Estimating degree rank in complex networks

Akrati Saxena, [1]✉

Email akrati.saxena@iitrpr.ac.in

Ralucca Gera, [2]

Email rgera@nps.edu

S. R. S. Iyengar, [1]

Phone +91-1881-242137
Email sudarshan@iitrpr.ac.in

[1] Department of CSE, Indian Institute of Technology Ropar, Rupnagar, India

[2] Department of Applied Mathematics, Naval Postgraduate School, Monterey, CA, 93943 USA

Received: 21 December 2017 / Accepted: 2 June 2018

Abstract

Identifying top-ranked nodes can be performed using different centrality measures, based on their characteristics and influential power. The most basic of all the ranking techniques is based on nodes degree. While finding the degree of a node requires local information, ranking the node based on its degree requires global information, namely the degrees of all the nodes of the network. It is infeasible to collect the global information for some graphs such as (i) the ones emerging from big data, (ii) dynamic networks, and (iii) distributed networks in which the

whole graph is not known. In this work, we propose methods to estimate the degree rank of a node, that are faster than the classical method of computing the centrality value of all nodes and then rank a node. The proposed methods are modeled based on the network characteristics and sampling techniques, thus not requiring the entire network. We show that approximately 1% node samples are adequate to find the rank of a node with high accuracy.

AQ1

AQ2

Keywords

Degree centrality
Ranking nodes
Social network analysis
Sampling techniques

1. Introduction

In complex networks, each node has some unique characteristics that define its importance in the given network. It is of interest to know how important/strong the node is in the given network. The rank of a node would answer that it can be computed using several centrality measures defined in the literature like degree centrality (Shaw 1954), semi-local centrality (Chen et al. 2012), closeness centrality (Sabidussi 1966), betweenness centrality (Freeman 1977), eigenvector centrality (Stephenson and Zelen 1989), Katz centrality (Katz 1953), PageRank (Brin and Page 1998), and so on. In the present work, we focus on degree ranking, i.e., based on the most basic centrality measure. The degree of a node *u* is denoted by d_u, which represents the number of neighbors of the node. The degree rank of a node *u* can be computed as, $R_{\text{act}}(u) = \sum_v X_{uv} + 1$, where

$$X_{uv} = \begin{cases} 1, & \text{if } d_v > d_u \\ 0, & \text{otherwise.} \end{cases}$$

It has been referred as actual degree rank throughout the paper. The node having the highest degree is ranked 1. All nodes having the same degree will have the same rank. The node who is interested in computing its degree rank is hereby referred to as the interested node. The estimated degree rank of the node *u* is denoted by $R_{\text{est}}(u)$.

The default method to compute the degree rank of a node collects the degree of all the nodes, and the degree of the interested node is compared with other nodes to compute its rank. This method requires the degree of all the nodes, which is not feasible to gather for large-scale dynamic networks. In dynamic networks, the rank of the nodes also keeps changing and the latest snapshot of the network is required to estimate the rank. Moreover, in some cases, such as online social networks, collecting the degree of all the nodes using API calls is not feasible due to the fixed number of calls allowed for collecting the data. Therefore, in real-life dynamic networks, there is a need for fast and efficient methods to estimate the degree rank of a node of interest.

We thus propose degree rank estimation methods that do not depend on having the entire network. The proposed methods are either based on sampling techniques or network characteristics, such as real-world networks following power law degree distribution (Barabási and Albert 1999). In power law degree distribution, there are a high number of nodes having lower degrees and a low number of nodes manage to acquire higher degrees. The first introduced method uses this characteristic to estimate the degree rank of a node in $O(1)$ time given the network parameters. The sampling-based methods collect a small sample of the network using different sampling techniques based on the random walk or its variations or uniform sampling. For the uniform sampling, we compute the local rank of the node in the collected samples and then extrapolate it to estimate its

global rank. This method works as the base case to compare the results of other techniques. The last two methods use classical random walk and metropolis-hastings' random walk to collect the samples for the rank estimation.

In this work, we extend our previous work on degree ranking (Saxena et al. 2017) and verify our proposed methodology on 20 real-world social networks as well as on synthetic networks. The detailed comparison of these methods is discussed in the results section. In our previous work, we compare the accuracy of the introduced methods using absolute error on the original six real networks and five synthetic ones; we now augment our analysis by also using the weighted error functions, which depends on the quartile the nodes fall in. This weighted error function decreases the error measure of the lower ranked nodes while emphasizing the error in higher ranked nodes; which is desired, since, in real-life applications, the top-ranked nodes are of interest. We further extend our work by studying the effect that the (1) network size and (2) sampling size have on the error. For this, we vary (1) the size of the network and (2) the size of the sample, and observe the change in both error functions, respectively.

These ranking methods are further extended to rank nodes in random networks. The efficiency of the proposed methods is verified on random networks of 100,000–500,000 nodes. The results are consistent with the scale-free networks, showing that the degree rank of a node can be estimated efficiently using a small (1% nodes) sample size.

As per the best of our knowledge, this is the first attempt to answer this question. The rest of this paper is organized as follows. Next, we discuss related work. In Sect. 3, we discuss methods that are used to estimate the required network parameters. In Sect. 4, all notations used in the paper are explained. Section 5 describes degree rank estimation methods for scale-free networks. Each of its subsection explains one method in depth. Section 6 explains data sets, error functions, and simulation results for all the proposed methods. Section 7 explains ranking methods for random

networks and their validation on Erdos–Renyi networks. The paper is concluded in Sect. 8. This project has a few interesting directions that can be explored further. These are also discussed in the conclusion.

2. Related work

Most real-world networks are highly dynamic and growing very fast with time. In many cases, they are stored in a decentralized way, and thus, it is not feasible to gather the whole network to review its characteristics, such as network size, average degree, clustering coefficient, and so on. Therefore, only a small snapshot of the network can be collected at any given time to study its characteristics. This has motivated researchers to use sampling-based techniques to study network parameters. While sampling, the main focus is that the collected sample should be a good representative of the complete data set.

The sampling techniques can be mainly categorized as node selection-based sampling techniques, edge selection-based sampling techniques (Leskovec and Faloutsos 2006), and graph traversal-based sampling techniques. In node selection or edge selection methods, nodes or edges are sampled uniformly at random from the network, respectively. Haralabopoulos and Anagnostopoulos proposed Enhanced Random Node Sampling method and used it to estimate network parameters like clustering coefficient, average degree, assortativity, and the number of components (Haralabopoulos and Anagnostopoulos 2014). The paper contains the comparison of its efficiency with already existing methods. Kurant et al. (2012) proposed a method called SafetyMargin that uses Induced Edges sampling techniques to estimate the network size. The proposed method outperforms state-of-the-art methods even using ten times smaller sample size.

The node or edge sampling methods are not feasible in real-world networks as the structure of social networks is not known in advance. Therefore, these networks can be sampled using graph traversal techniques

like breadth first search (BFS) (Even 2011), depth first search (DFS) (Even 2011), forest fire sampling (FFS) (Leskovec and Faloutsos 2006), snowball sampling (Goodman 1961), or random walk-based methods like simple random walk (RW) (Lovász 1993), Metropolis-Hastings random walk (MHRW) (Metropolis et al. 1953), re-weighted random walk (RWRW) (Hansen and Hurwitz 1943), respondent driven sampling (RDS) (Salganik and Heckathorn 2004), supervised random walk (Backstrom and Leskovec 2011), Modified TOpology Sampling (MTO) (Zhou et al. 2016), walk-estimate (Nazi et al. 2015), Frontier sampling (m-dimensional random walk) (Ribeiro and Towsley 2010), Rank Degree sampling based on edge selection (Voudigari et al. 2016), preferential random walk (Davis et al. 2016), and so on.

In 2013, Hardiman and Katzir proposed an efficient method to estimate the network size using random walk samples (Hardiman and Katzir 2013). This method is discussed in more detail in Sect. 3.1. They also proposed methods to compute average clustering coefficient and global clustering coefficient of the network using random walk. Other methods that estimate network size exist, such as (Cem and Sarac 2016b; Musco et al. 2016; Lucchese and Varagnolo 2015; Chen et al. 2016; Ye and Wu 2011).

Sampling techniques also have been used to identify degree-related properties like high degree nodes, average degree, or degree distribution of the network. Cooper et al. proposed a biased random walk method to identify high degree nodes in the scale-free networks (Cooper et al. 2012). Marchetti-Spaccamela proposed a method to estimate the degree of a node in directed networks (Marchetti-Spaccamela 1988). Dasgupta et al. proposed a method to estimate the average degree of the network using smooth random walk, that is discussed in Sect. 3.2 (Dasgupta et al. 2014). Eden et al. proposed an algorithm to estimate the average degree using $\tilde{O}(1)$ uniform sampling queries (Eden et al. 2016). There have been proposed some more methods to estimate the average degree (Cem and Sarac 2016b; Lu and Li 2012). Cem and Sarac proposed methods to estimate the size and the average degree of online social networks where

only one random neighbor of the node can be accessed using API calls (Cem and Sarac 2015). They further used ego-centric sampling and showed that the use of neighborhood information is not always beneficial to estimate network parameters like network size and average degree (Cem and Sarac 2016a).

Ribeiro et al. studied the mean square error while computing the degree distribution of the network (Ribeiro and Towsley 2012). They further compute the normalized mean square error for estimating the out-degree and in-deg distribution of the directed networks (Ribeiro et al. 2012). The proposed method uses Directed Unbiased Random Walk (DURW) that takes a random jump with a fixed probability depending on the degree of the node while taking the walk. The results show that the out-degree distribution can be estimated more efficiently and accuracy of the in-degree distribution is very less unless the graph is not symmetric.

In the present work, we use sampling techniques to estimate degree rank of the nodes. Degree centrality plays an important role in diverse domains of science. Fortunato et al. (2006) studied the correlation of in-degree with PageRank of the node. They show that the PageRank is directly proportional to the in-degree, modulo an additive constant, and it can be used to estimate the PageRank of a node. Ghoshal and Barabasi showed the dependence of super stable nodes on their degrees (Ghoshal and Barabási 2011). Degree centrality also has been combined with many other centrality measures to identify the influential nodes (Chen et al. 2012; Hou et al. 2012; Yu and Fan 2015). The current proposals of degree rank estimation methods based on local information will help to design faster influential nodes identification methods.

3. Estimation of network parameters

The ranking methods which we propose require different network parameters that one needs to compute during the pre-processing steps, such as: the network size, average degree, minimum degree, and maximum

degree. We will discuss ways to estimate these parameters from known methods in the literature.

3.1. Network size

To estimate the network size, we use the method proposed by Hardiman and Katzir (we refer it as HK method) (Hardiman and Katzir 2013). The proposed estimator is based on the concept of collision to count the total number of nodes, where samples are collected using the classical random walks. The information of the neighbors of the sampled nodes is used to detect the collision a step before it actually occurs. There is a high probability of collision among neighbors of the sampled nodes if the distance between nodes is small while the sampling. Therefore, a pair of nodes in random walk samples is considered to count the collision if their distance is large during the random walk. For our research, we consider a pair of nodes if their distance is more than 2.5% of sample size, as used by the authors in their simulations.

3.2. Average degree

The average degree of the network is used while estimating the rank of a node using power law degree distribution method. It is estimated using the method proposed by Dasgupta et al. (referred as AD method) (Dasgupta et al. 2014). The samples are collected using smoothed random walk with a distribution $D_{d,c}$, where the probability to sample a node u is directly proportional to $d_u + c$, and c is a smoothing parameter (constant during the entire random walk). The samples generated using this distribution are equivalent to samples generated from the network, where c self-loops are added to each node.

Dasgupta et al. propose two estimators: (1) Smooth, and (2) Guess and Smooth. The Guess and Smooth estimator is used to compute the optimal value of c. This c value is used while collecting the samples by Smooth estimator. These two estimators are combined to propose an estimator that

takes $O(\log U \cdot \log logU)$ samples to estimate average degree with high accuracy, where U is an upper bound on the maximum degree of the network. Thus, just a few samples are needed to estimate the average degree. The accuracy of the proposed estimator depends on the maximum, minimum, and average degree of the network.

