Supplementary Material for submission 75

1 Proofs of Lemmas

Lemma IV.1. Algorithm DegreeEst provides unbiased estimates of the degree distribution and the average degree of an undirected graph.

Proof. Given an undirected graph G, in Step I, since any NRP $\mathcal{A}_i$ is a random sample of the linked lists in $\mathcal{A}$, A_s that is constructed from a set of NRPs is a random sample of the linked lists in $\mathcal{A}$. Since V_o contains the owner vertices of the linked lists in A_s, V_o is a random sample of the vertices set V in G. In addition, the obtained degree values of the vertices in V_o are correct because the corresponding linked lists contain all neighbors. Thus, the degree values of vertices in V_o construct a uniform random sample of the degree distribution in the graph G.

Given that the degree values of vertices in V_o construct a uniform random sample of the degree distribution $f^{\deg}(x)$, the degree distribution $f_s^{\deg}(x)$ of V_o satisfies $\mathbb{E}\left[f_s^{\deg}(x)\right] = f^{\deg}(x)$, which means that $f_s^{\deg}(x)$ in Step II is an unbiased estimate of the degree distribution of G.

Given that the average of a random sample is an unbiased estimate of the average of population [18], deg_s^{avg} in Step III is an unbiased estimate of the average degree deg^{avg} of G. □

Lemma IV.2. Algorithm DegreeEst provides an asymptotic unbiased estimation for the power law exponent of a given undirected graph G.

Proof. Given an undirected graph G, in Step I, since any NRP $\mathcal{A}_i$ is a random sample of the linked lists in $\mathcal{A}$, A_s that is constructed from a set of NRPs is a random sample of the linked lists in $\mathcal{A}$. Since V_o contains the owner vertices of the linked lists in A_s, V_o is a random sample of the vertices set V in G. In addition, the obtained degree values of the vertices in V_o are correct because the corresponding linked lists contain all neighbors. Thus, the degree values of vertices in V_o construct a uniform random sample of the degree distribution in the graph G.

The Power Law Exponent(PLE) γ is estimated by maximin likelihood estimation(MLE) in Step IV of DegreeEst, and the estimated equation have been proved in previous work [13, 15, 1]. Since the MLE will provide asymptotic unbiased estimation [6] on a uniform random sample, γ_s in NRPEst is an asymptotic unbiased estimate, which means that as the number n_s of sampled vertices increases, the deviation of the estimated value of PLE from the correct value becomes small [6], i.e., for all $\epsilon > 0$, $\lim_{n_s \to N} \Pr\left[|\mathbb{E}\left[\gamma_s\right] - \gamma| \geq \epsilon\right] = 0$.
Noted: The Power Law Exponent(PLE) γ of a graph is defined as $f^{\deg}(x) = C \cdot x^{-\gamma}$, where C is a constant. Previous studies take its logarithm form $\log(f^{\deg}(x)) = C + \log(x) \cdot \gamma$ and use the least squares method to solve γ [4, 3, 14, 17, 15]. However, the least squares method may result in a large arbitrary error. MLE performs better than the least squares method [13, 15, 1], thus, we use MLE method to estimate. □

Lemma IV.3. Algorithm CCEst provides unbiased estimates of the clustering coefficient distribution and the average clustering coefficient of an undirected graph.

Proof. Given an undirected graph G, in Step I, since any NRP $\mathcal{A}_i$ is a random sample of the linked lists in $\mathcal{A}$, A_s that is constructed from a set of NRPs is a random sample of the linked lists in $\mathcal{A}$. Since V_o contains the owner vertices of the linked lists in A_s, V_o is a random sample of the vertices set V in G. Since Step II-IV retrieve the edges between the neighbors of each vertex in V_o, the local clustering coefficient value of each vertex in V_o in graph G_s is the same as that in graph G. Thus, the local clustering coefficient values of vertices in V_o construct a uniform random sample of the clustering coefficient distribution in the graph G.

Given that the local clustering coefficient values of vertices in V_o construct a uniform random sample of the clustering coefficient distribution $f^{\mathrm{lcc}}(x)$, the clustering coefficient distribution $f_s^{\mathrm{lcc}}(x)$ of V_o satisfies $\mathbb{E}\left[f_s^{\mathrm{lcc}}(x)\right] = f^{\mathrm{lcc}}(x)$, which means that $f_s^{\mathrm{lcc}}(x)$ in Step V is an unbiased estimate of the clustering coefficient distribution of G.

Given that the average of a random sample is an unbiased estimate of the average of population [18], acc_s in Step VI is an unbiased estimate of the average clustering coefficient acc of G. □

Lemma IV.4. Given an undirected graph G, algorithm SPEst provides unbiased estimations for the probabilities $f^{\mathrm{sp}}(1)$, $f^{\mathrm{sp}}(2)$, and $f^{\mathrm{sp}}(3)$.

