefficient measure, with which we measure the relative ordering between all pair of nodes in the two ranking lists $r_G(top-k)$ and $r_S(top-k)$.

4.2 Data and Sampling Setup

We evaluate the efficiency of $RD\,(max)$ as influential spreaders identifier in five real world datasets, two of small and three of medium graph size (Table 1). We restrict our analysis to undirected graphs, therefore we transform the directed graphs (*wiki-Vote* and *p2p-Gnutella30*) to undirected, by applying to each edge the symmetric one. In addition, we study the efficiency of FF - a well known sampling algorithm which, contrary to *RD*, inadequately identifies the most influential nodes, even if it

Table 1 Datasets

Graph	egoFacebook	wiki-Vote	CA-CondMat	p2p-Gnutella30	Email-Enron
Description	Ego-net	Wiki-net	Collaboration Net.	P2P Net.	Comm. Net.
Type	Undirected	Directed	Undirected	Directed	Undirected
# Nodes	4039	7115	23133	36682	36692
# Edges	88234	103689	93497	88328	183831

is producing representative samples with regard to some topological properties of the graph.

For each dataset and each method separately, we collect 40 samples, per sample size, where the sample sizes are $10\%,\ldots,50\%$. In all experiments, the number of initial seeds is defined by the 1% of the target sample size. For instance, for a given graph G with 2000 nodes and target sample size 10%, the number of initial seeds is 2. Moreover, we compute the OSim and Kendall tau for each top-k interval separately. Therefore, we define two top-k intervals, the small top-k, where $k \in [0.001, 0.01]$ (i.e. one per mill to one percent) as well as the medium top-k, where $k \in [0.01, 0.1]$ (i.e. 1% to 10%)

4.3 Results

4.3.1 Effectiveness of Rank Degree

Top-k similarity (OSim): For a given graph G, top-k and centrality measure, we calculate the OSim between the top-k nodes in G and the top-k nodes in each of the 40 samples separately.

Fig. 1 and Fig. 2 present the average OSim for $RD(max)$ samples, of the small and medium size graphs. Specifically, for each graph, for each top-k interval, and for each sample size, we plot the average OSim values of the 40 samples, for each centrality measure separately. The results for small and medium top-k (i.e. $k \in [0.001, 0.01]$ and $k \in [0.01, 0.1]$) are given in separate plots. For the sake of clarity, only the sample sizes 10% and 30% are shown.

We observe that, in *egoFacebook* the samples size 30% maintain at least the 80% of influential spreaders in terms of k-core and degree centrality for *small* top-k (Fig. 1(a)), while for *medium* top-k, the corresponding OSim values are larger than 90% (Fig. 1(b)). Moreover, from Fig. 1(c) (*wiki-Vote*), it is clear that all centrality OSim values are higher than 70% for all sample sizes. In medium top-k (Fig. 1(d)) and for samples size 30%, the degree centrality and k-core have the largest OSim values where in some cases are close to 100%.

In Fig. 2(a) (*CA-CondMat*), we can see that for small top-k, degree centrality and closeness centrality are close to 80% with betweenness and k-core following. The results are similar for medium top-k (Fig. 2(b)).

