Reconstructing degree distribution and triangle counts from edge-sampled graphs

Naomi A. Arnold, Raúl J. Mondragón, and Richard G. Clegg

Queen Mary University of London, London, E1 4NS, UK
n.a.arnold@qmul.ac.uk

Abstract. Often, due to prohibitively large size or to limits to data collecting APIs, it is not possible to work with a complete network dataset and sampling is required. A type of sampling which is consistent with Twitter API restrictions is uniform edge sampling. In this paper, we propose a methodology for the recovery of two fundamental network properties from an edge-sampled network: the degree distribution and the triangle count (we estimate the totals for the network and the counts associated with each edge). We use a Bayesian approach and show a range of methods for constructing a prior which does not require assumptions about the original network. Our approach is tested on two synthetic and two real datasets with diverse degree and triangle count distributions.

Keywords: network reconstruction, Bayesian statistics, sampling

1 Introduction

Analysis of complex networks remains a growing area and network data sets are more and more commonly available. However, some data sets are only a sample of the entire network. For very large networks, it may not be possible to work with complete data because of its size. Additionally, APIs can rate-limit the number of queries, meaning that not all nodes and edges are present [15]. A common example is the Twitter stream API which returns a 1% random sample of all tweets in real-time [21]. In the usual Twitter graph formulation where edges constitute 1:1 replies or retweets, this corresponds to uniform edge sampling of the full Twitter reply/retweet graph. Inferring even simple characteristics such as the true number of nodes or edges from a sample can be nontrivial [12, 6].

In this work we present a methodology for recovering the degree sequence and the triangle sequence (per edge) under a uniform edge-sampling scenario where for an undirected graph G, a sample is constructed by uniformly sampling each edge of G with probability p. First, we build on methods by Ganguly et al [11] who recover the degree distribution from node-sampled networks using a Bayesian approach and we extend this to edge-sampled networks. We address the problem of finding an appropriate prior degree distribution by proposing two different ways to construct a prior. We further extend this Bayesian approach to estimating the edge triangle count (the number of triangles associated with each edge) and the total triangle count.

We find that our Bayesian method outperforms the standard scale-up method at estimating the degree sequence, particularly in small p scenarios where as few as 10% of the edges remain. Moreover, the priors we use do not make any assumptions about the original degree distribution. For estimating the triangle per link count, in 3 out of the 4 network datasets we use, a Poisson prior achieves similar performance to a correct prior.

This paper is structured as follows. First, in section 3 we describe the edge sampling procedure and derive properties of graphs that have been sampled in this way. Then in section 4 we introduce the various estimators used for these properties, with section 5 showing how to construct a prior for the Bayes estimators. Finally in section 6 we present our results on recovering these properties on synthetic and real datasets. We discuss the implications in section 7.

2 Related Work

Sampling of complex networks in general is a well studied problem. One point of interest is how well sampling preserves different properties, such as node rankings in Twitter networks [15], temporal features [1] and scaling properties [14]. These works have aimed also at designing sampling schemes specifically to preserve a given quantity. Other works have used sampling to estimate quantities on graphs that are prohibitively large to work with in their entirety, with a focus on triangle counting [20, 2, 18] or other motifs [13, 5].

Two recent works studied the problem of recovering a network's degree distribution working from a small sample, first posed by Frank [10] in his PhD thesis in 1971. The first by Zhang et al [24] frames it as an inverse problem involving the vector of observed degree counts and a linear operator representing the sampling scheme. The second by Ganguly et al [11] uses a range of estimators for individual vertex degrees in node-sampled networks; simple scale-up estimators, risk minimisation estimators and Bayes posterior estimates. Antunes et al [2] whose work was on sampling methods for estimating the triangle distribution, studied the $n = 1$ sample size problem as restricted access scenario as a case study. Other than this, little attention has been given specifically to these restricted access problems, noted in [24].

A related problem is reconstructing network structure from unreliable or noisy data, such as social networks constructed from reported friendships, which are well known for having missing or spurious edges due to the different interpretations of "friendship" [23]. Young et al [23] address this using a Bayesian approach for finding posterior probabilities for an edge's existence given the measurements obtained. Newman [17] use a Bayesian approach involving the empirical false and true positive rates of observing an edge from the data.

