Sampling A Large Network: How Small Can My Sample Be?

Group 36: Dan Frank, Zhiheng Huang, Alvin Chyan

December 10, 2012

1 Introduction & Objective

The underlying problem we would like to investigate is that of a researcher who has developed a computationally intensive algorithm he wishes to run on a large graph. The algorithm is too slow to run on the large graph so instead, he would like to sample the large graph to get a smaller *similar* graph where he can feasibly run the algorithm. By *similar* we mean that the smaller graph preserves the properties of the original graph as much as possible. This is the scale-down goal of [1] which is also discussed in [2] from which we take our primary inspiration for this project, whereas [3] and [4] discuss uniform sampling of nodes via exploration and sampling from dynamically evolving graphs, respectively.

To build on the above work we would like to pose the problem of computing on sampled graphs explicitly in the framework of statistics. We will view the benchmark quantities referenced in [1] and [3] as random variables estimating the true quantities for the large graph, which depend on the structure of the original graph, the sampling technique, and the sampling percentage (i.e. the size of the sampled graph as a percentage of the original graph). Specifically, we will analyze the degree distribution (DD) and clustering coefficient (CC) distribution of our sampled graphs. We take our error in estimating these quantities to be the Komolgorov-Smirnov D-statistic between the true distribution (unsampled graph) and the sampled distribution, following [1]. Viewed this way the D-statistic we observe under sampling is a random variable following an unknown distribution (under infinite graph samples). The practical outcome of this project will be to investigate for each benchmark quantity the relation between sampling percentage and the spread of its D-statistic error (previously uninvestigated), as well as the expected value of the D-statistic error (previously investigated).

The spread of our error (the D-statistic in this case) will allow us to create approximate confidence intervals for the D-statistic of a sampled graph in cases when we cannot compute it explicitly. Such an analysis will be helpful even in the presence of highly accurate sampling schemes. With knowledge of this relationship, if a researcher were to say that his algorithm was able to handle a sampled graph with maximum Degree Distribution D-statistic value of $d_{\max}$, then we can say that with 95% confidence a α percent sample can be used and the D-statistic will be below $d_{\max}$ (i.e. when the true D-statistic is not computable).

2 Previous Work

2.1 Sampling from Large Graphs, Leskovec et al. 2006 [1]

Summary Sampling from Large Graphs introduces two goals of graph sampling. First, we may wish to take a large graph and get a scaled down version of it; second, we may wish to take a temporally evolving graph and revert to a previous version of the graph. The authors propose a set of 9 benchmark metrics for the scale-down goal and 5 for the back-in-time goal. The metrics are typically distributions themselves and are compared to the distribution in the original graph using the Kolmogorov Smirnov D-statistic.

The paper introduces 3 classes of sampling algorithms. (1) Node based: uniform, PageRank weighted, degree weighted; (2) Edge based: uniform, uniform node edge (pick a node at random and then pick one of its edges at random), hybrid of uniform edge and uniform node edge; and (3) Exploration based: random walk, random jump, forest fire, and random node neighbor. For each sampling technique, the D-statistic is calculated for all of the metrics and averaged to produce one score for each sampling procedure. The findings were that random walk performed the best for the scale-down goal, while the forest fire technique performed better for the back-in-time goal.

