Accuracy and Sampling Trade-offs for Inferring Internet Router Graph

Çiğdem Gündüz
Department of Computer Science
Rensselaer Polytechnic Institute, NY 12187
gunduz@cs.rpi.edu

Bülent Yener
Department of Computer Science
Rensselaer Polytechnic Institute, NY 12187
yener@cs.rpi.edu

***Abstract*—**

There has been an increasing interest on construction of router-level Internet graphs, using *traceroute* like measurement primitives. Furthermore, many metrics reflecting the properties of the Internet graph have been defined based on these measurements. However, many important questions remain: How much measurements (sampling) one needs to conduct over the Internet to get a good estimation of its properties? How accurately can these properties be computed? What is the impact of sampling techniques on the computation of these properties?

This paper shows that most of the metrics are *evasive*: their exact values cannot be determined without visiting all links in the Internet graph. This indicates a fundamental difficulty: due to its size and dynamics, a complete Internet graph cannot be obtained, thus the metrics are computed under incomplete information.

The contribution of this paper is three-fold: First it provides a *meta-metric* called (γ, σ)-evasiveness to determine if a metric can be estimated with at least 1-σ accuracy by sampling γ percentage of data. Second, it demonstrates an important relationship between the *kurtosis* of measured data and the σ value of sampling. Third, it provides a novel technique to compare different metrics and data collection methods by using the kurtosis of data and their σ values.

***Index Terms*— Graph theory, Internet measurements, statistics, topology inference.**

I. Introduction

Obtaining an accurate map of the Internet topology is a time consuming and non-trivial task. Inferring the Internet router graph is based on probing techniques such as traceroute. This graph has routers as its vertices (nodes) and there is an edge between a pair of nodes if they are one IP hop apart. Due to the immense size of the Internet, only a limited amount of data can be collected within a short period of time. Furthermore due to its dynamics, the data may become outdated in a short time. Therefore snapshots of the Internet in different time periods are taken and metrics are computed on them. As the size of the snapshots determines the values of metrics, one must determine (i) at least how much data must be sampled to draw meaningful conclusions, and (ii) what is the sampling technique to be used.

It is obvious that the metrics converge to their actual values as the size of data increases. However the complexity of collecting data from the Internet, the memory requirements to store this data, and the computational and space complexity of the metric evaluation also increase with the data size. For these reasons it is important to determine minimum amount of data that must be collected so that results will be generalized to the whole Internet.

In [1], Floyd and Paxson discuss the difficulties in simulating the Internet. Recently, in [2] authors showed that the measurement techniques based on traceroute-like probing may create a biased view of the Internet graph. This work continues with explaining the difficulties of measuring the Internet from a novel perspective. We show that most of the metrics of the Internet router graph are *evasive*: their exact values cannot be determined without visiting all edges in the Internet graph. There are several contributions of this paper.

First, it introduces a *meta-metric* called (γ, σ)-evasiveness, to answer the following questions: (i) can a metric be estimated with 1-σ accuracy by using only γ percentage of the total data? (ii) how do the values of γ and σ change for different metrics? (iii) is the sampling method important in this task?

Clearly a metric with high γ and σ cannot be used to distinguish between different topology generators since it is evasive and its value cannot be computed accurately. Thus, even if that metric captures an important property of the Internet, it cannot be useful because of its evasiveness. In contrast, metrics with low γ and low σ are better metrics since they reflect more "accurate" properties of the Internet. Second, it demonstrates a strong relationship between the kurtosis of data and the accuracy of a metric computed on them. This relationship determines whether a metric is computed accurately with only a small part of

data or not as well as it provides a measure to decide on when a data collection technique has a greater impact on the correctness of metrics. Third, the paper proposes to use σ and kurtosis values in the comparison of different metrics and different data collection methods. Finally, it shows that the values of metrics depend strongly on the type of sampling method deployed for measurements.

The rest of the paper is organized as follows: We define (γ, σ)-evasiveness in Section II. In Section II-B, we explore the relationship between (γ, σ)-evasiveness and property testing. The metrics reflecting different properties of the Internet are explained in Section III. We explain our experiments and give results in Sections IV and V. Section VI concludes the work and discusses possible future work.

II. Theoretical Foundations

This work uses *graph evasiveness* to evaluate different metrics and sampling techniques used for internet measurements. Graph evasiveness (also known as elusiveness) considers the following problem. Given an input graph G suppose we are to decide if G has a certain property P by asking, to an oracle $\mathcal{O}$, whether or not edge (u, v) belongs to G. In a graph with N nodes there are at most $N(N-1)/2$ edges that can be used as a query to the oracle $\mathcal{O}$. If the decision about P can only be made using *exactly* $N(N-1)/2$ queries then the property P is said to be *evasive*. In other words if P can only be decided by checking all the edges of G then it is an evasive property. Thus, evasiveness of graphs is used for determining the worst case complexity of computing some graph properties [4].

It is conjectured by Karp [5] and proven in [6] that every nontrivial monotone graph property is evasive. A property is *monotone* if insertion of new edges to a graph with property P does not destroy the property. P is *nontrivial* if it holds for some graphs with N nodes and it does not hold for some other with the same number of nodes. Planarity, 2-connectivity, connectivity are examples of such properties.

A. Does Sampling Size Matter? - (γ, σ)-Evasiveness

This work introduces a new concept called (γ, σ)-evasiveness by relaxing the strict or exact evasiveness definition as follows:

Definition 1: Given a graph G with vertex set V and edge set E, a property P is called (γ, σ)-evasive if it can be "computed" by making at least $\lceil \gamma |E| \rceil$ queries with an error margin of at most $\pm\sigma$ for $0 \leq \gamma \leq 1$ and $\sigma \geq 0$.

Note that in this definition, the graph is given but computation of a property has (γ, σ)-evasiveness. The (γ, σ)-evasiveness is an approximation to the exact evasiveness since in exact evasiveness (i) the next query can be chosen based on the current answer from the oracle, and (ii) there is no margin of error - the property P exists or not.