Proof. Given an undirected graph G, in Step I, since any NRP $\mathcal{A}_i$ is a random sample of the linked lists in $\mathcal{A}$, A_s that is constructed from a set of NRPs is a random sample of the linked lists in $\mathcal{A}$. Since V_o contains the owner vertices of the linked lists in A_s, V_o is a random sample of the vertices set V in G. Since V_o can be considered as a random sample of the vertices in G, all vertex pairs in V_o can be regarded as randomly sampled. Thus, the random variables of the shortest path lengths of all vertex pairs in V_o are independent and identically distributed [19].

Considering all vertex pairs in V_o, Figure 1 shows the cases when the shortest path length of a vertex pair is 1, 2, and 3 in G_s, respectively. Black dots represent the owner vertices in V_o and white dots denote the neighbors of the owner vertices. Note that the owner vertices could also be the neighbors of each other. For any vertex pair u, v in V_o, if the length of the shortest path between u and v is 1 or 2 in G_s, as shown in Figures 1(a) and 1(b), the length of the shortest path between u and v must be 1 or 2 in G. This is because G_s includes the neighbors of all vertices in V_o. If the length of the shortest path between u and v is 3 in G_s, as shown in Figure 1(c), they must be connected via two neighbors in G_s. If these two neighbors are not owner vertices and there exists an edge between them in G, Step II-III adds this edge to G_s. Since G_s includes the neighbors of all vertices in V_o, it is impossible for u and v to have a path shorter than 3. Thus, the length of the shortest path between u and v must be 3 in G. Therefore, the shortest paths of lengths 1, 2, and 3 are preserved correctly of vertex pairs of V_o in G_s.

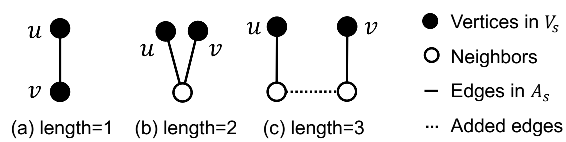


Figure 1: Vertex pairs with shortest path lengths 1, 2, and 3.

Let $x^l_{(u,v)}$ be a random variable, such that if the shortest path length between u and v is l, $x^l_{(u,v)} = 1$, otherwise, $x^l_{(u,v)} = 0$. Random variable $x^l_{(u,v)}$ obeys the binomial distribution, i.e., $x^l_{(u,v)} \sim b(1, f^{\mathrm{sp}}(l))$, where $f^{\mathrm{sp}}(l)$ is the probability of any vertex pair in G with the shortest path length l. Given $l = 1, 2, 3$, we have that

$$\mathbb{E}\left[f_s^{\mathrm{sp}}(l)\right] = \mathbb{E}\left[\frac{L_s^l}{\binom{|V_o|}{2}}\right] = \mathbb{E}\left[\frac{\sum_{\forall u,v\in V_o} x^l_{(u,v)}}{\binom{|V_o|}{2}}\right] = \frac{\sum_{\forall u,v\in V_o} \mathbb{E}\left[x^l_{(u,v)}\right]}{\binom{|V_o|}{2}} = \frac{\binom{|V_o|}{2}\cdot f^{\mathrm{sp}}(l)}{\binom{|V_o|}{2}} = f^{\mathrm{sp}}(l).$$

Hence, $f_s^{\mathrm{sp}}(l)$ in Step VI is an unbiased estimation of the probability of any vertex pair in G with the shortest path length l, $l = 1, 2, 3$. ☐

Lemma IV.5. Algorithm JDEst provides an unbiased estimation for the joint degree distribution of a given undirected graph G.

Proof. Given an undirected graph G, in Step I, since any NRP $\mathcal{A}_i$ is a random sample of the linked lists in $\mathcal{A}$, A_s that is constructed from a set of NRPs is a random sample of the linked lists in $\mathcal{A}$. Since V_o contains the owner vertices of the linked lists in A_s, V_o is a random sample of the vertices set V in G.