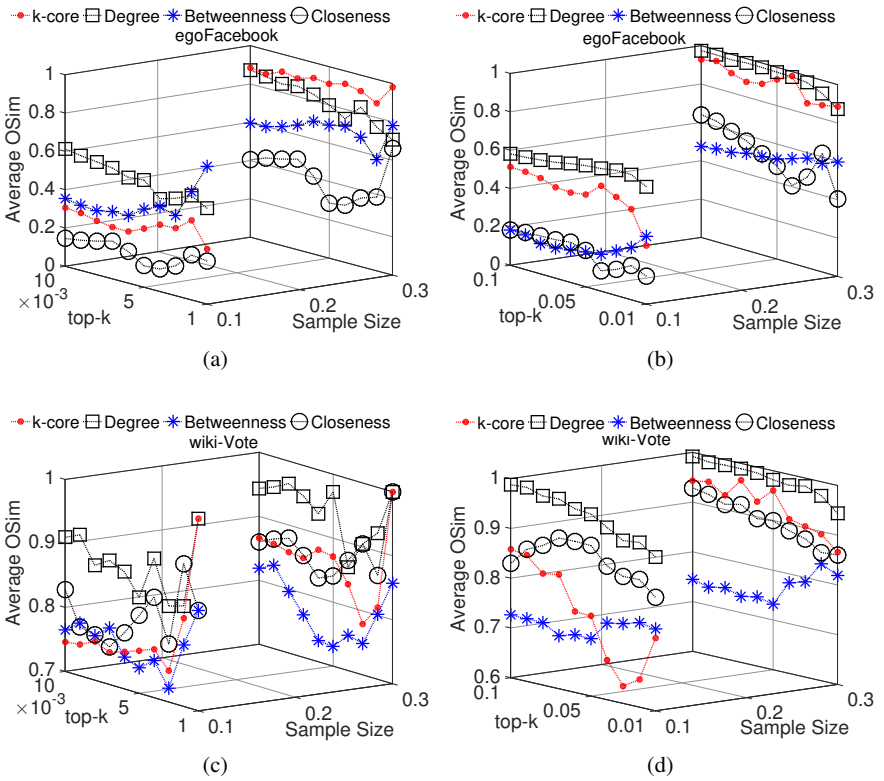

Fig. 1 Average OSim per top-k. Small size graphs

In the case of *p2p-Gnutella30* (Fig. 2(c)), *k*-core comes first for sample sizes 10% and 30% with closeness, degree centrality and betweenness following. For medium top-k, three out of four centrality measures have OSim values larger than 80% (Fig. 2(d)).

In *Email-Enron* and small top-k, three out of four centrality measures have OSim values larger than 80%. In almost all sample sizes and top-k intervals, the OSim for *k*-core is close to 100% (Fig. 2(e)). Finally, the results for medium top-k and samples size 30%, three out of four centrality measures have OSim values larger than 90% (Fig. 2(f)).

Ranking similarity (Kendall tau): For a given graph G, top-k and centrality measure, we apply the Kendall tau on the ranking values of the common nodes between the top-k nodes in the graph G and in a given sample S. Specifically, consider two ranking lists $r_G(top-k)$ and $r_S(top-k)$. First, we compute the intersection $R = r_G(top-k) \bigcap r_S(top-k)$. Then, we define the $R_G(top-k)$ and $R_S(top-k)$ which contain only the ranking values from $r_G(top-k)$ and $r_S(top-k)$ that corre-