3 Properties of edge sampled graphs

Let $G = (V, E)$ be an undirected simple graph with vertex set $V = \{v_1, \ldots, v_N\}$ and edge set $E = \{e_1, \ldots, e_M\}$. Consider a sampling regime where each edge

$e_l \in E$ is included in the sampled graph with probability $p \in [0,1]$, and each vertex $v_i \in V$ is included if any edge incident to it is included. Denote this sampled graph $G' = (V', E')$, where $V' \subseteq V$ and $E' \subseteq E$. Let the sizes of V' and E' be N' and M' respectively. This is known as *incident subgraph sampling* [14].

3.1 Degree

Let k_i, respectively k_i' denote the degree of a node $v_i \in G$ and G' respectively. Then k_i' follows a binomial distribution $k_i' \sim B(k_i, p)$, with conditional probability given by

$$\mathbb{P}(k_i' = k' | k_i = k) = \binom{k}{k'} p^{k'} (1-p)^{k-k'}, \tag{1}$$

with expectation $\mathbb{E}(k_i'|k_i = k) = kp$ and variance $\mathrm{Var}\,(k_i'|k_i = k) = kp(1-p)$. The probability that node $v_i \in V$ of degree k_i has degree 0 in G' is given by $\mathbb{P}(k_i' = 0) = (1-p)^{k_i}$.

Nodes in G that become isolated as part of the sampling process are invisible to observers of G' and should be considered removed. In this way, let δ_i be the indicator random variable representing the removal of node v_i from G, with probability $(1-p)^{k_i}$. Then, the expected number N_0 of removed nodes from G is given by $\mathbb{E}\left(N_0\right) = \sum_{i=1}^{N} \mathbb{E}\left(\delta_i\right) = \sum_{k \geq 1} (1-p)^k N_k$, where N_k is the number of vertices in G of degree k. This is dependent on the degree distribution; distributions with large numbers of low-degree nodes will experience higher numbers of nodes removed. This brings to mind the friendship paradox [9], where a node incident to a randomly chosen edge will on average have a higher degree than a randomly chosen node. From fig. 1 we see Barabási-Albert [4] networks experience more node removal than Erdős-Rényi [8] networks.

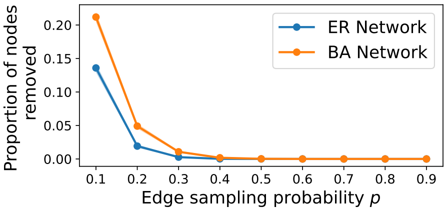


Fig. 1: Proportion of nodes removed in Erdős-Rényi and Barabási-Albert graphs, of size 1000 nodes and 10000 (ER), 9900 (BA) edges (see table 1), using the edge-sampling procedure.

The variance in the number of nodes removed is given by

$$\begin{aligned}\mathrm{Var}\left(N_0\right) = \mathrm{Var}\left(\sum_{i=1}^{N}\delta_i\right) &= \sum_{i=1}^{N}\mathrm{Var}\left(\delta_i\right) + \sum_{1\leq i\neq j\leq N}\mathrm{Cov}\left(\delta_i,\delta_j\right)\\ &= \sum_{k\geq 0}(1-p)^k\left[1-(1-p)^k\right]N_k\\ &\quad + \sum_{k,k'\geq 0}\left[(1-p^{k+k'-1}-(1-p)^{k+k'}\right]N_{k,k'}\end{aligned}$$

where $N_{k,k'}$ is the number of edges connecting vertices of degree k and k'.

3.2 Triangles

Let T_l be the number of triangles in G which include edge $e_l \in E$. Then the number of triangles in G, denoted by T, is given by

$$T = \frac{1}{3}\sum_{e_l\in E}T_l \tag{2}$$

where the factor of $\frac{1}{3}$ is present because each triangle in the sum is counted three times, once for each link.