Suppose that $z_{(x,y)}$ means the number of edges with the joint degree pair (x, y) in G and $\hat{z}_{(x,y)}$ means the estimated number of edges with the joint degree pair (x, y) in G. Though G is an undirected graph, the probability of edge contained in E' by firstly sampling an owner vertex with degree x then sampling a neighbor with degree y is different from firstly sampling an owner vertex with degree y then sampling a neighbor with degree x. Therefore, we define $c_{\langle x,y\rangle}$ as the number of edges in G obtained by firstly sampling an owner vertex with degree x then sampling a neighbor with degree y and $\hat{c}_{\langle x,y\rangle}$ as the number of the ordered pair $\langle x, y\rangle$ in JD. It is easy to know that $z_{(x,y)} = c_{\langle x,y\rangle} + c_{\langle y,x\rangle}$.

$$\begin{aligned}
\mathbb{E}\left[\hat{z}_{(x,y)}\right] &= \mathbb{E}\left[\frac{c_{\langle x,y\rangle}}{\pi_{\langle x,y\rangle}} + \frac{c_{\langle y,x\rangle}}{\pi_{\langle y,x\rangle}}\right] = \mathbb{E}\left[\frac{c_{\langle x,y\rangle}}{\pi_{\langle x,y\rangle}}\right] + \mathbb{E}\left[\frac{c_{\langle y,x\rangle}}{\pi_{\langle y,x\rangle}}\right] \\
&= \mathbb{E}\left[\sum_{\substack{deg(v_i)=x, deg(v_j)=y \\ (v_i,v_j)\in E'}} \frac{1}{\pi_{\langle x,y\rangle}}\right] + \mathbb{E}\left[\sum_{\substack{deg(v_i)=y, deg(v_j)=x \\ (v_i,v_j)\in E'}} \frac{1}{\pi_{\langle y,x\rangle}}\right] \\
&= \mathbb{E}\left[\sum_{(v_i,v_j)\in E'} \frac{I_{(v_i,v_j)}}{\pi_{\langle x,y\rangle}}\right] + \mathbb{E}\left[\sum_{(v_i,v_j)\in E'} \frac{I_{(v_i,v_j)}}{\pi_{\langle y,x\rangle}}\right] \\
&= c_{\langle x,y\rangle} + c_{\langle y,x\rangle} = z_{(x,y)}, \qquad (1)
\end{aligned}$$

where $I_{(v_i,v_j)}$ is a variable of edge (v_i, v_j). When $deg(v_i) = x$ and $deg(v_j) = y$ or $deg(v_i) = y$ and $deg(v_j) = x$, $I_{(v_i,v_j)} = 1$, otherwise $I_{i,j} = 0$. Equation 1 holds because of Horvitz-Thompson estimation theory [8]. Horvitz-Thompson estimation theory illustrate that if there is a population U contain all elements, each element i in S has a value y_i, then we want to estimate the total number T of all elements in U, $T = \sum_{i\in U} y_i$. Horvitz-Thompson estimator is an unbiased estimator to estimate the total of population T by sampling without replacement. If we have a sample $S \subset U$, each element in S is sampled independently, and T can be estimated as $\hat{T} = \sum_{i\in S} \frac{y_i}{\pi_i}$, where π_i is the probability of element i being sampled into S without replacement and $\mathbb{E}\left[\hat{T}\right] = T$.

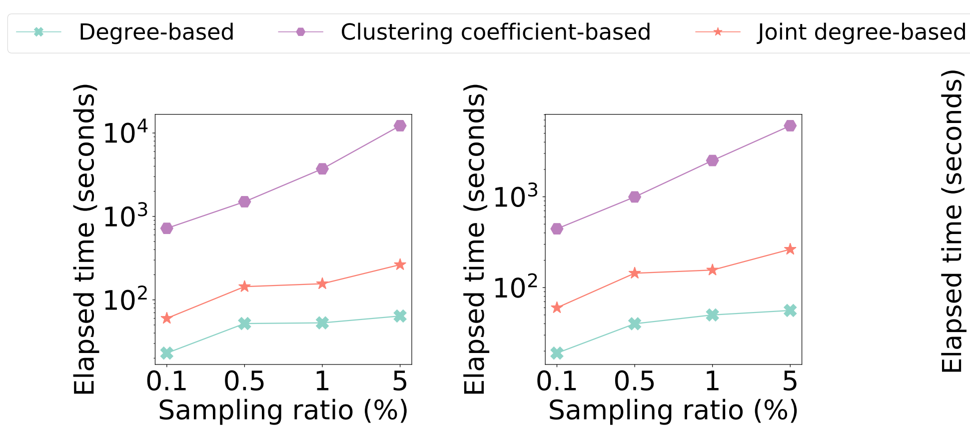
Because $\hat{z}_{(x,y)}$ is an unbiased estimator of $z_{(x,y)}$, thus, we can estimate the portion of joint degree in G, $f_s^{\text{jd}}(x,y) = \frac{\hat{z}_{(x,y)}}{\sum \hat{z}_{(x,y)}}, (v_x, v_y) \in E'$, which is unbiased.

□

2 Experiments for different sampling ratio

This section presents the evaluation results of the proposed NRPEst when varying the sampling ratio from 0.1% to 5%. The experiments are performed on a cluster consists of 11 computing nodes.

