Sampling networks by the union of m shortest path trees

Huijuan Wang*and Piet Van Mieghem†
Faculty of Electrical Engineering, Mathematics, and Computer Science
Delft University of Technology, P.O. Box 5031, 2600 GA Delft

October 21, 2009

Abstract

Many network topology measurements capture or sample only a partial view of the actual network structure, which we call the underlying network. Sampling bias is a critical problem in the field of complex networks ranging from biological networks, social networks and artificial networks like the Internet. This bias phenomenon depends on both the sampling method of the measurements and the features of the underlying networks. In RIPE NCC and the PlanetLab measurement architectures, the Internet is mapped as $G_{\cup_m spt}$, the union of shortest paths between each pair of a set $\mathcal{M}$ of m testboxes, or equivalently, m shortest path trees. In this paper, we investigate this sampling method on a wide class of real-world complex networks as well as on the weighted Erdös-Rényi random graphs. This general framework examines the effect of the set of testboxes on $G_{\cup_m spt}$. We establish the correlation between the subgraph $G_{\mathcal{M}}$ of the underlying network, i.e. the set $\mathcal{M}$ and the direct links between nodes of set $\mathcal{M}$, and the sampled network $G_{\cup_m spt}$. Furthermore, we illustrate that in order to obtain an increasingly accurate view of a given network, a higher than linear detection/measuring effort (the relative size m/N of set $\mathcal{M}$) is needed, where N is the size of the underlying network. Finally, when the relative size m/N of set $\mathcal{M}$ is small, we characterize the kind of networks possessing small sampling bias, which provides insights on how to place the testboxes for good network topology measurement.

Keywords: network sampling, sampling bias, shortest path tree, weighted and unweighted networks.

1 Introduction

Topologies of complex networks ranging from biological networks such as gene regulatory networks [1], metabolic networks [2], artificial networks like the Internet, the WWW to social networks, e.g. paper citations, collaboration networks etc. [3], have been accumulated by active investigation in recent years. However, many surveyed networks to date are, in fact, subnets of the actual network, which we call the "*underlying network*". For example, only a subset of the molecular entities in a cell have been sampled in protein interaction, gene regulation and metabolic networks. The topology of the Internet is inferred by aggregating paths, which reveals only a part of the whole Internet. Thus, these identified networks are sampled networks of the underlying networks according to different mapping or sampling methods.

In this work, we study the bias phenomenon of a sampling method that originated from the Internet. The topology of the Internet has typically been measured by the union of sampling traceroutes [4], which are approximately shortest paths. Mainly two sampling methods exist: (a) The topology is built from the union of traceroutes from a small set of sources to a larger set of destinations as in the CAIDA skitter project [5]. The sampled map can be modeled as the union of the spanning trees rooted at the sources. (b) The traceroute

*Email: h.wang@ewi.tudelft.nl
†Email: p.vanmieghem@ewi.tudelft.nl

measurements are carried out between each pair of a set $\mathcal{M}$ of m testboxes or testbeds. The sampled network, denoted as $G_{\cup_m spt}$, is the union of m shortest path trees $SPTs$, where each SPT is the union of shortest paths from the root $\in \mathcal{M}$ to the other $m-1$ testboxes $\in \mathcal{M}$. Equivalently, $G_{\cup_m spt}$ is the union of shortest paths between each node pair in the set $\mathcal{M}$ of m testboxes. The RIPE NCC [6] and the PlanetLab [7] measurement architectures are examples of this type. The methodology in (a) has been argued and even proved to introduce such intrinsic biases that statistical properties of the sampled topology may sharply differ from that of the underlying graph (see e.g. [8, 9, 10]). While most related works on Internet exploration have been devoted to the sampling method (a), we investigate the other sampling method (b). Although the number of destinations may be limited to the number m of measurement boxes, the spurious effects in (a), where nodes and links closer to the sources are more likely to be sampled than those surrounding the destinations, can be reduced.

With statistical and graph theory methodologies, we investigate this sampling method (m shortest path trees) on a wide class of networks: the weighted Erdös-Rényi random graphs, which represent dense and homogeneous networks, and the unweighted real-world complex networks which are generally sparse and inhomogeneous graphs. Various underlying networks are investigated, because network sampling is a generic problem residing in various disciplines and the actual underlying network topology is mostly uncertain. Here, we focus on the sampling bias (the incompleteness of the network mapping) introduced purely by the sampling method. Technical limitations in the topology measurements may also introduce significant sampling bias. For example, the network measured by traceroute represents the interconnections of IP addresses. The bias in mapping the router level Internet topology depends highly on the alias resolution technique, which maps IP addresses to the corresponding routers [11]. Such specific technical concerns, which vary in the measuring of different complex networks, are not explored in this paper.

The sampled network $G_{\cup_m spt}$ depends on the set $\mathcal{M}$ of m boxes as well as the underlying network. In this work, we focus on the effect of the testboxes, in particular, 1) the subgraph $G_{\mathcal{M}}$ of the underlying network, consisting of the set $\mathcal{M}$ and the direct links between nodes of set $\mathcal{M}$, and 2) the relative size m/N of set $\mathcal{M}$, where N is the size of the underlying network. With a given set of testboxes, the sampling bias varies for different networks. The kind of networks with small sampling bias will also be briefly mentioned in this paper.

The main contributions of this study can be summarized as follows.

1. Introduction of a general framework for network sampling on both weighted and unweighted complex networks.

2. Establishment of the correlation between the interconnections of set $\mathcal{M}$, i.e. the subgraph $G_{\mathcal{M}}$, and the sampled network $G_{\cup_m spt}$.

3. Illustration of the detection/measuring effort (the relative size m/N of set $\mathcal{M}$) to obtain an increasingly accurate view of a given network.

4. Characterization of networks bearing small sampling bias when m/N is small and the corresponding proposal of testbox placement for good network topology measurements.

2 Modeling the sampling process of large networks

Assuming that traceroutes used in RIPE NCC and the PlanetLab are shortest paths, the sampled topology is then the union $G_{\cup_m spt}$ of shortest paths between each pair of a small group of $m \ll N$ nodes, while the number of nodes in the underlying graph N is much larger. When $m = N$, the graph $G_{\cup_m spt}$ becomes $G_{\cup spt}$, the union of all shortest paths between any node pair. $G_{\cup spt}$ is thus the maximal measurable or observable part of a network by traceroute measurements [12]. It is also regarded as the "transport overlay network" [13]. In the Internet, for example, all the traffic is carried along the overlay $G_{\cup spt}$, a fraction of the links in the underlying network. An example to represent the relation between the sampled overlay network $G_{\cup_m spt}$, the overlay network $G_{\cup spt}$

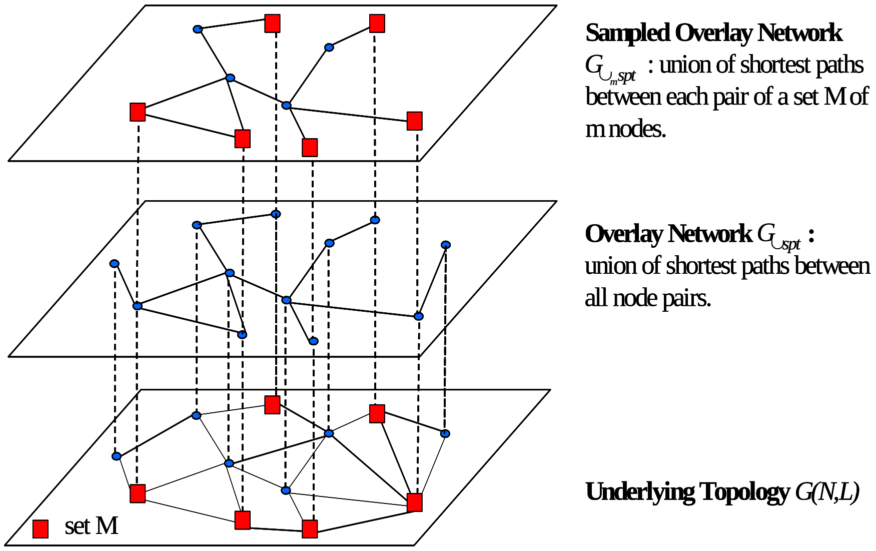
Sampled Overlay Network
$G_{\cup_m spt}$: union of shortest paths between each pair of a set M of m nodes.
Overlay Network $G_{\cup spt}$:
union of shortest paths between all node pairs.
Underlying Topology $G(N,L)$
set M

Figure 1: The relation between the sampled overlay network, the overlay network and the underlying graph.

and the underlying graph (or substrate) is shown in Figure 1. The robustness of networks, e.g. the persistence of epidemics [14] and the vulnerability to node failures and attacks [15] are depending on structural properties of $G_{\cup spt}$. Hence, the sampling bias refers to the difference between the sampled overlay $G_{\cup_m spt}$ and the overlay network $G_{\cup spt}$. We show in Section 4 that the sampling bias can be quantitatively characterized by $\frac{E[L_{mspt}]}{E[L_o]}$, where L_{mspt} and L_o are the number of links in $G_{\cup_m spt}$ and $G_{\cup spt}$. When the underlying graph is unweighted networks, the overlay network is equal to the underlying graph $G_{\cup spt} = G(N, L)$, because each link (i, j) in $G(N, L)$ is the shortest path between node i and j.