Let T'_l be the number of triangles which include edge e_l in the sampled graph G', defining $T'_l = 0$ if $e_l \notin E'$. In the case that edge e_l remains in the sampled network, then each triangle that includes e_l will remain in the sampled network if and only if the other two edges remain; this occurs with probability p^2. There are T_l such triangles, so the number of these which remain in the sampled network is binomially distributed with T_l trials and probability p^2. That is,

$$\mathbb{P}(T'_l = t'|T_l = t, e_l \in E') = \binom{t}{t'}p^{2t'}(1-p^2)^{t-t'}. \tag{3}$$

In the case that e_l does not remain in the sampled network, the following holds:

$$\mathbb{P}(T'_l = t'|T_l = t, e_l \notin E') = \delta_{0,t'} \tag{4}$$

where $\delta_{0,t'}$ is the Kronecker delta function, taking the value of 1 if $t' = 0$ and 0 otherwise (since we defined $T'_l = 0$ if $e_l \notin E'$).

We can use the law of total probability to remove the conditioning on e_l from eqs. (3) and (4) and find $\mathbb{P}(T'_l = t')$ as follows,

$$\begin{aligned}\mathbb{P}(T'_l = t'|T_l = t) &= \mathbb{P}(T'_l = t'|T_l = t, e_l \in E')\mathbb{P}(e_l \in E')\\ &\quad + \mathbb{P}(T'_l = t'|T_l = t, e_l \notin E')\mathbb{P}(e_l \notin E')\\ &= p\binom{t}{t'}p^{2t'}(1-p^2)^{t-t'} + \delta_{0,t'}(1-p).\end{aligned} \tag{5}$$

Therefore, the conditional probability mass function for T'_l given T_l is given by eq. (5).

The expected value of T'_l given T_l is given by

$$\begin{aligned}\mathbb{E}(T'_l|T_l = t) &= \sum_{t'=0}^{t} t'\mathbb{P}(T'_l = t'|T_l = t)\\ &= \sum_{t'=0}^{t} t'\left[p\binom{t}{t'}p^{2t'}(1-p^2)^{t-t'} + \delta_{0,t'}(1-p)\right]\\ &= p\sum_{t'=0}^{t} t'\binom{t}{t'}p^{2t'}(1-p^2)^{t-t'} = p \times p^2 t = p^3 t \qquad (6)\end{aligned}$$

where eq. (6) comes from noting that the sum in the lhs precisely evaluates the expected value of a binomial random variable with t trials and probability p^2.

Let T' be a random variable representing the triangle count of G', then

$$\mathbb{E}(T') = \mathbb{E}\left[\frac{1}{3}\sum_{e_l \in E} T'_l\right] = \frac{1}{3}\sum_{e_l \in E} p^3 T_l = p^3 T$$

where T' is the triangle count of the sampled network G'.

The variance of T'_l given T_l is then

$$\mathrm{Var}\,(T'_l|T_l = t) = p^3 t(1 - p^2 + p^2 t - p^3 t). \qquad (7)$$

An argument involving computation of the covariances $\mathrm{Cov}\,(T_j, T_l)$ shows that the variance of the expected total triangle count of G' given the individual triangle counts $T_1, \ldots, T_M$ is given by

$$\begin{aligned}\mathrm{Var}\,(T'|T_1,\ldots,T_M) = \frac{1}{9}\Bigg[&3p^3(1-p^2)T + (p^3 - p^2)\sum_{e_l \in E} T_l^2\\ &+ 6T(p^3 - p^6) + 8k(p^5 - p^6)\Bigg]. \qquad (8)\end{aligned}$$

where k is the number of triangles which share a link. Full derivations of eqs. (7) and (8) can be found in the first author's thesis [3] which also contains derivations for wedge counts and clustering coefficient. An expression for this variance conditioned on total triangle count T not edge triangle counts is given in [20].

The number of triangles per node T_i can be obtained from T_{e_ℓ} from $2T_i = \sum_k T_{e_\ell = (i,k) \in E}$ meaning the estimators of the edge-sampled network can be extended to evaluate vertex statistics eg local transitivity of the nodes $c_i = \sum_k T_{e_\ell=(i,k)\in E}/(k_i(k_i - 1))$.

4 Estimators for the degree sequence and triangle count

The previous showed how the distribution of a quantity X' in a sampled graph G' could be calculated as a conditional probability $P(X' = x'|X = x)$ given the unsampled measurement $X = x$. This section aims to estimate the true network quantity X given its sampled counterpart X'.

4.1 Method of moments estimators

Let X be a random variable associated with a statistic of G and let X' be that statistic on G' with expected value $\mathbb{E}(X') = f(X, p)$. A naive 'scale-up' estimator for X given observed value x' for X' is the solution $\hat{x}$ to the equation $x' = f(\hat{x}, p)$, provided a solution exists. Borrowing the terminology from [11], we will refer to these estimators as *method of moments estimators*(MME).