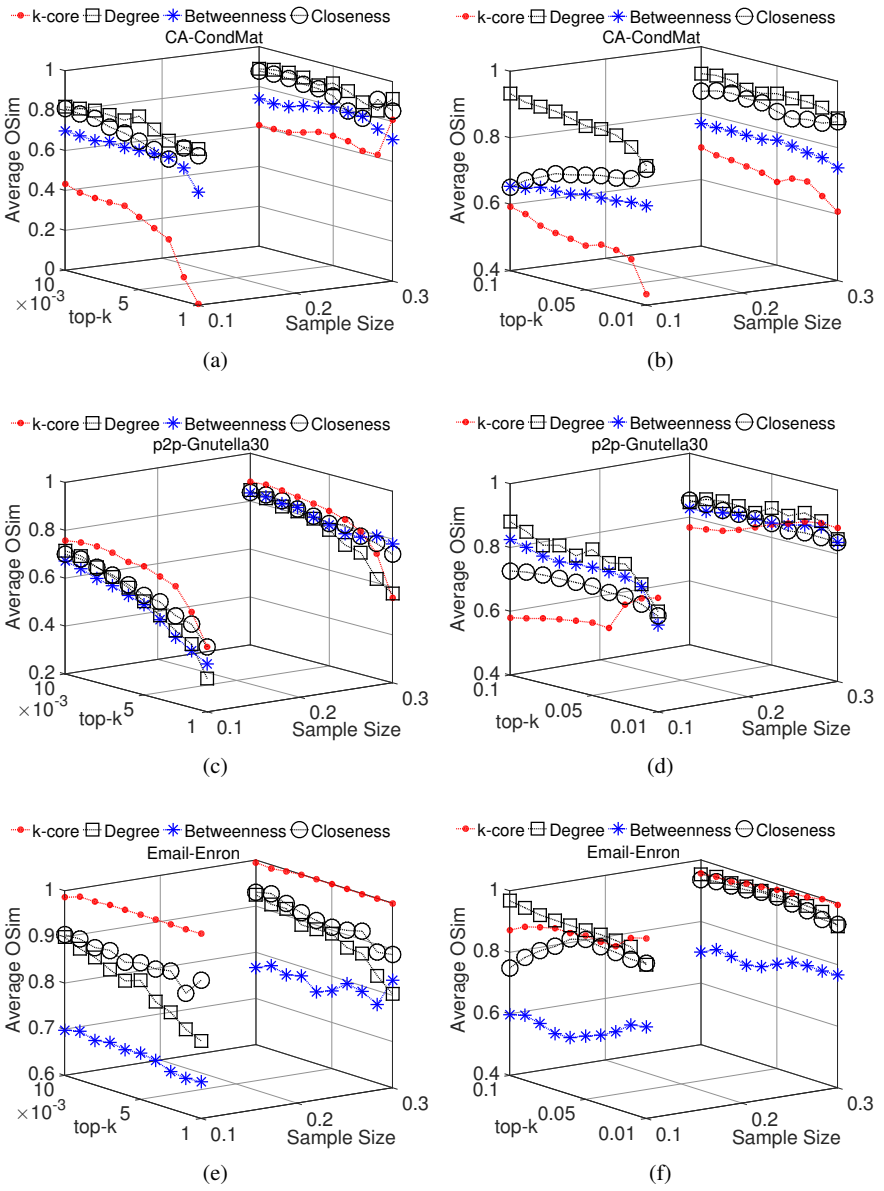

Fig. 2 Average OSim per top-k. Medium size graphs

spond to the nodes in R. Finally, we compute the Kendall tau of $R_G(top-k)$ and $R_S(top-k)$.

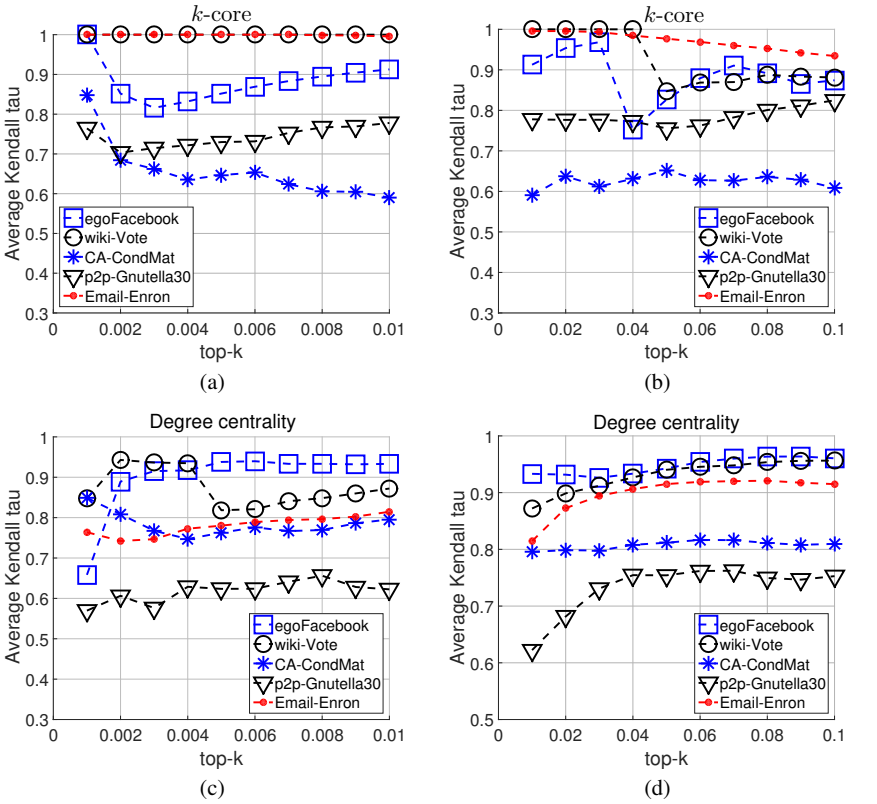

Fig. 3 Ranking similarity: Average Kendall tau per top-k. Samples size 30%

Fig. 3 presents the average Kendall tau values for k-core and degree centrality for small and medium top-k and samples size 30%.