2.2 Statistical Properties of Sampled Networks, Lee et al. 2009 [2]

Summary Statistical Properties of Sampled Networks also considers the problem of estimating various quantities of interest when we only have a sample from a large graph, specifically degree distribution, average path length(APL), betweenness centrality distribution(BC), assortativity, and clustering coefficient(CC). It considers these quantities on the Barabasi-Albert model as well as three real networks from various areas: protein interaction network(PIN), Internet at the autonomous systems level(AS) and e-print archive coauthorship network(arxiv.org). The sampling methods analyzed are uniform node sampling, uniform edge sampling, and snowball sampling. To evaluate the accuracy of degree distribution estimation, the authors assume that all these networks and their samples follow approximate power law distributions and calculate the power law exponents using a maximum likelihood estimate. This paper also derives an analytical expression for the degree distribution of a sampled graph under uniform node sampling or uniform edge sampling and shows that experimentally the analytical expression holds true. The power law exponents from the sampled graphs are compared against ground truth computed from the original graph in order to understand how the error changes with respect to sampling size. It is found that snowball sampling tends to underestimate this power law exponent due to bias towards high degree nodes. For APL, its observed that snowball underestimates it by introducing many edges but node sampling and edge sampling overestimates APL, but only heuristic arguments are given for why this might be the case. A similar analysis is done for BC and hypotheses are given for why the bias behaves differently under different graphs. Assortativity is then analyzed where we have an analytical expression under link sampling and empirically we see that snowball sampling underestimates assortativity and node/link sampling seem to estimate the assortativity well even for low sampling proportions. Finally, CC is estimated and we see that edge sampling reduces the CC.

2.3 Walking in Facebook: A Case Study of Unbiased Sampling of OSNs, Gjoka et al. 2010 [3]

Summary Traditional crawling techniques such as Breadth First Search(BFS) and Random Walk(RW) are known to be biased towards high degree nodes. The authors of this paper introduce two unbiased methods to uniformly sample nodes from the online social network Facebook, namely Re-weighted Random Walk(RWRW) and Metropolis-Hastings Random Walk(MHRW). The first method re-weights the random walk transition probabilities by estimating the bias using Markov Chain analysis. The second method also builds on random walk sampling. It applies Metropolis-Hastings algorithm to derive transition probability between nodes so that the resulting stationary distribution is uniform. Specifically, when we are about to transit to a node, we throw a dice; with some chance, we move forward or stay in the same node. Generally the probability of moving from a low degree node to a high degree one is very low. Finally, this paper chooses to use Geweke Diagnostic and Gelman-Rubin Diagnostic to decide burn-in period and total running time.

For the experiments, the authors were able to obtain a uniform sampling of 957K Facebook users. They evaluated the sampling method based on 3 metrics, degree distribution, assortativity and clustering coefficient. The conclusion is that RWRW and MHRW performs remarkably well compared to BFS or RW.

3 Data Used & Sampling Procedures

In order to get results that can be generalized to many graphs, we have analyzed the benchmark quantities against a couple reasonably sized undirected networks from SNAP[5], including Enron-email Network(Enron, 36k nodes), Condense Matter collaboration network(CondMat, 23k nodes) and DBLP collaboration network(DBLP, 317k nodes). We chose the networks to be sufficiently large such that we can down-sample significantly while maintaining some hope of estimating our benchmark quantities. At the same time, we kept the networks small enough that we can calculate our benchmark properties many times for different sampling percentages in order to estimate the distribution of the D-statistic. The Enron-email network is a communication network and features a long-tail degree distribution, the highest node degree is above 1000. The latter two networks are collaboration networks and share similar degree distribution, with the largest degree around 300. We note that all nodes in the DBLP network belong to one connected component and hence we do not analyze the SCC benchmark for DBLP. Besides these real world networks, we also explored generated graphs such as Erdos-Renyi (Gnm) network and Preferential Attachment(PA) network. Note that we generate our Gnm and PA network using SNAP functions and they are of size similar to the Enron-email network, i.e. 36692 nodes and 367662 edges for Gnm; 36692 nodes and each node introduces 10 edges for PA.

To keep the size of this project reasonable we only plan to perform our analysis with the Random Walk sampling technique identified as one of the best sampling algorithms by [1]. Specifically, our RW picks a node uniformly at random as start point and begins a sequence. At each step, with 0.85 probability it selects one node among neighbors of the current node with equal probability and moves to that node. If the neighboring node or the corresponding edge does not exist in the sample graph, they will be added to the graph; with 0.15 probability, we will fly back to the starting point. This ensures that the neighborhood of a selected node could be sufficiently explored. The higher the fly back probability, the more similar RW is to Breadth First Seach(BFS). To avoid being stuck in a node or loop, as in [1], we implemented a check-and-fly mechanism. We defined a period T and an expected expanding size M over the period. After each T iteration we check whether we have visited more than M new nodes. If not, we repick the starting node and begin a new sequence.