3.3. Maximum and minimum degree

In power law degree distribution, the frequency of highest degree node is almost 1, i.e., unique. The maximum degree is estimated as the available maximum degree in the samples, $d'_{\max} = \max\{d_u, \forall u \in S\}$, where S is the set of samples. The minimum degree is observed to be 1 or close to 1 in real-world networks, so we set it to $d'_{\min} = 1$. In BA and ER networks, the minimum degree is estimated as the available minimum degree in the samples. We estimate both of them using the samples collected for the network size estimator.

4. Notation

Let $\mathcal{G}(f)$ represent the set of networks having n nodes and following degree distribution f. All notations used in the paper are explained in Table 1.

Table 1

Notations

Notation	Description
G	A network, $G \in \mathcal{G}(f)$
n	Total number of nodes in the network
m	Total number of edges in the network
n'	Estimated number of nodes in the network
n_j	Total number of nodes having degree j in the network
u, v, w	Nodes in the network

Notation	Description
d_u	Degree of node u
$d_{\max}$	Maximum degree in the network
$d_{\min}$	Minimum degree in the network
d_{avg}	Average degree of the network
S	Set of sampled nodes
s	Sample size, $s = \|S\|$
$d'_{\max}$	Estimated maximum degree/maximum degree in S
$d'_{\min}$	Estimated minimum degree/minimum degree in S
d'_{avg}	Estimated average degree/average degree of S
$R_{\text{act}}(u)$	Actual rank of node u in the network
$R_{\text{est}}(u)$	Estimated rank of node u in the network
$R_{\text{local}}(u)$	Rank of node u in sample S
$R_{\text{G}}(u)$	A random variable that denotes the rank of node u in G
$R_{\text{S}}(u)$	A random variable that denotes the rank of node u in S

5. The main idea: estimating the degree rank

We now introduce and discuss our proposed methods to estimate the degree rank of a node.

5.1. Using power law degree distribution (PL Method)

Most real-world scale-free networks follow a power law degree distribution, i.e., the probability $f(j)$ of a node having degree j is given as $f(j) = cj^{-\gamma}$, where c and γ are constants for the given network (Barabási and Albert 1999). Thus, the networks have only a few hubs. It has also been observed in real-world scale-free networks, that the power law

exponent value varies as $2 < \gamma < 3$. We use the observed information to propose a method estimating these parameters, that can be further used to estimate the degree rank of the node.

Theorem 1 *In a scale-free network G* $(G \in \mathcal{G}(f))$*, the power law exponent of degree distribution can be computed as,* $\gamma \approx 2 + \frac{d_{\min}}{d_{\text{avg}} - d_{\min}}$*, where* $d_{\min}$ *and* d_{avg} *represent minimum and average degree of the network, respectively.*

Proof Let network *G* follows the power law degree distribution $f(j) = cj^{-\gamma}$. First, we derive an equation to estimate the value of *c*. The sum of probabilities of a node having degree *j* ($d_{\min} \leq j \leq d_{\max}$) is equal to 1. The probability function of degree distribution can be written as follows:

$$\sum_{j=d_{\min}}^{d_{\max}} f(j) = 1.$$

We switch to integration[1] to compute *c*:

$$\int_{d_{\min}}^{d_{\max}} f(j)dj = 1,$$

$$\int_{d_{\min}}^{d_{\max}} c \cdot j^{-\gamma} dj = 1.$$

After integration (as $\gamma > 2$), we obtain the value for *c*:

$$c \cdot \frac{(d_{\max})^{1-\gamma} - (d_{\min})^{1-\gamma}}{1-\gamma} = 1$$

$$c = \frac{1-\gamma}{(d_{\max})^{1-\gamma} - (d_{\min})^{1-\gamma}}. \quad (1)$$

To compute γ, the average degree of the network, (d_{avg}), is used. Using $f(j) = c \cdot j^{-\gamma}$, it can be computed as follows:

$$d_{\text{avg}} = \sum_{j=d_{\min}}^{d_{\max}} j \cdot f(j)$$

$$d_{\text{avg}} = \int_{d_{\min}}^{d_{\max}} j \cdot (c \cdot j^{-\gamma})\, dj.$$

After integration, we have that

$$d_{\text{avg}} = c \cdot \frac{d_{\max}^{2-\gamma} - d_{\min}^{2-\gamma}}{2-\gamma}.$$

Putting value of c from Eq. (1) in this equation:

$$d_{\text{avg}} = \frac{1-\gamma}{2-\gamma} \cdot \frac{d_{\max}^{2-\gamma} - d_{\min}^{2-\gamma}}{d_{\max}^{1-\gamma} - d_{\min}^{1-\gamma}}$$

$$d_{\text{avg}} = \frac{\gamma-1}{\gamma-2} \cdot \frac{d_{\max}^{\gamma-2} - d_{\min}^{\gamma-2}}{d_{\max}^{\gamma-1} - d_{\min}^{\gamma-1}} \cdot d_{\max} \cdot d_{\min},$$

where, $d_{\min} << d_{\max}$, and $2 < \gamma < 3$ for scale-free real networks (Barabási and Albert 1999):

$$d_{\text{avg}} \approx \frac{\gamma-1}{\gamma-2} \frac{d_{\max}^{\gamma-2}}{d_{\max}^{\gamma-1}} \cdot d_{\max} \cdot d_{\min}$$

$$d_{\text{avg}} \approx \frac{\gamma-1}{\gamma-2} \cdot d_{\min},$$

that is, $\gamma \approx 2 + \frac{d_{\min}}{d_{\text{avg}} - d_{\min}}$. □

We next present the expected degree rank of a node.

Theorem 2 *In a network G* $(G \in \mathcal{G}(f))$*, the expected degree rank of a node u can be computed as* $E[R_G(u)] \approx n\left(\frac{d_{\max}^{1-\gamma} - (d_u+1)^{1-\gamma}}{d_{\max}^{1-\gamma} - d_{\min}^{1-\gamma}}\right) + 1$*, where* γ *is the power law exponent of the degree distribution of network G.*

Proof In a given network *G*, the actual rank of a node *u* having degree d_u can be computed as follows:

$$R_{\text{act}}(u) = \sum_{j=d_u+1}^{d_{\max}} n_j + 1,$$

where n_j represents total number of nodes having degree *j* in network *G* $(G \in \mathcal{G}(f))$. Let N_j be a random variable that represents total number of nodes having degree *j* in *G*. Then, the expected value of N_j can be computed as $E[N_j] = n \cdot f(j)$. Thus, the expected degree rank of a node *u* can be computed as follows:

$$E[R_{\text{G}}(u)] = E\left[\sum_{j=d_u+1}^{d_{\max}} N_j + 1\right]$$

$$E[R_{\text{G}}(u)] = \sum_{j=d_u+1}^{d_{\max}} E[N_j] + 1$$

$$E[R_{\text{G}}(u)] = \sum_{j=d_u+1}^{d_{\max}} n \cdot f(j) + 1$$

$$E[R_{\text{G}}(u)] \approx n \int_{d_u+1}^{d_{\max}} f(j)dj + 1.$$

Since $f(j) = cj^{-\gamma}$, after the integration of $E[R_G(u)] \approx n \int_{d_u+1}^{d_{\max}} c \cdot j^{-\gamma} dj + 1$, we have

$$E[R_{\mathrm{G}}(u)] \approx nc \frac{d_{\max}^{1-\gamma} - (d_u + 1)^{1-\gamma}}{1-\gamma} + 1.$$

Replacing the value of c from Eq. (1), we obtain

$$E[R_{\mathrm{G}}(u)] \approx n\left(\frac{d_{\max}^{1-\gamma} - (d_u + 1)^{1-\gamma}}{d_{\max}^{1-\gamma} - d_{\min}^{1-\gamma}}\right) + 1,$$

as desired. □

And so, using Theorem 2 and given general estimators about the network, we can estimate the degree rank of nodes.

Corollary 1 *In a network G* $(G \in \mathcal{G}(f))$*, the degree rank of a node u can be estimated as follows*:

$$R_{\mathrm{est}}(u) = n' \left(\frac{(d'_{\max})^{1-\gamma} - (d_u + 1)^{1-\gamma}}{(d'_{\max})^{1-\gamma} - (d'_{\min})^{1-\gamma}} \right) + 1,$$

where $\gamma = 2 + \frac{d'_{\min}}{d'_{\mathrm{avg}} - d'_{\min}}$, *and* $n', d'_{\min}, d'_{\max}$, *and* d'_{avg} *denote the estimated value of network size, minimum degree, maximum degree, and average degree of the network, respectively*.

The proposed method estimates the rank with high accuracy for BA networks, but results are not good for real-world networks, as their degree distribution does not follow the perfect power law.

Next, we discuss sampling-based approaches that perform better on real-world networks.

5.2. Using uniform sampling (US Method)

In this section, the uniform sampling technique is used to collect a small sample of actual data set. In uniform sampling, the probability of sampling a node is equal to 1 / *n*, where *n* is the total number of nodes. Uniform samples preserve the characteristics of actual data set. Therefore, the collected samples follow similar degree distribution as observed in real-world networks. Here, we assume that the network *G* is generated using degree distribution f_1 and $G \in \mathcal{G}(f_1)$. First, we compute the local rank of the node in the collected samples. Then, Theorem 3 can be used to estimate the global rank of the node using its local rank.

Theorem 3 *In a network G* $(G \in \mathcal{G}(f_1))$*, if sample S is collected uniformly, the expected local rank of node u can be computed as,* $E[R_S(u)] \approx \frac{s}{n} E[R_G(u)]$*, where* $R_G(u)$ *and* $R_S(u)$ *are random variables that denote the rank of node u in network G and sample S, respectively.*

Proof We are interested in computing the rank of a node *u* having degree d_u. Let us take a random variable N_j, that denotes the number of nodes having degree *j* in the network. The expected value of N_j can be computed as $E[N_j] = n \cdot f_1(j)$.

The expected rank of node *u* in network *G* can be computed as follows:

$$E[R_{\mathrm{G}}(u)] = E\left[\sum_{j=d_u+1}^{d_{\max}} (N_j) + 1\right]$$

$$E[R_{\mathrm{G}}(u)] = \sum_{j=d_u+1}^{d_{\max}} (n \cdot f_1(j)) + 1. \qquad (2)$$

Now, we have a uniform sample *S* of size *s*. In network *G*, the probability *p* to sample a node *v* uniformly at random having degree greater than d_u

$(d_v > d_u)$ can be defined as follows:

$$p = \frac{\sum_{j=d_u+1}^{d_{\max}} (n \cdot f_1(j))}{\sum_{j=1}^{d_{\max}} (n \cdot f_1(j))}.$$

Using Eq. (2):

$$p = \frac{E[R_G(u)] - 1}{\sum_{j=1}^{d_{\max}} (n \cdot f_1(j))}$$

$$p = \frac{E[R_G(u)] - 1}{n \sum_{j=1}^{d_{\max}} (f_1(j))}.$$

Using the property of probability distribution $\sum_{j=1}^{d_{\max}} f_1(j) = 1$.

$$E[R_{\mathrm{G}}(u)] = p \cdot n + 1. \tag{3}$$

The expected value of local rank of node u in sample S can be computed as follows:

$$E[R_{\mathrm{S}}(u)] = \sum_{j=0}^{s} \binom{s}{j} p^j (1-p)^{(s-j)} j + 1$$

$$E[R_{\mathrm{S}}(u)] = s \cdot p + 1. \tag{4}$$

Using Eqs. (3) and (4):

$$E[R_{\mathrm{S}}(u)] = \frac{s}{n} E[R_{\mathrm{G}}(u)] + \frac{n-s}{n}, \tag{5}$$

where $0 \leq (n-s)/n < 1$, if $s \leq n$. Therefore

$$E[R_{\mathrm{S}}(u)] \approx \frac{s}{n} E[R_{\mathrm{G}}(u)].$$

□

In a network G, $R_{\text{act}}(u) \approx E[R_{\text{G}}(u)]$ and $R_{\text{local}}(u) \approx E[R_{\text{S}}(u)]$. $R_{local}(u)$ denotes the rank of node *u* in sample *S*, and $R_{\text{local}}(u) = \sum_{j=d_u+1}^{d'_{\max}} (n'_j) + 1$, where n'_j is the number of nodes having degree *j* in sample *S*. Using Theorem 3, the actual rank of node *u* can be computed as follows:

$$R_{\text{act}}(u) \approx \frac{n}{s} R_{\text{local}}(u). \tag{6}$$

Corollary 2 *In a network G, using uniform samples, degree rank of a node u can be estimated as* $R_{est}(u) = \frac{n'}{s} R_{local}(u)$, *where* n' *is the estimated network size*.