Degree For a node of degree k'_i in G', the MME for k_i is given by k'_i/p. This is an unbiased estimator with mean $\mathbb{E}\left(k'_i/p\right) = \frac{1}{p}k_i p = k_i$ and variance $\mathrm{Var}\left(k'_i/p\right) = \frac{1}{p}k(1-p)$. Nodes with the lowest possible degree (one) in the sampled graph are estimated as having degree $1/p$ in the unsampled graph so as p decreases, the estimation of low-degree nodes becomes poorer.

Triangle count The expected triangle count $\mathbb{E}(T'_l)$ of edge e_l is $p^3 T_l$. If in addition, e_l remains in G', its expected triangle count is given by $p^2 T_l$. Therefore the MME for T_l is $p^{-3}T'_l$ or $p^{-2}T_l$, without and with the conditioning respectively. Similar to the MME for degree, it provides poor estimates for edges that have a low triangle count, as it disallows any estimates of T_l in the range $(0, 1/p^2)$.

Similarly, an MME estimate proposed by Tsourakakis et al [20] for the total triangle count of a network is $p^{-3}T'$, which has expected value $\mathbb{E}\left(p^{-3}T'\right) = T$. They found that this estimator has variance $\frac{1}{p^6}\left((p^3 - p^6)T + 2k(p^5 - p^6)\right)$, where k is the number of pairs of triangles which share a link.

4.2 Bayes estimator

This estimator relies on Bayes theorem, giving

$$\mathbb{P}(X = x | X' = x') = \frac{\mathbb{P}(X' = x' | X = x)\mathbb{P}(X = x)}{\mathbb{P}(X' = x')}. \tag{9}$$

$\mathbb{P}(X' = x' | X = x)$ is the *likelihood* which is determined by the edge sampling procedure and is known. $P(X = x)$ is the *prior* function which will be denoted by $\pi(x)$; this is in general not known.

A posterior estimate for X given X' can then be given as the expected value

$$\mathbb{E}\left(X | X' = x'\right) = \frac{\sum_x x\mathbb{P}(X' = x' | X = x)\pi(x)}{\mathbb{P}(X' = x')}. \tag{10}$$

The immediate question arises of how to deal with the prior $\pi(x)$, as this may involve making assumptions about the structure of G. This will be discussed case by case for the degree and triangle count.

Degree Using the likelihood function for the degree from eq. (1), a posterior estimate for the degree of node v_i given it has degree k_i' in G' is

$$\mathbb{E}(k_i|k_i' = k') = \frac{\sum_{k=k'}^{\infty} k\binom{k}{k'}(1-p)^k \pi(k)}{\sum_{k=k'}^{\infty} \binom{k}{k'}(1-p)^k \pi(k)} \tag{11}$$

where $\pi(k)$ is a prior for the degree distribution $P(k)$ of G.

Triangle count Using the likelihood function from eq. (3), a posterior estimate for the triangle count of edge e_l in G given it remains in G' is

$$\mathbb{E}(T_l|T_l' = t', e_l \in G') = \frac{\sum_{t=t'}^{\infty} t\binom{t}{t'}(1-p^2)^t \pi(t)}{\sum_{t=0}^{\infty} \binom{t}{t'}(1-p^2)^t \pi(t)} \tag{12}$$

where $\pi(t)$ is a prior for the proportion of edges with triangle count t. [1]

To establish the total triangle count, summing the value of this estimator over the remaining edges in G' and dividing by 3, as in eq. (2), will provide an underestimate for the total triangle count of G, since there are potentially many missing edges in G'. To mitigate this, we scale this factor up to the estimated number of edges in G. That is, our estimate of the total triangle count becomes

$$\hat{T} = \frac{1}{3p} \sum_{e_l \in G'} \hat{T}_l.$$

5 Constructing a prior

The Bayes estimators for the degree and triangle per link sequences require the choice of a prior. In this section, we propose methods for constructing priors.

5.1 Degree distribution

A prior could be obtained from chosen family of distributions such as the Zipf distribution or a power law distribution, but this baked-in assumption may not be desirable. Furthermore, it has been shown that the distribution of a sampled network may not even follow the distribution of the true network [19]. Therefore, we propose two different methods of constructing a prior which do not make assumptions about the degree distribution of the true network.