Although our project mainly views the expense of sampling in terms of the sampling percentage, we also note here that due to the properties of Random Walk, the cost of increasing sample size is not the same over all percentages. Concretely, figure 1 shows that as we gets to higher sampling percentage, if we want to increase the sampling percent, say by 1 percent, the number of actual random walk steps it takes increase exponentially for real world graphs. On the other hand, this cost remains almost constant for Gnm network and Preferential Attachment Network.

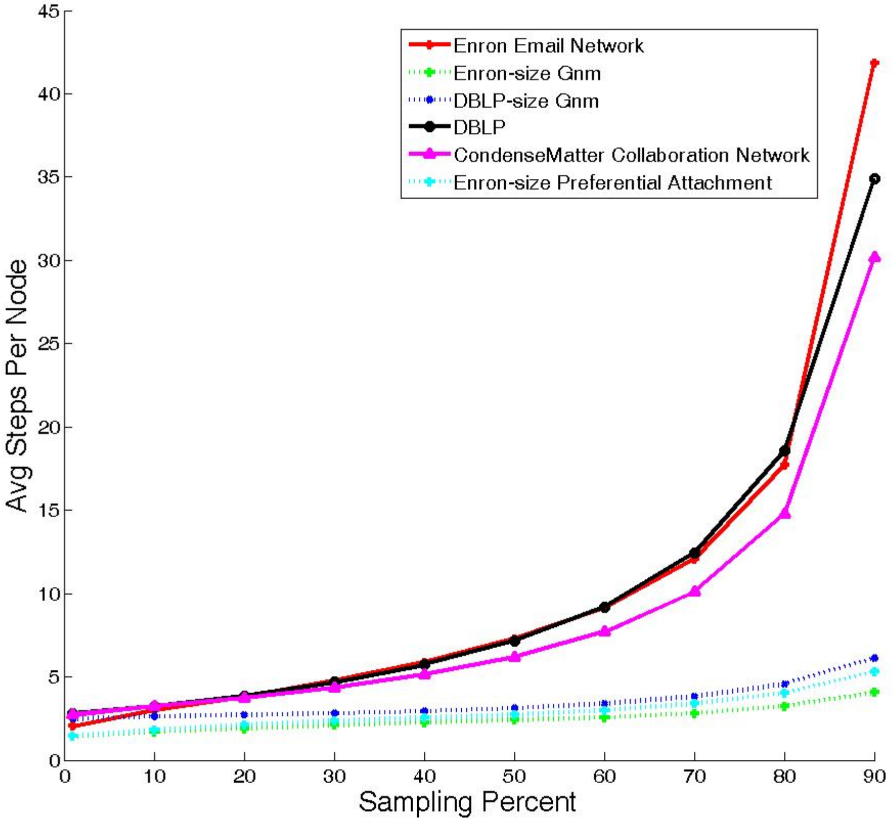


Figure 1: Steps taken to sample one new node as a function of sampling percent

4 Mathematical Background

To evaluate how close the sampled graph is to the original graph, we need to compare our benchmark metrics on the sampled graphs to the true metric on the unsampled graph, namely clustering coefficient and degree distribution. Following [1] we define our error as the D-statistic, which can be computed as follows. Given two cumulative distribution functions P and Q defined over $\mathcal{S}$

$$Dstat(P, Q) = \max_{x \in \mathcal{S}}(|(P(x) - Q(x))|) \tag{1}$$

The D-statistic captures the largest gap between the cumulative probability distributions P and Q and serves as a measure of distance between two distributions. In our case, the distributions are our benchmark metrics. Previous work has concentrated on how different sampling algorithms might influence the expected value of the D-statistic but paid little attention to its spread. We aim to explore the relationship between sampling percentage and the full distribution of the D-statistic. For example it may be the case that the expected value of the D-statistic varies linearly with sampling percentage whereas the spread of the D-statistic increases more quickly than this. Each time we sample our graph, we view the resulting value of the D-statistic of our benchmark quantities as an instance of a random variable which has some unknown distribution. Since we are instead interested in using *one* sampled graph as 'representative' of the original graph, we need to understand the D-statistic's spread.

