Pre-print version of the article: Soc. Netw. Anal. Min. (2017) 7: 24
The final publication is available at Springer via doi:10.1007/s13278-017-0441-6

Deterministic graph exploration for efficient graph sampling

Nikos Salamanos · Elli Voudigari · Emmanuel J. Yannakoudakis

Received: 22 January 2017 / Revised: 30 April 2017 / Accepted: 6 May 2017

Abstract Graph sampling is a widely used procedure in social network analysis, has attracted great interest in the scientific community and is considered as a very powerful and useful tool in several domains of network analysis. Apart from initial research in this area, which has proposed simple processes such as the classic *Random Walk* algorithm, *Random Node* and *Random Edge* sampling, during the last decade, more advanced graph sampling approaches have been emerged. In this paper, we extensively study the properties of a newly proposed method, the *Rank Degree* method, which leads to representative graph subgraphs. The *Rank Degree* is a novel graph exploration method which significantly differs from other existing methods in the literature. The novelty of the *Rank Degree* lies on the fact that its core methodology corresponds to a deterministic graph exploration; one specific variation corresponds to a number of parallel deterministic traverses that explore the graph. We perform extensive experiments on twelve real world datasets of a different type, using a variety of measures and comparing our method with *Forest Fire*, *Metropolis Hastings Random Walk* and *Metropolis Hastings*. We provide strong evidence that our approach leads to highly efficient graph sampling; the generated samples preserve several graph properties, to a large extent.

Nikos Salamanos ✉
Athens University of Economics and Business
76, Patission Str. GR10434 Athens, Greece
Tel.: +302108203 911
E-mail: salaman@aueb.gr

Elli Voudigari
Athens University of Economics and Business
76, Patission Str. GR10434 Athens, Greece
Tel.: +302108203 911
E-mail: elliv@aueb.gr

Emmanuel J. Yannakoudakis
Athens University of Economics and Business
76, Patission Str. GR10434 Athens, Greece
Tel.: +302108203 911
E-mail: eyan@aueb.gr

Keywords Graph sampling · Graph mining · Social network analysis

1 Introduction

During the last decades, the study of real world networks has attracted considerable attention from different scientific domains, including biology, bioinformatics, computer science, economics, engineering, mathematics, physics, sociology and statistics. One main topic of interest constitutes the estimation of topological features of large and complex networks particularly in recent years, with the vast growth of World Wide Web and the popularity of social networks. In these networks, large amount of information and data are shared and transferred among hundred of million or even billion of users. For instance, the Facebook, one of the most popular social networks, is consisting of 1.18 billion active users, daily and 1.79 billion active users, monthly (September 30, 2016). Other examples constitute co-author and citation networks in DBLP, the extended social media (e.g. Twitter, see Gabielkov et al (2014)) and the World Wide Web. In the area of statistics, a wide variety of proposed network models exist - ranging from models of largely mathematical interest to models designed for statistical data fitting.

Networks constitute a natural representation for a variety of data in real life. Apart from the web-based, there are networks of several domains where the interactions among the entities are represented by links (connections). For instance, in Bioinformatics, exhaustive research has been realized on protein networks in order to discover hidden information about the proteins' interactions with direct application in medicine. On the other hand, the natural problem of modeling networks based on their links has lead to the development of such mathematical models that are able to fit to data statistically (Krivitsky and Kolaczyk (2015)). There have also been developed generative network models that can produce data samples which exhibit the natural characteristics of real world networks (Robles-Granda et al (2016)) and may be used in many important tasks such as sensitivity analysis and benchmark testing (Gu et al (2005)).

In this paper, we focus on graph sampling in terms of efficient representation of a large network, which has gained considerable interest from the scientific community due to their remarkable size, in the last years. This fact has motivated our work to introduce *Rank Degree* (Voudigari et al (2016)), a novel graph exploration sampling method which differs significantly from the existing methods in the literature. The novelty of *Rank Degree* lies on the fact that its core methodology corresponds to a deterministic graph exploration method; one specific variation corresponds to m parallel deterministic traverses that explore the graph.

As we have proved in Voudigari et al (2016), the *Rank Degree* method is superior to other sampling methods, because the generated samples maintain the properties of the original graph - such as *Degree Similarity*, *Clustering Coefficient* and *Diameter*. Additionally, the generated samples maintain the nodes centralities - the important/-central nodes (often called as *influential spreaders*) of the original graph. It is in this direction that we performed an extensive study about the effectiveness of the *Rank Degree* as *influential spreaders* identifier (Salamanos et al (2016)). We were able to successfully identify the *influential spreaders* of 5 large real networks, using the *Rank Degree* method. The results demonstrated that in samples of size 30%, the *Rank Degree* is able to discover more than 80% of the influential nodes.

The contribution of this work is that we extend the analysis of *Rank Degree* on twelve different types of networks, applying new measures for the sampling evaluation and comparing our method with three more, well known, sampling methods - *Forest Fire*, *Metropolis Hastings Random Walk* and *Metropolis Hastings*. The experimental results denote that the generated samples preserve several graph properties, to a large extent.

The rest of this paper is organized as follows. Sect. 2 describes the related work. Sect. 3 presents the *Rank Degree* method. In Sects 4 and 5, we provide information on the measures and datasets used in the experimental analysis. Sect. 6 describes the experimental analysis and Sect. 7 concludes the paper.

2 Related Work

An extensive survey of network sampling is presented in Ahmed et al (2012), where the authors studied and tested network sampling methods not only in static, but also in streaming graphs. Hu and Lau (2013) presented an analytical survey on *traversal-based sampling* methods, such as *Snowball Sampling*, *Metropolis-Hastings Random Walk* and *Forest Fire*, among others. Additionally, they discussed several graph properties that the generated samples have to preserve.

Potamias et al (2009) introduces degree rank-based sampling method in order to estimate the *Landmarks*, in a large graph. The Rank Degree follows a completely different sampling approach from the work followed in Potamias et al (2009). In Potamias *et al.* the input of the method is the centralities values of all nodes in the original graph (global information). Their Landmarks selection method is based on the true nodes rankings/values; those rankings/values remain constant, as well as the graph, during the process of the algorithm. On the other hand, the Rank Degree is a graph exploration algorithm with no prior global information of the original graph. The Rank Degree, at each visited node, ranks only its friends nodes (local information) based on the remaining links, at that time step. Neither the nodes degree nor the graph are stable. At each time step, the traversed edges are added to the sample and at the same time, they are deleted from the original graph. Hence, the algorithm does not cross the same edge for a second time. This edges-elimination process alters the original graph and consequently, the nodes degree and ranking.

Returning to the central subject of this work, a variety of sampling techniques have been proposed. The well known algorithm *Forest Fire*, introduced in Leskovec and Faloutsos (2006), due to its effectiveness, is often used as a benchmark for the evaluation of new sampling methods. Stutzbach et al (2009) investigated the problem of collecting representative graph samples from unstructured peer-to-peer network and proposed *Metropolized Random Walk with Backtracking* as an efficient method which leads to unbiased samples. *Frontier Sampling* is introduced in Ribeiro and Towsley (2010) - a multidimensional Random Walks method - which is able to maintain basic properties of the original graph, even if it is collecting samples of small size. A hybrid sampling model is introduced in Xu et al (2014) which leads to unbiased graph samples incorporating the random walk approach along with the benefits of the random-jump based methods (random vertex sampling), in terms of producing uncorrelated samples. The problem of collecting unbiased samples has also been studied in Gjoka et al (2010), where the authors presented a crawling methodology for collecting uniform samples from Facebook graph.