First, it is possible to estimate the prior using a Monte Carlo method to minimise the ℓ_2 norm of the error. In this approach, we find a degree sequence $\{\kappa_i\}$ which minimises $\min\left(||p\kappa_i - k_i'||_2^2\right)$, with the restrictions that the degree is an integer number and the sum of the degrees is equal to twice the number

[1] In experimental runs, the native binomial functions introduced numerical inaccuracies for large powers of $(1-p)$. Therefore, an equivalent evaluation of binomial probabilities using the log-gamma function and laws of logs was used in practice.

of links. To do this, we start with $\kappa_i = \lfloor k'_i/p \rfloor$. If the sum $\sum_{v_i \in V'} \kappa_i$ of the estimated degrees is not equal to the estimated number of links $\lfloor 2M'/p \rfloor$ then we increment or decrement the degree of nodes chosen uniformly at random until equality holds. Then we rewire links at random for a large number of iterations (15000 in our case), accepting each proposed rewire if it decreases the ℓ_2 error. If $1/p$ is an integer, then the MME $\kappa_i = k'_i/p$ is a global minimum.

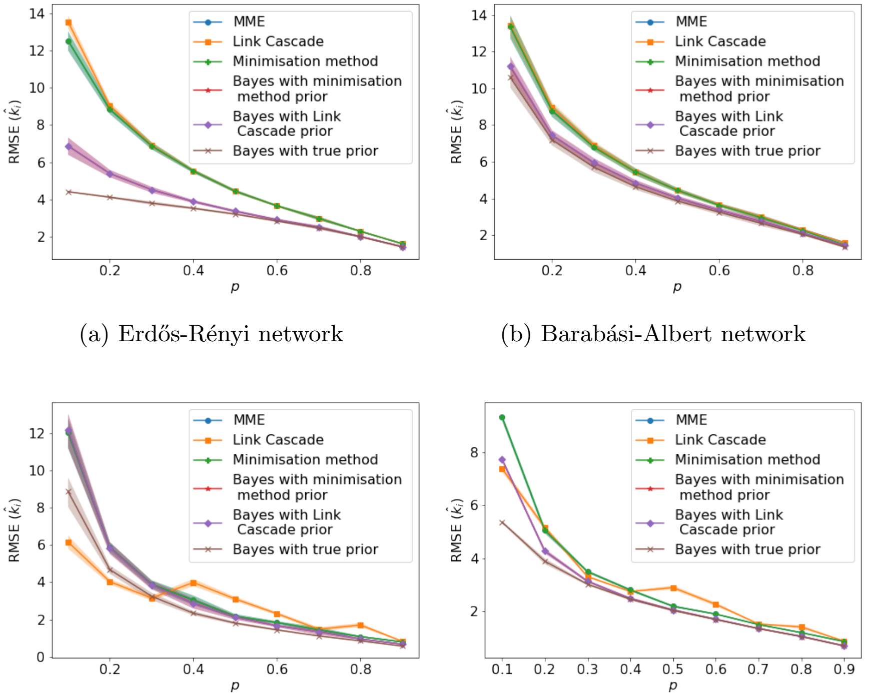
The MME (and hence sometimes the minimisation method) cannot estimate the degree $k_i \approx k'_i/p$ when $k'_i = 0$; that is, the lowest possible estimated degree is $1/p$. If a good estimate for the original number of nodes is known then another prior we for capturing these low degree nodes is constructed by "cascading" links from high degree to low degree nodes. More precisely, as with the minimisation method, we start with an estimated degree sequence $\kappa_i = \lfloor k'_i/p \rfloor$, redistributing links as before if the total estimated degree does not match twice the estimated number of links. Then, we place the nodes in descending order based on their estimated degree, with the knowledge of the original number of nodes in the network being used to append placeholder nodes which would have been removed by the sampling process. Finally, we pick the first occurring node in this list with degree zero, and increment this degree by simultaneously decrementing the degree of the node directly before it. This step is performed iteratively until there are no degree zero nodes. Finally, as a comparison point representing the best possible result achievable with the Bayes method, we use a true prior which is the degree frequencies of the original network as a probability distribution.