Classically, we analyze the D-statistic's variance using several graph samples (at each sampling percentage). This analysis should give us an intuitive idea of how the spread of the D-statistic is changing with sampling percentage. Further, we examine the relationship between a given $d_{\max}$ D-statistic tolerance (mentioned above) and the minimum sampling percentage possible such that the D-statistic is below $d_{\max}$ with 95% probability. This is the explicit relationship that we aimed to find in our objective. Since sampling at all percentages is infeasible, of the sampling percentages actually computed we pick the smallest one such that the probability of the D-statistic being greater than $d_{\max}$ is less than 95%. This is why figure 3 and 5 have staircase patterns. We do this by computing the percentage of D-statistics below $d_{\max}$ and find the smallest sampling percentage that has 95% of its D-statistics below $d_{\max}$. That is,

$$\alpha_{\min} = \arg\min_{\alpha} \frac{|\{D_{i,\alpha} \leq d_{\max}\}|}{n} \geq .95 \tag{2}$$

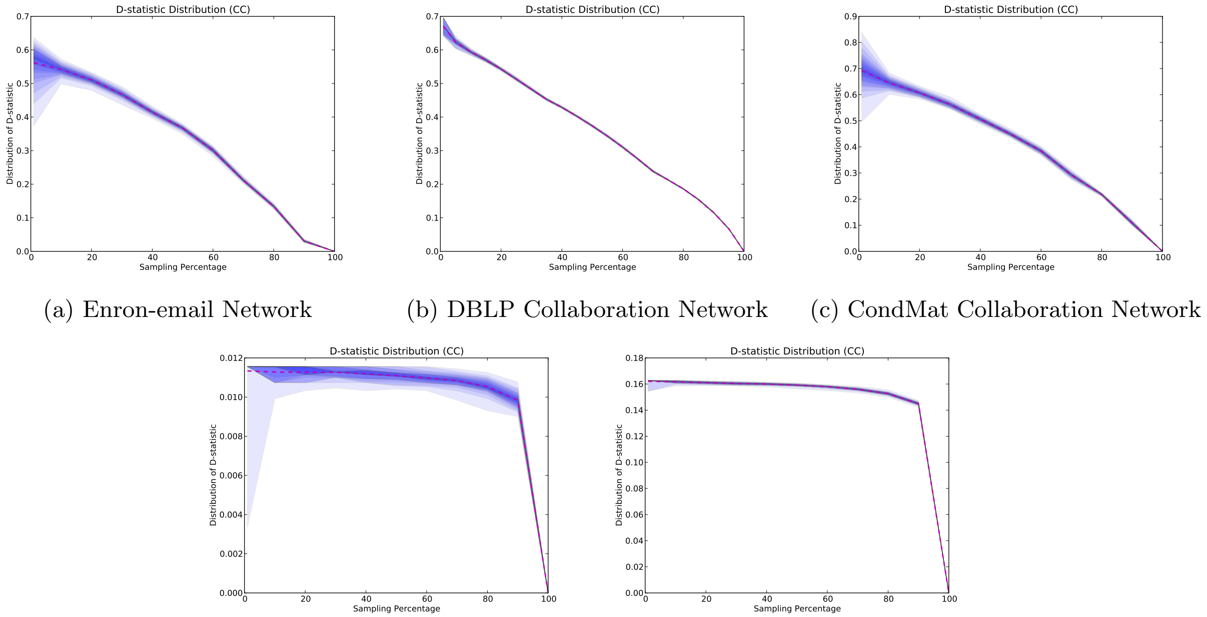
where $D_{i,\alpha}$ are different D-statistic values at sampling percentage α and n is the total number of samples at that sampling percentage.

