Bigfoot, Sasquatch, the Yeti and Other Missing Links: What We Don't Know About the AS Graph.

Matthew Roughan
University of Adelaide
SA, Australia

Jonathan Tuke
University of Adelaide
SA, Australia

Olaf Maennel
Tech. Universität Berlin
Deutsche Telekom Labs

{matthew.roughan,simon.tuke}@adelaide.edu.au olaf@maennel.net

ABSTRACT

Study of the Internet's high-level structure has for some time intrigued scientists. The AS-graph (showing interconnections between Autonomous Systems) has been measured, studied, modelled and discussed in many papers over the last decade. However, the quality of the measurement data has always been in question. It is by now well known that most measurements of the AS-graph are missing some set of links. Many efforts have been undertaken to correct this, primarily by increasing the set of measurements, but the issue remains: how much is enough? When will we know that we have enough measurements to be sure we can see all (or almost all) of the links. This paper aims to address the problem of estimating how many links are missing from our measurements. We use techniques pioneered in biostatistics and epidemiology for estimating the size of populations (for instance of fish or disease carriers). It is rarely possible to observe entire populations, and so sampling techniques are used. We extend those techniques to the domain of the AS-graph. The key difference between our work and the biological literature is that all links are not the same, and so we build a stratified model and specify an EM algorithm for estimating its parameters. Our estimates suggest that a very significant number of links (many of thousands) are missing from standard route monitor measurements of the AS-graph. Finally, we use the model to derive the number of monitors that would be needed to see a complete AS-graph with high-probability. We estimate that 700 route monitors would see 99.9% of links.

Categories and Subject Descriptors

C.2.3 [**Computer-Communications Networks**]: Network Operations—*network monitoring*; G.3 [**Probability and Statistics**]: Probabilistic algorithms

General Terms

Measurement

Keywords

Topology Inference

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
IMC'08, October 20–22, 2008, Vouliagmeni, Greece.
Copyright 2008 ACM 978-1-60558-334-1/08/10 ...$5.00.

1. INTRODUCTION

Internet topology has drawn interest from areas as diverse as physics and biology. The discovery of power-law degree in nodes [1] resulted in a large number of papers. Though the nature of this power-law has been disputed [2], a more problematic aspect is whether inadequate measurements of the topology artificially induce the power law [3]. The central issue is missing links. All studies reporting measurements of network topologies potentially have missing links. How do we know if they do? How do we know how many they miss? These questions seem to be about things we don't know, and for this reason appear unanswerable, but here we show that they are not.

We use techniques developed in biological research for estimating the population of say fish. The technique "capture-recapture" works as follows. One goes into the field and catches some fish. The captured fish are tagged and then released. Then, at some later point in time one repeats the study. The number of tagged fish that we recapture, along the number of captured fish allows us to make an estimate of the total population.

However, the simple models of capture-recapture assume that all "fish" are equally easy (or hard) to catch, so we sample at random from the population. To paraphrase George Orwell, "All Internet links are equal, but some are more equal than others." Some links are harder to see than others! Although this violates the assumptions underlying the simple capture-recapture model, there is a long literature extending the ideas to many other cases. We draw on this literature to develop a new model and estimation algorithm specific to our problem.

So this paper provides us with a way of estimating what we don't know, in this case the number of hidden or *missing* links in an AS graph (here AS stands for Autonomous System, not Abominable Snowman as you might guess from the title). The technique performs well against the best data we have for validation, supplying supporting evidence for these previous studies. In addition, this approach is easily applied (in contrast to previous papers where extensive efforts were required to clean and combine multiple data sources), and so we can use it to look at the size of the Internet over time. We use it to examine the number of links in the AS-graph over more than four years. From this extended dataset we look for trends, and our results indicate that the AS-graph is growing at 18.7 links per day.

Finally, by mapping this data to link types: peer-2-peer (p2p), customer-provider (c-p) and sibling-2-sibling (s2s) we can confirm the intuition that p2p links are much harder to see than c-p and s2s links. In fact, the classes in our model have a strong correspondence to the link types.