In (γ, σ)-evasiveness a property (metric) P is called monotone if for $\gamma' > \gamma$ property P still holds. In other words additional information will not change the fact that metric P can be computed with an error margin of at most $\pm\sigma$. Similarly, if the variance of a metric P, which is computed over say K graphs $G = (V, E)$ with $\gamma|E|$ queries with an error margin of at most $\pm\sigma$, is non-zero then P is said nontrivial. [1]

We show in this work that (γ, σ)-evasiveness can be used to evaluate the metrics as well as to compare the graph generators as reported in [3].

B. Property Testing and (γ, σ)-Evasiveness

It follows from the above description that (γ, σ)-evasiveness establishes a trade off between accuracy and cost. Thus, a relevant study to (γ, σ)-evasiveness is *property testing*.

Property testing concerns with the problem of determining whether an unknown function has a particular property or it is ϵ "far away" from it. Property testing of functions were first proposed in [7] and testing graph properties was initiated in [8]. For example, testing ρ-clique property for a graph with N nodes requires determining whether or not the graph has a clique of size ρN for $0 < \rho < 1$, or the graph is ϵ-far from the class of N-vertex graphs with a clique of size ρN. The tester will accept a property with confidence δ and will reject it also with probability δ. The parameter ϵ is a *distance* metric and depends on the functional description of the graph. For example, in an adjacency matrix representation, distance is the ratio of entries that are not the same in both matrices to total number of entries. A convenient representation for a simple graph is to use a symmetric Boolean function which takes a pair of vertices and outputs 1 if there is an edge between them, 0 otherwise [8]. Remark that computation of the function is similar to asking a query to the oracle $\mathcal{O}$ in graph evasiveness.

Property testing starts by choosing a small subset of nodes randomly and uniformly as a *sample* and checks how close this subset is having the property. If it requires more than ϵ operations to obtain the property, the test fails. There are three measures that determines the complexity of a property testing: (i) number of samples, (ii) number of queries and (iii) running time of the tester.

We note several differences and similarities between (γ, σ)-evasiveness and property testing. First, it is the

[1] For the rest of this paper evasiveness will refer to (γ, σ)-evasiveness unless a distinction is made explicitly.

type of the properties examined. The (γ, σ)-evasiveness considers the statistics of the distribution of certain properties. For example we examine the degree distribution and define a property to determine if the average node degree is δ or if the clustering coefficient of the graph is C which is also computed as an average over all the nodes. In contrast, property testing can be used to answer queries such as is the graph connected, is it planar, does it have a cut of size at least ρN^2 edges, does it have a clique of size ρN? However, similarities exists between (γ, σ)-evasiveness and property testing with respect to some local metrics/properties.

For example, if the testing for ρ-clique property is affirmative with accuracy δ then, with confidence σ, there is at least a node u with degree $d \leq \rho + \eta$ such that the clustering coefficient of u is at most 1 and at least $2(\rho + \eta)/(\rho + \eta)^2$ for $\eta \geq 0$.

In general, we conjecture that tests for first order graph properties (i.e., properties that are expressed using quantifiers on the vertices) may be relevant to measuring local metrics in the Internet.

Second, the approximation parameter ϵ in property testing is fundamentally different from the approximation parameters σ and γ in (γ, σ)-evasiveness. Parameter σ does not bound the number of operations needed to ensure the property and γ is not an accuracy measure. In property testing, ϵ operations ensures the property which enables property testing algorithms to get good approximations to NP hard problems.

Finally, testing graph properties by queries with uniform distribution is not trivial in the Internet. Some of the metrics (properties) of the Internet graph have long-tailed distributions in which rare and dominant events that reside in the tail. As a result, a random walk may not provide a uniform distribution for sampling (in contrast with regular graphs), in the Internet graph obtaining a uniform sampling is not trivial as observed by [2].

III. METRICS

Metrics can be distinguished in terms of the scope of the information they provide. Local metrics, which are extracted from individual nodes, give information about the individual nodes whereas global metrics reflect the properties of a whole network. A single global value can be extracted from the local values by using simple statistics.

A. Degree

Degree is the most trivial metric and it is defined as the number of the connections of a single node for an undirected graph. Although degree of a node is a local property, the statistics on it gives connectivity information of a whole graph. For example, the average node degree gives the number of connections that a typical node has.

In [9], it is stated that metrics based on the minimum, maximum and average values are not sufficient to describe the skewed distributed data and proposed to use the exponents of power laws as new metrics (these are the outdegree, rank, eigen, and hop plot exponents). The exponents measure the tendency of a property.

B. Clustering Coefficients

Clustering coefficients are the local metrics that reflect the connectivity information in the neighborhood environment of a node [10]. It can be also thought that they provide the transitivity information [11], since it controls whether two different nodes are connected, or if they are connected to the same node.

Clustering coefficient C_i is defined as the percentage of the connections between the neighbors of node i, and it is given as:

$$C_i = \frac{2 \cdot E_i}{k \cdot (k-1)} \tag{1}$$

where k is the number of neighbors of node i and E_i is the existing connections between its neighbors.

Clustering coefficient D_i is defined similar to C_i with an exception. It also considers node i and its connections in the computation of the clustering coefficient [10]. The formula of D_i is given as:

$$D_i = \frac{2 \cdot (E_i + k)}{k \cdot (k+1)} \tag{2}$$

The global clustering coefficients of C and D are computed as the averages of C_i and D_i respectively. Another global clustering coefficient, $C^{(2)}$, can be computed by taking the average over clustering coefficients of all the nodes, C_i, except the ones whose degrees are one [12].

C. Distance Between Nodes

The hop distance between nodes u and v is defined as the shortest path between them, taking the weight of each edge as a unit length. Note that, it is also possible to define distances in terms of physical distances between nodes [13]. This work takes the hop distance as the distance measure. The *diameter* of a graph is the maximum of minimum distances between any two nodes and it determines the effective size of a network.

Closeness is a local metric that measure the connectedness of a network [11]. The closeness of a node is the average of the distances between this node and the other ones. It reflects the centrality property of a single node

TABLE I

THE σ VALUES COMPUTED FOR $(0.2,\sigma)$-EVASIVENESS FOR THE METRICS, I.E., THE ERROR PERCENTAGES WHEN 20 PER CENT OF DATA ARE COVERED. THE $\pm$ VALUES INDICATE THE STANDARD DEVIATION. IT IS CLEAR THAT NONE OF THE METRICS ARE DETERMINED EXACTLY.