A sampling algorithm that has been proved to collect representative graph samples of small size is the *Metropolis-Hastings* (Hubler et al (2008)). Initially, a graph sample is selected at random and by incorporating the *Metropolis* algorithm, together with some predefined graph properties, the algorithm adds and removes nodes from the current sample until the graph properties in question are optimized. Moreover, in Lee et al (2012), the authors proposed a sampling technique which combines random walk with Metropolis-Hastings and produces better results than the *Metropolis-Hastings* ones. Furthermore, on a recent study (Li et al (2015)), the authors analyzed the limitations of *random walk* and *Metropolis-Hastings* based sampling methods and introduce the *rejection controlled Metropolis-Hastings* and *generalized maximum-degree random walk* algorithms.

Another important problem in graph sampling is whether the samples can preserve advanced properties of nodes relations. In this direction, Vattani et al (2011) introduced the notion of *personalized page rank value (PPV)* and proposed a method that produces subgraphs which maintain *PPV* values more effectively than the existing sampling methods. In Choudhury et al (2010), the authors evaluated several sampling methods on a large graph from Twitter, with nearly half a million nodes. They developed metrics for evaluating the quality of samples with respect to information diffusion and showed that the most effective methods are those that incorporate the graph structure along with users' activity or attributes. In Krivitsky and Kolaczyk (2015), the authors analyzed the problem of effective sample size in network modeling by considering popular methods developed in the corresponding area of research and they inferred that the requested size is strongly connected to the corresponding model used for its scaling. In Chiericetti et al (2016), the authors also studied the problem of sampling a large graph according to a prescribed distribution. The focus of their work is on the uniform distribution where each node can be selected with probability $\frac{1}{|V|}$ (where $|V|$ is the number of nodes) and they developed a methodology which uses random walk as the basic component.

On more advanced studies on the sampling properties and the evaluation of sampling techniques, as in Maiya and Berger-Wolf (2011), a detailed study is conducted on biases in network sampling strategies, as well as on how specific biases can be beneficial to sampling, with the final goal to define the best sampling strategies. Furthermore, in Çem et al (2013), the authors provide experimental evidence that the design of sampling techniques needs to be dependent on the network characteristics that aim to investigate. Finally, Kurant et al (2011) studied the degree bias of *Breadth First Search (BFS)* comparing the results for *BFS* with the degree bias observed in other popular methods such as *Forest Fire* and *Snowball Sampling*, among others. Then, they proposed a bias-correction technique which can improve the outcome of *BFS*.

3 The Rank Degree Method

The *Rank Degree (RD)* is a graph exploration sampling method based on a deterministic edge selection rule - the ranking of nodes according to their degree values.

Algorithm 1 Rank Degree

```
 1: Parameters: (i) $s$: number of initial seeds, (ii) top-$k$ strategy (see Step-10), (iii) target sample size $x$
 2: Input: undirected graph $G(V, E)$
 3: Output: sample of size $x$
 4: Initialization: $\{Seeds\} \leftarrow s\ nodes\ selected\ uniformly\ at\ random$
 5: $Sample \leftarrow \emptyset$
 6: while sample size < target size x do
 7:     $\{New\ Seeds\} \leftarrow \emptyset$
 8:     for $\forall w \in \{Seeds\}$ do
 9:         Rank w's friends based on their degree values
10:         Selection rule:
11:         (i) $RD(max, s)$: select the max degree (top-1) friend of $w$
12:         (ii) $RD(\rho, s)$: select the top-$k$ friends of $w$, where $k = \rho \cdot (\#friends(w))$, $0 < \rho \leq 1$
13:         Update the current sample with the selected edges $\{(w,\ \text{top-}k\ \text{friend})\}$ along with the symmetric ones
14:         Add to $\{New\ Seeds\}$ the top-$k$ friends of $w$
15:     end for
16:     Update graph: $G \leftarrow G \setminus \{Selected\ Edges\}$ (delete from the graph all the currently selected edges)
17:     $\{Seeds\} \leftarrow \{New\ Seeds\}$
18:     If $\{New\ Seeds\} = \emptyset$ (all nodes in Seeds were leaves (degree = 1))
19:     then repeat Step-4 (random jump)
20: end while
```

The RD process is presented by Algorithm 1. The method is based on three parameters (s, top-k strategy, x): (i) s is the number of initial seeds, i.e. the starting points of the algorithm. (ii) The top-k strategy controls the edge selection rule during the sampling process. It is defined either as top-1 or by a parameter $0 < \rho \leq 1$, hence the notation $RD(max, s)$ and $RD(\rho, s)$. (iii) x is the target sample size.

Initially, s nodes (seeds) are selected uniformly at random (Step 4). Then, at each step and for each node w in the current set of seeds, we do the following. We find the set of nodes that have an edge with w (the *friends* of w) and we rank $w's$ friends based on their degree values (Step 9). From the ranking list, we collect the top-k friends $v_1, v_2, \ldots, v_k$, i.e. the first k friends of w with the largest degree (Step 11 or Step 12). In the case that the node in question has degree less than k, we select one friend node (instead of none) - the friend with the largest degree. The current sample will be updated by the selected edges $(w, v_1), (w, v_2), \ldots, (w, v_k)$ along with the symmetric ones $(v_1, w), (v_2, w), \ldots, (v_k, w)$ (Step 13). Moreover, the end nodes $v_1, v_2, \ldots, v_k$ are added to the new set of seeds (Step 14).

At the end of this iterative process, the current set of seeds has been fully examined. Then, the selected edges from all seed nodes are removed from the graph (Step 16). Hence, the algorithm does not cross the same edges for a second time.

Observe that, from a current set of Seeds, it is possible to select the same edge twice. For instance, consider two seed nodes w_i, w_j which are connected by an edge and they both have the highest degree value for each other (at their neighborhood).

When the algorithm examines node w_i, it will add to the set of selected edges the edge (w_i, w_j) as well as the symmetric one (w_j, w_i). But since w_i, w_j are mutual friends, the algorithm will add the same pair of edges during the examination of the seed node w_j - hence the duplicate entries must be deleted.

Let us observe that the selection process - the phase where the algorithm collects the top-k friend-nodes from each node in the current set of seeds, is deterministic. Thus, the graph traverse and consequently the generated sample are uniquely determined by the initial seeds. Only the initialization phase (Step 4, initial seeds selection) as well as the extreme case where all the current seeds are leaves (nodes with degree equal to one) are randomized. In the latter case, the algorithm executes a random jump (Step 19) in order not to get trapped.

Another observation is that in the case of top-1 strategy, $RD(max, s)$, the algorithm corresponds to s parallel deterministic walkers that explore the graph and share information about all the edges that have been currently traversed. Moreover, since the selection rule is deterministic, if two walkers simultaneously visit the same seed, then, in the next time step, those two walks will collapse to one.

One may argue that the algorithm is biased on the large degree values - which is partially correct. First, the initial seeds are defined randomly, allowing the algorithm to start from different areas of the graph, where some nodes may be of low degree. Furthermore, the visited nodes remain present in the graph and only the selected edges are removed. Thus, any selected node can be visited multiple times, but not from the same paths. The last fact provides an insight into the ability of RD to collect samples which maintain the original graph properties.

In the rest of the paper and for the sake of clarity of presentation, instead of the notations $RD(max, s)$ and $RD(\rho, s)$ we use the simplified versions maxRD and rhoRD, respectively. Moreover, as referred to in Voudigari et al (2016), the rhoRD version generates the most representative samples for $\rho = 0.1$ i.e. when we select the top-10% from the ranking lists (Step 12). Hence, in this paper, we concentrate our analysis to maxRD and rhoRD for $\rho = 0.1$.