However, there are caveats on the results. For instance, it is possible that there exists a class of links which are never observed. In

absence of *any* data, our technique obviously cannot estimate the size of this class of links. Hence, it is still possible that our approach underestimates the number of links present in the Internet.

There are many possibilities for extending this work. At the moment we only examine the number of links in the AS-graph, but it is clearly interesting to examine subsets of this graph, for instance to examine the node degree problem, or to estimate the number of backup links. We could also incorporate other data sources, or an improved model. However, perhaps the greatest potential for these ideas is in their extension to other Internet measurement problems, such as estimating the number of hidden anomalies.

2. BACKGROUND

2.1 Capture-Recapture

Simple capture-recapture can be used as follows. Imagine an unknown population E, from which we capture two samples E_1 and E_2. The initial samples are tagged so that we can also measure E_{12} the number recaptured in the second experiment. Assuming that the two experiments are independent, "fish" are all the same, the population remains static, and that tags to not "drop off", then we can make an estimate of E using Petersen's formula [4, 5].

$$\hat{E} = \frac{E_1 E_2}{E_{12}}. \quad (1)$$

The approach can be generalized in various ways [4,5], for instance by introducing K measurements (often referred to as K-lists), or by allowing for dependencies between monitors. The typical approach seems to be to use regression on a K dimensional contingency table generated from the measurements. Typically these techniques don't scale well for large K (given the 2^K entries in the table). Hence we will adopt a different approach in this paper.

2.2 Network Topology Measurements

One of the key ways to look at a network is to examine its connectivity, which can be captured in a graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, where $\mathcal{N}$ is a set containing the nodes (or vertices) of the graph. These represent, for example, routers switches or even whole networks. $\mathcal{E}$ is a set containing the edges of the graph, i.e., the connections between nodes. We define the numbers of nodes and edges by $N = |\mathcal{N}|$ and $E = |\mathcal{E}|$.

For instance, consider the case of the AS-graph, where nodes correspond to Autonomous Systems (ASes). There are many important details of inter-AS connections (i.e., the nature of the relationships between connected ASes, the behavior of the Border Gateway Protocol, etc.). Although it is possible to benefit from retaining these details [6], the simplicity of a graph based model has its attractions.

A common mistake in using such data has been to treat the observed data as completely accurate. It isn't! For example, a major source of observations of the AS graph comes from BGP monitors [7, 8]. Such monitors participate in the BGP routing protocol. BGP propagates the AS-path of a route, and this path provides us with information about the links in the AS-graph. However, BGP is a path-vector protocol, which means that only "best" routes are propagated. In contrast to a link-state routing protocol, we do not see the whole topology of the network, only those routes which are propagated. Hence, a BGP route monitor gets an incomplete view of the topology.

There are other sources of data that can be used to infer the structure of the AS graph. In this paper we will concentrate on BGP routing data, as this is the most up-to-date source of data. It allows us to see routing changes as they occur, and this is important because it allows us to see route exploration as a new route is propagated, or an old one withdrawn. This route exploration process reveals links that might otherwise be hidden. In addition, BGP monitors are unlikely to see a non-existent link (in comparison to other data sources such as registries or traceroutes) and this is a useful property. In this paper we use data from RouteViews [7] from Jan 1st 2004 until March 31st 2008. The data consists of all table dumps, and routing updates seen from all peers of RouteViews, grouped into one month periods. One month represents a reasonable tradeoff between obtaining a more complete view (through seeing additional updates over time), and the desire to measure dynamic properties of the network such as growth rates.

Each update, or table entry provides us with an AS-path, from which we can determine a set of links (and nodes) that are *observed* or *visible*. There are also sets of links which are active, but unobserved, and it is these links which cause all the trouble. We refer to these links as *hidden* links. Note that for an AS to be reachable at all, it must appear in observations, and so with any substantial set of route monitors the set of hidden nodes should be almost empty [9]. There may be ASes which we don't observe (private ASes for instance) but as these are not routable from the general Internet they are of limited interest. There is also the possibility that incorrectly configured filters will restrict our view of some ASes [10], but the lack of visibility means these ASes are still partitioned from the rest of the Internet, and so will be unlikely to play a practical role in the Internet topology.