Metric	Hop sampling	Random walk	Biased walk
Average degree	0.791($\pm$0.064)	0.079($\pm$0.003)	0.660($\pm$0.077)
Clustering coefficient C	3.349($\pm$0.270)	0.212($\pm$0.018)	2.697($\pm$0.115)
Clustering coefficient $C^{(2)}$	2.836($\pm$0.241)	0.015($\pm$0.012)	1.984($\pm$0.064)
Clustering coefficient D	0.023($\pm$0.008)	0.059($\pm$0.001)	0.047($\pm$0.012)
Hop plot exponent	0.017($\pm$0.009)	0.145($\pm$0.001)	0.129($\pm$0.015)
Effective hop diameter	0.352($\pm$0.020)	0.018($\pm$0.003)	0.134($\pm$0.004)
Average path length	0.381($\pm$0.024)	0.004($\pm$0.003)	0.253($\pm$0.005)
Characteristic path length	0.375($\pm$0.017)	0.005($\pm$0.004)	0.249($\pm$0.007)
Minimum closeness	0.336($\pm$0.028)	0.022($\pm$0.004)	0.224($\pm$0.006)
Average eccentricity	0.398($\pm$0.039)	0.028($\pm$0.004)	0.247($\pm$0.002)
Hop diameter	0.581($\pm$0.094)	0.100($\pm$0.049)	0.269($\pm$0.047)

TABLE II

THE σ VALUES COMPUTED FOR $(0.5,\sigma)$-EVASIVENESS FOR THE METRICS, I.E., THE ERROR PERCENTAGES WHEN 50 PER CENT OF DATA ARE COVERED. THE $\pm$ VALUES INDICATE THE STANDARD DEVIATION. ALTHOUGH 50 PER CENT OF DATA ARE COVERED, THERE ARE STILL ERRORS IN THE VALUES OF THE METRICS (INTERNET METRICS ARE EVASIVE).

Metric	Hop sampling	Random walk	Biased walk
Average degree	0.225($\pm$0.007)	0.127($\pm$0.000)	0.225($\pm$0.011)
Clustering coefficient C	0.992($\pm$0.029)	0.345($\pm$0.007)	0.993($\pm$0.022)
Clustering coefficient $C^{(2)}$	1.161($\pm$0.055)	0.253($\pm$0.006)	1.080($\pm$0.032)
Clustering coefficient D	0.014($\pm$0.004)	0.023($\pm$0.000)	0.009($\pm$0.001)
Hop plot exponent	0.025($\pm$0.008)	0.054($\pm$0.000)	0.023($\pm$0.006)
Effective hop diameter	0.199($\pm$0.012)	0.039($\pm$0.001)	0.117($\pm$0.005)
Average path length	0.239($\pm$0.017)	0.068($\pm$0.001)	0.152($\pm$0.000)
Characteristic path length	0.221($\pm$0.011)	0.067($\pm$0.001)	0.147($\pm$0.000)
Minimum closeness	0.210($\pm$0.018)	0.055($\pm$0.001)	0.132($\pm$0.000)
Average eccentricity	0.274($\pm$0.026)	0.060($\pm$0.001)	0.157($\pm$0.001)
Hop diameter	0.493($\pm$0.116)	0.063($\pm$0.000)	0.215($\pm$0.017)

and smaller values indicate that the node resides close to the center of a network.

The *average path length* is one of the global metrics defined as the average of the closeness values for all nodes [14]. In [12], the *characteristic path length* is defined as the median of all closeness values. *Eccentricity* of a node is a local metric defined as the minimum number of hops required to reach at least 90 per cent of its reachable nodes. The average of the eccentricity of all nodes reflects the size of a network.

In [9] the number of node pairs within h hops (denoted by $P(h)$) is used to define the *hop-plot exponent* $\mathcal{H}$. It is stated that $P(h) \propto h^{\mathcal{H}}, h \ll \delta$, where δ is the diameter of a network. Given the hop plot exponent the *effective hop diameter* δ_{ef} is defined as:

$$\delta_{ef} = \frac{N^2}{N + 2 \cdot E}^{1/\mathcal{H}} \tag{3}$$

IV. EMPIRICAL STUDY

A. Data and Sampling Methods

In the experiments, we use the router level Internet data obtained by using the Mercator software [16][2]. This data consists of approximately 230K nodes and 320K links, and it is used by three sampling techniques. In the empirical study, the sampling simulates the data collection from the Internet, and the router level data represents the whole Internet topology. We compute the metrics on each sampled data and observe their values, depending on the size of the sampled data.

[2]The data are available at http://www.isi.edu/scan/mercator/maps.html.

1) Hop Sampling: The first sampling technique is the *hop sampling*. It creates a subnetwork of the original topology by randomly selecting an initial node and growing it from that node according to the specified hop count h. All nodes and edges visited within h hops are taken to form a subnetwork. This technique is called "ball growing" in [17]. In our experiments, we select hop count value starting with one and incrementing it until the sampled graphs consist of at least 90 per cent of its original topology.

2) Random Walk: The second technique is the *random walk* [18]. In the random walk, we start a randomly selected node and in each step, we select a neighbor of the current node i at random with probability $1/d_i$ where d_i is the degree of i and move to it. In this algorithm, in each step a new link is added, therefore a new subnetwork is formed. On the other hand, we compute the metrics on these subnetworks only at certain points, at the first time that the following percentages of the total nodes $10\%, 20\%, ..., 90\%$ are covered.

3) Biased Random Walk: We implement a modified version of the random walk as the last sampling technique. *Biased walk* is similar to the random walk, the only difference is in selecting a neighbor of the current node. A bias factor is added such that the nodes with larger degrees have higher selecting probabilities. A node is selected with a probability proportional to its degree, e.g., if the current node have neighbors with the degrees of 3 and 4, their selecting probabilities will be 3/7 and 4/7 respectively. In the biased walk, we favor the nodes with higher connectedness. As in the random walk, the metrics are computed at the first time when $10\%, 20\%, ..., 90\%$ are hit.

B. Metrics used

We select less costly metrics to make them efficiently computable in terms of time and memory. For example we discard the eigenvalue exponent since the memory and time complexity become too high with the amount of data we use. We do not compute the outdegree and the rank exponent, since they are defined on directed graphs and we keep our graphs as undirected.