5.3. Using Metropolis-Hastings random walk (MH Method)

In online networks, uniform sampling is not feasible due to reasons such as the network size and nodes' address are not known. These networks can be sampled using graph sampling techniques like breadth first traversal, random walk, etc. These sampling methods are biased towards higher degree nodes and fail to generate uniform samples. We use metropolis-hastings random walk that generates sample equivalent to uniform samples, that can be used for rank estimation.

Metropolis-Hastings Random Walk (MH): This technique was first proposed by Metropolis et al. (1953) in 1953. It modifies the probability function of random walk, so that the distribution of the collected sample set is the same as the actual distribution of the network. In MH, the crawler will move to the next node with probability *p* and will stay at the same node with probability $(1 - p)$. Therefore, the probability distribution is defined as follows:

$$P_{u\to v} = \begin{cases} \frac{1}{d_u} \cdot \min(1, \frac{d_u}{d_v}), \text{if } v \text{ is the neighbor of } u, \\ 1 - \sum_{w \neq u} P_{u\to w}, \textit{if } v = u, \\ 0, \text{otherwise.} \end{cases}$$

This probability distribution collects more samples of lower degree nodes and fewer samples of higher degree nodes, so the collected samples are not biased towards higher degrees. Gjoka et al. (2010) studied the samples collected using metropolis-hastings random walk and showed that MH can be used to study the network parameters. Once the samples are collected, use Corollary 2 to estimate the degree rank.

5.4. Using random walk (RW Method)

The classical random walk is a well-known method to collect the samples in large dynamic networks (Lovász 1993). In Random walk, a crawler starts from a randomly chosen node. It moves to the next node that is chosen uniformly at random among the neighbors of the current node. The probability to move to node *v* from node *u* is defined as follows:

$$P_{u\to v} = \begin{cases} \frac{1}{d_u}, & \text{if } v \text{ is a neighbor of } u, \\ 0, & \text{otherwise.} \end{cases}$$

In a random walk, the probability of a node being sampled converges to a stationary distribution, $p(u) = d_u/2m$. Therefore, the collected samples are biased towards high degree nodes. We propose Theorem 4 to estimate the degree rank using random walk samples.

First, notice that in a random walk, the probability of a node being sampled is directly proportional to its degree. These samples can be converted to uniform samples using a new probability distribution, where the probability of picking a node is inversely proportional to its degree $p(u) \propto 1/d_u$, known as re-weighted random walk sampling

technique (Hansen and Hurwitz 1943).

Theorem 4 In a network G $(G \in \mathcal{G}(f_1))$, using random walk sample S, the degree rank of a node u can be computed as $R_{\text{act}}(u) \approx \frac{n}{k} \cdot R_{\text{local}}(u) - \frac{n-k}{k}$, where $R_{\text{local}}(u) = \sum_{j=d_u+1}^{d'_{\max}} (q(j) \cdot k) + 1$, and k is a constant, $q(j)$ is the re-sampling probability function $q(j) = \frac{n'_j/j}{\sum_{i=d'_{\min}}^{d'_{max}} n'_i/i}$, and n'_j represents total number of nodes having degree j in sample S.

Proof The probability q to resample a j degree node can be computed as, $q(j) = \frac{n'_j/j}{\sum_{i=d'_{\min}}^{d'_{\max}} n'_i/i}$, where n'_j represents total number of nodes having degree j in sample S.

To estimate the degree rank of a node, collect $q(j) \cdot k$ samples of each degree j from S, where k is a constant. Therefore, the expected size of new sample set S' can be computed as $\sum_{j=d'_{\min}}^{d'_{\max}} (q(j) \cdot k) = k$. The expected rank of node u in S' can be computed as follows: $\sum_{j=d_u+1}^{d'_{\max}} (q(j) \cdot k) + 1$.

As the new sample set S' follows uniform distribution, the rank of node u can be computed using Eq. (5).

In the experiments, the value of k is chosen as $k = 1/\min(q)$, so that the regenerated samples also contain higher degree nodes and their rank is estimated with high accuracy.

Once we put the value of local rank $(R_{\text{local}}(u))$ in actual rank $(R_{\text{act}}(u))$ equation in Theorem 4, the degree rank can be estimated using corollary 3.

Corollary 3 In a network G, using random walk samples, the degree rank of a node u can be estimated as follows:

$$R_{est}(u) = n' \cdot \frac{\sum_{j=d_u+1}^{d'_{\max}} \left(\frac{n'_j}{j}\right)}{\sum_{i=d'_{\min}}^{d'_{\max}} \left(\frac{n'_i}{i}\right)} + 1,$$

where n'_j represents total number of nodes having degree j in sample S.

6. Simulation results

In this section, we will discuss the data sets, error functions, and simulation results.

6.1. Data sets

All proposed methods are simulated on both synthetic as well as on real-world scale-free networks. Synthetic networks are generated using Barabási–Albert (BA) model $G(n, k)$, where each new coming node makes k preferential connections with already existing nodes (Barabási and Albert 1999). The probability $p(u)$ to make a connection with an existing node u is directly proportional to the degree of node u, as $p(u) = d_u / \sum_v d_v$. Therefore, the nodes having higher degrees acquire more links over time and it gives birth to power law degree distribution. All data sets are explained in Table 2.

Table 2

Data sets

Network	Type	#Nodes	#Edges	References
BA1	Synthetic network	100000	999900	Barabási and Albert (1999)
BA2	Synthetic network	200000	1999900	Barabási and Albert (1999)
BA3	Synthetic network	300000	2999900	Barabási and Albert (1999)

Network	Type	#Nodes	#Edges	References
BA4	Synthetic network	400000	3999900	Barabási and Albert (1999)
BA5	Synthetic network	500000	4999900	Barabási and Albert (1999)
Actor	Collaboration network	374511	15014839	Barabási and Albert ((1999)
DBLP	Co-authorship network	317080	1049866	Yang and Leskovec, (2015)
Digg	Social network	261489	1536577	Hogg and Lerman (2012)
Eu-Email	Communication network	224832	339925	Leskovec et al., (2007)
Gowalla	Social network	196591	950327	Cho et al., (2011)
Gplus	Social network	107614	12238285	McAuley and Leskovec, (2012)
Catster	Social network	148826	5447464	Rossi and Ahmed, (2015)
YouTube	Social network	1134885	2987468	Zafarani and Liu, (2009a)

6.2. Error functions

The accuracy of all methods is evaluated using absolute and weighted error functions. These are discussed below:

1. *Absolute error:* Absolute error for a node u is computed as follows:

$$Err_{\text{abs}}(u) = |R_{\text{est}}(u) - R_{\text{act}}(u)|.$$

The percentage average absolute error can be computed as follows:

$$Err_{\text{paae}} = \frac{\text{average \; absolute \; error}}{\text{network \; size}} \times 100\%.$$

2. *Weighted error:* In real-life applications, the significance of the error depends on two important parameters: 1. rank of the node and 2. network size. The same rank difference has more impact for the higher ranked nodes than the lower ranked nodes. Similarly, the same error in the rank will be perceived higher in smaller networks than the larger networks. We consider both of these parameters and propose a weighted error function. It is defined as follows:

$$Err_{wtd}(u) = \frac{Err_{\text{abs}}(u)}{n} \cdot \frac{(n - R_{\text{act}}(u) + 1)}{n} \times 100\%,$$

where $\frac{(n-R_{\text{act}}(u)+1)}{n} \times 100$ denotes percentile of node u. The weighted error increases linearly with the percentile and decreases with the network size, if the absolute error is constant.

6.3. Results and discussion

In this section, we discuss simulation results of all proposed methods. We estimate the network parameters (size, average, maximum, and minimum degree) using the methods discussed in Sect. 3. Each experiment is repeated ten times and the average value is used for the next steps of the experiments.

To measure the performance of the proposed methods, we calculate the average error for each degree and average it over all degrees to compute

the overall error in the rank estimation. Each value (absolute and weighted errors) is computed by taking the average of 20 iterations of the experiment. Table 3 shows the results for US, MH, and RW methods when 1% nodes are sampled, validated on 20 real-world social networks. The detailed results are shown in Appendix 9.

Table 3

Average estimation error on 20 real-world social networks

Method	Err_{paae}	Err_{wtd}
PL	1.51	1.14
US	0.13	0.12
MH	0.50	0.41
RW	0.16	0.13

Results show that the US method performs best on real-world networks. As uniform sampling is not possible in real-world networks, RW method is the most feasible and accurate method. In random walk samples, the probability of sampling a node is directly proportional to its degree once the samples are stabled. Our results are shown with 1% samples, for which we have not removed samples before mixing time, making the proposed random walk method even faster. The results using MH method show larger error than both US and RW methods, because MH random walk samples are not perfectly equivalent to uniform samples for small sample size. The performance of PL method is poor on real-world networks as they do not follow the perfect power law.

For the RW method, we could reuse the samples to estimate the network size and degree rank, making it faster than the other sampling methods. RW method is also faster than PL method as it does not have to estimate the average degree using smoothed random walk. However, if the network parameters are already known and thus do not need to be estimated, the PL

method can be used to estimate the degree rank in $O(1)$ time.

The detailed results are discussed for the networks given in Table 2, and Table 4 shows their estimated parameters. The average error is shown using actual parameters (A.P.) as well as estimated parameters (E.P.) to observe the error caused by the estimation of network parameters.

Table 4

Estimated network parameters

Network	Number of nodes		Average degree	
	Actual	Estimated	Actual	Estimated
BA1	100000	106773	20.00	19.68
BA2	200000	199303	20.00	19.75
BA3	300000	292649	20.00	19.89
BA4	400000	406837	20.00	20.06
BA5	500000	500688	20.00	20.30
Actor	374511	417560	80.18	92.53
DBLP	317080	315587	6.62	7.20
Digg	261489	260435	11.75	17.00
Eu-Email	224832	223151	3.02	2.96
Gowalla	196591	199568	9.67	10.92
Gplus	107614	102456	227.45	278.00
Catster	148826	153075	33.02	33.82
YouTube	1134885	1136445	5.26	10.15

Figures 1 and 2 show the percentage average absolute error and average-weighted error for BA and real-world networks, respectively.