4 Methods and Measures

4.1 Sampling Methods

In Voudigari et al (2016), we compared the efficiency of RD with five sampling methods, *Forest Fire (FF)*, *Frontier Sampling (FS)*, *Random Walk (RW)*, *Random Node (RN)* and *Random Edge (RE)*. Performing extensive experimental evaluation we concluded that only the performance of *Forest Fire* was comparable to *Rank Degree*, while all the other methods showed lower performances. In this paper, we concentrate our attention to *Forest Fire*, *Metropolis Hastings Random Walk* and *Metropolis Hastings*. The first two methods perform a graph exploration, while the *Metropolis Hastings* takes as input a global property of the original graph, hence, it can be viewed as a ”centralized” algorithm.

Forest Fire (FF) algorithm starts from a randomly selected node (seed); 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 (seeds), a random number x is generated which is geometrically distributed with mean $p_f(1 - p_f)$ (see Leskovec et al (2005), Leskovec and Faloutsos (2006), Leskovec et al (2007)). The parameter p_f is called *forward burning probability* which is set to 0.7. Then, x outgoing edges are selected from node w outgoing edges. The end nodes of the selected edges constitute the next set of current nodes (seeds). 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.

Metropolis-Hastings Random Walk (MHRW) (see Stutzbach et al (2009), Gjoka et al (2011), Li et al (2015), Chierichetti et al (2016)) is an application of the Metropolis algorithm (Metropolis et al (1953)) for uniform sampling. It modifies the Random Walk algorithm as follows: (i) First, select a node x, uniformly at random. (ii) Select a neighbor y of x, uniformly at random. (iii) Find the degree of y. (iv) Generate a random number, p, uniformly between 0 and 1. If $p \leq \frac{degree(x)}{degree(y)}$, y is the next step. Otherwise, remain at x, as the next step. A condition of the algorithm is that the graph is connected. In this paper, we study the real world graphs as they are, hence, they may not be well connected. In this case, it is possible that the MHRW will get trapped in a small region of the graph. In order to avoid that, we increase the number of initial seeds from one to 1% (see also Sect. 5.2).

Metropolis-Hastings (MH) (Hubler et al (2008)) is a sampling algorithm that is able to collect representative graph samples of small size. Given a target sample size n, the algorithm initially selects n nodes, at random, the edges of which form an initial graph sample. Then, by incorporating the *Metropolis* algorithm, together with some predefined graph properties, the algorithm adds and removes nodes from the current sample, until the graph properties in question are optimized. Hubler et al (2008) proposed several sampling strategies; one of the best uses the degree distribution as the optimized property. This is the version of Metropolis-Hastings, implemented in this paper.

4.2 Sampling evaluation

4.2.1 Distance measures and correlation

Distances: We study the efficiency of the sampling methods using six distance measures:

1. **Δ(Degree)**: the distance between the sampled graph to the original graph with respect to the degree distribution. We use the *Total Variation Distance (TVD)* - a well known metric for measuring the distance between distributions, as it has been previously used in graph sampling (Bar-Yossef and Gurevich (2008), Li et al (2015)). Hence, in this paper we define $\varDelta(Degree)$ = TVD. The TVD distance between two probability distributions μ and ν on the same space $\varOmega$ is defined by (see, Levin et al (2008))
$$\|\mu - \nu\|_{TV} = \max_{A \subseteq \varOmega} |\mu(A) - \nu(A)| = \frac{1}{2} \sum_{x \in \varOmega} |\mu(x) - \nu(x)| = \sum_{\substack{x \in \varOmega \\ \mu(x) > \nu(x)}} |\mu(x) - \nu(x)|$$
.

Additionally, we computed the *two-sample Kolmogorov-Smirnov (K-S) test* on the cumulative distribution functions (CDF) of the two graphs, considering the degree as a random variable.

The result is the distance measure called *D-statistic* (see Leskovec and Faloutsos (2006), Hubler et al (2008), Maiya and Berger-Wolf (2011)), whose value is defined as the maximum difference (distance), $max_x|F_G(x) - F_S(x)|$, between the two CDFs, where x is the range of nodes degrees. F_G and F_S are the cumulative degree distributions of the original graph G and its sample S, respectively.

2. **Δ(LocalCC)**: the normalized absolute $L1$-norm distance $\dfrac{L1(CC_d(G), CC_d(S))}{|D|} = \dfrac{\sum_{d \in D} |C_d(G) - C_d(S)|}{|D|}$, where $CC_d(G)$ and $CC_d(S)$ are the *local clustering coefficient distributions* of the original graph G and of a given sample S, respectively and D is the domain of the degree values of G and S. The CC_d is a vector consisting of the values C_d - the average local clustering coefficient of the nodes with degree d. The notion of *local clustering coefficient distribution* has been previously used in Leskovec and Faloutsos (2006) and Hubler et al (2008).
3. **Δ(ACC)**: the relative error between the *average clustering coefficient (ACC)* of the original graph G and the ACC of a given sample S.
   Hence, $\varDelta(ACC) = \dfrac{|ACC_S - ACC_G|}{ACC_G}$.
4. **Δ(GCC)**: the relative error between the *global clustering coefficient (GCC)* of the original graph and the GCC of a given sample.
5. **Δ(Diameter)**: the relative error between the *diameter* of the original graph and the *diameter* of a given sample.
6. ***Total Distance***: It measures the overall quality of a given sample subgraph. It is the average value of the five distances $\varDelta(Degree)$, $\varDelta(LocalCC)$, $\varDelta(ACC)$, $\varDelta(GCC)$ and $\varDelta(Diameter)$ of the sample in question.

Correlation: Given a sample subgraph, we measure the *Pearson correlation coefficient* between the centrality values of the nodes in the sample, with the centrality values that the sample nodes have in the original graph.

We use four centrality measures: (i) *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, (ii) *degree centrality*, (iii) *betweenness centrality* and (iv) *closeness centrality* (Freeman (1978)).

4.2.2 top-k similarity

We compare the top-k nodes in the original graph with the top-k nodes in the samples, using the four centrality measures that we mentioned previously. The top-k nodes are often called in the literature as *influential spreaders* - the nodes that are more likely to spread information or a virus in a large part of the network (see Kitsak et al (2010), Chen et al (2012), Zeng and Zhang (2013)).

First, given a centrality measure, we rank the nodes in the original graph and in the samples based on the centrality values. Then, we compare the ranking lists applying the object similarity measure ***OSim*** (Haveliwala (2003)). In our case, the objects are the nodes on the ranking lists. The *OSim* is 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. Then, we rank the nodes accordingly (in descending order), creating the ranking lists r_G and r_S. Afterwards, 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.

5 Datasets and Setup

5.1 Datasets

We have used twelve datasets of different type (social networks, collaboration networks, location based social network etc.), previously used for graph mining (Leskovec and Krevl (2014)). We restrict our analysis to undirected graphs, therefore we transform the directed graphs to undirected ones, by applying the symmetric one to each edge, after removing the self loops, if any.

We study four large graphs (D9 - D12). (a) one very dense, the dataset D9 *flickr*, which is large in regard to the number of edges, (b) one large graph, the D10 (com-DBLP) and (c) two very large, the *com-Youtube* and *wiki-Talk* (D11, D12).

Table 1 presents their basic characteristics (undirected versions) which have been computed using the igraph R package (Csardi and Nepusz (2006)). Briefly:

D1 ***egoFacebook*****:** undirected graph of $4,039$ users' "friends-list", i.e. the ego-net, from Facebook (McAuley and Leskovec (2012)).
D2 ***wiki-Vote*****:** voting network - directed graph - from *Wikipedia* consisting of $7,115$ users (Leskovec et al (2010)).
D3 ***CA-CondMat*****:** scientific collaborations network - undirected - between $23,133$ authors with paper submitted to Condense Matter category (Leskovec et al (2007)).
D4 ***p2p-Gnutella30*****:** Gnutella $p2p$ network topology - directed - of $36,682$ nodes (hosts in the Gnutella network) (Ripeanu et al (2002), Leskovec et al (2007)).
D5 ***Email-Enron*****:** email communication network - undirected - of $36,692$ nodes (email addresses) (Leskovec et al (2009)).
D6 ***loc-Brightkite*****:** online location-based social network - undirected - of $58,228$ nodes (Cho et al (2011)).
D7 ***soc-Epinions1*****:** web of trust obtained from Epinions - directed - of $75,879$ nodes, members of www.epinions.com (Richardson et al (2003)).
D8 ***soc-Slashdot0922*****:** social network - directed - of $82,168$ nodes/users (Leskovec et al (2009)).
D9 ***Flickr*****:** contact network - undirected - (group membership) from Flickr of 80,512 nodes/users (Zafarani and Liu (2009)).