2.1 Substrates: networks to be sampled

We consider two classes of substrates: the weighted Erdös-Rényi random graph $G_p(N)$ and real-world complex networks that are unweighted.

The Erdös-Rényi random graphs $G_p(N)$ can be generated from a set of N nodes by randomly assigning a link with probability p to each pair of nodes. Besides their analytic tractability, the Erdös-Rényi random graphs [16] have also served as idealized structures for peer-to-peer networks [17], ad-hoc networks [18], gene networks and ecosystems [19]. Other network models, such as power law graphs [20], which are random graphs specified by a power law degree distribution $\Pr[D = i] = ci^{-\tau}$, are usually sparse. The sampling via $G_{\cup_m spt}$ of a sparse network is the same no matter whether this network is weighted or not, because paths between any node pair are likely unique. Hence, in the class of the weighted networks, we consider the Erdös-Rényi random graph $G_p(N)$, which is dense. We assign to each link an i.i.d. uniform link weight within $[0, 1]$. A link weight may represent e.g. the delay, the distance, the monetary cost, etc. Apart from being attractive in a theoretical analysis, the uniform distribution on $[0, 1]$ is the underlying distribution to generate an arbitrary other distribution and is especially interesting for computer simulations. Hence, this distribution appears most often in network simulations and deserves – for this reason alone perhaps – to be studied. Furthermore, the shortest path problem is mainly sensitive to the smaller link weights, especially in a dense network. Statistical properties of the shortest paths remain asymptotically the same when the network is equipped with i.i.d. regular[1] link weights [21], e.g. uniform or exponential distributed link weights, which may capture the link weight features in many real networks. Thus, the uniform distribution is much less restrictive than it appears at the first glance. All the links are assumed undirected.

[1] A regular link weight distribution $F_w(x) = \Pr[w \leq x]$ has a Taylor series epansion around $x = 0, F_w(x) = f_w(0)x + O(x^2)$, since $F_w(0) = 0$ and $F'_w(0) = f_w(0)$ exists. A regular link weight distribution is thus linear around zero.

We also consider the unweighted real-world networks which represent the topology of various complex systems. Some of these networks possess a power law degree distribution, a feature that is claimed in many complex networks. Most of the data sets we have used are available publicly. They are complex networks from a wide range of systems in nature and society:

- the Gnutella [22] snapshots (Crawl2) retrieved from firewire.com;
- the air transportation network representing the world wide airport connections, documented at the Bureau of Transportation Statistics (http://www.bts.gov) database, and the connection between United States airports [23];
- the Western States Power Grid of the United States[24];
- the coauthorship network [25] between scientists posting preprints on the High-Energy Theory E-Print Archive between Jan 1, 1995 and December 31, 1999;
- the citation network [26] created using the Web of Science database: Kohonen [27];
- the coauthorship network [28] of scientists working on network theory and experiment;
- the network representing soccer players association to Dutch soccer team [29];
- the adjacency network [28] of common adjectives and nouns in the novel David Copperfield by Charles Dickens.

A network is connected if there exists a path between each pair of nodes. We consider only the networks formed by the largest connected component of our real-world networks.

2.2 The overlay network $G_{\cup spt}$ on top of the weighted Erdös-Rényi random graph $G_p(N)$

A uniform recursive tree URT grows from its root and at each stage a new node is attached uniformly to one of the existing nodes. The overlay network $G_{\cup spt}$ is also the union of shortest path trees[2] $SPTs$ rooted at each node. In [30], a URT is shown to be asymptotically the SPT in the Erdös-Rényi random graph $G_p(N)$ with link density p above the disconnectivity threshold $p_c \sim \frac{\log N}{N}$ and with regular link weight distribution, e.g. uniform or exponential distribution. We first review an interesting result about the degree $D_{G_{\cup spt}}$ of an arbitrary node in the overlay $G_{\cup spt}$, which is derived from the URT modeling.

Theorem 1 *For large N, the degree distribution in the overlay $G_{\cup spt}$ on top of the Erdös-Rényi random graph $G_p(N)$ with link density p above the disconnectivity threshold p_c and equipped with i.i.d. regular link weights is*

$$\Pr[D_{G_{\cup spt}} = k] = \frac{(-1)^{N-1-k} S_{N-1}^{(k)}}{(N-1)!} \tag{1}$$

where $S_N^{(k)}$ is the Stirling number of the first kind [31].

Proof: See [12]. □

If a link in the underlying graph belongs to the overlay network $G_{\cup spt}$, it is said to be detected or observed in the overlay network.

[2] The shortest path tree is the union of shortest paths from the root to all the other nodes in the network.

Theorem 2 *In the Erdös-Rényi random graph $G_p(N)$ with link density p above the disconnectivity threshold p_c, large N and equipped with i.i.d. regular link weights, the probability of a link to be detected in the overlay $G_{\cup spt}$ is equal to*

$$\Pr[P^*_{i\to j} = i \to j] = \Pr[H_N = 1] = \frac{1}{N-1} \sum_{n=1}^{N-1} \frac{1}{n} \tag{2}$$

where $P^*_{i\to j}$ is the shortest path between i and j and H_N is the hopcount of a shortest path.

Proof: Any link $i \to j$ with link weight $w(i \to j)$ in the $G_{\cup spt}$ must be the shortest path between i and j because a link in the $G_{\cup spt}$ must belong to a shortest path and a subsection of a shortest path is also a shortest path. Reversed, if a link $i \to j$ is the shortest path between i and j, it must belong to the $G_{\cup spt}$, because the $G_{\cup spt}$ is the union of shortest paths between all possible source and destination pairs. Therefore, the event that a link $i \to j$ is observed in the $G_{\cup spt}$ is equivalent to the event $\{P^*_{i\to j} = i \to j\}$ that the link $i \to j$ is the shortest path $P^*_{i\to j}$ between i and j. Hence, $\Pr[P^*_{i\to j} = i \to j]$ is also the probability that a link can be detected in the overlay $G_{\cup spt}$.

The event $\{P^*_{i\to j} = i \to j\}$ is equal to the event $\{H_N = 1\}$ that the hopcount of the shortest path is 1. Hence, $\Pr[P^*_{i\to j} = i \to j] = \Pr[H_N = 1]$ and $\Pr[H_N = 1] = \frac{1}{N-1}\sum_{n=1}^{N-1}\frac{1}{n}$ has been derived in [21, Section 16.6.3]. □

The average number of links in $G_{\cup spt}$, or the average observable links via $G_{\cup spt}$ is

$$E[L_o] = \frac{N(N-1)}{2}\Pr[P^*_{i\to j} = i \to j] = \frac{N}{2}\sum_{n=1}^{N-1}\frac{1}{n} \simeq \frac{N}{2}(\ln N + \gamma) \tag{3}$$

where $\gamma = 0.57721...$ is the Euler constant.

3 Effect of $G_{\mathcal{M}}$ on the sampled overlay $G_{\cup_m spt}$

Recall that a network is mapped as $G_{\cup_m spt}$, the union of shortest paths between each pair of a set $\mathcal{M}$ of m testboxes. The overlay network $G_{\cup spt}$ is the union of the shortest paths between all node pairs. We examine first the effect of $G_{\mathcal{M}}$ on the sampled overlay $G_{\cup_m spt}$ when the underlying network or substrate is a weighted Erdös-Rényi random graph. As shown in Figure 2, the subgraph $G_{\mathcal{M}}$ of a underlying network $G(N, L)$ is the set $\mathcal{M}$ and the direct links between nodes of set $\mathcal{M}$. The maximal observable part of the subgraph $G_{\mathcal{M}}$ is the overlay network $G_{\cup spt}$ upon $G_{\mathcal{M}}$. It is now denoted as $G_{\cup spt}(m)$ to include the number of nodes in the overlay network and $G_{\cup spt}(m) \subset G_{\mathcal{M}}$. The sampled overlay $G_{\cup_m spt}$ and the overlay $G_{\cup spt}(N)$ are constructed based on the shortest paths computed in the underlying network $G(N, L)$ while the overlay $G_{\cup spt}(m)$ on the subgraph $G_{\mathcal{M}}$ is based on the shortest path computation in the subgraph $G_{\mathcal{M}}$. Similar to the overlay $G_{\cup spt}(m)$, the sampled network $G_{\cup_m spt}$ is also the union of shortest path between each node pair of the set $\mathcal{M}$, however, upon the underlying network $G(N, L)$ instead of upon the subgraph $G_{\mathcal{M}}$. We now examine the similarity or difference between $G_{\cup spt}(m)$ and $G_{\cup_m spt}$.