Fig. 1

Average error for BA networks

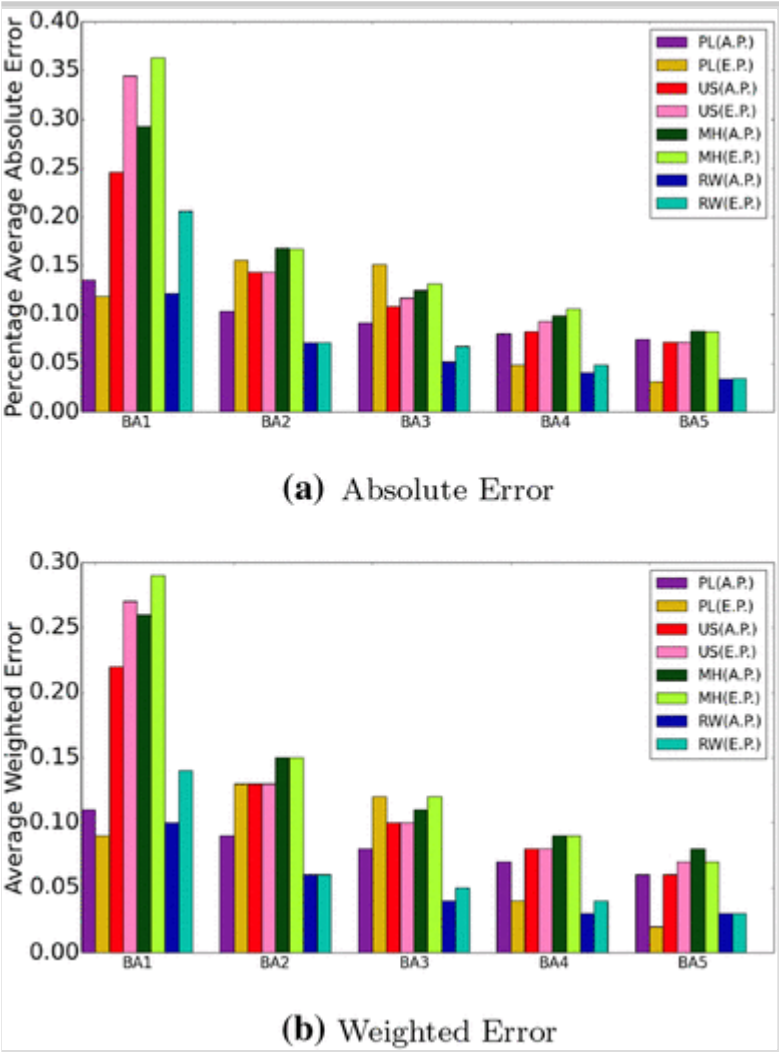
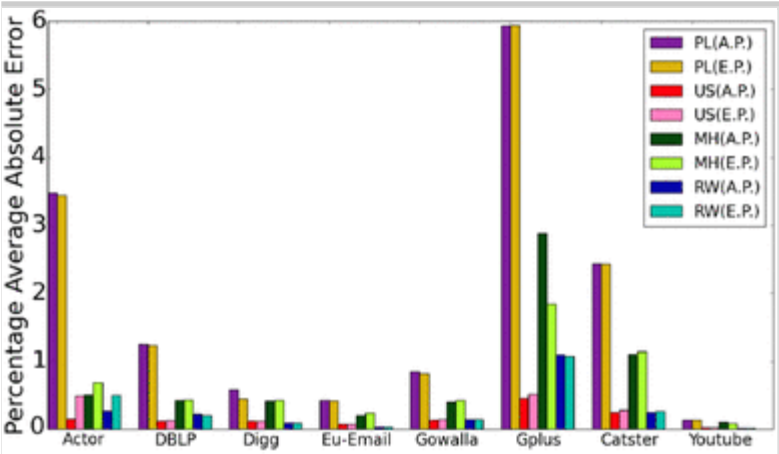
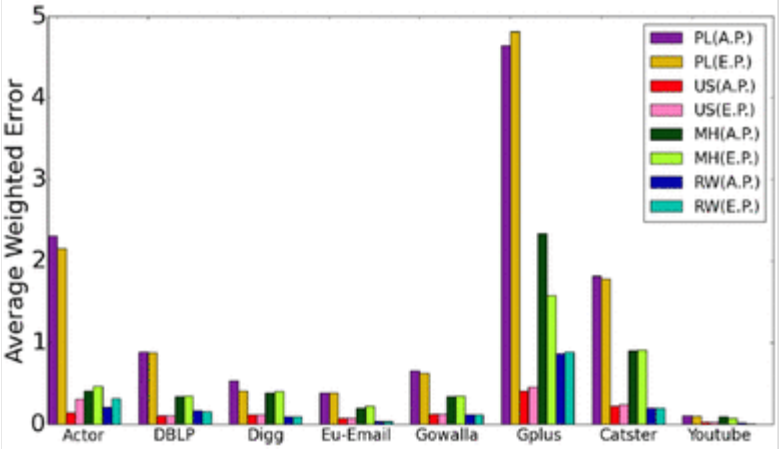


Fig. 2

Average error for real-world networks



(a) Absolute Error



(b) Weighted Error

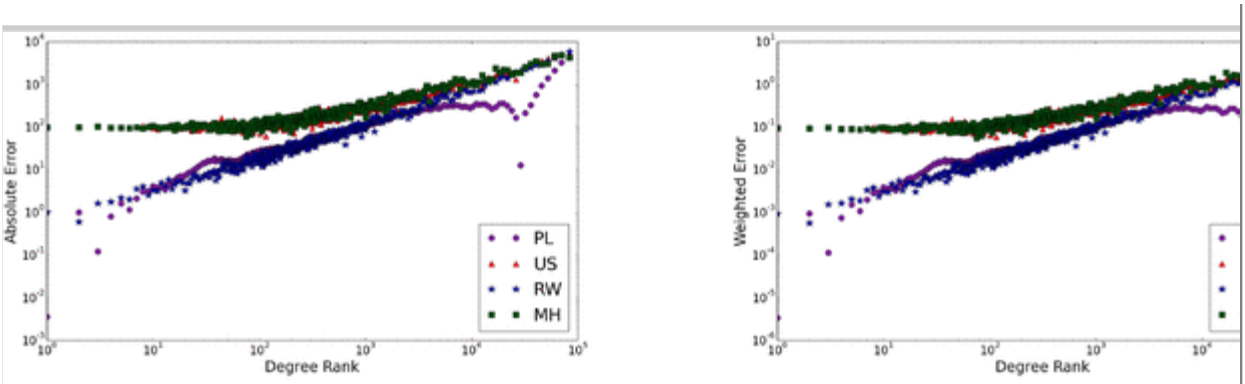
The first two bars show the average error of PL method using actual and estimated parameters, respectively. They are followed by US, MH, and RW methods. The results show that, in BA networks, RW method outperforms all other methods, while, in the real-world networks, they perform well but not best. This is because the accuracy of RW method depends on the density and structure of the network, and so our estimation has higher accuracy in sparse networks than the dense networks. The same pattern is observed in MH method, as it also collects samples using a walk over the network. In Fig. 2, it is observed that PL method gives a very high error for some networks like Actor, Gplus, and Catster, because, in these networks, the estimated degree distribution has the high difference from

the actual degree distribution. The actual and estimated number of nodes (using PL method) versus degrees for Actor network are shown in Fig. 5.

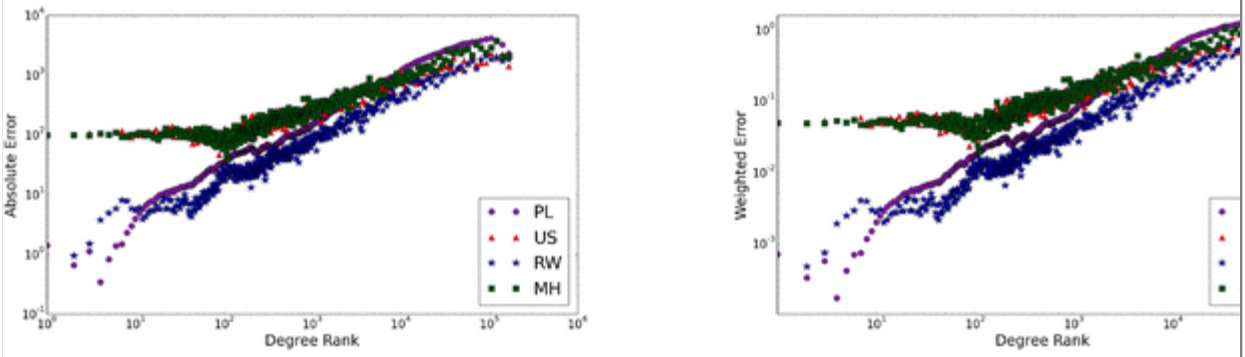
Figures 3 and 4 present the absolute and the weighted error versus degree ranking for BA and real-world networks, respectively. Notice that the absolute error increases with the rank for all the methods, and while the weighted error also increases with the rank, it decreases for the very low ranked nodes. This is because the weighted error depends on the node's rank, and as the rank value increases less importance is given to the error.

Fig. 3

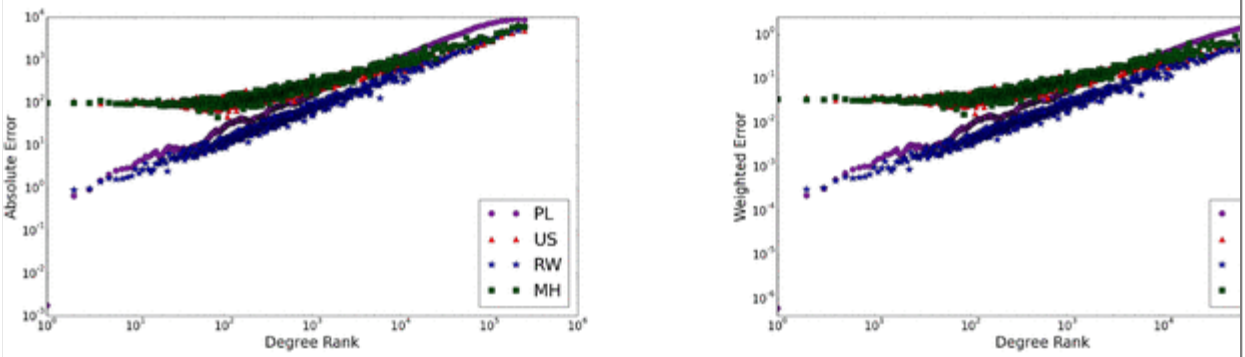
Average error versus degree rank for BA networks