The following local metrics are used by the empirical study: degree, clustering coefficient C, clustering coefficient D, closeness, and eccentricity of each individual node.

In addition to these local metrics, the hop plot values ($P(h)$) defined in [9] are computed to obtain the hop plot exponent.

These local metrics yield to computation of following global metrics to characterize the whole Internet graph: (i) the average node degree (ii) the average clustering coefficients C, $C^{(2)}$ and D (iii) the minimum closeness, the average path length and the characteristic path length (iv) the average eccentricity (v) the hop-plot exponent (vi) the effective hop diameter (vii) the hop diameter of a graph.

We run each sampling method 20 times, starting with different randomly selected initial nodes. We plot the values when $10\%, 20\%, ..., 90\%$ of data are covered. We use linear interpolation for the hop sampling values, since the data covered in each step differ in each run (unlike the random and biased walks). All results given in this paper are the averages of these 20 runs [3].

C. How Long are the Tails?- Kurtosis of the metrics

Kurtosis is the fourth moment and used to decide how long of tail a distribution has. It is a measure of the peakness or flatness of the distribution and computed as follows:

$$kurtosis = \frac{\sum_{i=1}^{N}(X_i - \bar{X})^4}{(N-1)s^4} - 3 \qquad (4)$$

where N is the size of data X, $\bar{X}$ is the mean and s is the standard deviation of the data.

Kurtosis ranges from -2 to plus infinity. Positive kurtosis value indicates a peaked distribution which is described as leptokurtic. The kurtosis of the normal distribution is zero. In leptokurtic distributions, observations have a tendency to cluster densely about some particular point far away from the average. Flat distributions described as platykurtic have negative kurtosis values. In platykurtic distributions, observations are distributed fairly uniformly across their ranges.

We have computed the kurtosis of the local metrics given in Section IV-B. We have also computed the kurtosis of the hop plot values ($P(h)$), defined as the number of pairs within each hop count [9]. Our objective is to learn about the shape charecteristics of the distribution of these metrics without actually computing their distributions.

V. Observations and Interpretations

For all global metrics, the explicit σ values for $(0.2, \sigma)$-evasiveness and $(0.5, \sigma)$-evasiveness are shown in Tables I and II respectively. The results in these tables are the averages, and the standard deviations are given in parentheses. The σ values indicate error percentages in computing the global metrics when 20 and 50 per cent of data are covered. The σ values for other γ values are plotted in Figures 1-6. Note that these results will help the interpretation of next subsections.

[3] We note that the number of runs are not large; however, standard deviations, shown in Tables I and II, are significantly low to justify for using limited number of runs for the submission.

TABLE III

RANKING OF METRICS FOR THREE DIFFERENT SAMPLING METHODS. THE VALUES IN THE PARENTHESIS INDICATE THE σ VALUES (I.E., PERCENTAGE OF THE ERROR) WHEN $\gamma = 0.20$ (I.E., 20 PER CENT OF DATA ARE USED). THEY ARE COMPUTED OVER 20 RUNS. EACH SAMPLING METHODS FAVORS DIFFERENT METRICS IN TERMS OF ACCURACY.

Hop sampling	Random walk	Biased walk
1. Hop plot exponent (1.7%)	1. Average path length (0.4%)	1. Clustering coeff. D (4.7%)
2. Clustering coeff. D (2.3%)	2. Characteristic path length (0.5%)	2. Hop plot exponent (12.9%)
3. Minimum closeness (33.6%)	3. Clustering coeff. $C^{(2)}$ (1.5%)	3. Effective hop diameter (13.4%)
6. Effective hop diameter (35.2%)	4. Effective hop diameter (1.8%)	4. Minimum closeness (22.4%)
4. Characteristic path length (37.5%)	5. Minimum closeness (2.2%)	5. Average eccentricity (24.7%)
5. Average path length (38.1%)	6. Average eccentricity (2.8%)	6. Characteristic path length (24.9%)
7. Average eccentricity (39.8%)	7. Clustering coeff. D (5.9%)	7. Average path length (25.3%)
8. Hop diameter (58.1%)	8. Average degree (7.9%)	8. Hop diameter (26.9%)
9. Average degree (79.1%)	9. Hop diameter (10.0%)	9. Average degree (66.0%)
10. Clustering coeff. $C^{(2)}$ (283.6%)	10. Hop plot exponent (14.5%)	10. Clustering coeff. $C^{(2)}$ (198.4%)
11. Clustering coeff. C (334.9%)	11. Clustering coeff.C (21.2%)	11. Clustering coeff. C (269.7%)

A. Evasiveness of the Metrics

In Tables I and II, it is clear that the metrics are not determined exactly, in other words there are always errors in the values of metrics unless all the graph is visited. Moreover, σ values differ for each metric as well as for each sampling method (each sampling method favors a different set of metrics, by yielding more accurate results).

We propose to use evasiveness values of the metrics to compare them in terms of accuracy. For example, for the hop sampling, the average degree and the global clustering coefficient D are $(0.2, 0.791)$ and $(0.2, 0.023)$-evasive respectively. It means with 20 per cent of data, the global clustering coefficient D gives more accurate results than the average degree. It makes the clustering coefficient D a better metric compared to the average degree, in terms of accuracy. For each sampling method, the rankings of the metrics are reported in Table III. The values in parentheses indicate the average of the σ values when $\gamma = 0.2$.

The σ values for the same metric provide a measure to compare different sampling methods. Smaller σ values indicate better sampling methods, e.g., the average path length is (0.2,0.381), (0.2,0.004), and (0.2,0.253)-evasive for the hop sampling, random walk and biased walk respectively. Thus the average path length "ranks" the sampling methods as the random walk, biased walk, and hop sampling when $\gamma = 0.2$ (in Table I). On the other hand the hop plot exponent ranks them as the hop sampling, biased walk, and random walk when 20 per cent of data are visited. Our experiments show that for most of the γ values the random walk is the best technique (see Figures 1-6).