Table 1: Datasets

Dataset	Type	# Nodes	# Edges	ACC	GCC	Diameter
D1: egoFacebook	Ego-net Undirected	4,039	88,234	0.6055	0.5192	8
D2: wiki-Vote	Wiki-net Directed	7,115	100,762	0.1409	0.1255	7
D3: CA-CondMat	Collabor. Net. Undirected	23,133	93,439	0.6334	0.2643	15
D4: p2p-Gnutella30	P2P Net. Directed	36,682	88,328	0.0063	0.0052	11
D5: Email-Enron	Comm. Net. Undirected	36,692	183,831	0.4970	0.0853	13
D6: loc-Brightkite	loc. social net. Undirected	58,228	214,078	0.1723	0.1106	18
D7: soc-Epinions1	Social net. Directed	75,879	405,740	0.1378	0.0657	15
D8: soc-Slashdot0922	Social net. Directed	82,168	504,230	0.0603	0.0241	13

Large

Dataset	Type	# Nodes	# Edges	ACC	GCC	Diameter
D9: Flickr	Social net. Undirected	80,512	5,899,882	0.1652	0.1875	6
D10: com-DBLP	Collabor. Net. Undirected	317,080	1,049,866	0.6324	0.3067	23
D11: com-Youtube	Social net. Undirected	1,134,890	2,987,624	0.0808	0.0062	24
D12: wiki-Talk	Comm. Net. Directed	2,934,385	4,659,568	0.0526	0.0022	11

D10 ***com-DBLP***: scientific co-authorship undirected network of 317,080 nodes/authors (Yang and Leskovec (2012)).

D11 ***com-Youtube***: Youtube friendship undirected network of 1,134,890 nodes/users (Yang and Leskovec (2012)).

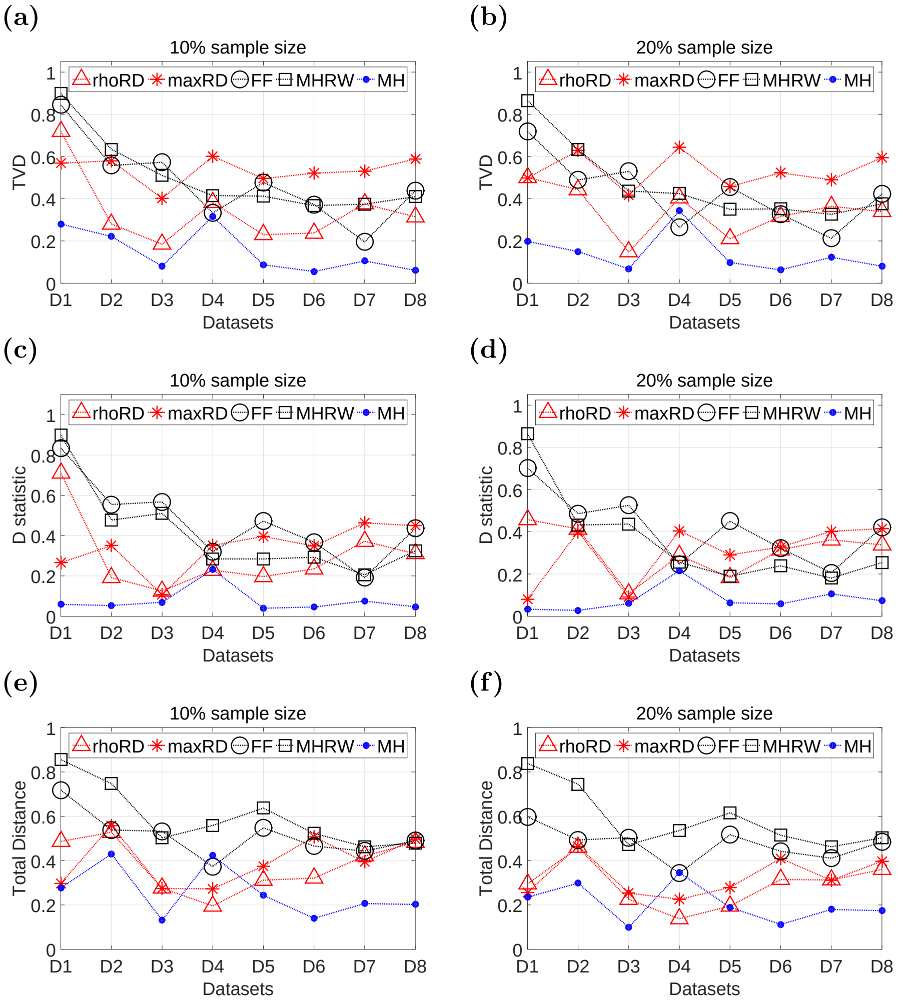
D12 ***wiki-Talk***: Wikipedia users directed network - 2,934,385 nodes/users -, where an edge from node i to node j represents that user i at least once edited a talk page of user j (Leskovec et al (2010)).

5.2 Experimental Setup

We simulate two top-*k* strategies of Rank Degree, (a) the top-1 that we call as maxRD and (b) the rhoRD, top-10% (i.e. $\rho = 0.1$). (see Algorithm 1, Steps 11 & 12). For both versions, the number of initial seeds is defined as the 1% over the number of nodes in the original graph. Moreover, in all datasets, the number of initial seeds is equal to one and 1% for FF and MHRW, respectively. Finally, we study the MH in the first eight datasets.

For each method and for each of the D1 - D8 datasets (see Table 1), we collected 300 samples, with final target sample size, the 20% of the graph. For the RD, FF and MHRW, in each sampling trial, we track the actual sample size and we also store the samples size 10%. For the MH, we collect the samples of 10% and 20%, independently. Furthermore, in each sampling trial, the iterations of MH are set to

Fig. 1: Degree distribution similarity and total distance. **a**, **b** TVD. **c**, **d** D-statistic. **e**, **f** Total Distance. Sample size 10% (plots on the left) and 20% (plots on the right)



20,000 and 30,000 for the datasets D1 - D5 and D6 - D8, respectively. In the large datasets D9 - D12, the first four methods have been studied by collecting 40 samples per method and dataset.

Finally, the experiments have been implemented in MATLAB and R. The standard graph properties of the samples and the original graphs have been computed using the igraph R package (Csardi and Nepusz (2006)). The code of the measures (see Sect. 4.2) is publicly available[1].

[1] https://drive.google.com/file/d/0BxezFIUuAvc3a29vY0Fqd3J5RUk

6 Results

6.1 Degree distribution similarity and Total distance

Figure 1 presents the average values of TVD, D-statistic as well as Total Distance over the 300 samples, for samples size 10% and 20%, respectively. We consider the TVD as a more accurate measure than the D-statistic, hence, the Total Distance values have been computed using the TVD values, along with the other four distance measures (see Sect. 4.2.1).

Generally, in all datasets, the MH has the best performance in terms of degree distribution similarity and Total Distance. This is expected, since the MH is a centralized algorithm which takes as input the global property of the original graph - the original degree distribution - and running for thousands iterations, it performs a degree distribution approximation. In real world large networks, this approach is not realistic, because the original networks are usually unknown.

