Supplementary Material for submission 75

1 Proofs of Lemmas

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' is a non-uniform random sample of E 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 [1] to construct unbiased estimators $\hat{z}_{x,y}$ from E', which can be used to construct an unbiased estimation of $f_s^{\text{jd}}(X)$.

For joint degree distribution property, let operation ψ_r be Step II-IV of JDEst, which 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 [1]. 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 $\mathbb{E}\left[\hat{T}\right] = T$.

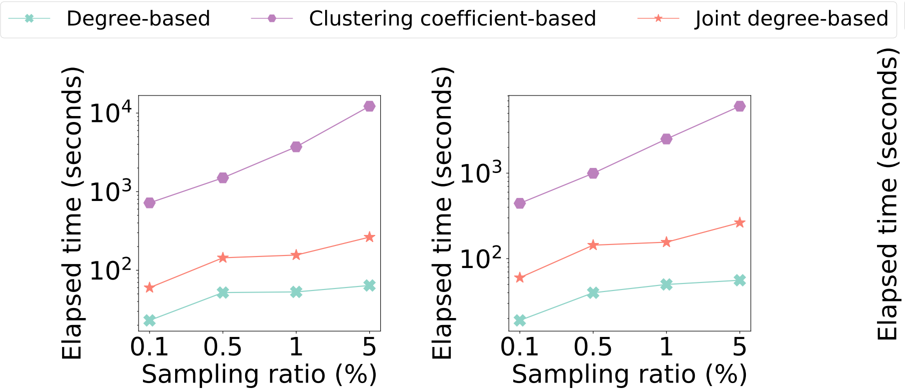
Because $\hat{z}_{(x,y)}$ is an unbiased estimator of $z_{(x,y)}$, thus, we can estimate the portion of joint degree in G, $f_s^{\text{jd}}(x, y) = \frac{\hat{z}_{(x,y)}}{\sum \hat{z}_{(x,y)}}$, $(v_x, v_y) \in E'$, which is unbiased.

□

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 1: Time cost when using different sampling ratios.

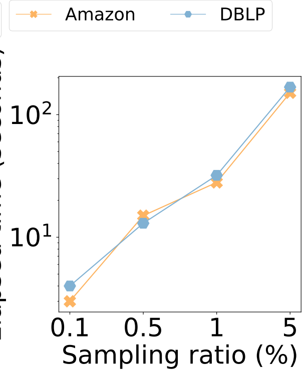
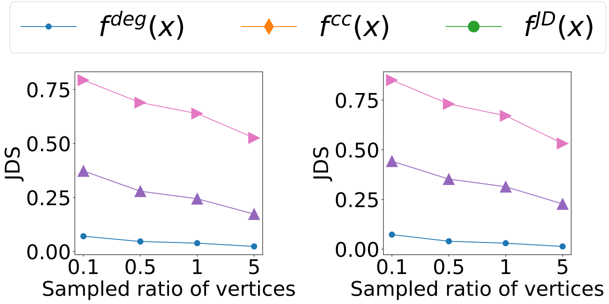


Figure 2: Time cost when using different sampling ratios on shortest path-based.

Figures 1 and 2 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.

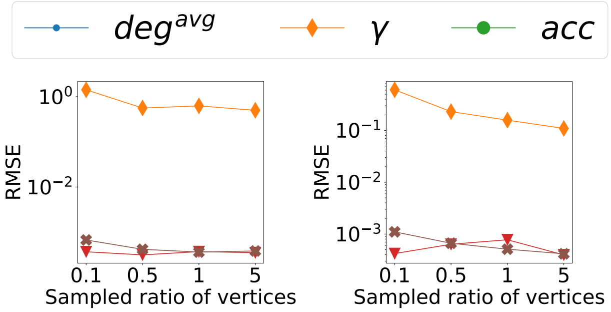
2.2 Accuracy



(a) UKDomain.

(b) Friendster.

Figure 3: Distribution properties.

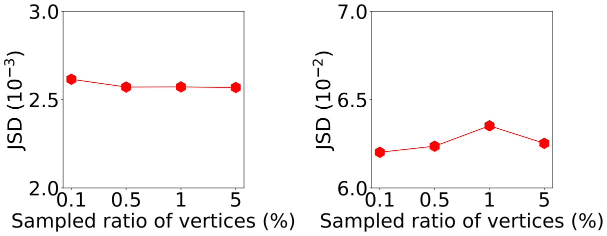


(a) UKDomain.

(b) Friendster.

Figure 4: Single-valued properties.

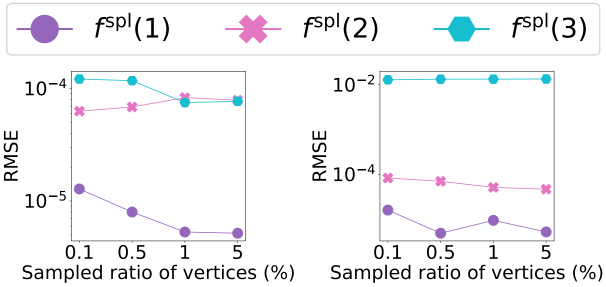
Figures 3 and 4 show the accuracy of the degree-based, clustering coefficient-based, and joint degree-based property estimation on datasets UKDomain and Friendster with different sampling ratios. As the sampling ratio increases, the accuracy of the estimated joint degree distribution is getting better (the value of *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,



(a) UKDomain.

(b) Friendster.

Figure 5: Shortest path distribution.



(a) UKDomain.

(b) Friendster.

Figure 6: The probability of shortest path l, $l = 1, 2, 3$.

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 5 and 6 show the accuracy of the shortest path-based property estimation on datasets Amazon and DBLP with different sampling ratios. Figure 6 shows that the accuracy of the estimated $f^{\text{spl}}(1)$ and $f^{\text{spl}}(3)$ 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%.

References

[1] 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.