2.1 Efficiency



(a) UKDomain.

(b) Friendster.

Figure 2: Time cost when using different sampling ratios.

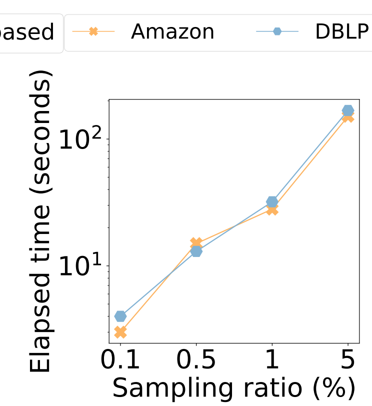
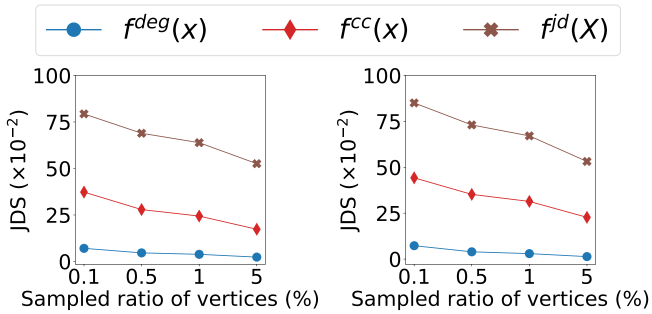


Figure 3: Time cost when using different sampling ratios on shortest path-based properties.

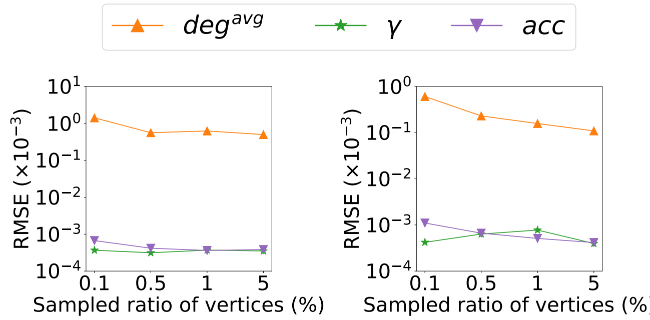
Figures 2 and 3 show the elapsed time of NRPEst when varying the sampling ratio for estimating degree-based, clustering coefficient-based, joint degree-based, and shortest path-based properties, respectively. As the sampling ratio increases, the time costs of the NRPEst increase for all graph properties. Estimating degree-based properties is more efficient than estimating the other properties. The time cost of estimating shortest path-based properties increases more rapidly than that of the other properties.

2.2 Accuracy



(a) UKDomain. (b) Friendster.

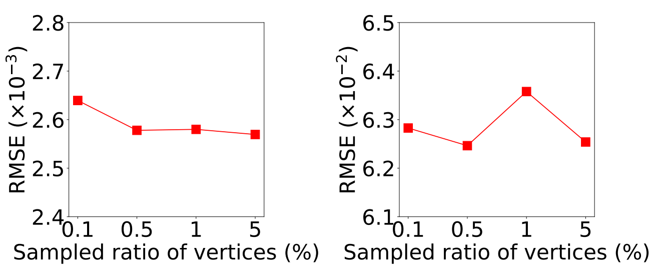
Figure 4: Distribution properties.



(a) UKDomain. (b) Friendster.

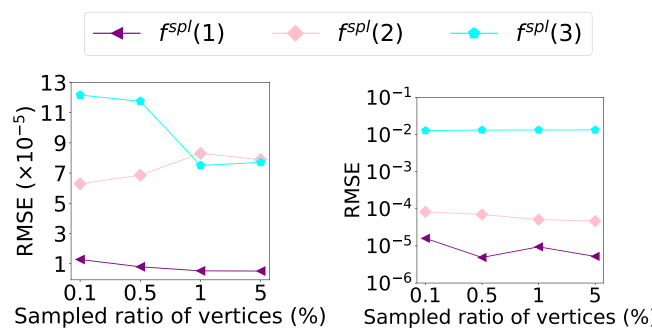
Figure 5: Single-valued properties.

Figures 4 and 5 show the accuracy of the degree-based, clustering coefficient-based, and joint degree-based property estimations on datasets UKDomain and Friendster with different sampling ratios. As the sampling ratio increases, the accuracy of the estimated joint degree distribution is getting better (the value of JSD becomes smaller), the accuracy of the estimated clustering coefficient distribution is improved slightly, and the accuracy of the estimated degree distribution does not change much. As the sampling ratio increases, the accuracy of the estimated average degree becomes better and the accuracy of the other estimated single-valued properties do not change much. In general, the accuracy of the estimated values is expected to become better as the sampling ratio increases. Our method provide unbiased estimation for degree-based and clustering coefficient-based properties, so that samples of size 0.1%–5% are enough to achieve accurate estimations on datasets UKDomain and Friendster.