5.2 Triangle per link distribution

The two methods for prior construction of the degree distribution do not immediately translate to an analogue for triangles, and little is known about the triangle per edge distribution as a starting place for selecting a prior. As an initial approach therefore, we use a Poisson distribution $\mathrm{Po}(\lambda)$ with $\lambda = 3\hat{T}/M'$, the average number of triangles per link in the MME estimator. As with the degree distribution, we include a result with a true prior as a comparison point.

6 Results

To test the capability of our estimators of degree sequence and triangle count, we consider four different starting networks: an Erdős–Rényi $G(N, M)$ network [8] with $N = 1000$ and $M = 10000$, a Barabási–Albert network [4] of approximately the same size and density, a real collaboration network from authors who submitted to the ArXiV high-energy theoretical physics category [16] (henceforth Hep-Th for brevity) and an Internet autonomous systems topology dataset (henceforth AS) [7]. A quick reference of some summary statistics can be found in table 1. These datasets were chosen to represent a heterogeneous selection of network types. The ER network has a Poisson degree and edge triangle count distribution and an overall low number of triangles for the density of the network. The BA network has a theoretically power law degree distribution, and a low



(a) Erdős-Rényi network

(b) Barabási-Albert network

(c) AS Topology network

(d) ArXiv Hep-Th collaboration network

Fig. 2: Error in estimation of the true degree sequence. Each value is averaged over 10 experiments with the shaded error bars representing standard deviation. The MME overlays the minimisation method in all plots.

triangle count for its density. The Hep-Th and AS networks have a heavy-tailed degree distribution but have very different degree correlations and the Hep-Th has a higher clustering than the AS network. For each of these datasets modelled as a graph G, we take an edge-sampled network G' with edge sampling probability p, for $p = 0.1, 0.2, \ldots, 0.9$ and from this, reconstruct the degree sequences, edge triangle counts and total triangle counts using our estimators. In the degree distribution experiment, we reconstruct the degree k of nodes in V' using our chosen estimators $\hat{k}$, and compute the root mean squared error of the degree sequences as $\mathrm{RMSE}(\hat{k}) = \sqrt{\frac{1}{N'}\sum_{v_i \in V'}(k_i - \hat{k}_i)^2}$. These results are shown in fig. 2, showing the mean and s.d. error over 10 experiments. In all but the AS topology network, the Bayes estimator with true prior has the lowest error, though this is included only to show the best possible result that could be obtained with the Bayes method since the true prior is unknowable. The next approaches that do well at this task are the “link cascade” method and the Bayes estimator using the link cascade as a prior. This method assumes knowledge of the number of nodes in G (i.e. the number of nodes pruned by the edge-sampling) so performs better at estimating low degree nodes. This is particularly evident in fig. 2(c),

Dataset	N	M	$\bar{C}$	ρ	$\bar{T}_l$
Erdős-Rényi	1000	10000	0.019	0.021	0.39
Barabási-Albert	1000	9900	0.063	-0.038	1.91
Cit-Hep-Th Collaborations	5835	13815	0.506	0.185	2.31
AS Topology	11174	23409	0.296	-0.195	2.55

Table 1: Original statistics of network datasets used prior to sampling. Shown is the number of nodes N, number of edges M, average node clustering coefficient $\bar{C}$ [22], degree assortativity ρ and average number of triangles per edge $\bar{T}_l$.

performing better than the Bayes approach with true prior. The Monte Carlo minimisation method on its own in many cases overlays the MME, due to the restriction that the degree sequence is an integer (c.f. section 5).

In the triangle count experiment, we estimate the triangle per edge count $\hat{T}_l$ for edges $e_l \in E'$ and compute the mean squared error as $\mathrm{RMSE}(\hat{\mathbf{T}}) = \left[\frac{1}{M'}\sum_{e_l \in E'}(T_l - \hat{T}_l)^2\right]^{\frac{1}{2}}$. In addition, we estimate the total number of triangles as described in section 4 and calculate the mean squared error over the 10 experiments performed. These are shown in fig. 3 with the triangle per link error on the left column and total triangle error on the right. In all experiments, the Bayes estimator with Poisson prior overlays the MME for total number of triangles; this is because the λ used in the Poisson distribution is the MME estimate of the average number of triangles per link. However, in all but the AS topology, the Poisson prior improves the estimate of triangles per link especially in the small p scenario. In the AS topology dataset, the Poisson is an inappropriate prior, performing poorly even with large sample sizes.