B. How Much to Measure?

Another interesting question is what amount of measurement (sampling) is necessary to estimate the value of a metric within a particular error range. The answer is not obvious for such a large graph like the Internet. However, (γ, σ)-evasiveness for sampled graphs gives some clues in the answer of this question. In Table IV, the average amount of necessary data to obtain 90 and 80 per cent accuracy is given. In this table, we see that, the hop sampling and biased walk fail to estimate most of the metrics within at most 20 per cent of error unless at least half of the nodes are visited. This is an important drawback for these sampling methods, since they cannot produce appropriate samples to estimate most of the metrics without visiting half of the graph. This table also shows that the random walk is successful in estimating a metric with 10 per cent error when at most 20 per cent of the graph is visited (the only exception is the hop plot exponent). With this sampling the amount of data will decrease if we tolerate 20 per cent of error.

C. How Much Do Local Properties Differ from Global Ones?

In this work, we observe that distributions of local metrics and the σ values of global ones are closely related. Before computing it, we can get some clues about the reliability of a global metric by using the local metric distributions on sampled graphs. This is important for the Internet graph, since we will estimate a metric by using only a small part of the graph.

In Figures 1-6, σ and kurtosis values as a function of γ are shown. In each figure, we focus on a different local metric. In these figures, x-axis indicates how much data are visited (i.e., γ value). σ values in y-axis indicate the error percentage of a global metric ($1 - computed\ metric\ value/real\ metric\ value$) whereas

TABLE IV

γ VALUES: AMOUNT OF MEASUREMENT (SAMPLING) NECESSARY FOR 0.90 AND 0.80 ACCURACIES. THESE ACCURACIES CORRESPOND TO $\sigma = 0.10$ AND $\sigma = 0.20$ RESPECTIVELY. RESULTS ARE THE AVERAGES OVER 20 RUNS AND STANDARD DEVIATIONS ARE GIVEN IN PARENTHESES. IT IS EVIDENT THAT SAMPLING TECHNIQUE PLAYS AN IMPORTANT ROLE.

	$\sigma = 0.1$			$\sigma = 0.2$		
Metric	Hop	Biased	Random	Hop	Biased	Random
Average degree	0.78(0.04)	0.80 (0.00)	0.10 (0.00)	0.60(0.00)	0.60 (0.00)	0.10 (0.00)
Clustering coefficient C	0.97(0.05)	1.00 (0.00)	0.10 (0.00)	0.90(0.00)	0.90 (0.00)	0.10 (0.00)
Clustering coefficient $C^{(2)}$	1.00(0.00)	1.00 (0.00)	0.20 (0.00)	0.90(0.00)	1.00 (0.00)	0.20 (0.02)
Clustering coefficient D	0.12(0.04)	0.10 (0.00)	0.10 (0.00)	0.10(0.00)	0.10 (0.00)	0.10 (0.00)
Hop plot exponent	0.11(0.03)	0.30 (0.00)	0.40 (0.00)	0.10(0.00)	0.10 (0.00)	0.20 (0.00)
Effective hop diameter	0.79(0.02)	0.60 (0.00)	0.13 (0.04)	0.55(0.05)	0.10 (0.00)	0.10 (0.00)
Average path length	0.87(0.06)	0.70 (0.00)	0.17 (0.05)	0.66(0.08)	0.40 (0.00)	0.10 (0.00)
Characteristic path length	0.82(0.05)	0.70 (0.00)	0.16 (0.05)	0.58(0.04)	0.40 (0.00)	0.10 (0.00)
Minimum closeness	0.88(0.06)	0.70 (0.00)	0.20 (0.00)	0.57(0.06)	0.30 (0.00)	0.11 (0.02)
Average eccentricity	0.90(0.00)	0.70 (0.00)	0.20 (0.00)	0.75(0.11)	0.40 (0.00)	0.11 (0.03)
Hop diameter	0.95(0.12)	1.00 (0.00)	0.12 (0.05)	0.87(0.27)	1.00 (0.00)	0.10 (0.00)

the kurtosis value in y-axis is the measure that characterizes the peakness or flatness of the distribution of local metrics.

In Figures 1 and 2, the kurtosis values are positive (they are greater than 5) which indicate leptokurtic distributions. This means observations of the node degree and the local clustering coefficient C are clustered far away from their averages. Thus it is hard to estimate the average values defined on these local metrics. The experiments indicate that sampling technique is becoming more important to estimate such metrics. In our case, the random walk is successful to sample a subgraph on which the average degree and the global clustering coefficients are estimated accurately even with small portion of data. On the other hand, the hop sampling and biased walk fail to sample a graph successfully regarding to these metrics.

In Figures 3 and 4, γ-kurtosis curves show that the local clustering coefficient D and the hop plot values $P(h)$ come from flat distributions. All sampling techniques successfully can sample the whole graph, resulting in smaller σ values even when a small part of data is visited. For such metrics, sampling technique is not as important as for the metrics with leptokurtic distributions. These metrics are stronger in terms of accuracy compared to those with leptokurtic distributions.

Figures 5 and 6 show local and global metrics related to the closeness and eccentricity. For both properties, local and global metrics are similar. Their kurtosis values approximately range from 0 to 3. Although their distributions are leptokurtic, the kurtosis values are not as high as the node degree or the clustering coefficient C. Their kurtosis are close to that of the normal distribution. This feature causes global metrics being more predictable. The σ values are between the platykurtic and leptokurtic ones. Like the average degree and global clustering coefficient C and $C^{(2)}$, the sampling method is important but the other sampling methods are still tolerable.

D. Does Measurement Size Matter?

In [3], we conducted our own measurements using traceroute servers to construct an Internet router graph with $7,000$ nodes from 10 different measurements. In order to validate the results reported in [3] for a small measurement size (i.e., 7K), we compare the properties of $7K$-node measurement graph with $230K$-node Mercator one. Note that both measurement techniques use traceroute-like primitives for prompting the Internet. The comparison, tabulated in Table I, shows remarkable similarity. Although quite counter intuitive (since one would expect to see a change in the measured properties of the Internet as the measurement size increases) this comparison suggests the existence of scale-free property [15] for the Internet router graph.