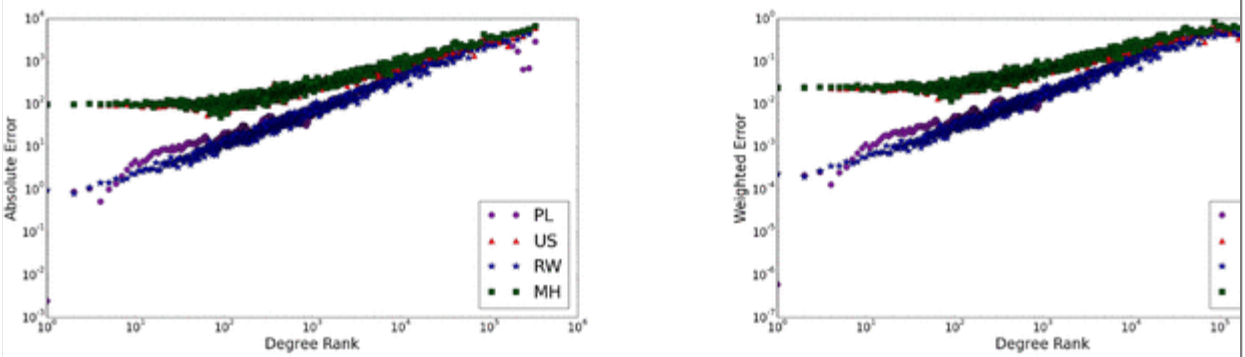
(a) BA1 Network



(b) BA2 Network



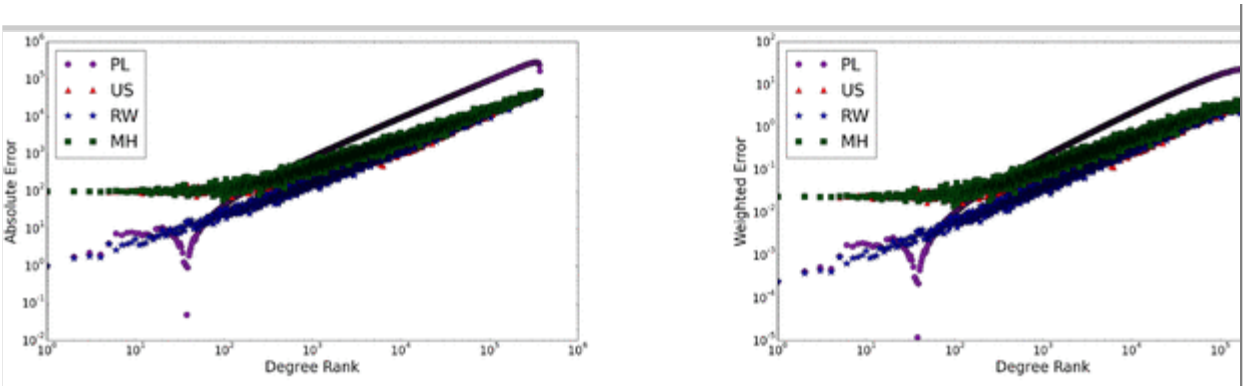
(c) BA3 Network



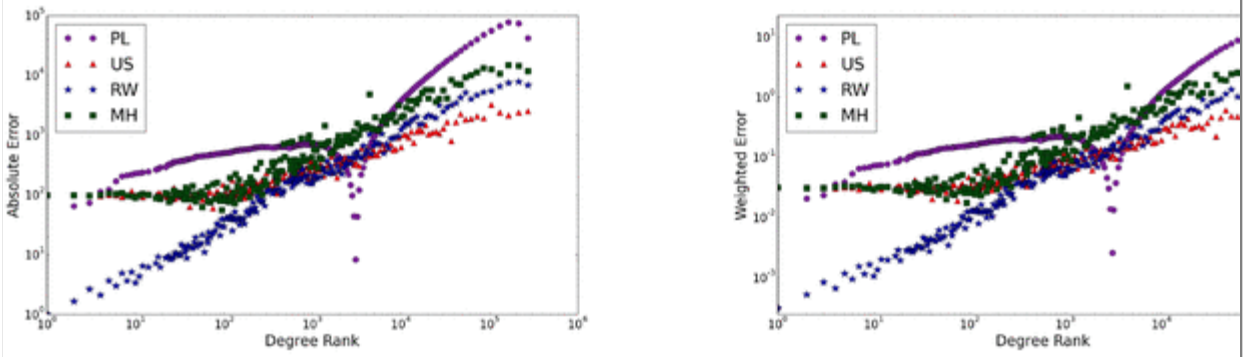
(d) BA4 Network

Fig. 4

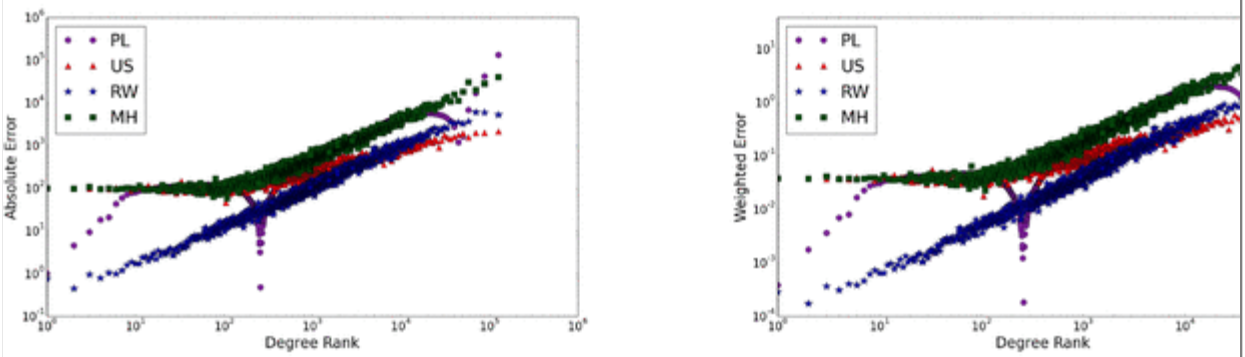
Average error versus degree rank for real-world networks



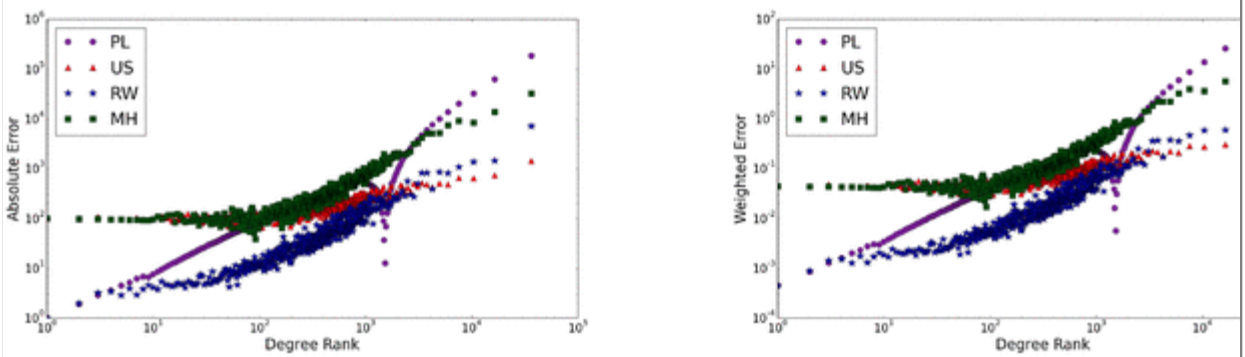
(a) Actor Network



(b) DBLP Network



(c) Digg Network



(d) Eu-Email Network

The results of Figs. 3 and 4 also show that for high ranked nodes, RW method outperforms all other methods. The US method gives more error for high ranked nodes, as the linear extrapolation technique starts assigning the rank from n / s, for sample size s. If α percent nodes are sampled, it will start ranking nodes from $100/\alpha$, that will induce more error for high ranked nodes. We observe that PL method works well for BA networks, but it gives a huge error for real-world networks. This error is driven by the fact that (1) real-world networks do not follow the perfect power law used in estimation, and (2) the rank of a degree d node is computed by integrating the probability distribution function from $d + 1$ to $d_{\max}$, and some degrees might not be present in real networks.

Moreover, we can observe in Figs. 3 and 4 that the rank estimation error in the PL method first increases and then decreases to near zero before it picks up again and it keeps increasing. This happens due to the error in the estimated slope of the degree distribution. The error drops close to 0 when the number of estimated nodes is close to the number of actual nodes having degree greater than a degree d. Figure 5 presents the actual and estimated number of nodes versus degrees, for two sampled networks, the Actor and DBLP network.

Fig. 5

Actual and estimated number of nodes versus degree

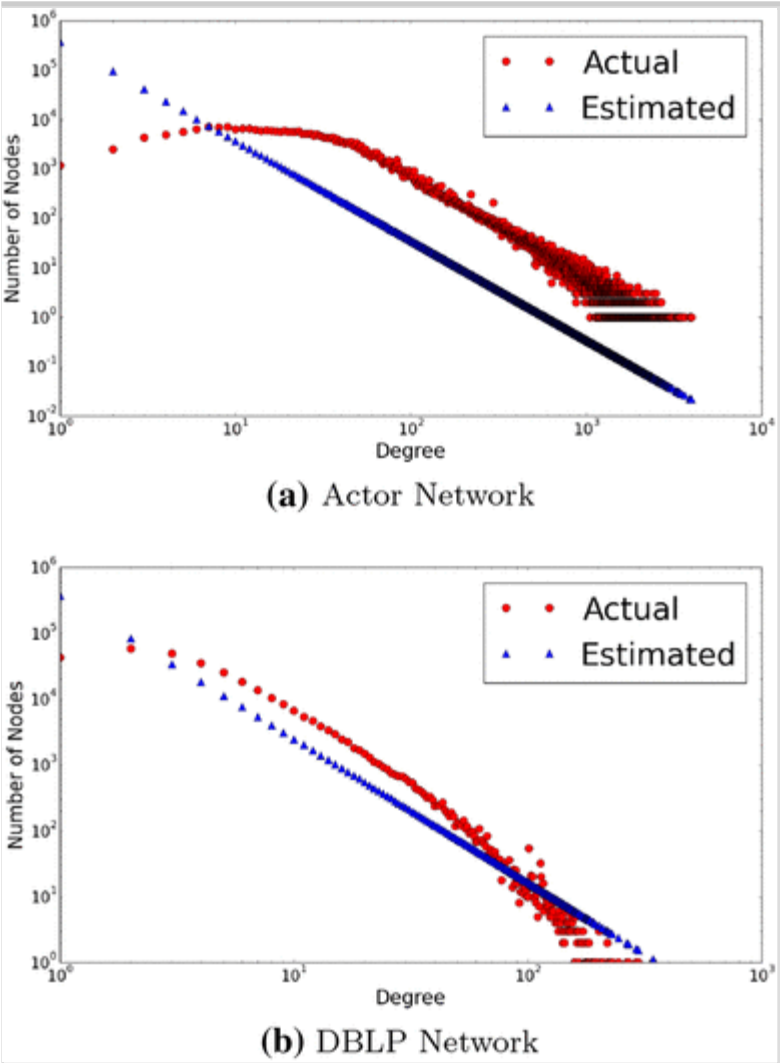


Fig. 6

Average estimation error versus network size for BA networks

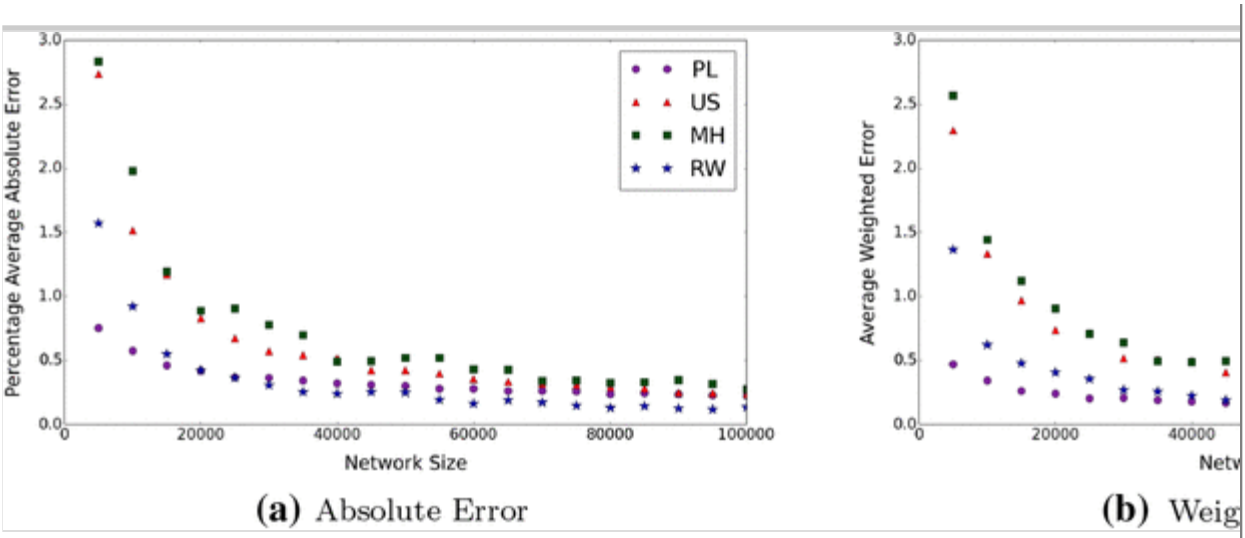
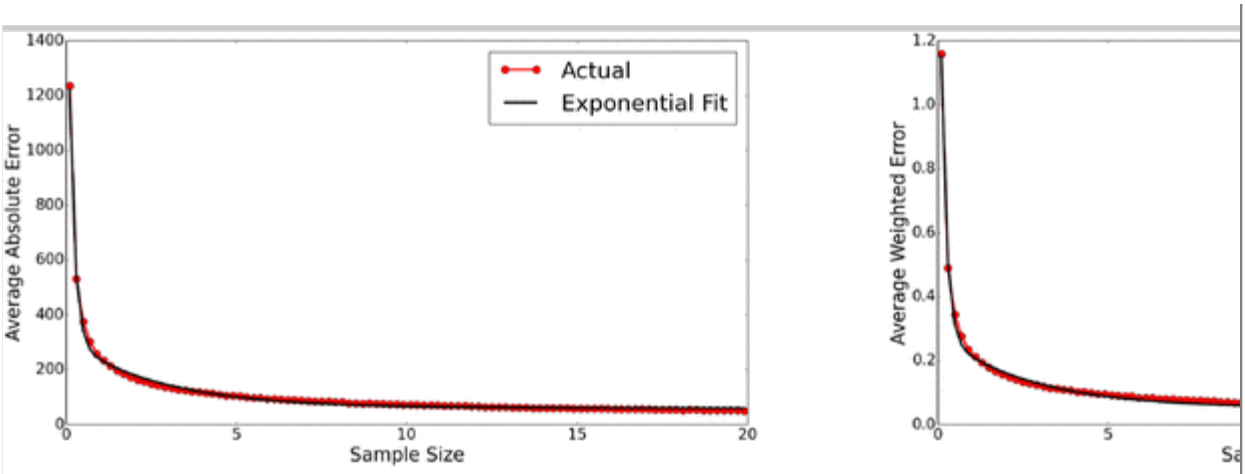