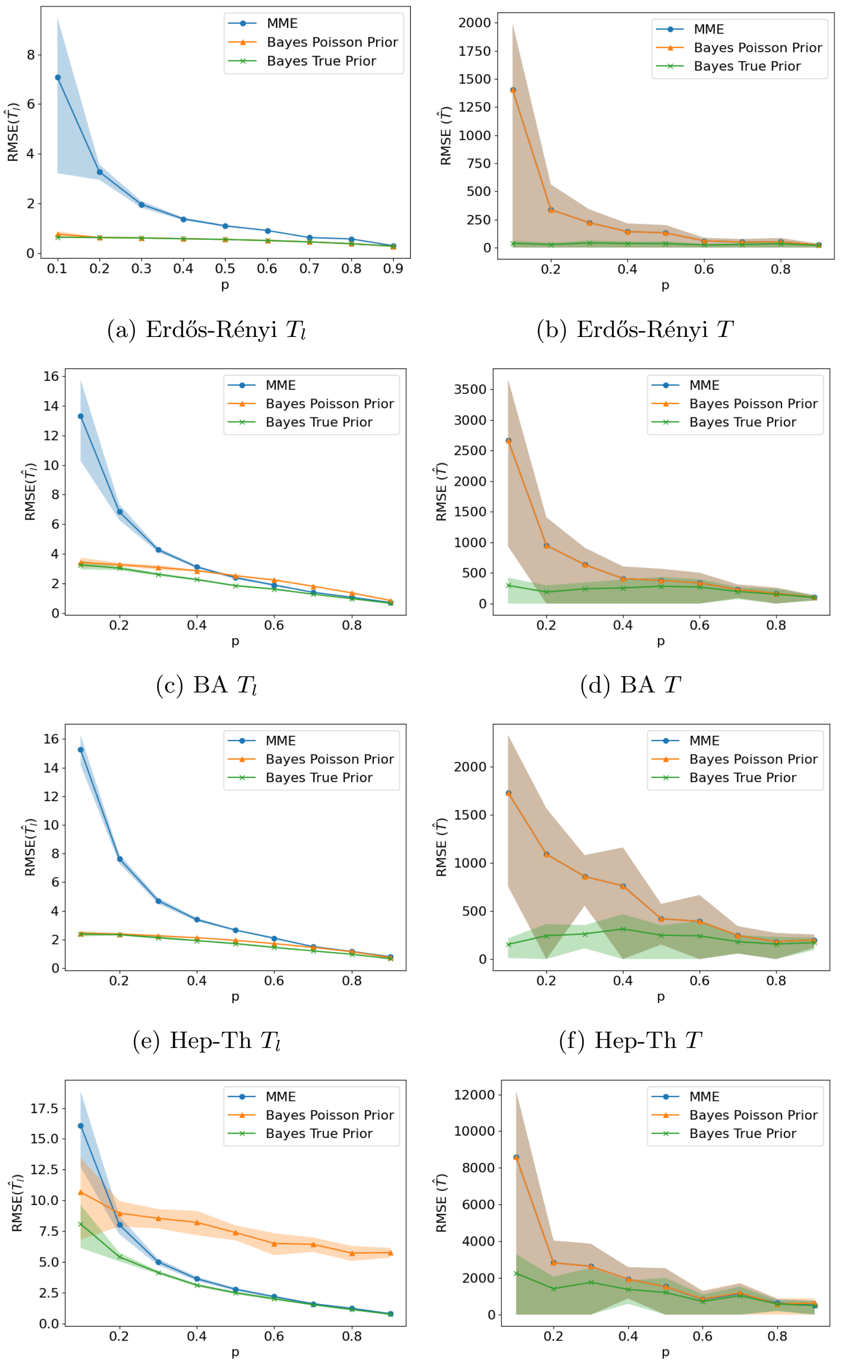
7 Conclusion

This paper provided methods for recovering statistics from networks sampled via uniform edge sampling such as graphs limited to a sample by the Twitter API. Our results show that our Bayesian estimators perform much better than standard approaches on the degree sequence even when the priors were constructed without knowledge of distributions for the original network. For the triangle count per edge, we showed that while the Bayes estimates do not always improve upon the MME for total triangle counts, they provide a markedly better estimate of triangles per link in the small p scenario. However, an inappropriate choice of prior can lead to a bias even when the sample size is large.