Of course, we aknowledge that it is entirely possible that the relationship between D-statistic and sampling size might depend strongly on the structure of the original graph. In fact this is what we are investigating. If this is true we would first have to determine which 'type' of graph we are using in order to use the corresponding model. See secion 5 for more discussion.

5 Analysis

We ran RW on all 5 graphs(Enron-email [6], DBLP [7], CondMat [8], Gnm and PA) with sampling percentage $1\%, 10\%, \ldots 90\%$ for each. For each sampling percentage, RW was run 100 times so that we could examine the distribution of the D-statistic. Figure 2 and figure 4 show the D-statistic distribution for clustering coefficient and degree distribution, respectively as a function of sampling percentage. The colors represent deciles of the D-statistic distribution and the dotted line represents the mean. As expected, the D-statistic increases as the sampling percentage decreases, and we see that the increase is roughly linear. The variance in our real world networks, on the other hand, stays mostly constant until very low sampling percentages where we see sharp increases. This suggests that past a certain point, even if we were able to keep the expected value of the D-statistic low, sampling a representative graph would be impossible. The distributions for Gnm and PA show other behavior described below.

5.1 Clustering Coefficient



(a) Enron-email Network

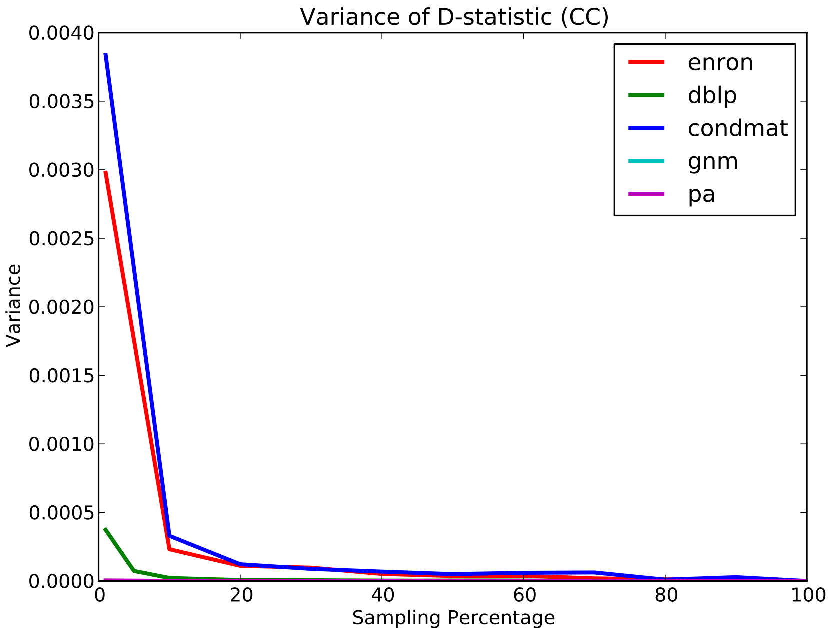
(b) DBLP Collaboration Network

(c) CondMat Collaboration Network

(d) Gnm Network

(e) PA Network

Figure 2: D-Statistic Spread Graphs for Clustering Coefficient



(a) D-Statistic Variance for Clustering Coefficient

For the Gnm graph, clustering coefficients are low and roughly uniform. Even with a small sample of nodes, clustering coefficients will still be low, since the probability of selecting many nodes from the same cluster is small. That's why it is possible to use a small sample size and maintain a low D-statistic. As for the preferential attachment graph, even though there will be some high degree nodes, the neighbors of a high degree node are not necessarily connected to each other, so the clustering coefficient is still small, leading to a similar plot as the Gnm graph. These plots are in sharp contrast to that of the real world graphs, which have high clustering. Hence in the real world graph, we witness a linear relationship between an accepted D-statistic threshold and the required sampling size whereas in Gnm and PA we witness sharp cutoffs.