(a) Amazon. (b) DBLP.

Figure 6: Shortest path distribution.

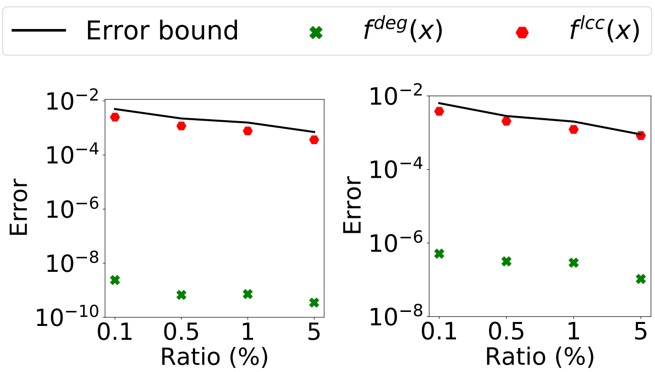


(a) Amazon. (b) DBLP.

Figure 7: The probability of shortest path l, $l = 1, 2, 3$.

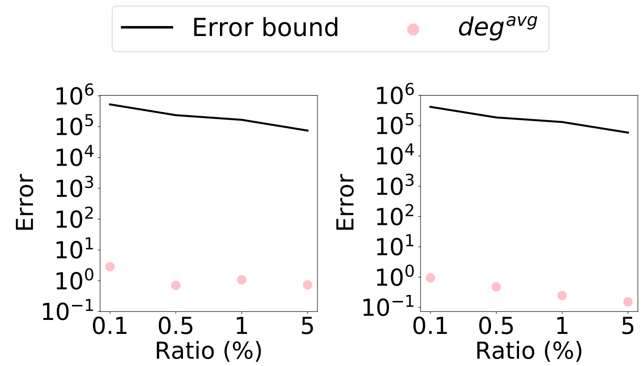
Figures 6 and 7 show the accuracy of the shortest path-based property estimation on datasets Amazon and DBLP with different sampling ratios. Figure 7 shows that the accuracy of the estimated $f^{\rm spl}(1)$ and $f^{\rm spl}(3)$ have decreasing trend as the sampling ratio increases on dataset Amazon, and for the other cases, the accuracy fluctuates slightly.

To summarize, our NRPEst achieve accurate and efficient estimation when the sampling ratio is 1%.



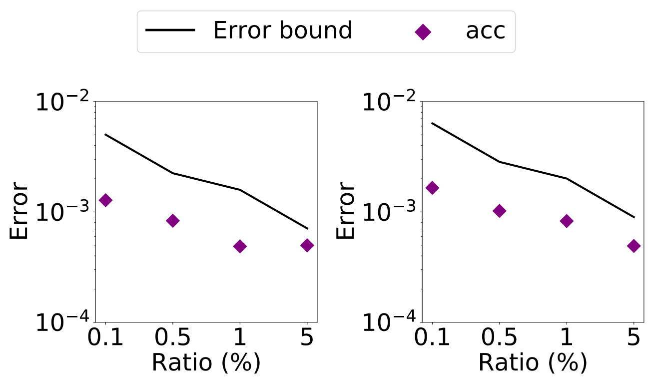
(a) UKDomain. (b) Friendster.

Figure 8: Degree and clustering coefficient distribution.



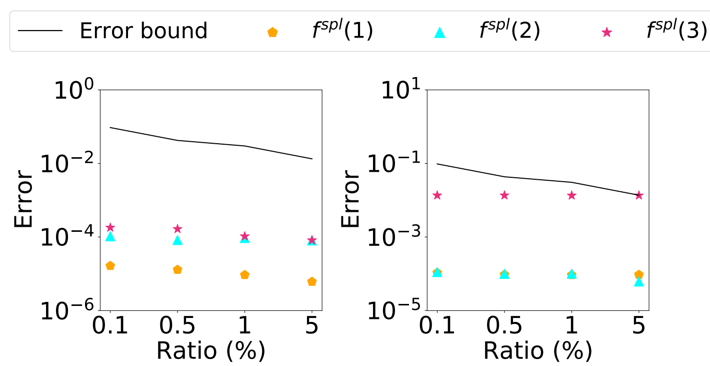
(a) UKDomain. (b) Friendster.

Figure 9: Average degree.