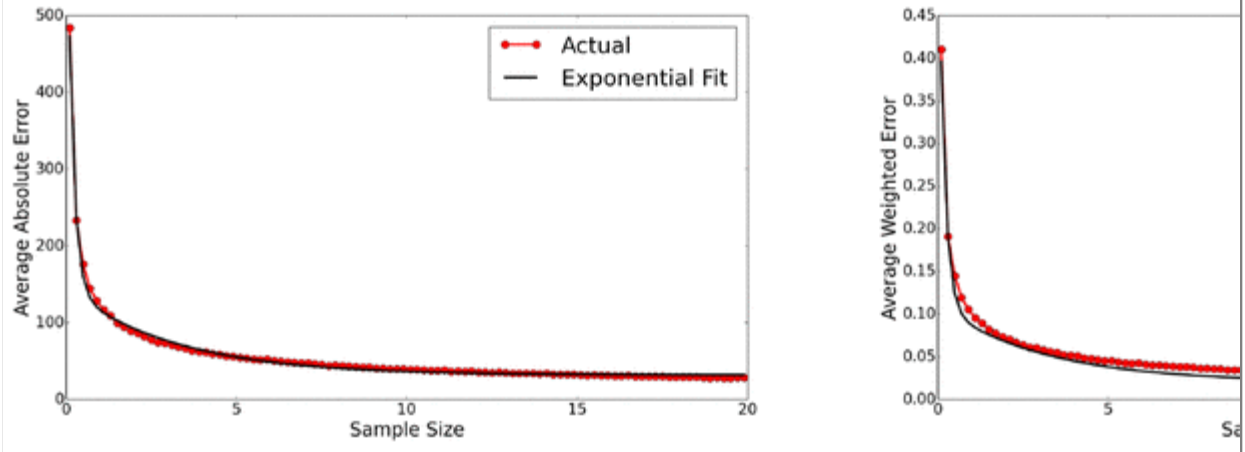
Fig. 7

Error versus sample size for BA1 network



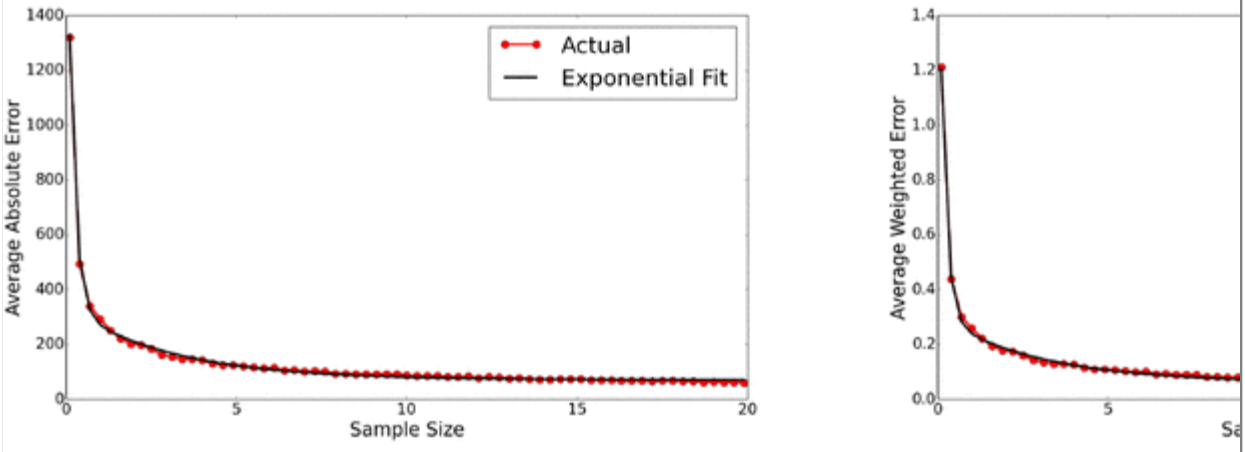
(a) Absolute Error (b) Wei

(a) US Method



(c) Absolute Error (d) Wei

(b) RW Method



(e) Absolute Error (f) Wei

(c) MH Method

Next, we study the behavior of the estimated error as the networks' size change. The synthetic networks are ideal for such an analysis, and we use them by increasing the BA network's size while maintaining the density. Figure 6 presents the percentage average absolute error and average-weighted error, respectively, against the network size. Plots show that the error decreases as the network size increases, which is desired, since no estimation is needed for small networks. Notice that RW method outperforms other methods as the network size increases, building towards our conclusion of recommending RW over the other sampling methods.

Figure 7 shows the error for our sampling methods (US, RW, and MH, respectively) versus sample size for one of the BA examples (BA1 network). The *x*-axis shows the sample size as the percentage of the network, and *y*-axis shows the estimated error. The results plotted in red show the actual error, and the black ones show the exponentially decaying fitted curve. The best fit curve is plotted using the scaled Levenberg–Marquardt algorithm (Moré 1978) with 1000 iterations and 0.0001 tolerance. Observe that both absolute and weighted error follow exponential decay with the sample size, as desired.

7. Random networks

We now discuss the degree rank estimation methods for random networks. In 1959, Erdős and Rényi proposed a model to generate random networks, called Erdős and Rényi (ER) model (Erdős and Rényi 1960). In the ER model, we create a network on n nodes by adding an edge between each pair of nodes with some fixed probability p. The degree distribution g of random networks follows Poisson law. Thus, the probability of a node having degree j can be approximated as $g(j) \to \frac{(d_{\text{avg}})^j e^{-d_{\text{avg}}}}{j!}$ as $n \to \infty$, where d_{avg} is average degree of the network.

In this section, we compare the additional degree ranking method based on Poisson law degree distribution to rest of the methods: US, MH, and RW.

All the sampling methods can be applied to random networks as they have no dependence on the type of the degree distribution function. We estimate the network size and average degree as previously done in this paper.

7.1. Poisson degree distribution method (PD Method)

This method uses Poisson degree distribution of random networks to estimate degree rank of a node.

Lemma 1 *If G is a random network* $(G \in \mathcal{G}(g))$*, then the expected degree rank of a node u is as follows*:

$$E[R_{\mathrm{G}}(u)] = n \cdot e^{-d_{\mathrm{avg}}} \sum_{j=d_u+1}^{d_{vmax}} \frac{(d_{\mathrm{avg}})^j}{j!} + 1.$$

Proof In a given network *G* that follows Poisson degree distribution, the actual rank of a node *u* is as follows:

$$R_{\mathrm{act}}(u) = \sum_{j=d_u+1}^{d_{\max}} n_j + 1,$$

where n_j represents total number of nodes having degree *j* in network *G*.

Let N_j be a random variable that represents the total number of nodes having degree *j* in the network *G* $(G \in \mathcal{G}(g))$. The expected value of N_j is $E[N_j] = n \cdot g(j)$. Then, the expected degree rank of a node *u* can be computed as follows:

$$E[R_{\mathrm{G}}(u)] = E\left[\sum_{j=d_u+1}^{d_{\max}} N_j + 1\right]$$

$$E[R_{\mathrm{G}}(u)] = \sum_{j=d_u+1}^{d_{\max}} E[N_j] + 1$$

$$E[R_{\mathrm{G}}(u)] = \sum_{j=d_u+1}^{d_{\max}} n \cdot g(j) + 1.$$

As we know $g(j) \to \frac{(d_{\mathrm{avg}})^j e^{-d_{\mathrm{avg}}}}{j!}$ as $n \to \infty$, so to compute the expected rank we use $g(j) = \frac{(d_{\mathrm{avg}})^j e^{-d_{\mathrm{avg}}}}{j!}$:

$$E[R_{\mathrm{G}}(u)] = n \cdot \sum_{j=d_u+1}^{d_{\max}} \frac{(d_{\mathrm{avg}})^j e^{-d_{avg}}}{j!} + 1$$

$$E[R_{\mathrm{G}}(u)] = n \cdot e^{-d_{\mathrm{avg}}} \sum_{j=d_u+1}^{d_{\max}} \frac{(d_{\mathrm{avg}})^j}{j!} + 1,$$

as desired. □

Corollary 4 *If G is a random network* $(G \in \mathcal{G}(g))$*, then the degree rank of a node u can be estimated as follows*:

$$R_{\mathrm{est}}(u) = n' \cdot e^{-d'_{\mathrm{avg}}} \sum_{j=d_u+1}^{d'_{\max}} \frac{(d'_{\mathrm{avg}})^j}{j!} + 1,$$

where n', $d'_{\max}$, and d'_{avg} are estimated network size, and maximum and average degree of the network, respectively.

7.2. Results for PD, US, MH, and RW on random networks

Table 5 shows the summary of the generated ER networks and their estimated parameters.

Table 5

Estimated parameters for Erdős and Rényi Networks

Network	Number of nodes		Average degree	
	Actual	Estimated	Actual	Estimated
ER1	100000	99874	11.50	11.24
ER2	200000	202731	12.34	12.08
ER3	300000	300503	12.71	12.49
ER4	400000	398168	12.99	12.81
ER5	500000	505675	13.19	13.07

Figure 8 shows the percentage average absolute and average-weighted error using actual and estimated parameters. For the PD method, the estimated parameters produce a high error, as the rank depends on the average degree: the number of nodes of degree *j* is directly proportional to $(d_{\text{avg}})^j$.

Fig. 8

Average error for ER networks

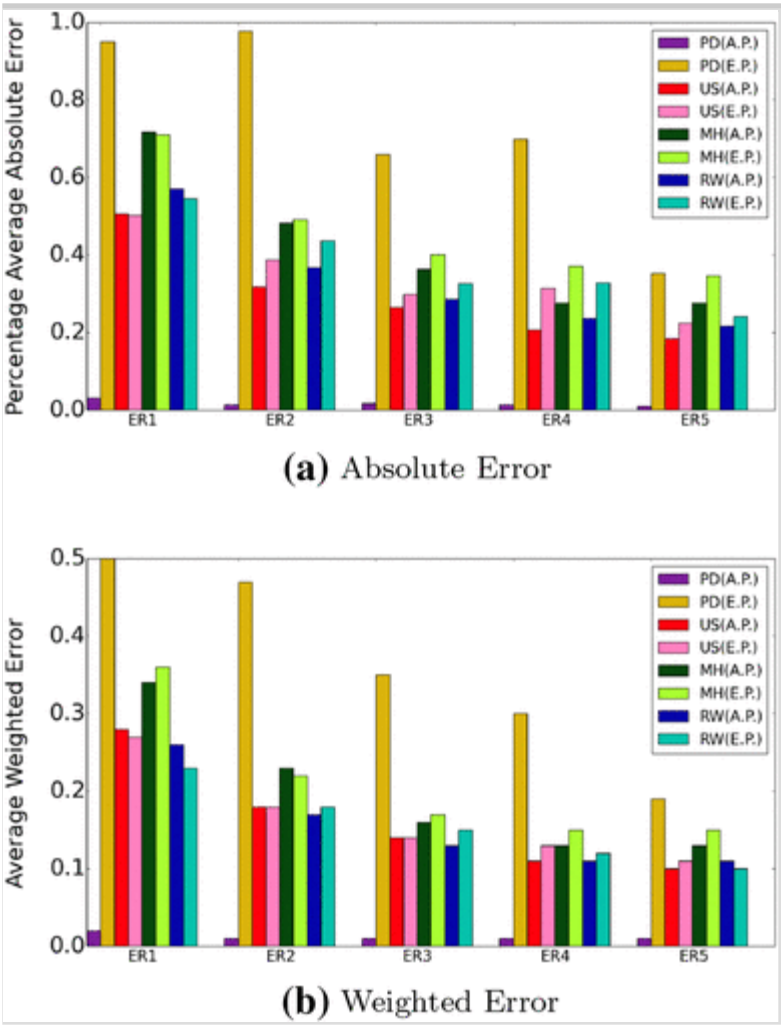
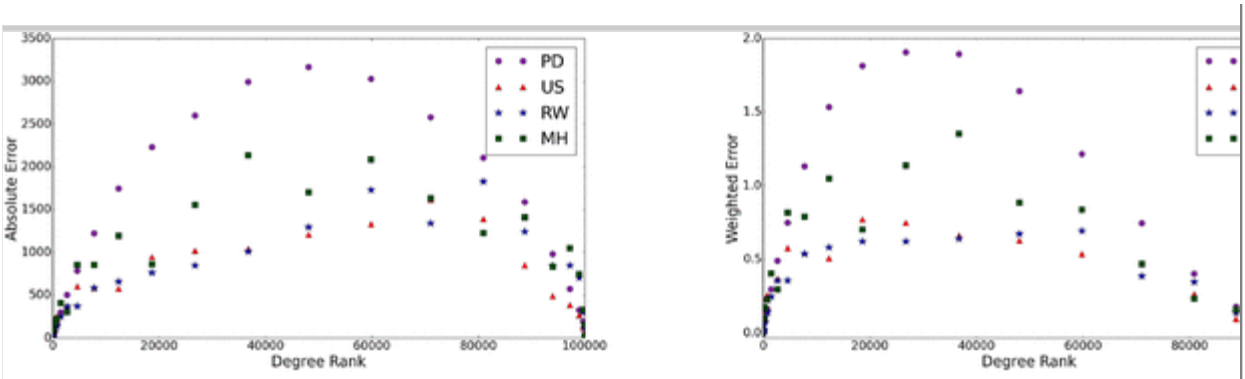
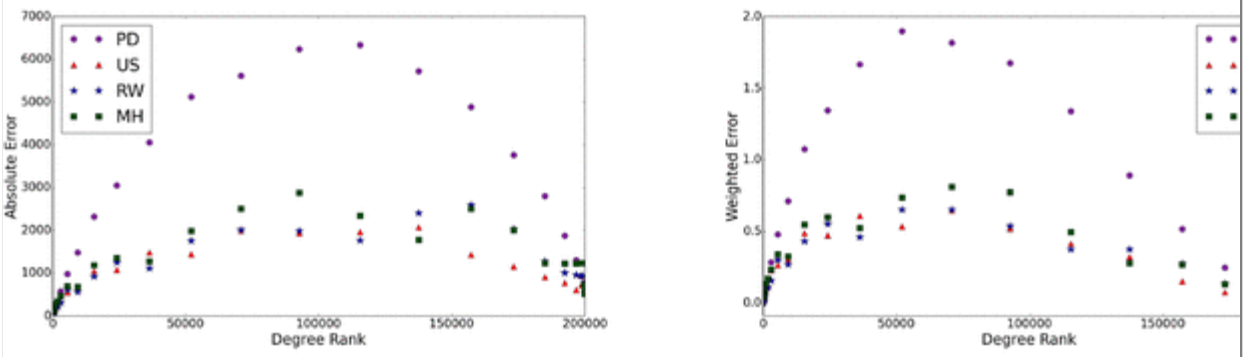