Each simulation on Erdös-Rényi random graphs consists of 10^4 iterations. Within each iteration, a set $\mathcal{M}$ of $m = 40$ nodes is uniformly chosen out of the generated substrate $G_{0.6}(200)$ and an *i.i.d.* uniform link weight is assigned to each link. Shortest paths are computed by the Dijkstra's algorithm [32]. We construct three networks (a) the sampled overlay $G_{\cup_m spt}$ and (b) the overlay $G_{\cup spt}(N)$ on top of the underlying graph $G(N, L)$ and (c) the overlay $G_{\cup spt}(m)$ on the subgraph $G_{\mathcal{M}}$. The degree distributions of these three networks are displayed in Figure 3. We denote $D_{\mathcal{M}}$ as the degree of set $\mathcal{M}$ in the sampled overlay $G_{\cup_m spt}$. The degree distribution of $D_{\mathcal{M}}$ is much closer to the degree distribution of the overlay $G_{\cup spt}(m)$ on top of $G_{\mathcal{M}}$ than that of the overlay $G_{\cup spt}(N)$. Beside the set $\mathcal{M}$, the other nodes in the sampled overlay $G_{\cup_m spt}$ belong to set $\mathcal{I}$. The degree distribution $D_{\mathcal{I}}$ of set $\mathcal{I}$ performs even worse to represent the overlay $G_{\cup spt}(N)$ as compared to set $\mathcal{M}$.

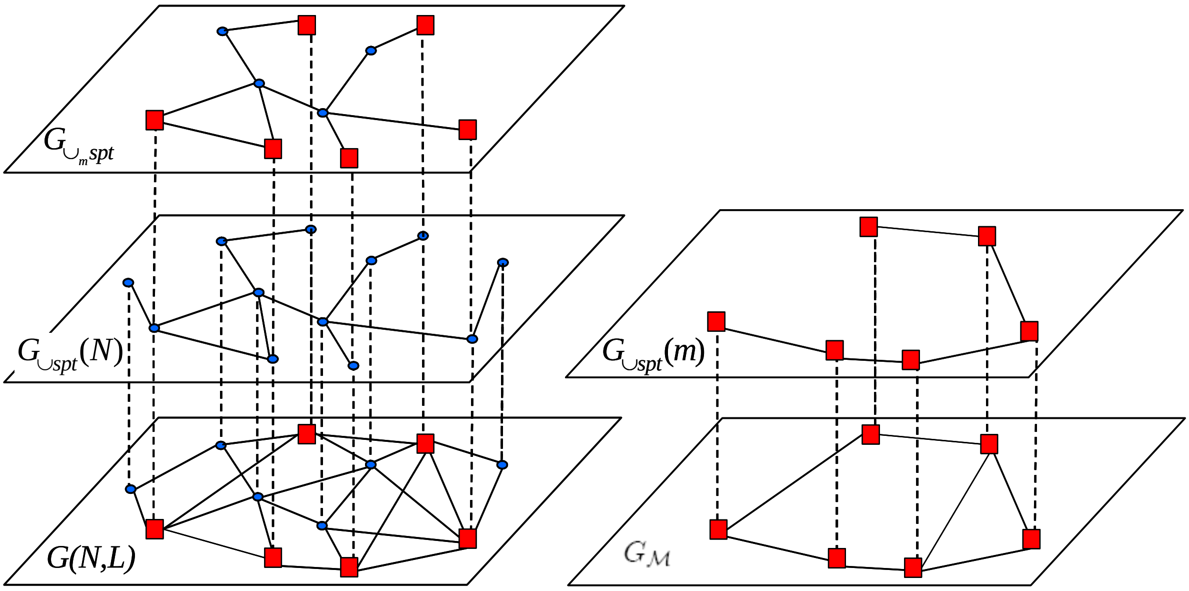

Figure 2: Sketch of the sampled overlay $G_{\cup_m spt}$ and the overlay $G_{\cup spt}(N)$ on top of the underlying graph $G(N, L)$ and the overlay $G_{\cup spt}(m)$ on the subgraph $G_{\mathcal{M}}$.

We further investigate the resemblance in degree distribution between $D_{\mathcal{M}}$ and the overlay $G_{\cup spt}(m)$ on the subgraph $G_{\mathcal{M}}$ over more Erdös-Rényi random graphs: $G_{0.2}(400)$ and $G_{0.2}(800)$ with different size m of the set $\mathcal{M}$. Figure 4 illustrates that the set $\mathcal{M}$ in the sampled overlay $G_{\cup_m spt}$ and the overlay $G_{\cup spt}(m)$ upon $G_{\mathcal{M}}$ possess almost the same degree distribution. The degree distribution of the overlay $G_{\cup spt}(m = 10, 20, 50)$ upon $G_{\mathcal{M}}$ is calculated based on Theorem 1, using $\Pr[D_{G_{\cup spt}}(m) = k] = \frac{(-1)^{m-1-k} S_{m-1}^{(k)}}{(m-1)!}$. It seems that $\Pr[D_{\mathcal{M}} = k] = \Pr[D_{G_{\cup spt}}(m) = k]$. The degree distribution of the set $\mathcal{M}$ in the sampled overlay $G_{\cup_m spt}$ is independent of the size N of the underlying network: the set $\mathcal{M}$ follows a same degree distribution in $G_{\cup_m spt}(N = 400)$, $G_{\cup_m spt}(N = 800)$ and $G_{\cup_m spt}(N = m) = G_{\cup spt}(m)$. Hence, we claim the following conjecture:

Conjecture 3 *Consider the sampled overlay graph $G_{\cup_m spt}$ on top of an Erdös-Rényi random graph $G_p(N)$ with link density p above the disconnectivity threshold p_c and equipped with i.i.d. regular link weights. The degree distribution of $D_{\mathcal{M}}$ of set $\mathcal{M}$ in the sampled overlay graph $G_{\cup_m spt}$ is independent of the size N of the network and*

$$\Pr[D_{\mathcal{M}} = k] = \Pr[D_{G_{\cup spt}}(m) = k] = \frac{(-1)^{m-1-k} S_{m-1}^{(k)}}{(m-1)!}$$

As presented in Appendix A, two extreme cases $\Pr[D_{\mathcal{M}} = 1]$ and $\Pr[D_{\mathcal{M}} = m - 1]$ can be proved. The Conjecture 3 states that the degree distribution of the set $\mathcal{M}$ is independent of the size of the underlying topology, but only of the number m of measurement nodes in $\mathcal{M}$. This "intermediate node invariant" degree property could be used, in principle, to reduce or infer $G(N, L)$ and the link weight structure. In other words, if the so measured $G_{\cup spt}(m)$ statistically has the same degree distribution as the set $\mathcal{M}$ of $G_{\cup_m spt}$, the network is possibly homogeneous and equipped with i.i.d. regular link weights.

On top of a dense homogeneous network equipped with i.i.d. regular link weights, the set $\mathcal{M}$ of the sampled overlay network well reflects the local overlay $G_{\cup spt}(m)$ on top of a subgraph $G_{\mathcal{M}}$ in the degree distribution, although $m \ll N$. It seems that the testboxes, i.e. the subgraph $G_{\mathcal{M}}$ (or, equivalently, $G_{\cup spt}(m)$ upon the subgraph $G_{\mathcal{M}}$) do effect the sampled overlay $G_{\cup_m spt}$ in the degree distribution of set $\mathcal{M}$. The Erdös-Rényi random graph is homogenous and so is the subgraph $G_{\mathcal{M}}$. Hence, the resemblance in degree distribution between $D_{\mathcal{M}}$ and the overlay $G_{\cup spt}(m)$ may originate from the fact that both $G_{\cup spt}(m)$ and $G_{\cup_m spt}$ take into account the union of $m(m - 1)/2$ shortest paths.