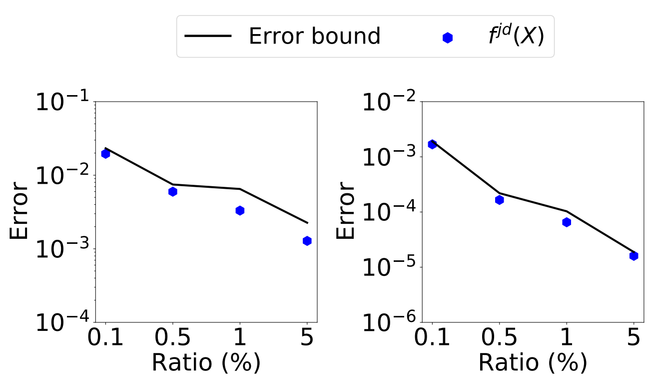
(a) UKDomain. (b) Friendster.

Figure 10: Average clustering coefficient.



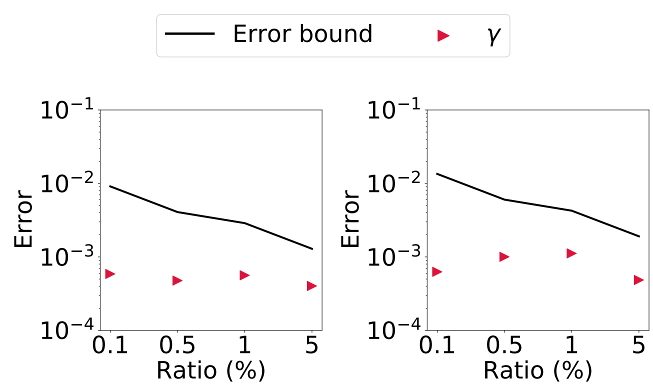
(a) Amazon. (b) DBLP.

Figure 11: Shortest path length distribution.



(a) UKDomain. (b) Friendster.

Figure 12: Joint degree distribution.



(a) UKDomain. (b) Friendster.

Figure 13: Power law exponent.

2.3 Error Bounds vs. Empirical Errors

Figures 8 – 13 give the comparison between estimation guarantee with the actual estimation error of NRPEst under different sampling ratios for different properties. The JSD and RMSE shown in Section 2.2 illustrate the average accuracy for 5 trials under a given sampling ratio, and the accuracy shown in this section is the worse case of

NRPEst. Obviously, if the the worse error can be bounded, though the estimation is trapped in the worse case, the estimation result will not be unacceptable. All the black lines in Figures 8 – 13 is the theoretical guarantee and points is the worse error for 5 trials under a given sampling ratio on different property estimations.

Degree and clustering coefficient distribution estimation. According to Dvoretzky–Kiefer–Wolfowitz–Massart inequality, the error bound of the estimated cumulative distribution function $\kappa_s(x)$ of property κ is

$$\Pr\left(\max_x |\kappa_s(x) - \kappa(x)| > \varepsilon\right) \le C \cdot e^{-2n_s\varepsilon^2}, \tag{2}$$

where $n_s = |V_s|$ is the number of sampled vertices and C is a constant, which can be calculated by sample ratio multiplying the number of vertices of the graph, and $\kappa_s(x)$ is the estimated cumulative distribution function of property κ. For Equation 2, define $\alpha = C \cdot e^{-2n_s\varepsilon^2} \in (0, 1)$ and we have

$$\varepsilon = \sqrt{-\frac{1}{2 \cdot n_s} \cdot \ln\left(\frac{\alpha}{C}\right)}.$$

We can construct the confidence interval for the estimated distribution property. For any x, we have the equation below with probability at least $1 - \alpha$

$$\kappa(x) - \varepsilon \le \kappa_s(x) \le \kappa(x) + \varepsilon. \tag{3}$$

Here we choose $\alpha = 0.01$, which guarantee that the probability that Equation 3 holds is higher than 99%, i.e., the estimation can be bounded by the the ε with the probability at least 99%. Since $C = 2k$, where k is the dimension of the distribution [11, 12], and the dimension of degree and clustering coefficient distribution is $k = 1$, the error bound of estimation under different sampling ratios can be calculated as black line shown in Figure 8. As Figure 8 shows that all estimation of degree distribution and clustering coefficient distribution in the NRPEst can be bounded by the Equation 2 (points of under the black lines).

