Supplementary Material for submission ICDE

1 Proofs of Lemmas

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

Proof. The PLE γ of a graph is defined as $f^{\deg}(x) = C \cdot x^{-\gamma}$. Previous studies take its logarithm form $\log(f^{\deg}(x)) = C + \log(x) \cdot \gamma$ and use the least squares method to solve γ [2, 1, 6, 8, 7]. However, the least squares method may result in a large arbitrary error and MLE performs better than the least squares method [5, 7]. Hence, Step VI in algorithm DegreeEst applies the MLE method to estimate γ based on vertex set V_s.

For power-law exponent property, let operation ψ_r be Step VI of DegreeEst. Obviously, ψ_r is polynomial. DegreeEst computes unbiased degree distribution as Lemma IV.1. shows, and according to the MLE method, as the number n_s of sampled vertices increases, the deviation of the estimated value of PLE from the correct value becomes small [3], i.e., $\lim_{n_s \to N} P\left[|E(\gamma_s) - \gamma| \geq \epsilon\right] = 0$. Therefore, DegreeEst offers asymptotic unbiased estimation of PLE with high probability. □

Lemma IV.3. Algorithm CCEst provides unbiased estimations for the clustering coefficient distribution and the average clustering coefficient of a given undirected graph G.

Proof. We prove Lemma IV.3. by showing that subgraph $G_s = (V_s, E_s)$ in Step IV is a graph random sample w.r.t. clustering coefficient distribution and the average clustering coefficient, because $V_o \subseteq V_s$ is a uniform sample w.r.t. the local clustering coefficient distribution of G. Given the labeled adjacency list $\mathcal{A}$ of an undirected graph $G = (V, E)$, any NRP $\mathcal{A}_i$ is a random sample of the linked lists in $\mathcal{A}$. Then, in Step I of CCEst, A_s constructed from a set of NRPs is a random sample of the linked lists in $\mathcal{A}$. Since Step IV retrieves the edges between the neighbors of the vertices in V_o, for each vertex in V_o, its local clustering coefficient in graph G can be calculated,i.e., V_o is a uniform random sample of the vertices in V with correct local clustering coefficient.

For clustering coefficient distribution property, let operation ψ_r be Step V of CCEst. Obviously, ψ_r is polynomial. Step V derives the clustering coefficient distribution $f_s^{\text{lcc}}(x)$ based on the uniform random sample. Easily, we have $\mathbb{E}\left[f_s^{\text{lcc}}(x)\right] = f^{\text{lcc}}(x)$. Hence, $f_s^{\text{lcc}}(x)$ is an unbiased estimation of the clustering coefficient distribution of G.

For average clustering coefficient property, let operation ψ_r be Step VI of CCEst. Obviously, ψ_r is polynomial. Step VI estimates the average clustering coefficient as $acc_s = \sum_{x=0}^{lcc_s^{max}} f^{\text{lcc}}(x) \cdot x$. Then, we have

$$\mathbb{E}\left[acc_s\right] = \mathbb{E}\left[\sum_{x=0}^{lcc_s^{max}} f_s^{\text{lcc}}(x) \cdot x\right] = \sum_{x=0}^{lcc^{max}} f^{\text{lcc}}(x) \cdot x = acc.$$

Hence, acc_s in Step VI is an unbiased estimation of the average clustering coefficient of G. □

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

Proof. 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_s are independent and identically distributed [9].

Considering all vertex pairs in V_s, 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_s and white dots denote the neighbors of the owner vertices. Note that the owner vertices could also be the neighbors of each other.

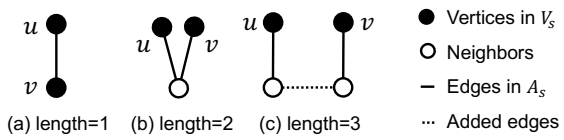


Figure 1: Vertex pairs with shortest path lengths 1, 2, and 3.

For any vertex pair u, v in V_s, 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_s. 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 III adds this edge to G_s. Since G_s includes the neighbors of all vertices in V_s, 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 in G_s.

For shortest path distribution property, let operation ψ_r be Step II-V of SPEst. Obviously, ψ_r is polynomial. 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. According to Steps V and VI, given $l = 1, 2, 3$, we have that

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

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. We prove Lemma IV.4. by showing that subgraph $G_s = (V_s, E_s)$ in Step II is a graph random sample w.r.t. joint degree distribution, because $E' \subseteq V_s$ is a random sample w.r.t. the joint degree distribution of G. Although edges in E' are sampled without replacement independently in unequal probabilities, Step III uses the Horvitz-Thompson estimator [4] to construct unbiased estimators $z_{x,y}$ from E', which can be used to construct an unbiased estimation of $f_s^{\mathrm{jd}}(X)$.

For joint degree distribution property, let operation ψ_r be Step II-IV of JDEst. Obviously, ψ_r is polynomial. 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 $z_{(x,y)} = c_{\langle x,y \rangle} + c_{\langle y,x \rangle}$.