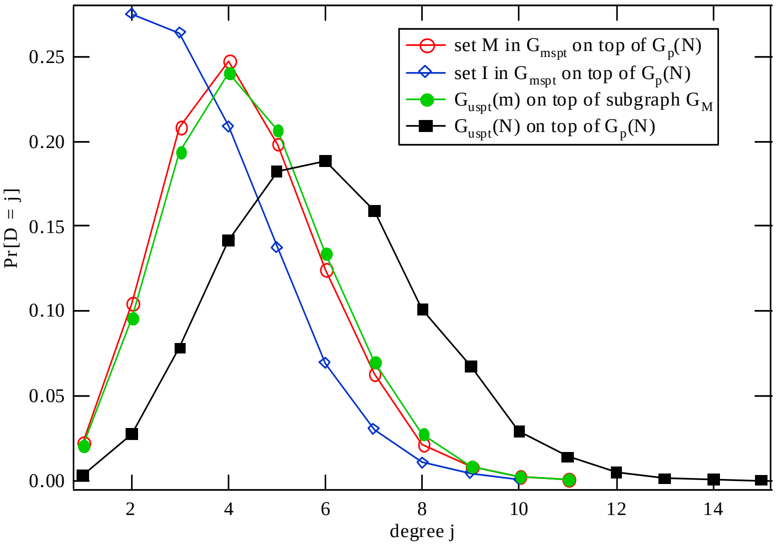

Figure 3: Degree distribution of (a) the sampled overlay $G_{\cup_m spt}$ upon $G_{0.6}(200)$ (b) overlay $G_{\cup spt}(N)$ upon $G_{0.6}(200)$ and (c) overlay $G_{\cup spt}(m)$ upon the subgraph $G_{\mathcal{M}}$, where $m = 40$.

In a real-world unweighted network, the overlay network $G_{\cup spt}(N)$ is equal to the substrate $G(N, L)$ and the overlay network $G_{\cup spt}(m)$ on top of subgraph $G_{\mathcal{M}}$ is $G_{\mathcal{M}}$ itself. For unweighted networks, we have

$$G_{\mathcal{M}} = G_{\cup spt}(m) \subset G_{\cup_m spt} \subset G_{\cup spt}(N) = G(N, L)$$

where $G_{\cup spt}(m) \subset G_{\cup_m spt}$ is due to the fact that any link (i, j) in an unweighted graph is the shortest path between its end nodes i and j. The structure of $G_{\cup_m spt}$ varies between $G_{\mathcal{M}}$ and the substrate $G(N, L)$. Hence, the subgraph $G_{\mathcal{M}}$ is correlated with the sampled network $G_{\cup_m spt}$, in the sense that $G_{\mathcal{M}} = G_{\cup spt}(m) \subset G_{\cup_m spt}$. As a larger proportion of the substrate is observed, the sampled overlay $G_{\cup_m spt}$ resembles the underlying network $G_{\cup spt}(N) = G(N, L)$ more.

4 Effect of the relative size m/N of the testboxes on the sampling bias

In this section, we first explain why $E[L_{mspt}]/E[L_o]$ quantifies the sampling bias well. Then, we investigate the effect of the relative size m/N of the testboxes on the sampling bias. Given the ratio m/N, the sampling bias differs for various networks depending on their topologies. We will briefly discuss which type of network tends to possess small sampling bias.

4.1 Characterizing the sampling bias by $E[L_{mspt}]/E[L_o]$

The sampling bias refers to the difference between the sampled overlay $G_{\cup_m spt}$ and the overlay network $G_{\cup spt}$. The relation $G_{\cup_m spt} \subset G_{\cup spt}(N)$ holds for both weighted Erdös-Rényi random graphs and unweighted networks. Hence, the ratio of the average number of links in the $G_{\cup_m spt}$ and $G_{\cup spt}$, $E[L_{mspt}]/E[L_o]$ can reasonably well characterize[3] the sampling bias of a network, where $E[L_o] = L$ in case the network is unweighted.

First, Figure 7 in Appendix B shows that the probability distribution of the normalized number of links $L^*_{mspt} = \frac{L_{mspt} - E[L_{mspt}]}{\sigma[L_{mspt}]}$ and the normalized number of nodes $N^*_{mspt} = \frac{N_{mspt} - E[N_{mspt}]}{\sigma[N_{mspt}]}$ in $G_{\cup_m spt}$ are both

[3] $E[L_{mspt}]/E[L_o]$ is a statistical property which takes into account different realizations of the set $\mathcal{M}$ selection as well as the link weight assignment.

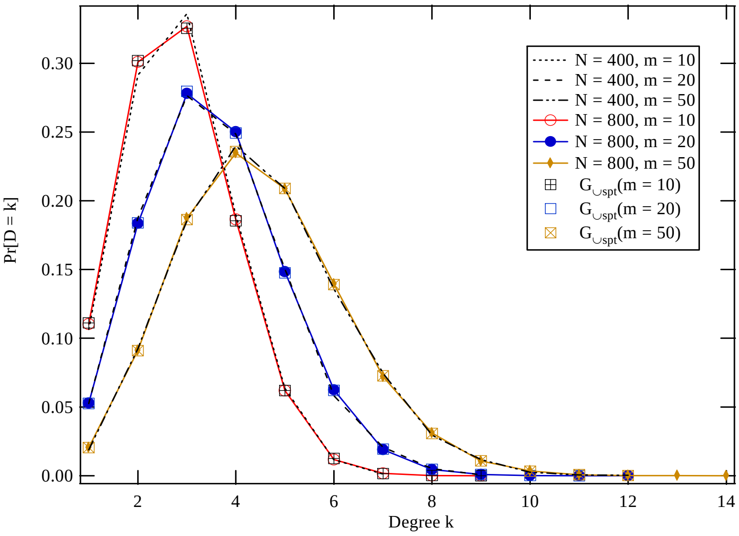

Figure 4: Degree distribution $D_{\mathcal{M}}\left(N_{mspt}\right)$ of set $\mathcal{M}$.

close to the Gaussian distribution $N(0,1)$. Moreover, their average and standard deviation, which determine the distribution, follow $\sigma(L_{mspt}) \ll E[L_{mspt}]$ and $\sigma(N_{mspt}) \ll E[N_{mspt}]$ as illustrated in Figure 8 and 9 in Appendix B. Hence, the random variables L_{mspt} and N_{mspt} are close to their mean $E[L_{mspt}]$ and $E[N_{mspt}]$, which are thus the appropriate quantities to be studied.

Furthermore, we investigate the sampling bias via $E[L_{mspt}]/E[L_o]$ instead of the number of nodes $E[N_{mspt}]/N$. The relation between $E[N_{mspt}]$ and $E[L_{mspt}]$ follows from the basic law of the degree:

$$\sum_{j=1}^{m} d_{j\in\mathcal{M}} + \sum_{j=m+1}^{N_{mspt}} d_{j\in\mathcal{I}} = 2L_{mspt}$$

Taking the expectation yields

$$m \cdot E[D_{\mathcal{M}}] + E[\sum_{j=m+1}^{N_{mspt}} d_{j\in\mathcal{I}}] = 2E[L_{mspt}]$$

Assume that N_{mspt} and $d_{j\in\mathcal{I}}$ are only weakly dependent such that we may apply Wald's identity [21, Chapter 1],

$$2E[L_{mspt}] \simeq m \cdot E[D_{\mathcal{M}}] + (E[N_{mspt}] - m) \cdot E[D_{\mathcal{I}}\left(N_{mspt}\right)]$$

or

$$E[L_{mspt}] \simeq \frac{1}{2}E[D_{\mathcal{I}}\left(N_{mspt}\right)] \cdot E[N_{mspt}] + \frac{m}{2}(E[D_{\mathcal{M}}] - E[D_{\mathcal{I}}\left(N_{mspt}\right)]) \tag{4}$$

Under the assumption of weak dependence between N_{mspt} and $d_{j\in\mathcal{I}}$, a linear relation exists between $E[L_{mspt}]$ and $E[N_{mspt}]$ with slope equal to $E[D_{\mathcal{I}}\left(N_{mspt}\right)]/2$, where $E[D_{\mathcal{I}}\left(N_{mspt}\right)]$ is a function of m. For example, we consider the substrate $G_{0.2}(800)$ equipped with i.i.d. uniformly distributed link weights. The left and right sides of (4) are shown to be almost the same in the table below, which justifies the weak dependency assumption.

m	10	20	30	40	50	60	100	300
left side of (4)	124.4	308.3	479.6	630.6	762.6	881.4	1242.6	2111.7
right side of (4)	124.6	308.6	479.6	630.6	763.4	881.1	1244	2117.2

4.2 Sampling of the weighted Erdös-Rényi random graph

The average number of links in the SPT rooted at a source to m uniformly chosen nodes in the complete graph K_N, or approximately in $G_p(N)$, with uniform link weights is given in [21, Chapter 17],

$$g_N(m) = \frac{mN}{N-m} \sum_{k=m+1}^{N} \frac{1}{k} \simeq \frac{mN}{N-m} \log \frac{N}{m} \tag{5}$$

Hence, the number of links in each of the m $SPTs$ of $G_{\cup_m spt}$ is, on average, equal to $g_N(m-1)$. The maximum number of links that can be detected in case $m = N$ via $G_{\cup spt}$ is $E\left[L_o\right]$ given by (3). Since L_{mspt} is not decreasing in m, we have that

$$g_N(m-1) \leq E[L_{mspt}] \leq E\left[L_o\right]$$

and

$$E[L_{mspt}] \leq m \cdot g_N(m-1)$$

Hence, for large N,

$$\frac{(m-1)N}{N-m+1} \log \frac{N}{m-1} \leq E[L_{mspt}] \leq \frac{N}{2}\left(\gamma + \ln N\right)$$