Future work will investigate generalising methods we used for constructing a degree distribution prior for constructing a prior for triangle counts per link. One can also consider other sampling regimes and network properties for which a likelihood can be calculated.

References

1. N. K. Ahmed, J. Neville, and R. Kompella. Network sampling: From static to streaming graphs. *ACM Transactions on Knowledge Discovery from Data*, 2013.



(a) Erdős-Rényi T_l

(b) Erdős-Rényi T

(c) BA T_l

(d) BA T

(e) Hep-Th T_l

(f) Hep-Th T

(g) AS Topology T_l

(h) AS Topology T

Fig. 3: Error in estimation of the triangles per link sequence using our different approaches, and total triangles. Each value is averaged over 10 experiments with the shaded error bars representing standard deviation. The MME overlays the Bayes estimator with Poisson prior in the right hand column.

2. N. Antunes, T. Guo, and V. Pipiras. Sampling methods and estimation of triangle count distributions in large networks. *Network Science*, 2021.
3. N. Arnold. *Studying Evolving Complex Networks*. PhD thesis, Queen Mary University of London, 2021.
4. A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 1999.
5. B. B. Bhattacharya, S. Das, and S. Mukherjee. Motif estimation via subgraph sampling: The fourth-moment phenomenon. *The Annals of Statistics*, 50(2):987–1011, 2022.
6. G. Bianconi. Grand canonical ensembles of sparse networks and Bayesian inference. *Entropy*, 2022.
7. Q. Chen, H. Chang, R. Govindan, and S. Jamin. The origin of power laws in internet topologies revisited. In *Proc. IEEE Comp. and Comm. Societies*, 2002.
8. P. Erdős, A. Rényi, et al. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 1960.
9. S. L. Feld. Why your friends have more friends than you do. *American Journal of Sociology*, 1991.
10. O. Frank. *Statistical inference in graphs*. PhD thesis, Foa Repro Stockholm, 1971.
11. A. Ganguly and E. D. Kolaczyk. Estimation of vertex degrees in a sampled network. In *Asilomar Conference on Signals, Systems, and Computers*, 2017.
12. L. Katzir, E. Liberty, and O. Somekh. Estimating sizes of social networks via biased sampling. In *Proc. Int. conf. on World wide web*, 2011.
13. J. M. Klusowski and Y. Wu. Counting motifs with graph sampling. In *Conference On Learning Theory*, pages 1966–2011. PMLR, 2018.
14. J. Leskovec and C. Faloutsos. Sampling from large graphs. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, 2006.
15. F. Morstatter, J. Pfeffer, H. Liu, and K. Carley. Is the sample good enough? comparing data from Twitter's streaming api with Twitter's firehose. In *Proc. of the International AAAI Conference on Web and Social Media*, 2013.
16. M. E. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 2001.
17. M. E. Newman. Network structure from rich but noisy data. *Nature Physics*, 14(6):542–545, 2018.
18. L. D. Stefani, A. Epasto, M. Riondato, and E. Upfal. Triest: Counting local and global triangles in fully dynamic streams with fixed memory size. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2017.
19. M. P. Stumpf, C. Wiuf, and R. M. May. Subnets of scale-free networks are not scale-free: sampling properties of networks. *PNAS*, 2005.
20. C. E. Tsourakakis, U. Kang, G. L. Miller, and C. Faloutsos. Doulion: counting triangles in massive graphs with a coin. In *Proceedings international conference on Knowledge discovery and data mining*, 2009.
21. Twitter. Stream Tweets in real-time: developer documentation. https://developer.twitter.com/en/docs/tutorials/stream-tweets-in-real-time, 2022.
22. D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world'networks. *Nature*, 1998.
23. J.-G. Young, G. T. Cantwell, and M. Newman. Bayesian inference of network structure from unreliable data. *Journal of Complex Networks*, 2020.
24. Y. Zhang, E. D. Kolaczyk, and B. D. Spencer. Estimating network degree distributions under sampling: An inverse problem, with applications to monitoring social media networks. *The Annals of Applied Statistics*, 2015.