We observe that in four out of five datasets the average Kendall tau values are large, at least 0.7. Thus, there is a large positive correlation between the ordering of the top-k nodes in the samples and the top-k nodes in the original graph.

For instance, in *wiki-Vote* and *Email-Enron*, for small top-k and top-k in $[0.01, 0.4]$, the Kendall's tau values are almost equal to one (Fig. 3(a) and Fig. 3(b)). Moreover, in every top-k, the samples from all datasets except *CA-CondMat* preserve strongly the relative ordering of the top-k nodes.

In the case of degree centrality, the results are similar. For instance, in four out of five datasets and for any interval of medium top-k, the average Kendall values are at least 0.8 (Fig. 3(d)).

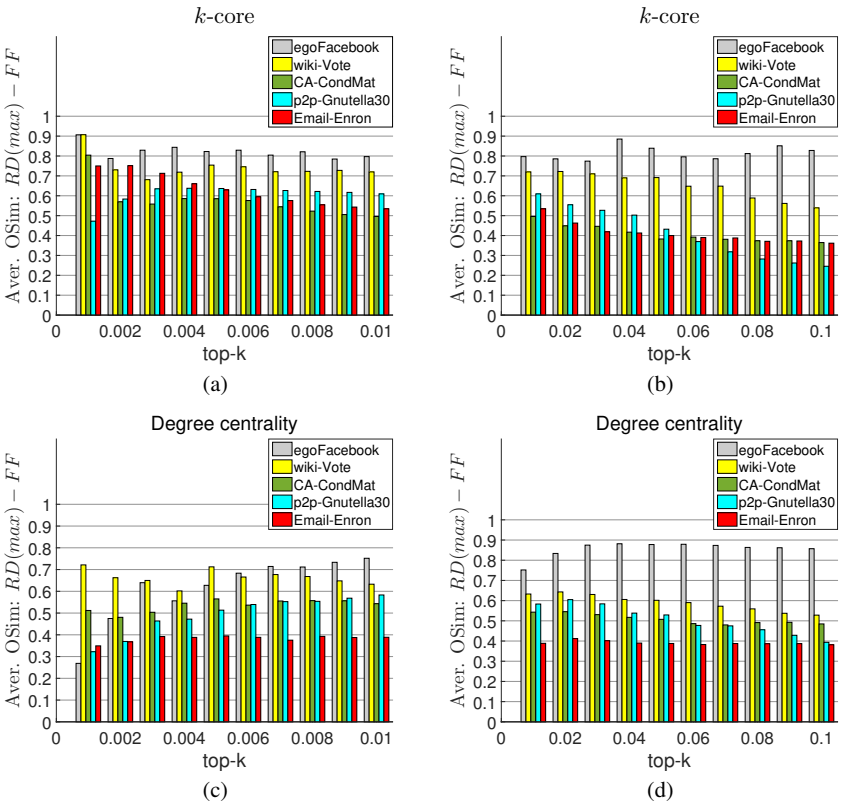

Fig. 4 Comparison of $RD(max)$ and FF: average OSim $RD(max)$ minus average OSim FF per top-k. Samples size 30%

4.3.2 Rank Degree vs Forest Fire

We conclude the analysis comparing our method with the Forest Fire (FF). For each top-k and for each sample size, we compute the difference between the average OSim of RD(max) and the average OSim of FF. We present the results only for k-core and degree centrality, as well as for samples size 30%. The results for the other sample sizes and centrality measures are similar, hence we omit the plots.

Observing the Fig. 4 and taking into account Fig. 1 and Fig. 2, where we present the average OSim between the original graph and all 40 samples, we conclude the following.

In both small and medium datasets and for every top-k, the difference of OSim values in terms of k-core and degree centrality is always positive. The range of difference is roughly between 0.3 to 0.9 which shows that RD is more efficient than FF as an influential nodes identifier.