The ratio $E[L_{mspt}]/E[L_o]$ quantifies the sampling bias, while the ratio $E[L_{mspt}]/(m \cdot g_N(m-1))$ reflects the extent of overlap between these m $SPTs$. As shown in Figure 5, for the substrate $G_{0.2}(800)$ and $m = 60$, 30% of links in $G_{\cup spt}$ have already been observed. For $m = 120$, about 40% links are discovered. Indeed, for any network, the larger m is, the smaller the sampling bias is, because $\lim_{m \to N} G_{\cup_m spt} = G_{\cup spt}(N)$. For $N = 800$, the ratio $E[L_{mspt}]/E[L_o] = O(m^{\beta})$ with $\beta \approx 0.23$, which implies that "the discovering rate of new links" decreases with m. In other words, to obtain an increasingly accurate view of the network, a higher detection/measuring effort is needed, in fact, much higher than proportional. Since $E[L_{mspt}]/E[L_o] = O(m^{\beta})$, we found via simulation that the exponent β increases with N. When $A = \frac{m}{N} \to 0$, the shortest paths between nodes of set $\mathcal{M}$ seldom overlap,

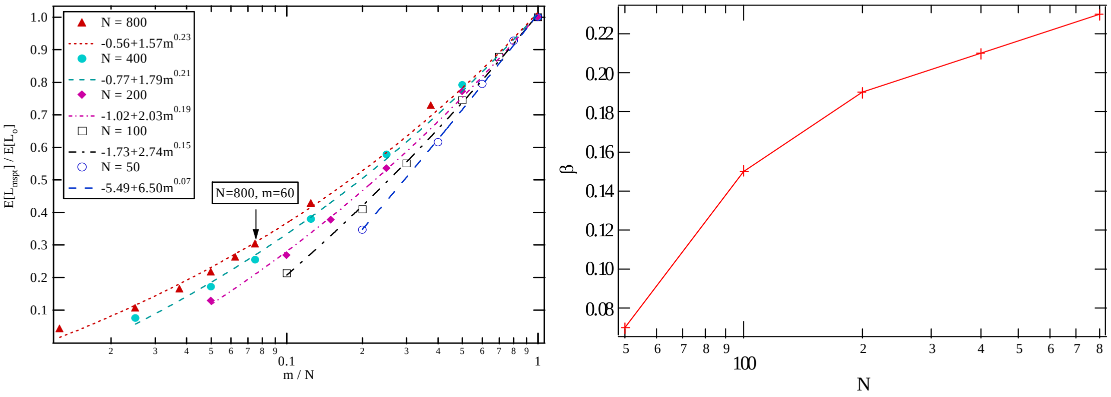

Figure 5: The ratio $E[L_{mspt}]/E[L_o]$ and the power exponent β in the corresponding curve fitting $E[L_{mspt}]/E[L_o] = a + bm^{\beta}$, where the substrate is $G_{0.2}(N)$.

$$E[L_{mspt}] \simeq \binom{m}{2} E[H_N] = \frac{A^2 N^2}{2} E[H_N]$$

Using (3) and [21, Section 16.3], we have

$$\frac{E[L_{mspt}]}{E[L_o]} \simeq \frac{\frac{A^2N^2}{2} E[H_N]}{\frac{N(N-1)}{2} p_o} \simeq \frac{\frac{A^2N^2}{2} (\ln N + \gamma)}{\frac{N}{2} (\ln N + \gamma)} = A^2 N$$

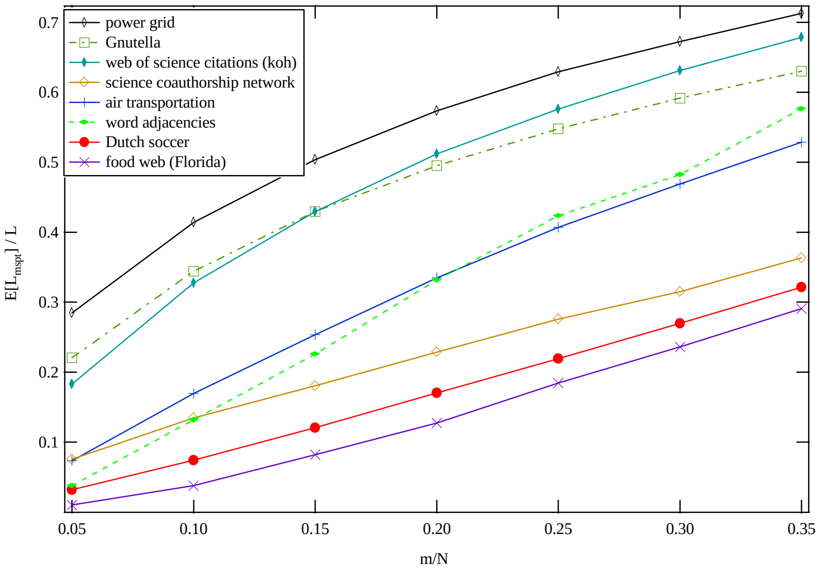

Figure 6: The average proportion of links $E[L_{mspt}]/L$ discovered via $G_{\cup_m spt}$ as a funtion of the relative size m/N of set $\mathcal{M}$.

where $\gamma = 0.57721..$ and p_o is the link density of the overlay $G_{\cup spt}$. Hence, for a small m/N, large networks tend to have a small sampling bias or large $E[L_{mspt}]/E[L_o]$. Moreover, a sparse overlay network characterized by a small p_o tends to have a small sampling bias, as observed in real-world complex network sampling in Section 4.3.

4.3 Sampling of the real-world complex networks

On top of each real-world network mentioned in Section 2.1, we increase the size of the set $\mathcal{M}$ from $\frac{m}{N} = 5\%$ to $\frac{m}{N} = 35\%$ with a step size of 5%. Given $\frac{m}{N}$, each simulation consists of 40 realizations[4] of the random selection of set $\mathcal{M}$. The average proportion of links $E[L_{mspt}]/L$ discovered in the corresponding sampled overlay $G_{\cup_m spt}$ is plotted as a function of $\frac{m}{N}$ in Figure 6. Similar to the weighted Erdös-Rényi random graph, to obtain an increasingly accurate view of the network, a higher than linear detection/measuring effort m/N is needed[5].

With a given proportion m/N of uniformly distributed testboxes in a network, the sampling bias $E[L_{mspt}]/L$ depends purely on the topology of the network. We compare the topology features of each real-world complex network in Table 1 to see which kind of network tends to possess a small sampling bias. We computed the following topological metrics for each network, which are considered relevant in the networking literature [33]:

- The number of nodes N and links L.
- Average degree $E[D] = 2L/N$ and link density $p = \frac{L}{\binom{N}{2}}$.
- The average hopcount (in number of links) and the largest hopcount $h_{\max}$ of the shortest paths between all node pairs. The latter $h_{\max}$ is also referred to as the diameter of a graph. Actually, we assign independently to each link a unit link weight plus a small uniform random variable within $[-\frac{1}{N}, \frac{1}{N}]$, such that a unique shortest path is found between each node pair.
- The *clustering coefficient* of a node $c_G(v)$ characterizes the density of connections in the environment of a node v and is defined as the ratio of the number of links y connecting the $d_v > 1$ neighbors of v over the total

[4] 20 or 10 iterations are carried out for large networks with $N > 3000$.

[5] This holds for the most examined networks except for networks with a high link density, such as the Dutch soccer and food web networks. Most complex networks are considered to be sparse.

possible $\frac{d_v(d_v-1)}{2}$, thus $c_G(v) = \frac{2y}{d_v(d_v-1)}$. The clustering coefficient $C(G)$ of a graph is the average of the clustering coefficient of nodes whose degree is larger than 1, given as $C(G) = \frac{1}{N-|\mathcal{N}^{(1)}|}\sum_{v\in\mathcal{N}-\mathcal{N}^{(1)}} c_G(v)$, where $\mathcal{N}$ is the set of all nodes and $\mathcal{N}^{(1)}$ is the set of degree 1 nodes.