VI. CONCLUSION

For analyzing the Internet and understanding its topological properties, it is important to define proper metrics. The metrics converge to their actual values as the data size increases. But it is not possible to collect complete Internet data and the cost of the computation gets higher as the data size increases. Thus the exact values of the metrics must be estimated by using only a small portion of data. In this work, we define (γ, σ)-evasiveness to assess whether it is possible to estimate the real value of a metric in $\pm\sigma$

TABLE V

COMPARISON OF 7K-NODE AND 230K-NODE (MERCATOR) MEASUREMENT GRAPHS. THE METRICS HAVE REMARKABLE CLOSE VALUES IN SPITE OF THE DIFFERENCE IN THE MEASUREMENT SIZE. THE VALUES IN PARENTHESES ARE THE STANDARD DEVIATIONS OF 10 MEASUREMENTS OF SIZE 7K NODES.

Metric	7K	230K
Average degree	3.881 (± 0.077)	2.807
Clustering coefficient C	0.012 (± 0.008)	0.026
Clustering coefficient $C^{(2)}$	0.012 (± 0.009)	0.062
Clustering coefficient D	0.501 (± 0.003)	0.804
Hop diameter	34.300 (± 4.057)	32.000
Hop plot exponent	3.242 (± 0.020)	4.940
Effective hop diameter	9.493 (± 0.135)	9.274
Characteristic path length	9.017 (± 0.171)	9.339
Average path length	9.285 (± 0.195)	9.515
Average eccentricity	11.768 (± 0.262)	11.706
Minimum closeness	6.532 (± 0.272)	5.527

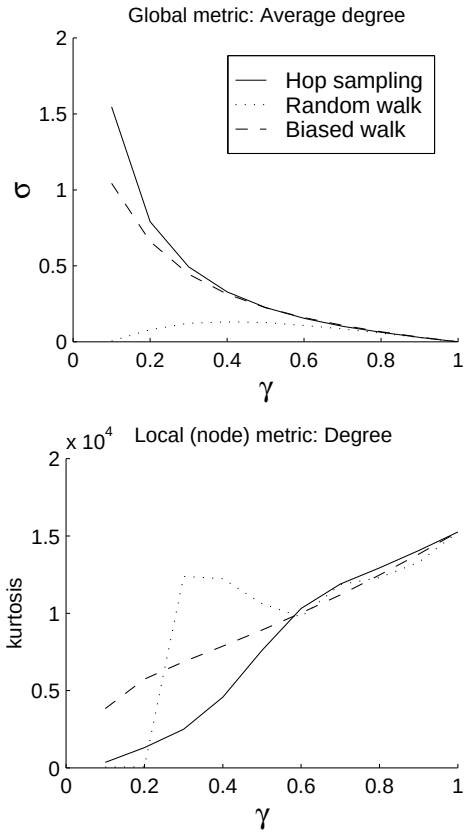

Fig. 1. γ-σ and γ-kurtosis curves. σ is computed on the average degree, which is a global metric, and kurtosis is computed on the node degrees, which is a local metric. It is obvious that the random walk more successfully sample the graph than the hop sampling and biased walk with respect to the accuracy of the average degree.

error margin or not. The γ value indicates the percentage of data and it gives us a measure to compare different metrics and sampling methods. The (γ, σ)-evasiveness is a novel relaxation of *evasiveness* and we discussed how it relates to *property testing*. The (γ, σ)-evasiveness helps us answer to an important question: how much measurement do we need to do to obtain accurate metric results? This is an essential question, because due to the immense size of the Internet, it is only possible to sample its small part.

We conducted an extensive study of (γ, σ)-evasiveness on the Internet router graph with 230K nodes, collected by the Mercator software [16]. Some of the results are as follows:

• We obtain different γ and σ values for different metrics and sampling methods. Table IV shows that, with less than half of the nodes, 10 per cent of error is only tolerable for the random walk. The other two sampling methods cannot tolerate this error with this amount of data except for the global clustering coefficient D and the hop plot exponent (they are the metrics with negative kurtosis). As the per cent of error margin increases (becomes 20 per cent in this table), with less than half of the nodes, the biased walk is starting to tolerate 20 per cent of error for the metrics whose kurtosis values are negative or close to that of the normal distributed data. With this much data, the hop sampling still make more than 20 per cent of error for the metrics with positive kurtosis values. As a result of this table, we can "rank" the sampling methods as the random walk, biased walk, and hop sampling in descending order. This table also summarizes which metrics can be estimated accurately by using a small part of data.

• There is a statistical relationship between the kurtosis of local metrics and the error margin in estimating corresponding global metrics. The σ values are small for the local metrics with platykurtic distribution independent of the sampling method. For leptokurtic distributed local metrics, the sampling method has a great impact on the value of σ. Increasing the value of kurtosis increases the importance of the sampling method. When kurtosis is closer to that of the normal distributed data, σ values are considered tolerable but not as good as for the platykurtic data for all sampling methods.

• (γ, σ)-evasiveness is a *meta metric* which can be used to compare different metrics and sampling methods. It shows that efficient metrics are the ones that converge their exact values quickly and accurately. That means they have smaller σ and γ values in (γ, σ)-evasiveness.

We think that analyzing the (γ, σ)-evasiveness of metrics and kurtosis values of the subgraphs will help us to come up with better sampling methods which increases the success of a metric estimation.

REFERENCES

[1] S. Floyd and V. Paxson, "Difficulties in Simulating the Internet", *IEEE/ACM Transactions on Networking*, vol. 9, pp. 392–403, 2001.

[2] A. Lakhina, J.W. Byer, M. Crovella, P. Xie, "Sampling Biases in IP Topology Measurements," in *Proceedings of the IEEE INFOCOM*, 2003

[3] C. Gunduz, M. Balman, and B. Yener, "Evasiveness of Internet Topology", Technical Report TR 03-02, Rensselaer Polytechnic Institute, 2003. (avaiable at http://www.cs.rpi.edu/ yener/research.html).

[4] L. Lovasz and N. Young, "Lecture notes on evasiveness of graph properties", Technical Report TR 317-91, Princeton University, 1994.

[5] A. L. Rosenberg, "On the Time Required to Recognize Properties of Graphs: a Problem,", *SIGACT News*, vol. 5, pp. 15–16, 1973.

[6] R. Rivest and J. Vuillemin, " A Generalization and Proof of the Aanderaa-Rosenberg Conjecture," in *Proceedings of the 7th Annual ACM Symposium on the Theory of Computing*, pp. 6-11, 1975.