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

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 estimator [4]. If there is a population U contain all research 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

$$
\hat{T} = \sum_{i\in S} \frac{y_i}{\pi_i},
$$

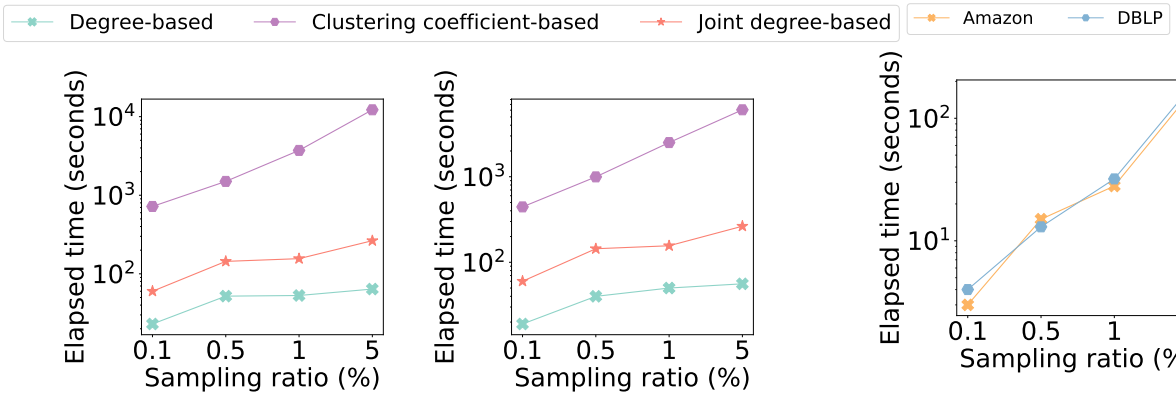
where π_i is the probability of element i being sampled into S without replacement and $\mathbb{E}\left[\hat{T}\right] = T$.

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

□

2 Experiments for different sampling ratio

2.1 Efficiency



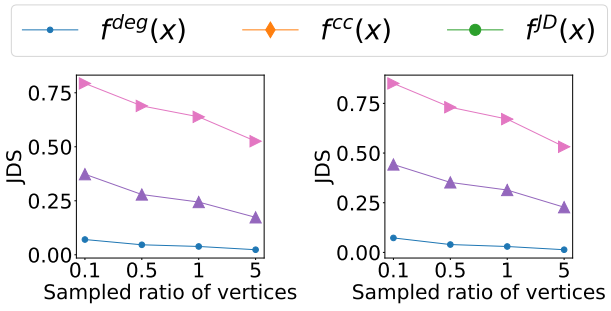
(a) UKDomain. (b) Friendster.

Figure 2: Time cost when using different sampling ratios.

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

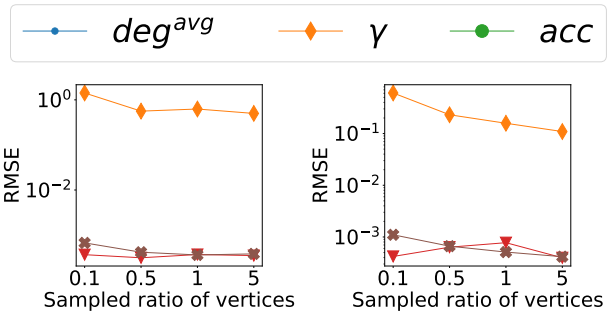
Figures 2 and 3 show the elapsed time of NRPEst when varying the sampling ratio for estimating degree-based, clustering coefficient-based, joint degree-based, and shortest path-based properties, respectively. As the sampling ratio increases, the time costs of the NRPEst increase for all graph properties. 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



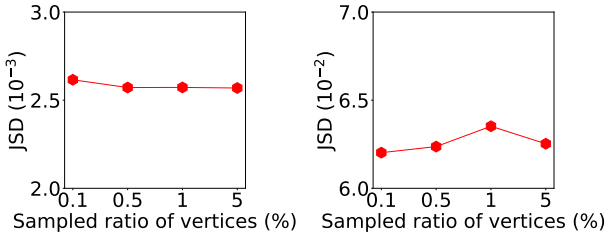
(a) UKDomain. (b) Friendster.

Figure 4: Distribution properties.



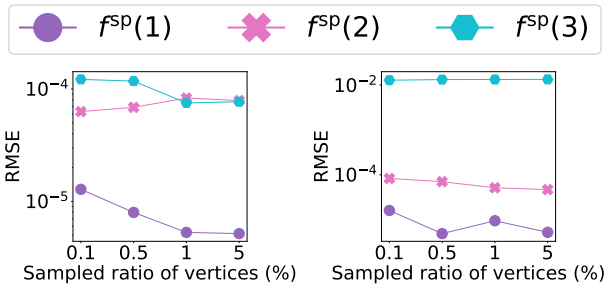
(a) UKDomain. (b) Friendster.

Figure 5: Single-valued properties.



(a) UKDomain. (b) Friendster.

Figure 6: Shortest path distribution.



(a) UKDomain. (b) Friendster.

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

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. Figures 4 and 5 show the accuracy of the degree-based, clustering coefficient-based, and joint degree-based property estimation 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 *JDS* 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.

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 $f^{\text{sp}}(1)$ and $f^{\text{sp}}(3)$ 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%.

References

[1] 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.

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

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

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

[9] 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.