Table 1	N	L	C	$E[H_N]$	$h_{\max}$	$E[D]$	p
Power grid	4941	6594	0.11	18.99	46	2.67	0.00054
Gnutella Crawl2	1568	1906	0.04	6.10	21	2.43	0.0016
Web of Science Citations(koh)	3704	12673	0.30	3.67	12	6.84	0.0018
Science coauthorship network	379	914	0.80	6.03	17	4.82	0.0128
Air Transportation	2179	31326	0.59	3.02	8	28.75	0.0132
Word adjacencies	112	425	0.19	2.51	5	7.59	0.068
Dutch soccer	685	10310	0.75	4.45	11	30.10	0.044
Food web(Florida)	128	2075	0.33	1.76	3	32.42	0.26

Table 1 presents the topological metrics of the real complex networks, in the decreasing order of their corresponding $E[L_{mspt}]/L$ at $m/N = 5\%$ as shown in Figure 6. Recall that a larger proportion $E[L_{mspt}]/L$ of the substrates observed via $G_{\cup_m spt}$ implies a lower sampling bias. Figure 6 and Table 1 show that a network tends to have a small sampling bias if its link density p is low and the average hopcount $E[H_N]$ is large, especially for small m/N. Indeed, when $A = \frac{m}{N} \to 0$, the shortest paths between the set $\mathcal{M}$ seldom overlap and

$$\frac{E[L_{mspt}]}{L} \simeq \frac{\frac{A^2N^2}{2}E[H_N]}{\frac{N(N-1)}{2}p} \simeq \frac{A^2E[H_N]}{p} \tag{6}$$

In fact, for any m, the proportion of observed links $\frac{E[L_{mspt}]}{L}$ can be upper bounded by (6). When m is larger, the shortest paths between set $\mathcal{M}$ overlap more, and $\frac{E[L_{mspt}]}{L}$ is far smaller than its upper bound (6). Therefore, the sampling bias of these networks may have a different order for large m/N. No clear correlation between the sampling bias and other metrics have been found.

In summary, in both the weighted Erdös-Rényi random graph and unweighted real-world networks, to obtain an increasingly accurate view of the network, a higher than linear detection/measuring effort m/N is needed. When m/N is small, the sampling bias depends purely on the average hopcount $E[H_N]$ and the link density of p (or p_o) of an unweighted network (or of the overlay $G_{\cup spt}$ upon a weighted network). Indeed, a larger average hopcount $E[H_N]$ and a small p or p_o imply a small sampling bias. For small m/N, the sampling bias of the weighted Erdös-Rényi random graph is positively correlated with N.

5 Conclusions

In this paper, we study a network sampling method originated from the Internet, namely $G_{\cup_m spt}$ the union of m shortest path trees, or equivalently, the union of shortest paths between each pair of a set M of m testboxes. The analysis covers a wide class of networks, ranging from real-world unweighted complex networks to weighted Erdös-Rényi random graphs.

The interconnections of set $\mathcal{M}$, i.e. the subgraph $G_{\mathcal{M}}$, are correlated with the sampled network $G_{\cup_m spt}$ as follows: When the underlying network is a real-world unweighted network $G(N, L)$, $G_{\mathcal{M}}$ is a subgraph of the sampled overlay $G_{\cup_m spt}$. Surprisingly, when the underlying network is an Erdös-Rényi random graph equipped with i.i.d. regular link weights, the set $\mathcal{M}$ in the sampled overlay graph $G_{\cup_m spt}$ follows the same degree distribution as the overlay $G_{\cup spt}(m)$ upon $G_{\mathcal{M}}$. The degree distribution of $D_{\mathcal{M}}$ of the set $\mathcal{M}$ in the sampled overlay graph $G_{\cup_m spt}$ is independent of the size N of the network.

To obtain an increasingly accurate view of a given network, a higher detection/measuring effort (the size m of set $\mathcal{M}$) is needed, in fact, higher than proportional.

When m/N is small, as in RIPE NCC and the PlanetLab measurement where the number m of testboxes (hundreds) is much smaller the number of routers in the Internet (hundreds of thousands), the sampling bias tends to be small if the average hopcount $E[H_N]$ is large and the link density p, or link density p_o of the overlay network $G_{\cup spt}$, is small. Hence, a large number of testboxes randomly placed far from each other is preferable for good network topology measurements. Furthermore, the sampled overlay network consists of a large number, m, of shortest paths that either start or end at each testbox. Links connected to the testboxes are more likely to be sampled than the other links. Hence, placing testboxes at hubs (nodes with a high degree in the underlying network) may contribute to a small sampling bias. In the sampled overlay $G_{\cup_m spt}$, the set of m textboxes tend to possess a higher average degree than the other (intermediate) nodes, if the underlying network is dense[6], as observed in Figure 3.

6 Acknowledgement

This research was supported by the Netherlands Organization for Scientific Research (NWO) under project number 643.000.503.

References

[1] N.M. Luscombe et al., “Genomic analysis of regulatory network dynamics reveals large topological changes”, Nature 431, pp. 308 (2004).

[2] H. Jeong et al., ”The large-scale organization of metabolic networks”, Nature 407, pp. 651 (2000).

[3] A.-L Barabasi, *Linked, The new science of networks*, Perseus, Cambridge, MA, 2002.

[4] W. Richard Stevens, TCP/IP Illustrated, volume 1, The Protocols. Addsion Wesley, Reading , Massachusetts, 1994.

[5] http://www.caida.org.

[6] Ripe test traffic measurements. http://www.ripe.net/ripencc/mem-services/ttm/.

[7] http://www.planet-lab.org.

[8] A. Lakhina, J. Byers, M. Crovella and P. Xie, "Sampling Biases in IP Topology Measurements", Proc. of IEEE INFOCOM, San Francisco, CA, 2003.

[9] D. Achlioptas, A. Clauset, D. Kempe and C. Moore, On the bias of traceroute sampling: or, power-law degree distributions in regular graphs, Proc. of the thirty-seventh annual ACM symposium on Theory of computing, Baltimore, MD, USA, 2005.

[10] A. Clauset and C. Moore, “Accuracy and Scaling Phenomena in Internet Mapping”, Phys. Rev. Lett. 94, 018701 (2005).

[11] R. Sherwood, A. Bender and N. Spring, "DisCarte: A Disjunctive Internet Cartographer", ACM SIGCOMM’08, Washington, USA, 2008.

[12] P. Van Mieghem and H. Wang, "Properties of the Observable Part of a Network", IEEE/ACM Transaction on Networking, Vol. 17, No. 1, pp. 93-105 (2009).

[6] When the underlying network is sparse, the uniformly distributed testboxes tend to possess a small degree in the underlying network, which limites the number of links incident to the testboxes to be sampled. On the other hand, those few high degree nodes in the underlying network are likely to appear in the sampled overlay as the intermediate nodes. Hence, in the sampled overlay network, the average degree of the intermediate nodes may be higher than that of the testboxes.

[13] H. Wang, J. Martin Hernandez and P. Van Mieghem, "Betweenness Centrality in Weighted Networks", Phys. Rev. E 77, 046105 (2008).

[14] A. Ganesh, L. Massoulie and D. Towsley, "The effect of network topology on the spread of epidemics", Proc. IEEE Infocom, 2005.

[15] L. Zhao, K. Park, and Y.-C. Lai, "Attack vulnerability of scale-free networks due to cascading breakdown", Phys. Rev. E 70, 035101(R) (2004).

[16] B. Bollobás, *Random Graphs,* Cambridge University Press, Cambridge, 2001.

[17] M. Castro, M. Costa, and A. Rowstron, "Should we build Gnutella on a structured overlay", ACM SIGCOMM Computer Communications Review 34(1), pp. 131-136 (2004).

[18] R. Hekmat and P. Van Mieghem, "Connectivity in wireless ad hoc networks with a log-normal radio model. Mobile Networks and Applications", Special Issue: Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks, Vol. 11, No. 3, pp. 351-360 (2006).

[19] R. M. May, "Stability and Complexity in Model Ecosystems ", Princeton University Press, Princeton, 1973.

[20] R. Albert, A. Barabasi, "Emergence of scaling in random networks", Science 286, 509-512 (1999).

[21] P. Van Mieghem, *Performance Analysis of Communications Systems and Networks*, Cambridge University Press, Cambridge, 2006.

[22] A. Beygelzimer, G. Grinstein, R. Linsker and I. Rish, "Improving network robustness", Physica A, 357(3-4), 593–612 (2005).

[23] V. Colizza, R. Pastor-Satorras and A. Vespignani, "Reaction–diffusion processes and metapopulation models in heterogeneous networks", Nature Physics 3, 276 - 282 (2007).

[24] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks", Nature 393, 440-442 (1998).

[25] M. E. J. Newman, "Scientific collaboration networks: I. Network construction and fundamental results", Phys. Rev. E 64, 016131 (2001).

[26] V. Batagelj and A. Mrvar (2006): Pajek datasets. <URL: http://vlado.fmf.uni-lj.si/pub/networks/data/>.

[27] T. Kohonen, *Self-Organizing Maps*, Springer, Berlin, 2001.

[28] M. E. J. Newman, "Finding community structure in networks using the eigenvectors of matrices", Phys. Rev. E 74, 036104 (2006).

[29] A. Jamakovic, R.E. Kooij, P. Van Mieghem and E.R. van Dam, "Robustness of networks against viruses: the role of the spectral radius", Proceedings of the 13th Annual Symposium of the IEEE/CVT Benelux, Liége, 2006.