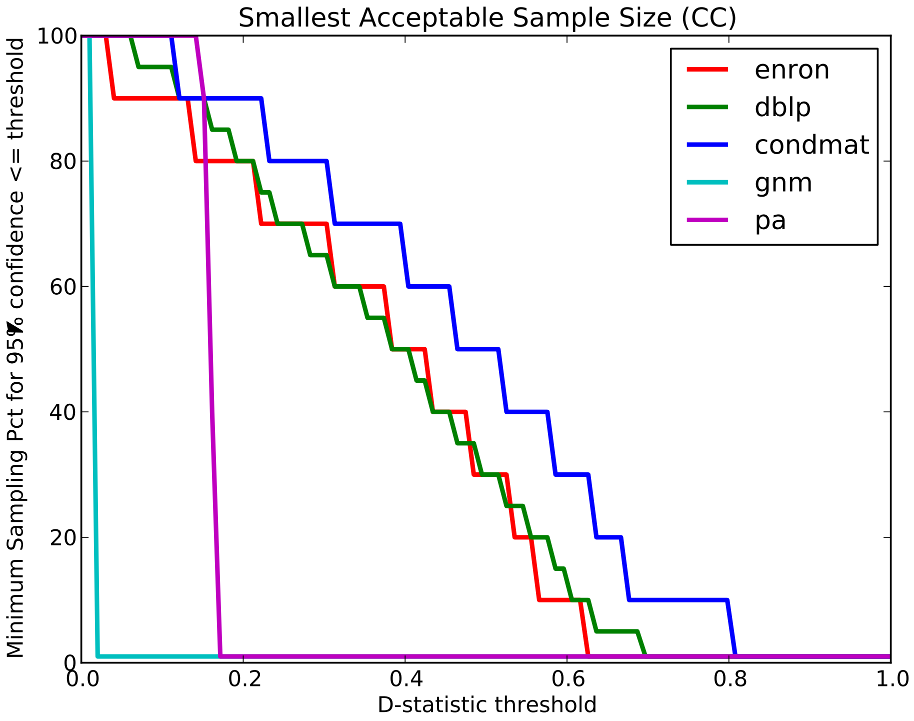
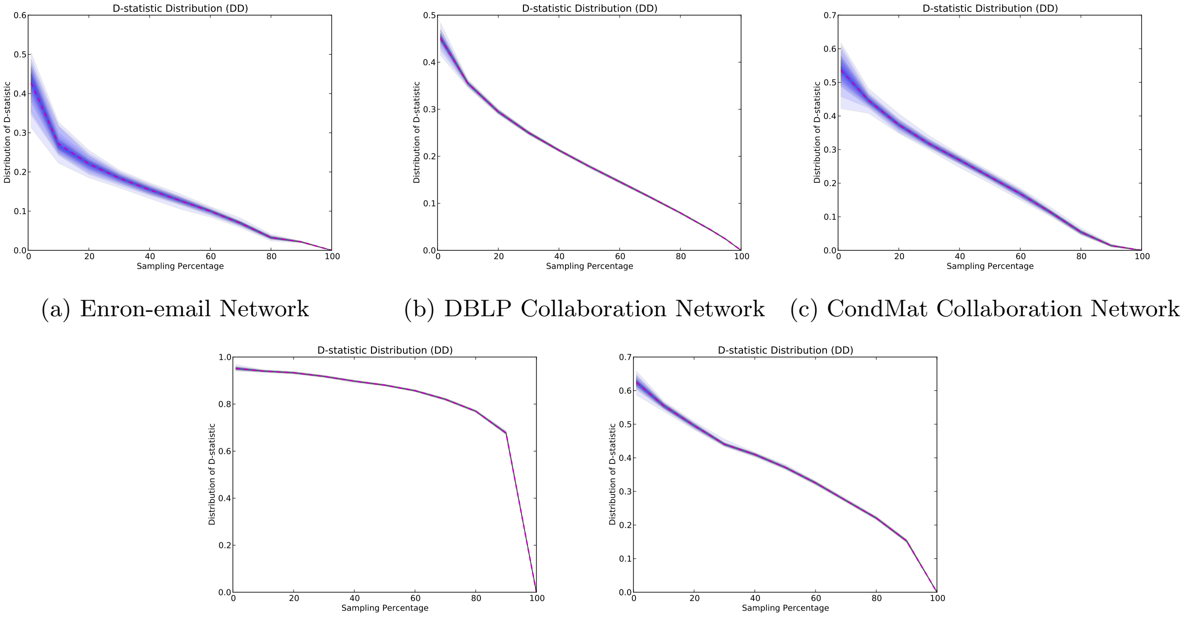