[7] R. Rubinfield and M. Sudan, "Robust Characterization of Polynomials with Applications to Program Testing," *SIAM Journal on Computing*, vol. 25, no.2, pp–252-271, 1996.

[8] O. Goldreich, S. Goldwasser, and D. Ron, "Property Testing and Its Connections to Learning and Approximation," *Journal of the ACM*, vol.45, no.4, pp. 653-750, 1998.

[9] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On Power-Law Relationships of the Internet Topology", in *Proceedings of ACM SIGCOMM*, pp. 251–262, 1999.

[10] S. N. Dorogovtsev and J. F. F. Mendes, "Evolution of Networks", *Advances in Physics*, cond–mat/0106144, 2002.

[11] M. E. J. Newman, "Who is the Best Connected Scientist? A Study of Scientific Coauthorship Networks", *Phys.Rev.*, cond–mat/0011144, 2001.

[12] T. Bu and D. Towsley, "On Distinguishing between Internet Power Law Topology Generators", in *Proceedings of the IEEE INFOCOM*, 2002.

[13] E. W. Zegura, K. L. Calvert, and M. J. Donahoo, "A Quantitative Comparison of Graph–based Models for Internet Topology", *IEEE/ACM Transactions on Networking*, vol. 5, pp. 770–783, 1997.

[14] A. Medina, I. Matta, and J. Byers, "On the Origin of Power Laws in Internet Topologies", *ACM Computer Communications Review*, vol. 30, no. 2, pp. 18–28, 2000.

[15] A. L. Barabasi and R. Albert, "Emergence of Scaling in Random Networks," *Science 286*, pp. 509–512, 1999.

[16] R. Govindan and H. Tangmunarunkit, "Heuristics for Internet Map Discovery", in *Proceedings of the IEEE INFOCOM*, 2000.

[17] H. Tangmunarunkit, R. Govindan, S. Jamin, S. Shenker, and W. Willinger, "Network Topology Generators: Degree-Based vs. Structural.", in *Proceedings of the ACM SIGCOMM*, 2002.

[18] L. Lovasz, "Random Walks on Graphs: A Survey", *Combinatorics: Paul Erdos is Eighty*, vol. 2, pp. 353–398, 1996.

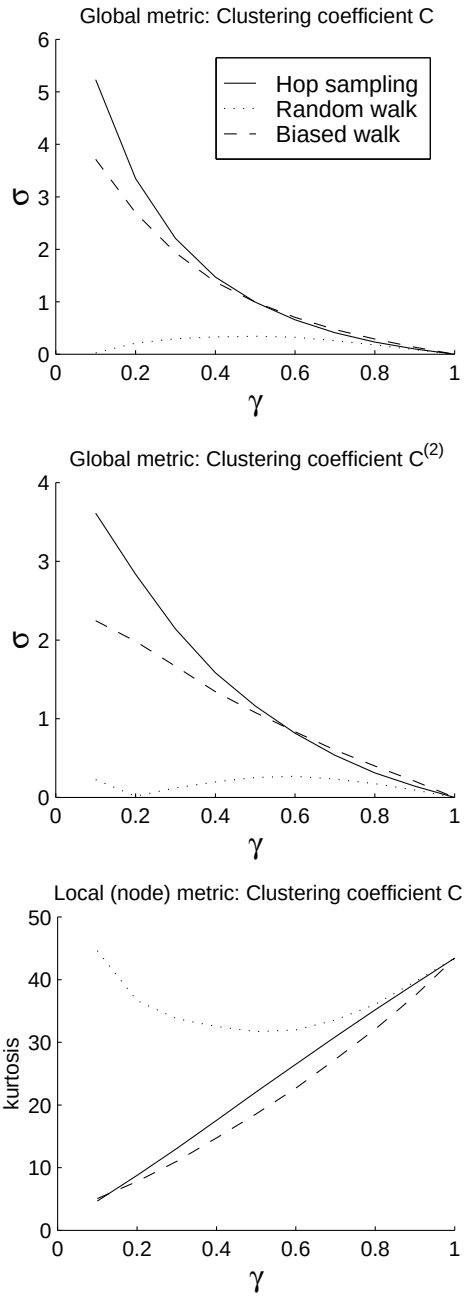

Fig. 2. γ-σ and γ-kurtosis curves. σ is computed on the clustering coefficients C and $C^{(2)}$, which are the statistics of local clustering coefficient C. Kurtosis is computed on the local clustering coefficient C, defined for each node. On leptokurtic distributed data, some sampling methods better sample the graph than the others.

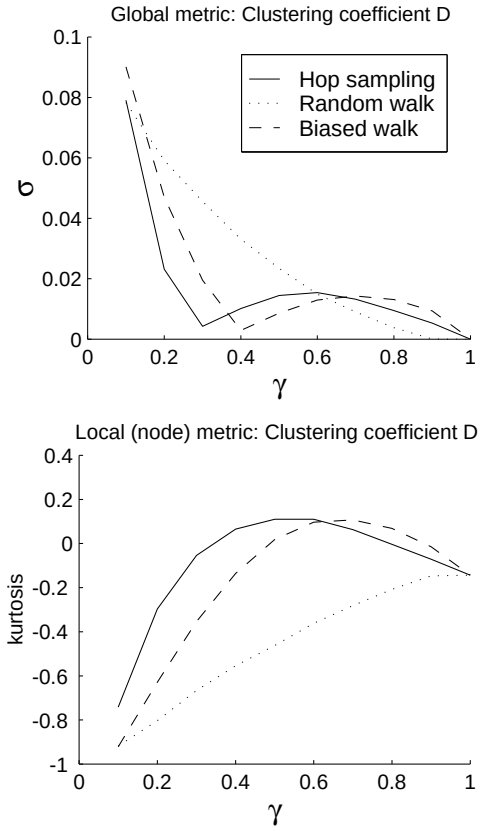

Fig. 3. γ-σ and γ-kurtosis curves. σ is computed on the clustering coefficient D, which is the average of the local clustering coefficient D. The kurtosis of the local clustering coefficient D is given in the second graph. Platykurtic distributions decrease the importance of the sampling methods. It is obvious that all sampling methods are successful.