[30] R. van der Hofstad, G. Hooghiemstra, and P. Van Mieghem, "First passage percolation on the random graph", Probability in the Engineering and Informational Sciences (PEIS), Vol. 15, pp. 225-237 (2001).

[31] M. Abramowitz and J. A. Stegun, *Handbook of Mathematical Functions*, Dover Publications Inc., New York, 1999.

[32] E. W. Dijkstra, "A note on two problems in connection with graphs", Num. Math., 1:269–271. (1959).

[33] P. Mahadevan, D. Krioukov, M. Fomenkov, B. Huffaker, X. Dimitropoulos, K. Claffy and A. Vahdat, “The Internet AS-Level Topology: Three Data Sources and One Definitive Metric”, ACM SIGCOMM Computer Communications Review 36(1), 17-26 (2006).

[34] R. van der Hofstad, R, G. Hooghiemstra and P. Van Mieghem, 2005, "Size and Weight of Shortest Path Trees with Exponential Link Weights", Combinatorics, Probability and Computing, vol. 15, pp. 903-926 (2006).

A Proof of extreme cases of conjecture 3

To simplify the proof, instead of $D_{\mathcal{M}}$, we use $D_N(m)$ to denote the degree of set $\mathcal{M}$ in the overlay $G_{\cup_m spt}$, where N is the number of nodes in the underlying graph and m is the number of testboxes.

A.1 Proof of the Corollary for $k = 1$

Firstly, we prove the conjecture for $\Pr[D_N(m) = 1]$. Van der Hofstad et al. [34] have shown that $p_n(i) = \frac{n-i}{ni}$ is the probability that the paths from the root to i uniformly chosen nodes *that may include the root* in a *URT* of size n share a common link. If one of the i nodes equals the root, there is no link in common because there is no path from the root to itself. Denote by $A_{\text{No Root}}$ the event that the paths from the root to m uniformly chosen nodes *that do not include the root* in a URT of size n share a common link and by A_{Root} the event that the paths from the root to m uniformly chosen nodes *that may include the root* in a *URT* of size n share a common link. The probability that the root is one of the m nodes is $\Pr[\text{root}] = \frac{m}{n}$. Then

$$\Pr[A_{\text{No Root}}] = \Pr[A_{\text{Root}}|\text{No root}] = \frac{\Pr[A_{\text{Root}} \cap \{\text{No root}\}]}{\Pr[\text{No root}]}$$

If one of the m nodes is the root, there is no link in common. That event is not included in A_{Root}, which means that

$$\Pr[A_{\text{Root}} \cap \{\text{No root}\}] = \Pr[A_{\text{Root}}] = p_n(m)$$

and that

$$\Pr[A_{\text{No Root}}] = \frac{p_n(m)}{1-\frac{m}{n}} = \frac{\frac{n-m}{n \cdot m}}{1-\frac{m}{n}} = \frac{1}{m} = p_n^*(m)$$

Finally, we arrive at $p_n^*(m)$, the probability that the paths from the root to m uniformly chosen nodes *that do not include the root* in a URT of size n share a common link. If these paths share a link, then the number of links connected to the root and traversed by these paths must be one. Therefore, the probability $\Pr[D = 1]$ of the set $\mathcal{M}$ in a underlying graph with N nodes is

$$\Pr[D_N(m) = 1] = p_N^*(m-1) = \frac{1}{m-1}$$

In the *URT* with m nodes, according to (1) the probability $\Pr[D_{G_{\cup spt}} = 1] = \frac{1}{m-1}$, which explain the match of the first node in Figure 4. □

A.2 Proof of the Corollary for $k = m - 1$

The extreme case $\Pr[D_N(m) = m-1]$ is proved by using the URTs separation theorem [21, Theorem 16.2.1] and considering Figure 18.3 in [21]. A URT of size N can be separated in a URT T_1 of size k and a URT T_2 of size $N-k$ that incorporates the root (see Figure 18.3 in [21, Theorem 16.2.1]). The maximum degree of the root is achieved in two cases: (a) there is precisely 1 node of $\mathcal{M}$ in T_1 and $m-2$ in T_2 or (b) there is none in T_1 and all $m-1$ are in T_2. If there is more than 1 node of $\mathcal{M}$ in T_1, the degree of the root $D_N(m)$ is smaller than $m-1$, because we need to have $m-1$ separate clusters attached to the root that each contain precisely one node of $\mathcal{M}$. Thus,

$$\begin{aligned}
\Pr\left[D_N\left(m\right)=m-1\right] &= \sum_{k=1}^{N-1}\Pr\left[D_{N-k}\left(m-1\right)=m-2\right]\frac{\binom{k}{1}\binom{N-k-1}{m-2}}{\binom{N-1}{m-1}}\Pr\left[T_1=k\right]+ \\
&+\sum_{k=1}^{N-1}\Pr\left[D_{N-k}\left(m\right)=m-1\right]\frac{\binom{k}{0}\binom{N-k-1}{m-1}}{\binom{N-1}{m-1}}\Pr\left[T_1=k\right]
\end{aligned}$$

because the number of ways to distribute $m-1$ nodes over $N-1$ places that are different from the root such that there is 1 of the m in T_1 and the other $m-2$ in T_2 is $\binom{k}{1}\binom{N-k-1}{m-2}$ and there are $\binom{N-1}{m-1}$ ways to distribute $m-1$ nodes over $N-1$ places. Further, the URTs separation theorem implies that $\Pr\left[T_1=k\right]=\frac{1}{N-1}$. This gives the recursion,

$$\begin{aligned}
\Pr\left[D_N\left(m\right)=m-1\right] &= \frac{1}{(N-1)\binom{N-1}{m-1}}\sum_{k=1}^{N-1}\left\{k\Pr\left[D_{N-k}\left(m-1\right)=m-2\right]\binom{N-k-1}{m-2}\right. \\
&\left.+\Pr\left[D_{N-k}\left(m\right)=m-1\right]\binom{N-k-1}{m-1}\right\} \\
&= \frac{1}{(N-1)\binom{N-1}{m-1}}\sum_{q=m-1}^{N-1}\left\{(N-q)\Pr\left[D_q\left(m-1\right)=m-2\right]\binom{q-1}{m-2}\right. \\
&\left.+\Pr\left[D_q\left(m\right)=m-1\right]\binom{q-1}{m-1}\right\}
\end{aligned}$$

where, in the last line, we have incorporated that $\Pr\left[D_q\left(m-1\right)=m-2\right]=0$ if $q<m-1$. From (1), the initial condition is $\Pr\left[D_m\left(m\right)=m-1\right]=\frac{1}{(m-1)!}$.

Further,

$$\begin{aligned}
(N-1)\binom{N-1}{m-1}\Pr\left[D_N\left(m\right)=m-1\right] &= (N-1)\sum_{q=m-1}^{N-1}\Pr\left[D_q\left(m-1\right)=m-2\right]\binom{q-1}{m-2} \\
&+\sum_{q=m-1}^{N-1}\left\{\Pr\left[D_q\left(m\right)=m-1\right]\binom{q-1}{m-1}\right. \\
&\left.-(q-1)\Pr\left[D_q\left(m-1\right)=m-2\right]\binom{q-1}{m-2}\right\}
\end{aligned}$$

After substitution of $N\to N+1$ in the above and subtracting the above yields, for the left-hand side,

$$L=N\binom{N}{m-1}\Pr\left[D_{N+1}\left(m\right)=m-1\right]-(N-1)\binom{N-1}{m-1}\Pr\left[D_N\left(m\right)=m-1\right]$$

and the right-hand side

$$\begin{aligned}
R &= Q+\binom{N-1}{m-1}\Pr\left[D_N\left(m\right)=m-1\right] \\
&-(N-1)\binom{N-1}{m-2}\Pr\left[D_N\left(m-1\right)=m-2\right]
\end{aligned}$$

with

$$\begin{aligned}
Q &= N\sum_{q=m-1}^{N}\Pr\left[D_q(m-1)=m-2\right]\binom{q-1}{m-2}-(N-1)\sum_{q=m-1}^{N-1}\Pr\left[D_q(m-1)=m-2\right]\binom{q-1}{m-2}\\
&= N\left[\sum_{q=m-1}^{N}\Pr\left[D_q(m-1)=m-2\right]\binom{q-1}{m-2}-\sum_{q=m-1}^{N-1}\Pr\left[D_q(m-1)=m-2\right]\binom{q-1}{m-2}\right]\\
&\quad+\sum_{q=m-1}^{N-1}\Pr\left[D_q(m-1)=m-2\right]\binom{q-1}{m-2}\\
&= N\Pr\left[D_N(m-1)=m-2\right]\binom{N-1}{m-2}+\sum_{q=m-1}^{N-1}\Pr\left[D_q(m-1)=m-2\right]\binom{q-1}{m-2}
\end{aligned}$$