Figure 3: Minimum Sampling Percent v.s. D-Statistics Threshold for Clustering Coefficient

5.2 Degree-Distribution



(a) Enron-email Network

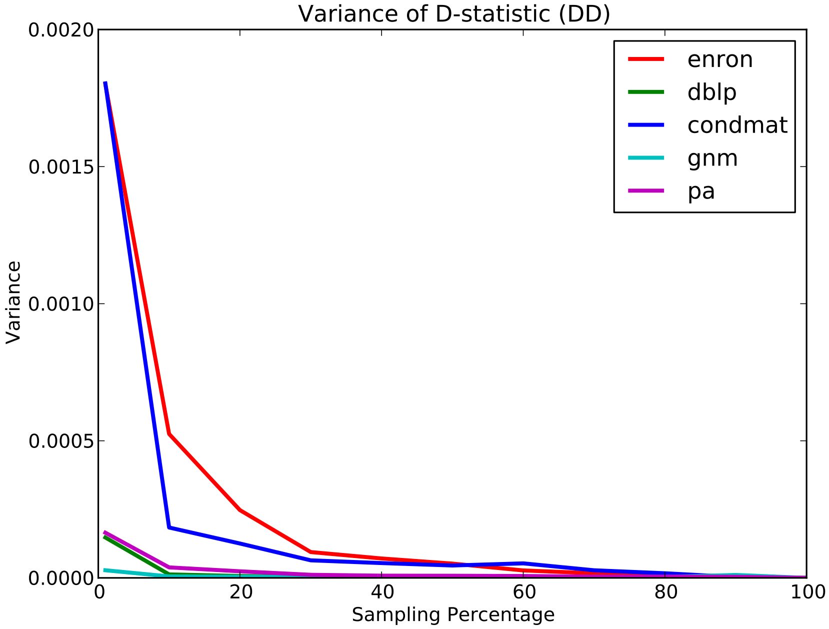
(b) DBLP Collaboration Network

(c) CondMat Collaboration Network

(d) Gnm Network

(e) PA Network

Figure 4: D-Statistic Spread Graphs for Degree Distribution



(a) D-Statistic Variance for Degree Distribution

Because there is no central node nor high clustering coefficient in a Gnm graph, a random walk can stray far from the starting node instead of exploring one neighborhood in detail. This makes it difficult to sample Gnm and maintain the integrity of the degree distribution. The plot reflects this observation, in that a high sampling percentage is required even for a high D-statistic threshold. The preferential attachment graph, in comparison, does have central nodes, so one neighborhood can be more adequately explored, leading to a more accurate degree distribution. This is reflected in the fact that the PA curve is very similar to our real world graphs. Our real world graphs have high clustering and some central nodes, so a random walk will have the best chance of fully exploring a neighborhood to reproduce the degree distribution with higher fidelity.