The rhoRD overall has very good performance, and is, in general, the second best method in regard to degree distribution and Total Distance. This fact demonstrates the importance of the ranking based selection rule, which is the fundamental part of the RD methodology.

Comparing the two variations of RD with the FF and MHRW, we conclude that the rhoRD is superior to maxRD as well as to FF and MHRW. Specifically:
For samples size 10%, in six out of eight datasets, the TVD, D-statistic and Total Distance values of rhoRD are lower than those of FF and MHRW (Fig. 1a, c & e). In the dataset D4, (*p2p-Gnutella30*), the TVD and D-statistic values of the three methods almost coincide (Fig. 1a & c). Only in the dataset D7 (*soc-Epinions1*), the TVD and D-statistic values of FF are lower than those of rhoRD (Fig. 1a & c).

From Fig. 1e, we observe that in all datasets, the Total Distance values of rhoRD and maxRd are equal or lower than those of FF and MHRW and in three out of eight datasets are very close to the MH values.

Similar results we have for samples of size 20%. The TVD and D-statistic values of rhoRD are always equal or lower of those of FF and MHRW. Only in the datasets D4 and D7, the FF TVD values are lower (Fig. 1b). Moreover, in regard to the Total Distance, the rhoRD and maxRD clearly outperform the FF and MHRW. Moreover, in D1, D4 and D5 datasets, the Total Distances of rhoRD and maxRD are at least equal to MH (Fig. 1f).

6.2 Distances

Figure 2 presents in details the average values of the four distances over the 300 samples of each sampling method, dataset and samples size.

Generally, in all the four metrics and almost in all datasets, rhoRD and maxRD have larger performances than FF and MHRW and in some metrics are comparable to MH. Specifically:

The $\Delta(LocalCC)$ values of RD, FF, MHRW and MH are very close for both sample sizes and in all datasets (Fig. 2a & b).

Fig. 2: Average Distances: **a**, **b** clustering coefficient distribution C_d. **c**, **d** average clustering coefficient (ACC). **e**, **f** global clustering coefficient (GCC). **g**, **h** Diameter. Samples size 10% (plots on the left) and 20% (plots on the right)

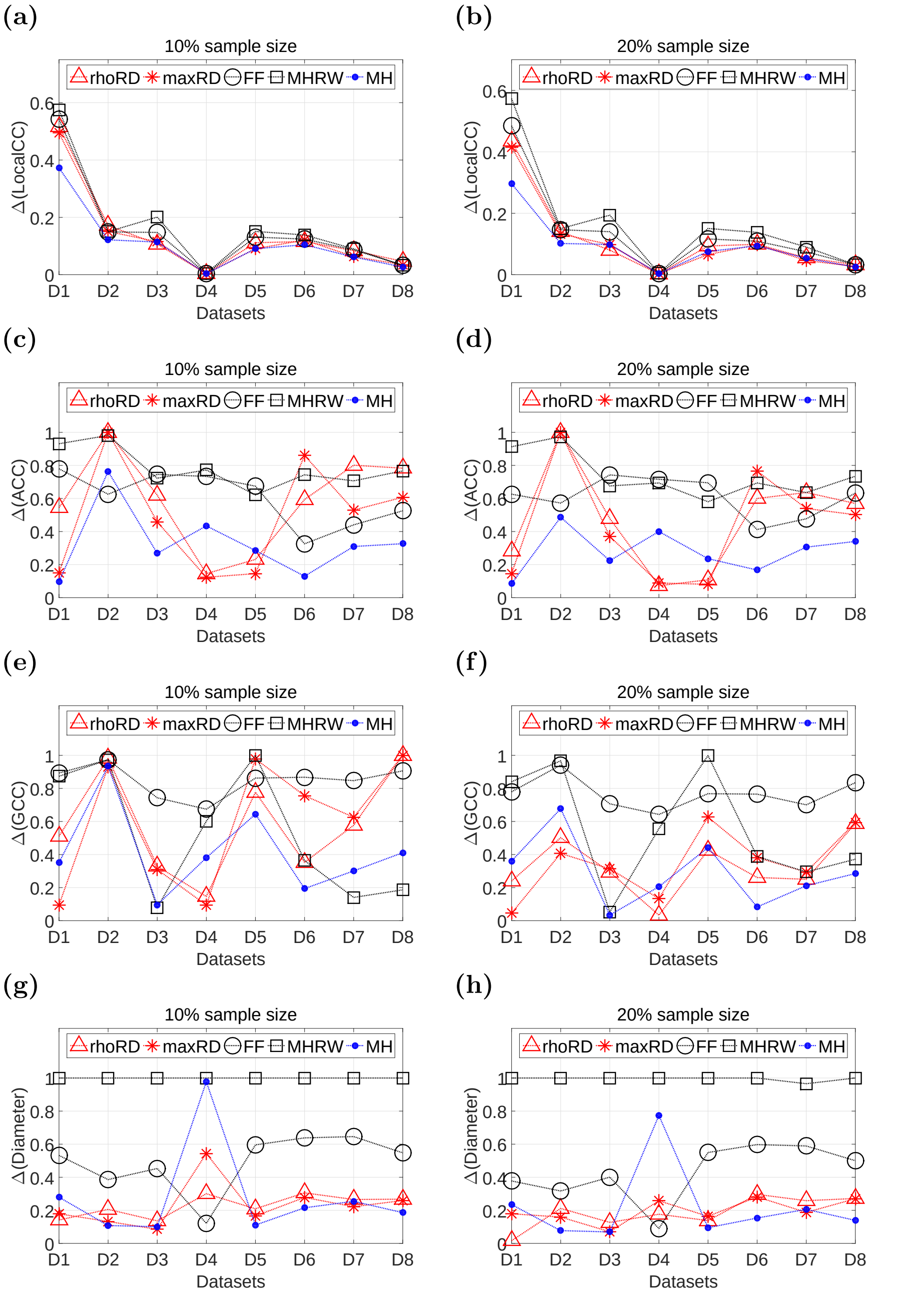
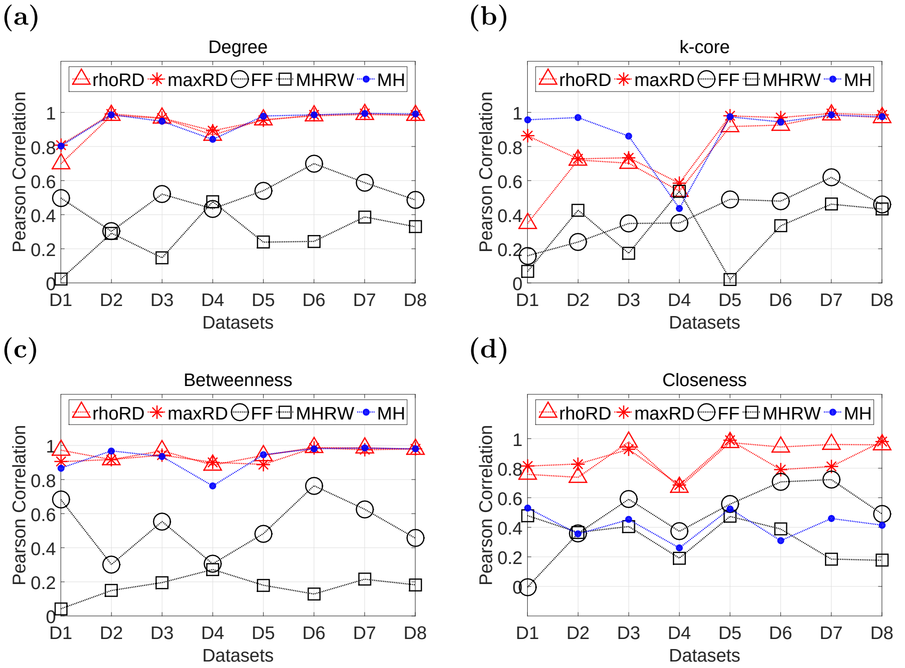


Fig. 3: Pearson correlation coefficient between the centrality values of the nodes in the samples and in the original graph. **a** Degree, **b** k-core, **c** Betweenness and **d** Closeness. Samples size 20%



As regards the $\varDelta(ACC)$ distance, the situation is differentiated between RD and FF, MHRW. The RD has better performance than those methods, in at least five out of eight datasets, in both sample sizes (Fig. 2c & d). Moreover, in D4 and D5 datasets the RD outperforms MH.