Fig. 9

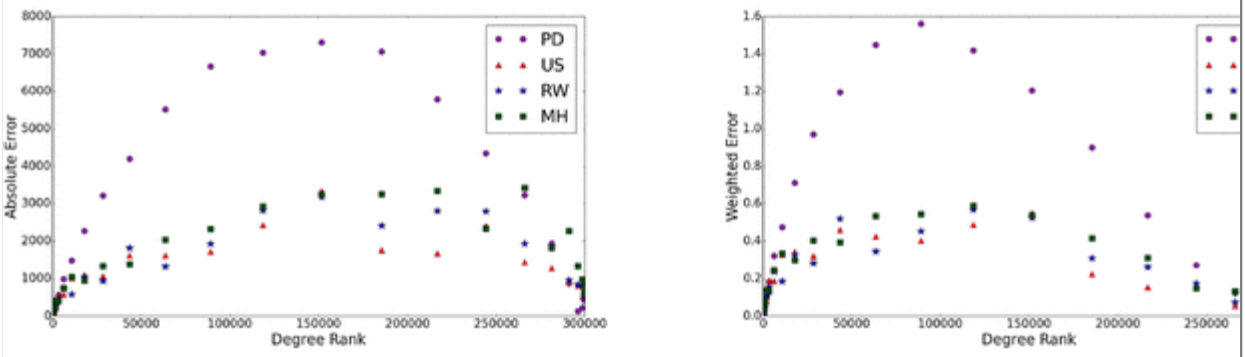
Error versus rank for random networks



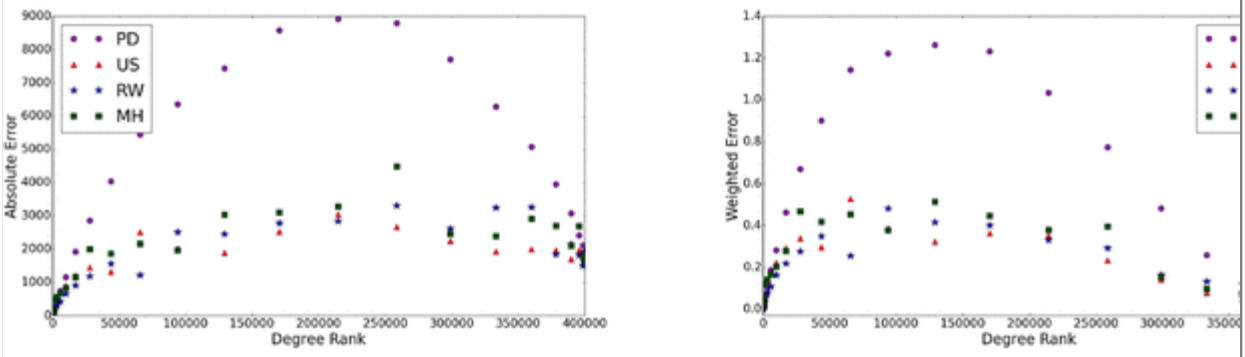
(a) ER1 Network



(b) ER2 Network



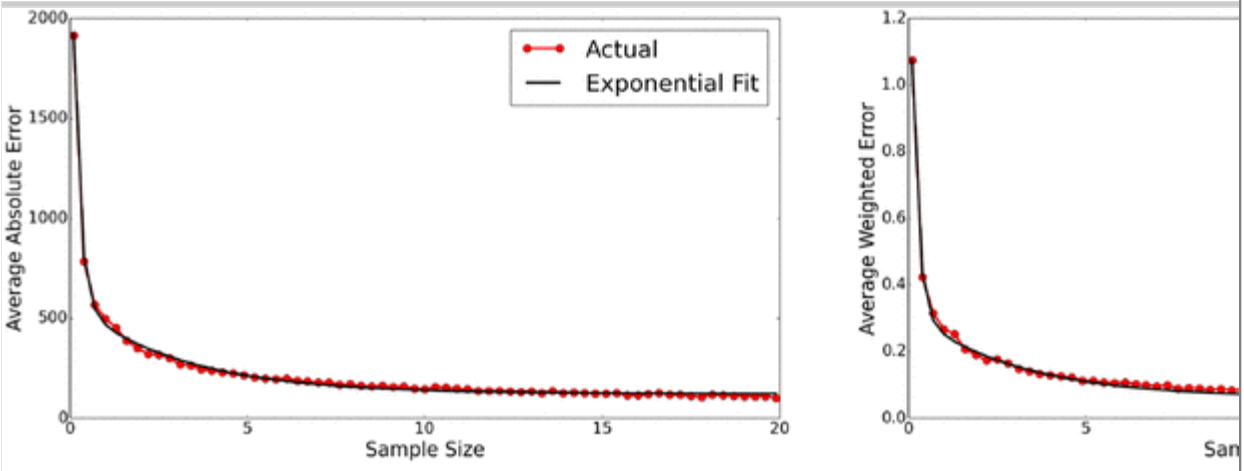
(c) ER3 Network



(d) ER4 Network

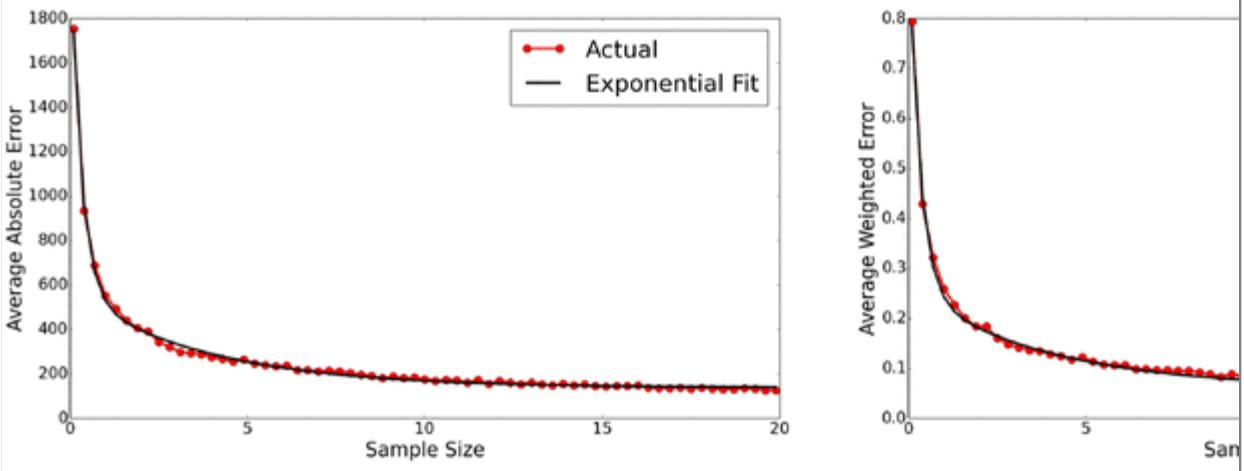
Fig. 10

Error versus sample size for ER1 network



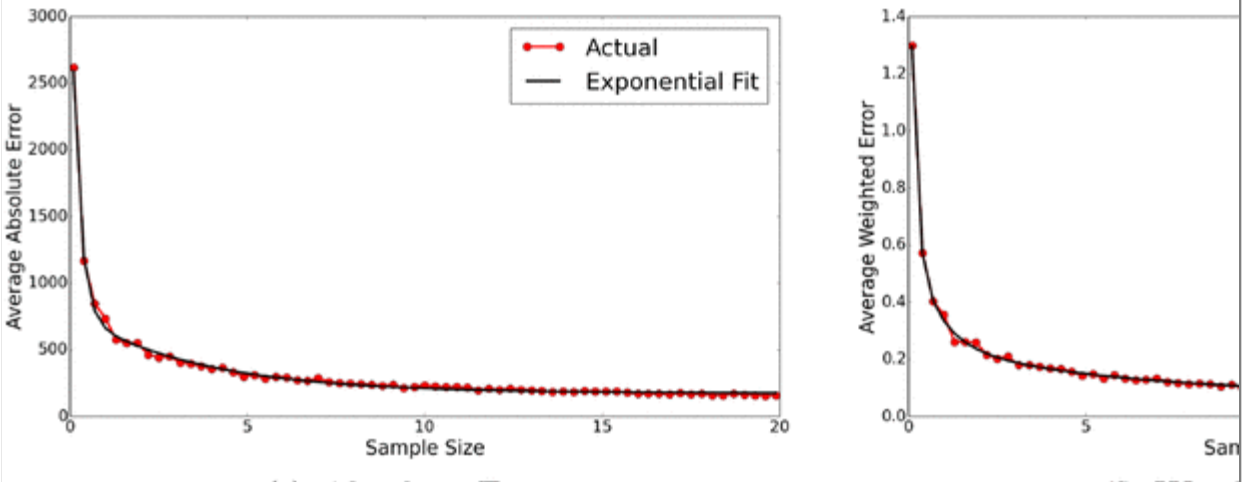
(a) Absolute Error (b) Weig

(a) US Method



(c) Absolute Error (d) Weig

(b) RW Method



The results for ER networks are consistent with the scale-free networks' results. The average error of RW method is very close to US method, and it can be efficiently used for large-size random networks.

Figure 9 presents the error versus degree rank. The results show a different behavior: the estimation error first increases with the rank and then decreases due to the Poisson degree distribution.

To see the behavior of the error as the sample size changes, we refer to the plots for one ER network displayed in Fig. 10, for each of US, MH, and RW methods, respectively. The red plot depicts the actual error and the black one is the exponential decaying fitted curve. The best fit curve is plotted using scaled Levenberg–Marquardt algorithm (Moré 1978) with 1000 iterations and 0.0001 tolerance. As expected, larger sizes produce more accurate results, both the absolute and weighted error following exponential decay as a function of the sample size.