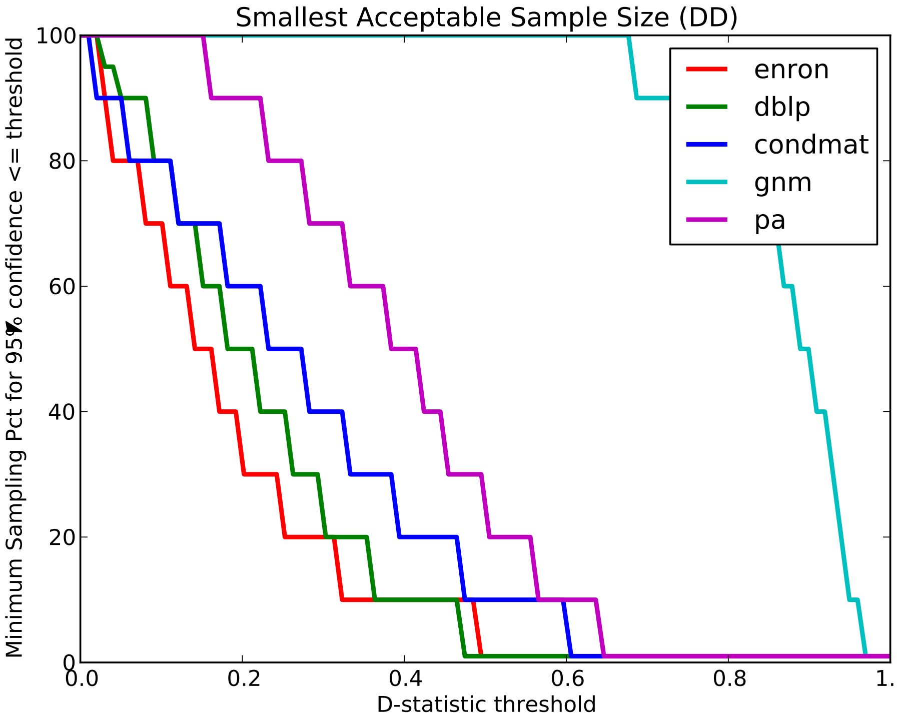


Figure 5: Minimum Sampling Percent v.s. D-Statistics Threshold for Degree Distribution

6 Conclusion & Further Work

From our experiments with a Gnm graph, preferential attachment graph, and several real world graphs, we find that the minimum sampling size required to accurately capture a graph metric does indeed depend heavily on the graph structure and the metric. Our generated graphs above show particular behavior which contrasts with the behavior of real world graphs. The real world graphs show linear increases with D-statistic expected value as sampling percentage decreases, as expected. However, our real world graphs tend to not have strong relationship with D-statistic spread until very low sampling percentages, at which point the variance increases rapidly suggesting that we have reached a sampling percentage below which representative graph sampling is not possible. Empirically, this breaking point seems to fall at around 15%.

Based on our comparison with real world and generated graphs, we have shown that graph structure strongly affects the relationship between sampling size and required D-statistic thresholds (or more generally D-statistic distributions). Further studies might try to examine the critical factors affecting this relationship or in addition examine other sampling techniques.

References

[1] Leskovec, Jure and Faloutsos, Christos. *Sampling from large graphs.* ACM SIGKDD 2006.

[2] Sang Hoon Lee, Pan-Jun Kim, Hawoong Jeong. *Statistical properties of sampled networks.*Phys. Rev. E 73, 016102 (2006)

[3] Minas Gjoka, Maciej Kurant, Carter T Butts, Athina Markopoulou. *Walking in Facebook: A Case Study of Unbiased Sampling of OSNs.*IEEE INFOCOM '10, San Diego, March 2010.

[4] Daniel Stutzbach and Reza Rejaie. *Sampling Techniques for Large, Dynamic Graphs.*in Global Internet Symposium 2006.

[5] Jure Leskovec*Stanford Network Analysis Platform (SNAP)*http://snap.stanford.edu/snap/index.html

[6] J. Leskovec, K. Lang, A. Dasgupta, M. Mahoney. *Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters.* Internet Mathematics 6(1) 29–123, 2009.

[7] J. Yang and J. Leskovec. *Defining and Evaluating Network Communities based on Ground-truth.* ICDM, 2012.

[8] J. Leskovec, J. Kleinberg and C. Faloutsos. *Graph Evolution: Densification and Shrinking Diameters.* ACM Transactions on Knowledge Discovery from Data (ACM TKDD), 1(1), 2007.