The RD clearly outperforms FF and MHRW in regard to $\varDelta(GCC)$. The performances of RD and MH are comparable. For instance, in samples 20%, the rhoRD values are lower than 0.4, in five datasets - almost half of the correspondent FF and MHRW values (Fig. 2e & f).

The outcome is much clearer in the case of $\varDelta(Diameter)$. The results for RD and MH are clearly comparable. Moreover, RD outperforms FF and MHRW. Especially the MHRW fails completely to estimate the correct diameter. Only in the dataset D4, (*p2p-Gnutella30*), the values of RD and FF methods are close enough. In the rest of the datasets, the RD values are smaller than half of the FF values (Fig. 2g & h).

6.3 Correlation

For a given graph G and centrality measure, we compute the *Pearson correlation coefficient* between the centrality values of the nodes in the sample with the centrality values that the sample nodes have in the original graph. For the sake of clarity, only the samples sizes 20% are presented.

In Figure 3, we present, for each dataset and sampling method, the average Pearson correlation coefficient values over the 300 samples. It is clear that the values of rhoRD and maxRD are very large in almost all datasets and centrality measures - in some cases larger than those of MH - while the Pearson values for the FF and MHRW are significantly lower.

For *degree centrality* and *betweenness centrality*, the average Pearson correlation of rhoRD and maxRD is always larger than 0.8, while the correspondent values of FF and MHRW are significantly lower (Fig. 3a & c). Similar are the results for *k*-core, in six out of eight datasets (Fig. 3b). Finally, for *closeness centrality*, the two versions of RD outperform all the other methods (Fig. 3d).

6.4 top-k Similarity

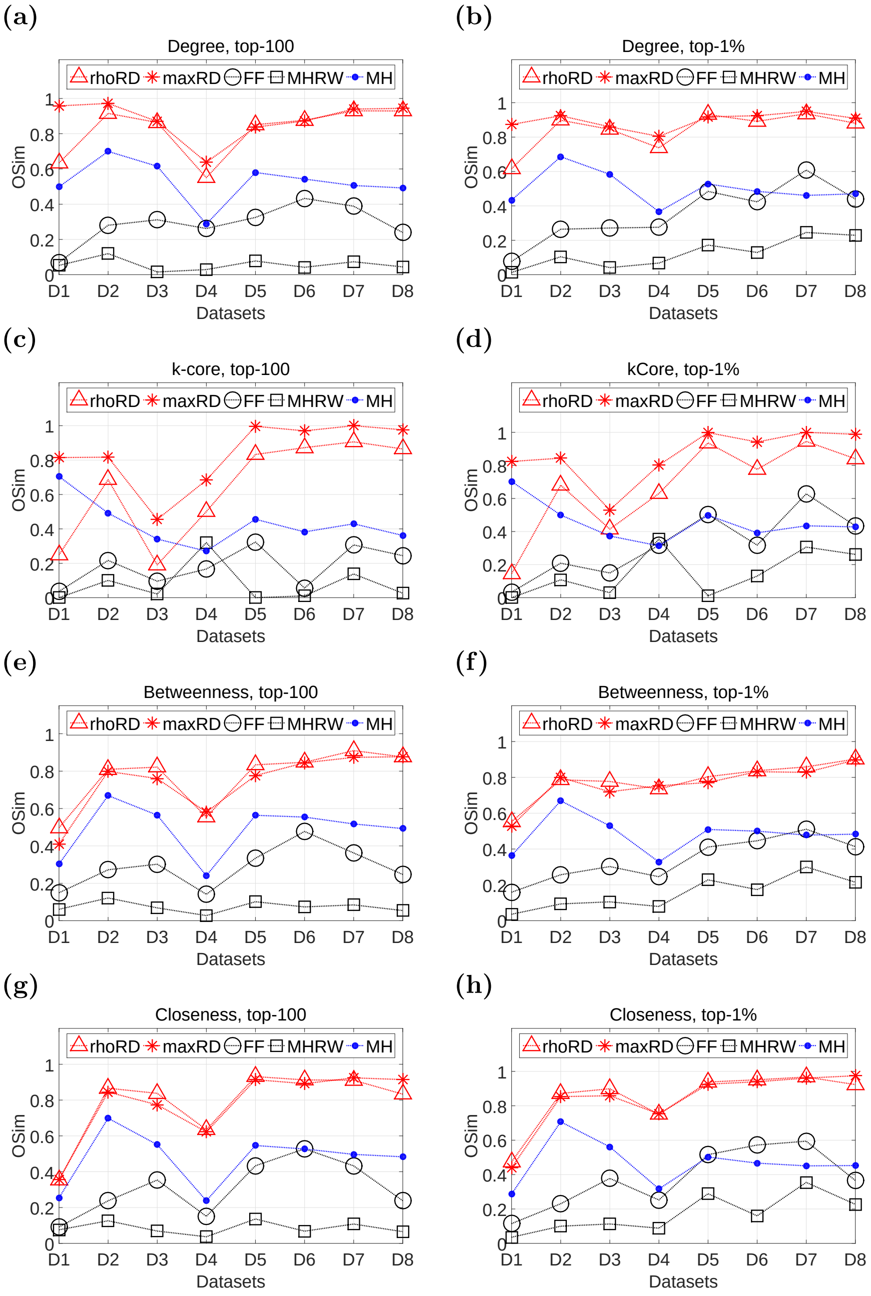
Any effective sampling method has to generate samples which can serve as a thumbnail of the original graph. The question we address here is whether the samples contain a large fraction of the central nodes of the original graph. In this direction, we apply OSim as a measure of the effectiveness of the sampling methods, with respect to four centrality measures: degree, *k*-core, closeness and betweenness.

Table 2: Average OSim for top-10. Samples size 20%

Datasets	D1	D2	D3	D4	D5	D6	D7	D8
				Degree				
rhoRD	0.6903	0.8983	**0.9883**	0.3100	0.7007	0.8993	**0.9000**	**0.7003**
maxRD	**0.8000**	**0.9000**	0.9763	**0.4050**	**0.7010**	**0.9000**	**0.9000**	0.7000
FF	0.1563	0.1843	0.3913	0.2597	0.3630	0.7177	0.4800	0.3617
MHRW	0.0050	0.0243	0.0017	0.0037	0.0117	0.0100	0.0120	0.0097
MH	0.4627	0.6460	0.6723	0.2243	0.5550	0.5953	0.5277	0.4703
				k-core				
rhoRD	0.1027	0.7267	0.0027	0	0.8827	0.8837	0.9590	0.8287
maxRD	**0.8887**	**0.8003**	0.2643	0	**1**	**0.9003**	**1**	**0.9913**
FF	0.4433	0.1543	**0.5777**	0.0270	0.2620	0.0387	0.3137	0.2073
MHRW	0	0.1530	0	0	0	0	0	0
MH	0.7177	0.5447	0.2780	**0.0560**	0.5137	0.4037	0.4277	0.3607
				Betweenness				
rhoRD	**0.8200**	**0.8917**	**0.8790**	0.4920	**0.7290**	**0.9010**	**0.9767**	**0.9230**
maxRD	0.5280	0.7873	0.8363	**0.6253**	0.6783	0.9000	0.9000	0.9000
FF	0.2327	0.2120	0.3723	0.0500	0.2490	0.6547	0.4150	0.3063
MHRW	0.0130	0.0210	0.0183	0.0030	0.0247	0.0140	0.0100	0.0033
MH	0.5150	0.6537	0.6590	0.2073	0.5373	0.5873	0.5297	0.4960
				Closeness				
rhoRD	**0.7113**	0.8977	**0.8973**	**0.6447**	**0.9007**	**0.9717**	**0.9537**	0.7067
maxRD	0.3123	**0.9000**	0.8000	0.6290	0.8997	0.9210	0.9000	**0.8977**
FF	0.1587	0.1993	0.3473	0.0473	0.3147	0.5457	0.3820	0.1997
MHRW	0.0083	0.0180	0.0157	0.0043	0.0217	0.0093	0.0087	0.0050
MH	0.3510	0.6713	0.5997	0.1277	0.5540	0.5797	0.5177	0.4910