Average degree, average clustering coefficient and the probability of vertex pairs with distance l, $l = 1, 2, 3$. According to Hoeffding's inequality, the error bound of a bounded estimator κ is

$$\Pr\left(|\kappa_s - \kappa| > \varepsilon\right) \le 2e^{-\frac{2n_s\varepsilon^2}{(b-a)^2}}, \quad \kappa \in [a, b]. \tag{4}$$

For equation 4, the same as mentioned above, define $\alpha = 2e^{-\frac{2n_s\varepsilon^2}{(b-a)^2}} \in (0, 1)$ and we have

$$\varepsilon = \sqrt{-\frac{(b-a)^2}{2 \cdot n_s} \cdot \ln\left(\frac{\alpha}{2}\right)}.$$

We can construct the confidence interval for the estimated single-valued property, then we have the equation below with probability at least $1 - \alpha$

$$\kappa - \varepsilon \le \kappa_s \le \kappa + \varepsilon. \tag{5}$$

The same, here we choose $\alpha = 0.01$, which guarantee that the probability that Equation 5 holds is higher than 99%, i.e., the estimation can be bounded by the the ε with the probability of 99%. The bounded estimated single value properties in this paper include average degree deg^{avg}, average clustering coefficient acc and the probability of shortest path $f^{\mathrm{spl}}(l)$ with length l, $l = 1, 2, 3$. We know $acc, f^{\mathrm{spl}}(1), f^{\mathrm{spl}}(2), f^{\mathrm{spl}}(3) \in [0, 1]$ and $deg^{avg} \in (0, N)$, where N is the number of vertices in the graph. The error bound of estimation of each property can be calculated as black lines shown in Figures 9 – 11. As Figures 9 – 11 show that all estimation of single-value properties in the NRPEst can be bounded by the error bound calculated by equation 5 (points of maximin error under the black lines).

Joint degree distribution. Note that the joint degree distribution $f^{\mathrm{jd}}(X)$ is estimated using the Horvitz-Thompson estimator where each edge is sampled with unequal probability, so that Equation 2 is not applicable. Therefore, the error bound of $f^{\mathrm{jd}}(x)$ is derived by using the Central Limit Theorem [2] to construct the confidence interval [10]. For each given X_0, $f^{\mathrm{jd}}(X_0)$ is a single value need to be estimated. Define a random variable $I^{X_0}_{(u,v)}$ for a edge (u, v), if the joint degree value of (u, v) is X_0, $I^{X_0}_{(u,v)} = 1$, otherwise $I^{X_0}_{(u,v)} = 0$. Then, $I^{X_0}_{(u,v)}$ obeys Bernoulli distribution, $I^{X_0}_{(u,v)} \sim b(M, f^{\mathrm{jd}}(X_0))$, where M is the number of edges of the graph. $f^{\mathrm{jd}}_s(X_0)$ is the estimated value of $f^{\mathrm{jd}}(X_0)$, and according to Central Limit Theorem [2], $f_s^{\mathrm{jd}}(X_0)$ can be approximated to obey to the normal distribution. Therefore, we have the equation below with probability at least $1-\alpha$

$$f^{\mathrm{jd}}(X_0) - \varepsilon_{X_0} \leq f_s^{\mathrm{jd}}(X_0) \leq f^{\mathrm{jd}}(X_0) + \varepsilon_{X_0}, \quad \text{where } \varepsilon_{X_0} = z_{\alpha/2} \cdot \frac{\sigma_{X_0}}{\sqrt{m_s}}, \tag{6}$$

where $z_{\alpha/2}$ is the z-value in Standard Normal Distribution Table [9], and m_s is the number of sampled edges. Here we choose $\alpha = 0.01$, which guarantee that the probability that Equation 6 holds is higher than 99%, i.e., the estimation can be bounded by the the ε_{X_0} with the probability at least 99%. Since the edges used for estimation in JDEst are obtained by sampling any neighbor of the owner vertex, the number of sampled edges m_s used for estimation is equal to the number of sampled owner vertices n_s. When $\alpha = 0.01$, $z_{\alpha/2} = 2.58$. σ_{X_0} is Standard Deviation of $f^{\mathrm{jd}}(X_0)$ and it can be approximated by Bootstrap method [16, 7].

Combined with the above analysis, we can calculate the error bound of $f^{\mathrm{jd}}(X_0)$. Moreover, the maximum value of ε_{X_0}, $\forall X_0$, can be the error bound of $f^{\mathrm{jd}}(X)$, which can be define as ε. For any X, we have the equation below with probability at least $1-\alpha$

$$f^{\mathrm{jd}}(X) - \varepsilon \leq f_s^{\mathrm{jd}}(X) \leq f^{\mathrm{jd}}(X) + \varepsilon, \tag{7}$$

where $\varepsilon = \max\limits_{X_0} \left\{\varepsilon_{X_0}\right\} = \max\limits_{X_0} \left\{2.58 \cdot \frac{\sigma_{X_0}}{\sqrt{m_s}}\right\}$ The error bound of estimation can be calculated as shown in Figure 12 black line, which means the estimation can be bounded by the the ε with the probability of 99%. As Figure 12 shows that all estimation of joint degree distribution in the JDEst can be bounded by the error bound calculated by Equation 7.