Simplified,

$$\begin{aligned}
L\&R &= N\binom{N}{m-1}\Pr\left[D_{N+1}(m)=m-1\right]-N\binom{N-1}{m-1}\Pr\left[D_N(m)=m-1\right]\\
&= \binom{N-1}{m-2}\Pr\left[D_N(m-1)=m-2\right]\\
&\quad+\sum_{q=m-1}^{N-1}\Pr\left[D_q(m-1)=m-2\right]\binom{q-1}{m-2}
\end{aligned}$$

Repeating the same procedure to remove the last remaining sum gives, for the left hand side,

$$\begin{aligned}
L &= (N+1)\binom{N+1}{m-1}\Pr\left[D_{N+2}(m)=m-1\right]-(N+1)\binom{N}{m-1}\Pr\left[D_{N+1}(m)=m-1\right]\\
&\quad-N\binom{N}{m-1}\Pr\left[D_{N+1}(m)=m-1\right]+N\binom{N-1}{m-1}\Pr\left[D_N(m)=m-1\right]\\
&= (N+1)\binom{N+1}{m-1}\Pr\left[D_{N+2}(m)=m-1\right]-(2N+1)\binom{N}{m-1}\Pr\left[D_{N+1}(m)=m-1\right]\\
&\quad+N\binom{N-1}{m-1}\Pr\left[D_N(m)=m-1\right]
\end{aligned}$$

The right hand side becomes,

$$\begin{aligned}
R &= \binom{N}{m-2}\Pr\left[D_{N+1}(m-1)=m-2\right]-\binom{N-1}{m-2}\Pr\left[D_N(m-1)=m-2\right]\\
&\quad+\sum_{q=m-1}^{N}\Pr\left[D_q(m-1)=m-2\right]\binom{q-1}{m-2}-\sum_{q=m-1}^{N-1}\Pr\left[D_q(m-1)=m-2\right]\binom{q-1}{m-2}\\
&= \binom{N}{m-2}\Pr\left[D_{N+1}(m-1)=m-2\right]-\binom{N-1}{m-2}\Pr\left[D_N(m-1)=m-2\right]\\
&\quad+\binom{N-1}{m-2}\Pr\left[D_N(m-1)=m-2\right]\\
&= \binom{N}{m-2}\Pr\left[D_{N+1}(m-1)=m-2\right]
\end{aligned}$$

Combining both sides gives,

$$\begin{aligned}
\binom{N}{m-2}\Pr\left[D_{N+1}(m-1)=m-2\right] &= (N+1)\binom{N+1}{m-1}\Pr\left[D_{N+2}(m)=m-1\right]\\
&\quad-(2N+1)\binom{N}{m-1}\Pr\left[D_{N+1}(m)=m-1\right]\\
&\quad+N\binom{N-1}{m-1}\Pr\left[D_N(m)=m-1\right]
\end{aligned}$$

By defining

$$r\left[N,m\right]=\binom{N-1}{m-1}\Pr\left[D_N\left(m\right)=m-1\right]$$

we arrive at the recursion,

$$r\left[N+1,m-1\right]=\left(N+1\right)r\left[N+2,m\right]-\left(2N+1\right)r\left[N+1,m\right]+Nr\left[N,m\right] \tag{7}$$

with initial condition

$$r\left[m,m\right]=\frac{1}{\left(m-1\right)!}$$

What we claim is that $\Pr\left[D_N\left(m\right)=m-1\right]=\Pr\left[D_m\left(m\right)=m-1\right]$ for all N, which means that

$$r\left[N,m\right]=\binom{N-1}{m-1}\Pr\left[D_m\left(m\right)=m-1\right]=\binom{N-1}{m-1}r\left[m,m\right]=\binom{N-1}{m-1}r\left[m,m\right]$$

Introduced in (7) gives

$$\begin{aligned}\binom{N}{m-2}r\left[m-1,m-1\right] &= \left(N+1\right)\binom{N+1}{m-1}r\left[m,m\right]-\left(2N+1\right)\binom{N}{m-1}r\left[m,m\right]\\ &\quad +N\binom{N-1}{m-1}r\left[m,m\right]\end{aligned}$$

or

$$\binom{N}{m-2}\left(m-1\right)=\left(N+1\right)\binom{N+1}{m-1}-\left(2N+1\right)\binom{N}{m-1}+N\binom{N-1}{m-1}$$

The relation is, indeed, an identity. □

B Number of links and nodes in $G_{\cup_m spt}$

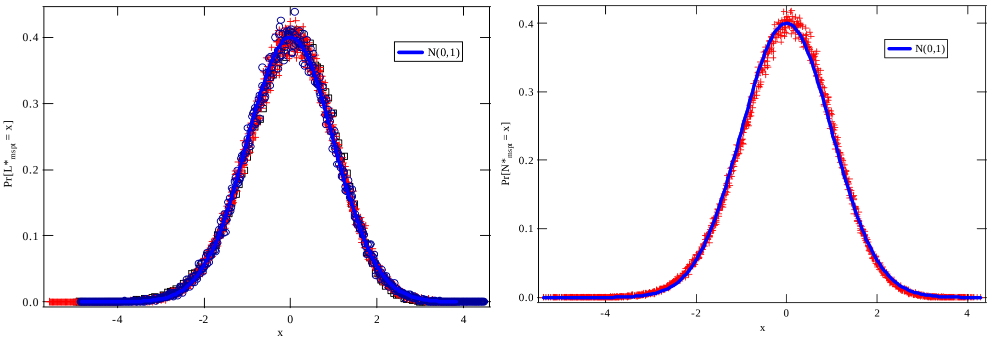

Figure 7: Probability distribution of the normalized number of links(left) L^*_{mspt} and nodes(right) N^*_{mspt} in $G_{\cup_m spt}$ on top of $G_{0.2}(800)$ and $m=10,20,...,60$.

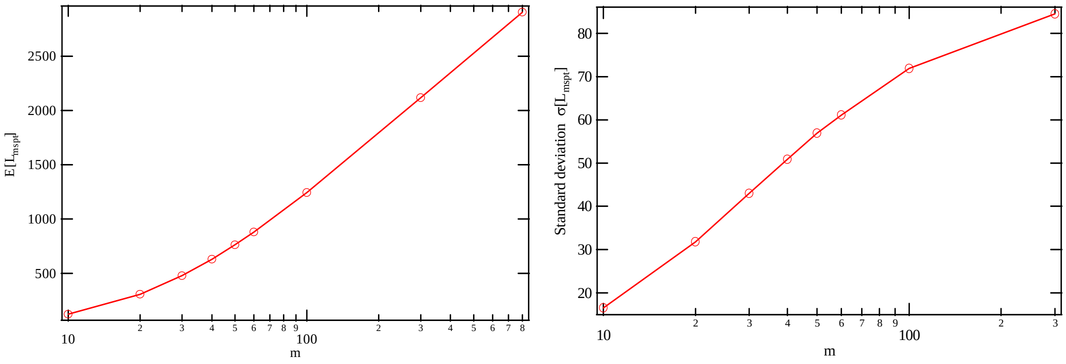

Figure 8: Average and standard deviation of the number of links in $G_{\cup_m spt}$ on top of $G_{0.2}(800)$.

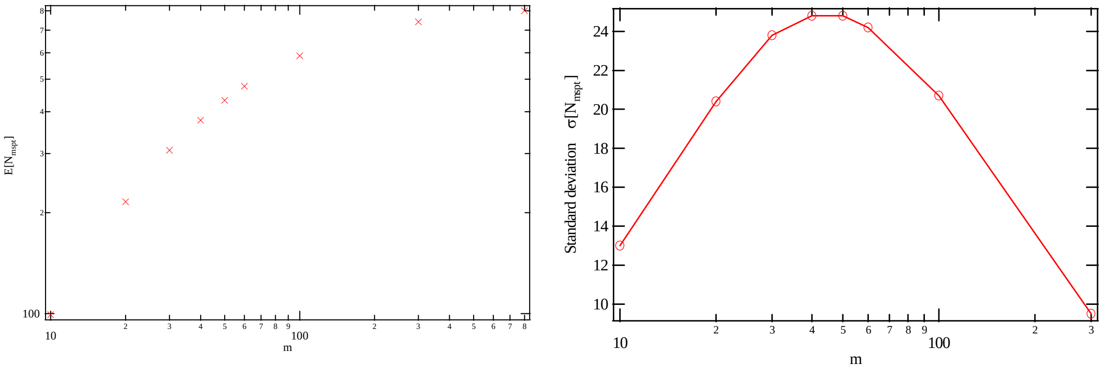

Figure 9: The average and standard deviation of the number of nodes N_{mspt} in $G_{\cup_m spt}$ on top of $G_{0.2}(800)$.