8. Conclusion

In real-life applications, the entire network is not available to be studied due to their big size. Therefore, it is not feasible to have the global properties of real-world networks around as input for research. Of interest to us is to identify the degree rank of a node that denotes how high a node ranks in a network (rather than its degree). We thus propose methods to estimate this degree rank of nodes using network characteristics and sampling techniques.

The first method exploits the degree distribution characteristic of networks to estimate the rank of a node. The sampling-based methods collect a small sample set using sampling methods: Uniform Sampling (US), Random Walk (RW), Metropolis-Hastings Random Walk (MH). Our results conclude that the proposed methods estimate the rank of higher degree nodes better than the lower degree nodes, which is desired, since, both in theory and in practice, the focus is on the top-ranked nodes. Moreover, we

show that the rank estimates are better as the network's size increases, which is also ideal for practical applications. Our experiments also show that the estimation error decreases exponentially with the sample size which is an important point to note for the practitioners.

Based on our analysis, our recommendation is to sample using RW; the accuracy of RW method follows closely with the US method which is as we observed is the best of all the methods discussed. The accuracy of the proposed methods is evaluated using absolute and weighted error functions, the latter depending on the quartile the sampled node is in.

As an extension, one can consider finding methods to estimate degree rank in other types of networks like weighted networks, directed networks, multilayer networks, and so on. Naturally, one can ask a similar question of quickly estimating the rank of a node based on other centrality measures like closeness centrality, betweenness centrality, Katz centrality, PageRank, coreness, and so on. The complexity to compute these global centrality measures is high, and so local algorithms to approximate the global rank of the nodes are of great interest. It has been shown in the literature that hybrid centrality measures perform well to identify the influential nodes. Therefore, these methods can be further extended to estimate the hybrid centrality rank.

Acknowledgements

Gera would like to thank the DoD for sponsoring this work. Saxena and Iyengar would like to thank IIT Ropar HPC committee for providing the resources to perform the experiments.

9. Appendix

10. Results on real-world scale-free networks

See Table 6.

Table 6

Absolute and weighted error in the estimated degree rank on real-world social networl

Network	Type	Ref	Nodes	Edges	Avg Deg	]]
Friendster	Social	Rossi and Ahmed (2015)	5689498	14067887	4.95	(
Academia	Social	Fire et al. (2011)	200167	1022440	10.22	1
Dogster	Social	Rossi and Ahmed (2015)	426485	8543321	40.06	1
Facebook1	Social	Traud et al. 2012)	3097165	23667394	15.28	(
Gowalla	Social	Cho et al. (2011)	196591	950327	9.67	(
Hyves	Social	Zafarani and (Liu 2009a)	1402673	2777419	3.96	(
Foursquare	Social	Zafarani and Liu (2009a)	639014	3214985	10.06	1
Last.fm	Social	Konstas et al. (2009)	1191805	4519330	7.58	(
Livemocha	Social	Zafarani and Liu (2009b)	104103	2193082	42.13	2
Delicious	Social	Rossi and Ahmed (2015)	536108	1365961	5.10	(
Douban	Social	Rossi and Ahmed (2015)	154908	327162	4.22	1
Actor	Collaboration	Barabási and Albert (1999)	374511	15014839	80.18	3
DBLP	Collaboration	Yang and Leskovec (2015)	317080	1049866	6.62	1

Network	Type	Ref	Nodes	Edges	Avg Deg	]
						]
Digg	Social	De Choudhury et al. (2009)	261489	1536577	11.75	(
Eu-Email	Communication	Leskovec et al. (2007)	224832	339925	3.02	(
Gplus	Social	McAuley and Leskovec (2012)	107614	12238285	227.45	5
Catster	Social	Rossi and Ahmed (2015)	148826	5447464	73.21	2
YouTube	Social	Zafarani and Liu (2009a)	1134885	2987623	5.27	(
Pokec	Social	Rossi and Ahmed (2015)	1632803	22301964	27.32	2
Hollywood	Collaboration	Boldi and Vigna (2004)	1069126	56306653	105.33	2
Summary						1

References

Backstrom L, Leskovec J (2011) Supervised random walks: predicting and recommending links in social networks. In: Proceedings of the fourth ACM international conference on Web search and data mining, ACM, pp 635–644

Barabási AL, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–512

Boldi P, Vigna S (2004) The WebGraph framework I: Compression techniques. In: Proc. of the Thirteenth International World Wide Web Conference (WWW 2004), ACM Press, Manhattan, USA, pp 595–601

Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. In: Seventh international world-wide web conference (www 1998), april 14-18, 1998, brisbane, australia. Brisbane, Australia

Cem E, Sarac K (2015) Estimating the size and average degree of online social networks at the extreme. In: Communications (ICC), 2015 IEEE International Conference on, IEEE, pp 1268–1273

Cem E, Sarac K (2016a) Average degree estimation under ego-centric sampling design. In: Computer Communications Workshops (INFOCOM WKSHPS), 2016 IEEE Conference on, IEEE, pp 152–157

Cem E, Sarac K (2016b) Estimation of structural properties of online social networks at the extreme. Comput Netw 108:323–344

Chen D, Lü L, Shang MS, Zhang YC, Zhou T (2012) Identifying influential nodes in complex networks. Physica Stat Mech Appl 391(4):1777–1787

Chen L, Karbasi A, Crawford FW (2016) Estimating the size of a large network and its communities from a random sample. In: Advances in Neural Information Processing Systems, pp 3072–3080

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, pp 1082–1090

Cooper C, Radzik T, Siantos Y (2012) A fast algorithm to find all high degree vertices in power law graphs. In: Proceedings of the 21st International Conference on World Wide Web, ACM, pp 1007–1016

Dasgupta A, Kumar R, Sarlos T (2014) On estimating the average degree. In: Proceedings of the 23rd international conference on World

wide web, ACM, pp 795–806

Davis B, Gera R, Lazzaro G, Lim BY, Rye EC (2016) The marginal benefit of monitor placement on networks. In: Complex Networks VII, Springer, pp 93–104
AQ3

De Choudhury M, Sundaram H, John A, Seligmann DD (2009) Social synchrony: Predicting mimicry of user actions in online social media. In: Computational Science and Engineering, 2009. CSE'09. International Conference on, IEEE, vol 4, pp 151–158

Eden T, Ron D, Seshadhri C (2016) Sublinear time estimation of degree distribution moments: The arboricity connection. arXiv preprint arXiv:160403661

Erdős P, Rényi A (1960) On the evolution of random graphs. Publ Math Inst Hungar Acad Sci 5:17–61

Even S (2011) Graph algorithms. Cambridge University Press, Cambridge

Fire M, Tenenboim L, Lesser O, Puzis R, Rokach L, Elovici Y (2011) Link prediction in social networks using computationally efficient topological features. In: IEEE Third International Confernece on Social Computing (SocialCom), IEEE, pp 73–80
AQ4

Fortunato S, Boguñá M, Flammini A, Menczer F (2006) Approximating pagerank from in-degree. In: Algorithms and models for the web-graph, Springer, pp 59–71

Freeman LC (1977) A set of measures of centrality based on betweenness. Sociometry pp 35–41

AQ5

Ghoshal G, Barabási AL (2011) Ranking stability and super-stable nodes in complex networks. Nat Commun 2:394

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, IEEE, pp 1–9

Goodman LA (1961) Snowball sampling. The annals of mathematical statistics pp 148–170

AQ6

Hansen MH, Hurwitz WN (1943) On the theory of sampling from finite populations. Ann Math Stat 14(4):333–362

Haralabopoulos G, Anagnostopoulos I (2014) Real time enhanced random sampling of online social networks. J Netw Comput Appl 41:126–134

Hardiman SJ, Katzir L (2013) Estimating clustering coefficients and size of social networks via random walk. In: Proceedings of the 22nd international conference on World Wide Web, International World Wide Web Conferences Steering Committee, pp 539–550

Hogg T, Lerman K (2012) Social dynamics of digg. EPJ Data Sci 1(1):1–26

Hou B, Yao Y, Liao D (2012) Identifying all-around nodes for spreading dynamics in complex networks. Phys A Stat Mech Appl 391(15):4012–4017

Katz L (1953) A new status index derived from sociometric analysis. Psychometrika 18(1):39–43

Konstas I, Stathopoulos V, Jose JM (2009) On social networks and collaborative recommendation. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, ACM, pp 195–202

Kurant M, Butts CT, Markopoulou A (2012) Graph size estimation. arXiv preprint arXiv:12100460

Leskovec J, Faloutsos C (2006) Sampling from large graphs. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp 631–636

Leskovec J, Kleinberg J, Faloutsos C (2007) Graph evolution: densification and shrinking diameters. ACM Trans Knowl Discov Data (TKDD) 1(1):2

Lovász L (1993) Random walks on graphs: A survey. Comb Paul erdos is eighty 2(1):1–46

Lu J, Li D (2012) Sampling online social networks by random walk. In: Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research, ACM, pp 33–40

Lucchese R, Varagnolo D (2015) Networks cardinality estimation using order statistics. In: American Control Conference (ACC), 2015, IEEE, pp 3810–3817

Marchetti-Spaccamela A (1988) On the estimate of the size of a directed graph. In: International Workshop on Graph-Theoretic Concepts in Computer Science, Springer, pp 317–326

McAuley JJ, Leskovec J (2012) Learning to discover social circles in ego networks. NIPS 2012:548–56

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of state calculations by fast computing machines. J Chem Phys 21(6):1087–1092

Moré JJ (1978) The levenberg-marquardt algorithm: implementation and theory. In: Numerical analysis, Springer, pp 105–116

Musco C, Su HH, Lynch N (2016) Ant-inspired density estimation via random walks. In: Proceedings of the 2016 ACM Symposium on Principles of Distributed Computing, ACM, pp 469–478

Nazi A, Zhou Z, Thirumuruganathan S, Zhang N, Das G (2015) Walk, not wait: faster sampling over online social networks. Proc VLDB Endow 8(6):678–689

Ribeiro B, Towsley D (2010) Estimating and sampling graphs with multidimensional random walks. In: Proceedings of the 10th ACM SIGCOMM conference on Internet measurement, ACM, pp 390–403

Ribeiro B, Towsley D (2012) On the estimation accuracy of degree distributions from graph sampling. In: Decision and Control (CDC), 2012 IEEE 51st Annual Conference on, IEEE, pp 5240–5247

Ribeiro B, Wang P, Murai F, Towsley D (2012) Sampling directed graphs with random walks. In: INFOCOM, 2012 Proceedings IEEE, IEEE, pp 1692–1700

Rossi RA, Ahmed NK (2015) The network data repository with interactive graph analytics and visualization. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, http://networkrepository.com

Sabidussi G (1966) The centrality index of a graph. Psychometrika 31(4):581–603

Salganik MJ, Heckathorn DD (2004) Sampling and estimation in hidden populations using respondent-driven sampling. Sociol Methodol 34(1):193–240

Saxena A, Gera R, Iyengar S (2017) Observe locally rank globally. In: Advances in Social Networks Analysis and Mining (ASONAM)

Shaw ME (1954) Some effects of unequal distribution of information upon group performance in various communication nets. J Abnorm Soc Psychol 49(4):547–553

Stephenson K, Zelen M (1989) Rethinking centrality: methods and examples. Soc Net 11(1):1–37

Traud AL, Mucha PJ, Porter MA (2012) Social structure of Facebook networks. Phys A 391(16):4165–4180

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

Yang J, Leskovec J (2015) Defining and evaluating network communities based on ground-truth. Knowl Inf Syst 42(1):181–213

Ye S, Wu SF (2011) Estimating the size of online social networks. Int J Soc Comput Cyber Phys Syst 1(2):160–179

Yu Y, Fan S (2015) Node importance measurement based on the degree and closeness centrality. J Inf Commput Sci 12:1281–1291
AQ7

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

AQ8

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

Zhou Z, Zhang N, Gong Z, Das G (2016) Faster random walks by rewiring online social networks on-the-fly. ACM Trans Database Syst (TODS) 40(4):26

[1] Here, discrete probability values are considered as continuous probability density function, as this introduces a very small error.