Power law exponent. Note that the power law exponent γ is estimated by maximin likelihood estimation which is asymptotically unbiased, so that Equation 4 is not applicable. Therefore, we will use the theory of the MLE method to construct confidence interval. In the case of large sample, the distribution of the MLE estimator approximately obeys a normal distribution [5] $\gamma_s \sim \mathcal{N}\left(\gamma, \left(\sqrt{n_s} \cdot \mathcal{I}(\gamma)\right)^{-1}\right)$, where γ_s is the MLE of power law exponent, γ is the ground truth of power law exponent, and $\mathcal{I}(\gamma)$ is the Fisher information matrix. Therefore, we have the equation below with probability at least $1-\alpha$

$$\gamma_{-}\varepsilon \leq \gamma_s \leq \gamma + \varepsilon, \quad \text{where } \varepsilon_{X_0} = z_{\alpha/2} \cdot \sqrt{\left(\sqrt{n_s} \cdot \mathcal{I}(\gamma)\right)^{-1}}, \tag{8}$$

where $z_{\alpha/2}$ is the z-value in Standard Normal Distribution Table [9]. Here we choose $\alpha = 0.01$, which guarantee that the probability that Equation 8 holds is higher than 99%, i.e., the estimation can be bounded by the the ε_{X_0} with the probability at least 99%. When $\alpha = 0.01$, $z_{\alpha/2} = 2.58$. The error bound of estimation can be calculated as shown in Figure 13 black line, which means the estimation can be bounded by the the ε with the probability of 99%. As Figure 13 shows that all estimation of power law exponent in the DegreeEst can be bounded by the error bound calculated by Equation8.

References

[1] Bauke, H. Parameter estimation for power-law distributions by maximum likelihood methods. *The European Physical Journal B 58* (2007), 167–173.

[2] Fischer, H. *A history of the central limit theorem: from classical to modern probability theory*, vol. 4. Springer, 2011.

[3] Gale, J. F., Laubach, S. E., Marrett, R. A., Olson, J. E., Holder, J., and Reed, R. M. Predicting and characterizing fractures in dolostone reservoirs: Using the link between diagenesis and fracturing. *Geological Society, London, Special Publications 235*, 1 (2004), 177–192.

[4] Guerriero, V. Power law distribution: Method of multi-scale inferential statistics. *Journal of Modern Mathematics Frontier 1*, 1 (2012), 21–28.

[5] Hansen, B. E., and Horowitz, J. L. Handbook of econometrics, vol. 4. *Econometric Theory 13*, 1 (1997), 119–132.

[6] Heijmans, R. D., and Magnus, J. R. Consistent maximum-likelihood estimation with dependent observations: The general (non-normal) case and the normal case. *Journal of Econometrics 32*, 2 (1986), 253–285.

[7] Horowitz, J. L. Bootstrap methods in econometrics. *Annual Review of Economics 11*, 1 (2019), 193–224.

[8] Horvitz, D. G., and Thompson, D. J. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association 47*, 260 (1952), 663–685.

[9] Huber, F. *A logical introduction to probability and induction.* Oxford University Press, 2018.

[10] Illowsky, B., and Dean, S. Introductory statistics. *OpenStax College, Texas* (2015).

[11] Massart, P. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The annals of Probability* (1990), 1269–1283.

[12] Naaman, M. On the tight constant in the multivariate dvoretzky–kiefer–wolfowitz inequality. *Statistics & Probability Letters 173* (2021), 109088.

[13] Newman, M. E. Power laws, pareto distributions and zipf's law. *Contemporary physics 46*, 5 (2005), 323–351.

[14] Ortega, O. J., Marrett, R. A., and Laubach, S. E. A scale-independent approach to fracture intensity and average spacing measurement. *AAPG bulletin 90*, 2 (2006), 193–208.

[15] Reitan, T., and Petersen-Øverleir, A. Existence of the frequentistic estimate for power-law regression with a location parameter, with applications for making discharge rating curves. *Stochastic Environmental Research and Risk Assessment 20* (2006), 445–453.

[16] Tibshirani, R. J., and Efron, B. An introduction to the bootstrap. *Monographs on statistics and applied probability 57*, 1 (1993), 1–436.

[17] Umemoto, D., and Ito, N. Power-law distribution in an urban traffic flow simulation. *Journal of Computational Social Science 1*, 2 (2018), 493–500.

[18] Underhill, L., and Bradfield, D. *Introstat.* Juta and Company Ltd, 1996.

[19] Ye, Q., Wu, B., and Wang, B. Distance distribution and average shortest path length estimation in real-world networks. In *ADMA* (2010), pp. 322–333.