Bold values are the highest values which according to the definition of OSim (Sect. 4.2.2) represent the most efficient method

Fig. 4: Average OSim values for top-100 (plots on the left) and top-1% (plots on the right). **a**, **b** Degree. **c**, **d** k-core. **e**, **f** Betweenness. **g**, **h** Closeness. Samples of size 20%



Towards this direction, we calculate the OSim between the top-*k* nodes of the original graph with the correspondent top-*k* nodes in each sample, where $k \in \{10, 100, 1\%\}$. In the case of top-1%, the 1% is referring to the number of nodes N of the original graph, hence, $k = 0.01 * N$.

In Figure 4, we present, for samples size 20%, the average $OSim$ values (over the 300 samples), for top-100 and top-1%. Additionally, in Table 2 we show the results for top-10, highlighting the highest values which according to the definition of OSim (Sect. 4.2.2) depict the most efficient method.

The results provide clear evidence that RD collects samples which contain a very large part of the central nodes of the original graph and in most cases, with almost double accuracy compared to other methods.

Specifically, in almost all centrality measures, top-*k* and for every dataset, the RD values are larger than those of FF, MHRW and MH. For instance, in datasets D6 to D8 - the largest three of the first eight datasets - RD identifies almost the 90% of the central nodes for every top-k and centrality measure.

An exception is the case of the top-10 of D4 (*p2p-Gnutella30*) in regard to *k*-core (Table 2). Although the RD fails to identify the 10 best nodes of the graph, it nevertheless identifies more than 60% of the top-100 nodes (see maxRD, Fig. 4c).

6.5 Large Graphs

We conclude the evaluation of the sampling methods by repeating the previous analysis in four large datasets. As we mentioned in Sect. 5.2, in large graphs, we analyze the RD, FF and MHRW by collecting, for each method, 40 samples of 10% size per dataset.

In Figure 5, we plotted the average distances for the first three, D9-D11, large datasets and in addition, in Table 3, we present the results for the *wiki-Talk*, the largest dataset (D12). In Table 3, we have highlighted the lowest values which according to the measures definition (Sect. 4.2.1) correspond to the best method. In two out of four datasets, the TVD and D-statistic values of rhoRD and maxRD are significantly lower than those of FF and MHRW (Fig. 5a & b).

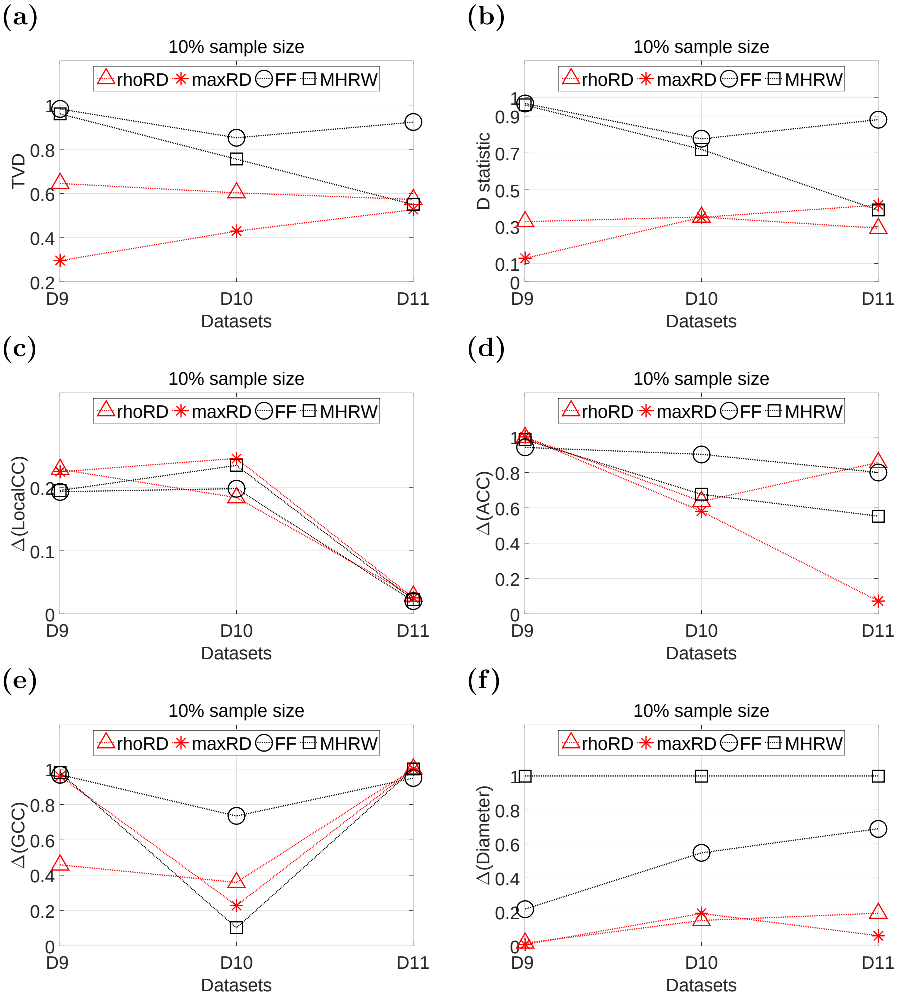
An important result of our method is that even the small samples of 10% have a diameter very close to the original graph diameter. We recall that the $\Delta(Diameter)$ is simply the relative error and for our method is always lower than 0.2 in D9-D11 (Fig. 5f) and only 0.14 in *wiki-Talk* (Table 3).

Table 3: Large graphs: wiki-Talk, dataset D12, samples size 10%

	TVD	D-statistic	Δ(LocalCC)	Δ(ACC)	Δ(GCC)	Δ(Diameter)	Total Dist.
maxRD	**0.7478**	**0.6183**	0.0250	**0.6726**	1	**0.1409**	**0.5173**
FF	0.9671	0.9507	**0.0175**	0.8040	0.8261	0.4705	0.6170
MHRW	0.7761	0.6747	0.0186	0.8440	**0.4705**	0.7000	0.5618

Bold values are the lowest values which according to the measures definition (Sect. 4.2.1) correspond to the best method

Fig. 5: Average distances for the large graphs and samples size 10%: **a** TVD, **b** D-statistic, **c** clustering coefficient distribution, **d** ACC, **e** GCC and **f** Diameter



7 Conclusion

In this paper, we have presented the *Rank Degree*, a deterministic graph exploration method, which produces representative samples - subgraphs - from complex real world networks. We performed an extensive experimental analysis on twelve real world datasets of different types, using several evaluation metrics and have found that the experimental results support the effectiveness of the method. In future, we intend to focus our analysis in the investigation of the fundamental reasons that make *Rank Degree* so effective, as well as to extend the applications and analysis of the method to the problem of *influential spreaders* identification in a real world network.