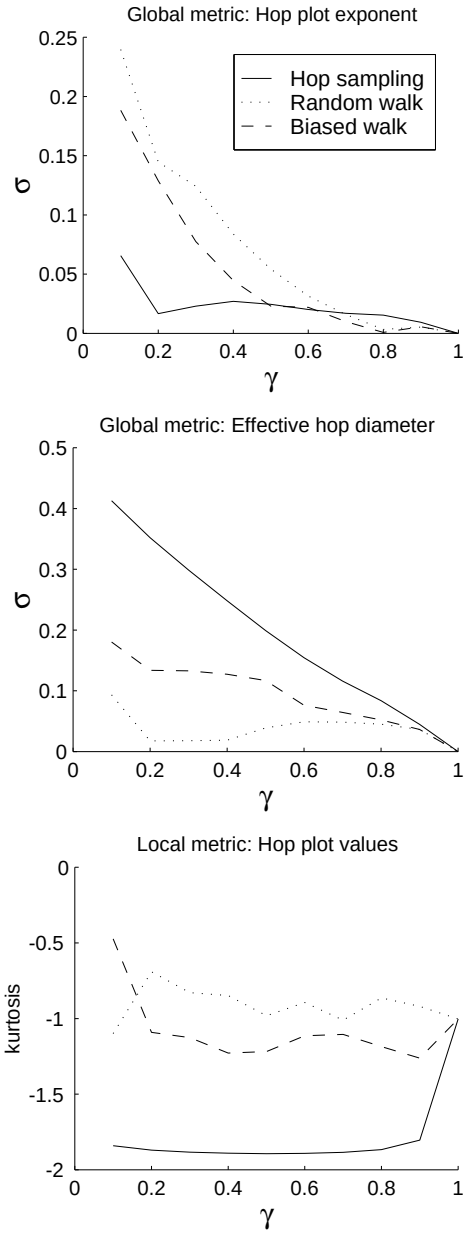

Fig. 4. γ-σ and γ-kurtosis curves. σ is computed on the hop plot exponent and the effective hop diameter, which are the global metrics, and kurtosis is computed on the hop plot values for each hop count. Note that this local metric is different than the others, since it is computed on hop plot values whereas the other local metrics belong to individual nodes.

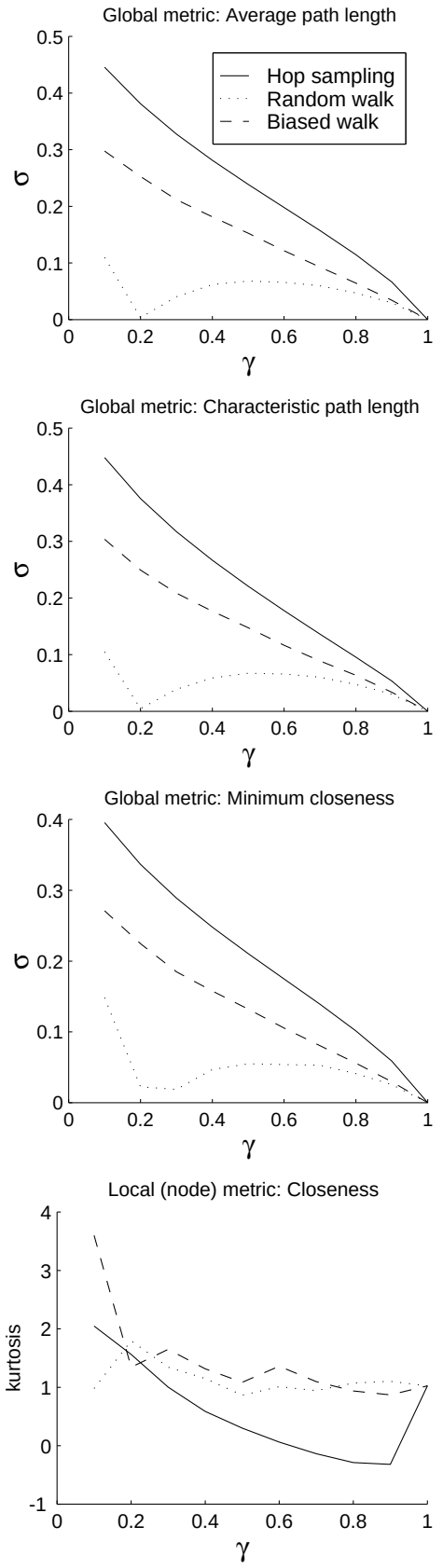

Fig. 5. γ-σ and γ-kurtosis curves. σ values are computed on the global closeness metrics whereas kurtosis is computed on the local closeness. The results are very similar to those in Figure 6

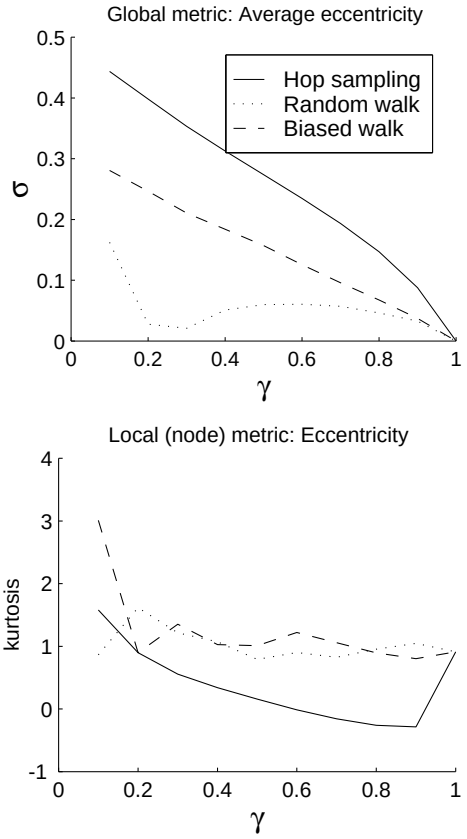

Fig. 6. γ-σ curve of the average eccentricity for the whole graph and γ-kurtosis curve for the node eccentricities. The local eccentricities as well as the local closeness values (Figure 5) are leptokurtic, but their kurtosis values are smaller compared to those in Figures 1 and 2. The selection of the sampling method is still important but it reduces according to the average degree and the global clustering coefficients C and $C^{(2)}$.