5 Conclusion

In this paper, we proposed a graph sampling approach to the problem of identifying the influential spreaders in a complex network. Our approach is based on graph sampling and specifically on Rank Degree, an efficient graph exploration sampling algorithm. We experimentally analyzed the proposed method using several centrality measures and studying five real world networks. The analytical experiments demonstrate that our method can identify, with high accuracy, a large fraction of the most influential nodes along with their original ranking in the whole graph. In future, we intend to extend our analysis applying the *SIR* and *SIS* epidemic models that will serve as ground truth information on the spreading efficiency of nodes. More specifically, we will investigate the correlation between the centrality measures and the spreading efficiency of nodes, as defined by the epidemic models in the original graph and in Rank Degree samples.

Acknowledgements We thank Kyriaki Chryssaki for her helpful comments on the final manuscript.

References

1. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems **30**(1), 107 – 117 (1998)
2. Chen, D., Lu, L., Shang, M.S., Zhang, Y.C., Zhou, T.: Identifying influential nodes in complex networks. Physica A: Statistical Mechanics and its Applications **391**(4), 1777 – 1787 (2012)
3. Chen, D.B., Gao, H., Lu, L., Zhou, T.: Identifying influential nodes in large-scale directed networks: The role of clustering. PLoS ONE **8**(10), 1–10 (2013)
4. Costenbader, E., Valente, T.W.: The stability of centrality measures when networks are sampled. Social Networks **25**(4), 283–307 (2003)
5. Freeman, L.C.: Centrality in social networks conceptual clarification. Social Networks **1**(3), 215 – 239 (1978)
6. Haveliwala, T.H.: Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. IEEE Trans. on Knowl. and Data Eng. **15**(4), 784 – 796 (2003)
7. Kempe, D., Kleinberg, J., Tardos, E.: Maximizing the spread of influence through a social network. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '03, pp. 137–146. ACM, New York, NY, USA (2003)
8. Kendall, M.G.: A new measure of rank correlation. Biometrika **30**(1-2), 81–93 (1938)
9. Kitsak, M., Gallos, L.K., Havlin, S., Liljerosand, F., Muchnik, L., Stanley, H.E., Makse, H.A.: Identification of influential spreaders in complex networks. Nature Physics (2010)
10. Leskovec, J., Faloutsos, C.: Sampling from large graphs. In: Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, August 20-23, 2006, pp. 631–636 (2006)
11. Leskovec, J., Kleinberg, J.M., Faloutsos, C.: Graphs over time: densification laws, shrinking diameters and possible explanations. In: Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, Illinois, USA, August 21-24, 2005, pp. 177–187 (2005)
12. Li, Q., Zhou, T., Lu, L., Chen, D.: Identifying influential spreaders by weighted LeaderRank. Physica A: Statistical Mechanics and its Applications **404**, 47 – 55 (2014)
13. Linyuan, L., Zhang, Y.C., Yeung, C.H., Zhou, T.: Leaders in social networks, the delicious case. PLoS ONE **6**(6), 1–9 (2011)

14. Ribeiro, B.F., Towsley, D.F.: Estimating and sampling graphs with multidimensional random walks. In: Proceedings of the 10th ACM SIGCOMM Internet Measurement Conference, IMC 2010, Melbourne, Australia - November 1-3, 2010, pp. 390–403 (2010)
15. Seidman, S.B.: Network structure and minimum degree. Social Networks **5**(3), 269 – 287 (1983)
16. Shai, C., Shlomo, H., Scott, K., Yuval, S., Eran, S.: From the Cover: A model of Internet topology using k-shell decomposition. PNAS **104**(27), 11,150–11,154 (2007)
17. Valente, T.W., Coronges, K., Lakon, C., Costenbader, E.: How correlated are network centrality measures? Connections (Toronto, Ont.) **28**(1), 16–26 (2008)
18. Voudigari, E., Salamanos, N., Papageorgiou, T., Yannakoudakis, E.J.: Rank degree: An efficient algorithm for graph sampling. In: 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), August 18-21, 2016, San Francisco, CA, USA (2016)
19. Zeng, A., Zhang, C.J.: Ranking spreaders by decomposing complex networks. Physics Letters A **377**(14), 1031 – 1035 (2013)