Acknowledgements We thank Kyriaki Chryssaki for her helpful comments on the final manuscript. We also wish to acknowledge the Athens University of Economics and Business for the technical equipment provision.

References

Ahmed NK, Neville J, Kompella RR (2012) Network sampling: From static to streaming graphs. CoRR abs/1211.3412

Bar-Yossef Z, Gurevich M (2008) Random sampling from a search engine's index. J ACM 55(5):24:1–24:74

Çem E, Tozal ME, Saraç K (2013) Impact of sampling design in estimation of graph characteristics. In: IEEE 32nd International Performance Computing and Communications Conference, IPCCC 2013, San Diego, CA, USA, December 6-8, 2013, pp 1–10

Chen D, Lu L, Shang MS, Zhang YC, Zhou T (2012) Identifying influential nodes in complex networks. Physica A: Statistical Mechanics and its Applications 391(4):1777 – 1787

Chiericetti F, Dasgupta A, Kumar R, Lattanzi S, Sarlós T (2016) On sampling nodes in a network. In: Proceedings of the 25th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, WWW '16, pp 471–481

Cho E, Myers SA, Leskovec J (2011) Friendship and mobility: User movement in location-based social networks. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, KDD '11, pp 1082–1090

Choudhury MD, Lin Y, Sundaram H, Candan KS, Xie L, Kelliher A (2010) How does the data sampling strategy impact the discovery of information diffusion in social media? In: Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010, Washington, DC, USA, May 23-26, 2010

Csardi G, Nepusz T (2006) The igraph software package for complex network research. InterJournal Complex Systems:1695, URL http://igraph.org

Freeman LC (1978) Centrality in social networks conceptual clarification. Social Networks 1(3):215 – 239

Gabielkov M, Rao A, Legout A (2014) Sampling online social networks: an experimental study of twitter. In: ACM SIGCOMM 2014 Conference, SIGCOMM'14, Chicago, IL, USA, August 17-22, 2014, pp 127–128

Gjoka M, Kurant M, Butts CT, Markopoulou A (2010) Walking in facebook: A case study of unbiased sampling of osns. In: INFOCOM, 2010 Proceedings IEEE, pp 1–9

Gjoka M, Kurant M, Butts CT, Markopoulou A (2011) Practical recommendations on crawling online social networks. IEEE Journal on Selected Areas in Communications 29(9):1872–1892

Gu Y, McCallum A, Towsley D (2005) Detecting anomalies in network traffic using maximum entropy estimation. In: Proceedings of the 5th ACM SIGCOMM Conference on Internet Measurement, USENIX Association, Berkeley, CA, USA, IMC '05, pp 32–32

Haveliwala TH (2003) Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. IEEE Trans on Knowl and Data Eng 15(4):784 – 796

Hu P, Lau WC (2013) A survey and taxonomy of graph sampling. CoRR abs/1308.5865

Hubler C, peter Kriegel H, Borgwardt K, Ghahramani Z (2008) Metropolis algorithms for representative subgraph sampling. In: In Data Mining, 2008. ICDM08. Eighth IEEE International Conference on, IEEE, pp 283–292

Kitsak M, Gallos LK, Havlin S, Liljerosand F, Muchnik L, Stanley HE, Makse HA (2010) Identification of influential spreaders in complex networks. Nat Phys 6:888–893 doi:10.1038/nphys1746

Krivitsky PN, Kolaczyk ED (2015) On the question of effective sample size in network modeling: An asymptotic inquiry. Statist Sci 30(2):184–198

Kurant M, Markopoulou A, Thiran P (2011) Towards unbiased BFS sampling. IEEE Journal on Selected Areas in Communications 29(9):1799–1809

Lee C, Xu X, Eun DY (2012) Beyond random walk and metropolis-hastings samplers: why you should not backtrack for unbiased graph sampling. In: ACM SIGMETRICS/PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems, SIGMETRICS '12, London, United Kingdom, June 11-15, 2012, pp 319–330

Leskovec J, Faloutsos C (2006) 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

Leskovec J, Krevl A (2014) SNAP Datasets: Stanford large network dataset collection. http://snap.stanford.edu/data

Leskovec J, Kleinberg JM, Faloutsos C (2005) 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

Leskovec J, Kleinberg JM, Faloutsos C (2007) Graph evolution: Densification and shrinking diameters. TKDD 1(1):1–40

Leskovec J, Lang KJ, Dasgupta A, Mahoney MW (2009) Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. Internet Math 6(1):29–123

Leskovec J, Huttenlocher D, Kleinberg J (2010) Signed networks in social media. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, New York, NY, USA, CHI '10, pp 1361–1370

Levin DA, Peres Y, Wilmer EL (2008) Markov Chains and Mixing Times. American Mathematical Society (AMS)

Li RH, Yu JX, Qin L, Mao R, Jin T (2015) On random walk based graph sampling. In: 2015 IEEE 31st International Conference on Data Engineering, pp 927–938

Maiya AS, Berger-Wolf TY (2011) Benefits of bias: towards better characterization of network sampling. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, August 21-24, 2011, pp 105–113

McAuley JJ, Leskovec J (2012) Learning to discover social circles in ego networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ (eds) Advances in Neural

Information Processing Systems 25 (NIPS 2012), Curran Associates, Inc., pp 539–547

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of state calculations by fast computing machines. The Journal of Chemical Physics 21(6):1087–1092, DOI 10.1063/1.1699114

Potamias M, Bonchi F, Castillo C, Gionis A (2009) Fast shortest path distance estimation in large networks. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, ACM, New York, NY, USA, CIKM ’09, pp 867–876

Ribeiro BF, Towsley DF (2010) 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

Richardson M, Agrawal R, Domingos P (2003) Trust Management for the Semantic Web, Springer Berlin Heidelberg, Berlin, Heidelberg, pp 351–368

Ripeanu M, Foster IT, Iamnitchi A (2002) Mapping the gnutella network: Properties of large-scale peer-to-peer systems and implications for system design. CoRR cs.DC/0209028

Robles-Granda P, Moreno S, Neville J (2016) Sampling of attributed networks from hierarchical generative models. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016, pp 1155–1164

Salamanos N, Voudigari E, Yannakoudakis EJ (2016) Identifying influential spreaders by graph sampling. In: Proceedings of the 5th International Workshop on Complex Networks and their Applications, Milan, Italy, November 30 - December 02, 2016

Stutzbach D, Rejaie R, Duffield N, Sen S, Willinger W (2009) On unbiased sampling for unstructured peer-to-peer networks. IEEE/ACM Trans Netw 17(2):377–390

Vattani A, Chakrabarti D, Gurevich M (2011) Preserving personalized pagerank in subgraphs. In: Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011, pp 793–800

Voudigari E, Salamanos N, Papageorgiou T, Yannakoudakis EJ (2016) Rank degree: An efficient algorithm for graph sampling. In: 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016, San Francisco, CA, USA, August 18-21, 2016, pp 120–129

Xu X, Lee C, Eun DY (2014) A general framework of hybrid graph sampling for complex network analysis. In: 2014 IEEE Conference on Computer Communications, INFOCOM 2014, Toronto, Canada, April 27 - May 2, 2014, pp 2795–2803

Yang J, Leskovec J (2012) Defining and evaluating network communities based on ground-truth. In: Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics, ACM, New York, NY, USA, MDS ’12, pp 3:1–3:8

Zafarani R, Liu H (2009) Social computing data repository at ASU. URL http://socialcomputing.asu.edu

Zeng A, Zhang CJ (2013) Ranking spreaders by decomposing complex networks. Physics Letters A 377(14):1031–1035