Combining all of the observations of one monitor will reveal a set of edges $\mathcal{E}_{\text{obs}}^{(i)}$ where we drop the *obs* subscript where it is implied. Given K monitors, we observe $\{\mathcal{E}^{(k)}\}_{k=1}^{K}$. The typical approach to estimate the observed links is to take

$$\mathcal{E}_{\text{obs}} = \cup_{k=1}^{K} \mathcal{E}^{(k)}. \quad (2)$$

Implicit in (2) is the belief that $\mathcal{E}^{(k)}$ contains few false positives, but many false negatives (missing links). BGP measurements generally fall into this category because the information provided corresponds to real routes through the Internet. Exceptions arise where the protocol is abused, for instance by hackers seeking to hijack address space. More importantly, when measurements are taken over a period of time, they may include links which are not alive for the entire measurement period. However, it is commonly assumed that the number of false positives introduced in this way is small.

2.3 A quick and dirty refutation

We can make a quick estimate of Petersen's formula for the RouteViews data. A typical monitor for RouteViews (in October 2007) sees of the order of 45,000 links. The typical intersection between a pair of such monitors is somewhere around 40,000 links. We can easily calculate Petersen's formula (1) to be $\simeq 45,000 * 9/8 \simeq 50,000$ links. In fact, equation (2) indicates that at least 57,000 links exist. Though the calculation above is only rough, it is representative of the real results. Given the estimates fall well below a known lower-bound for the number of links, we know there is something wrong with this approach as it stands.

We also considered the simple K-list approach [4, 5] for this problem. Taking K to be the number of monitors produced a problem with an unrealistically large number of table entries ($\sim 2^{40}$) to estimate, but smaller values still showed very high variance in the estimates. For instance, we found that taking three monitors at a time produced estimates ranging from 10's of thousands up to millions of links depending on which three monitors were chosen. Given the flaws in the above approaches, we seek a better model.

3. A MULTI-CLASS MODEL

The simple capture-recapture approaches described above clearly fail. How can this be? Petersen's assumptions are

1. independence (between measurements)
2. homogeneity (across links)
3. the population is static (between measurements).
4. tags do not drop off.

Clearly at least one of these is violated. The forth assumption is valid for our measurements because we do not physically tag links – we simply use their unique identifier (the nodes they connect). The third assumption is also valid, because unlike a typical capture-recapture experiment, our measurements are all taken at the same time using different monitors. So the problem lies in the first two assumptions, which are in fact closely coupled (heterogeneity will introduce correlations).

It has often been postulated, and at least been partially confirmed [9], that some links are harder to see than others. For instance, a link that connects a stub-AS (a non-transit providing, single homed AS) to the Internet will always be visible (whenever the AS itself is visible). It must appear on any observed path originated by that AS. On the other hand, it has been postulated that the majority of the missing links are "peering" links. The simple c-p/p2p model of AS relationships, along with the resulting valley-free routing policy means that peering links between lower-tier ASes will only be visible from a small subsection of the AS-graph. Peering links are therefore considered harder to observe.

Our approach to include these facts is to use a stratified model. We assume that there are multiple classes of links, with different observational properties. This immediately violates the homogeneity assumption, and hence invalidates the independence assumption by introducing correlations between measurements dependent on class. We are left with a substantially weaker set of assumptions

1. conditional independence
2. homogeneity (across monitors)

The conditional independence assumes that we incorporate the majority of the correlation structure through the stratification of links into different classes. We do remove monitors with clear dependencies (for instance those monitoring the same AS) to avoid the majority of graph-structure related dependencies. Other dependencies such as related to "tiering" of the AS-graph, or geographic bias (because RouteViews is focussed in North America) are incorporated through the different classes of links. The second assumption is quite different from the earlier homogeneity condition in that it says that any monitoring point has the same probability of observing a particular link. We also remove any monitors which substantially violate this condition, e.g., monitors without a default-free feed. We are left with between about 30 and 40 route monitors (see Figure 4 for exact numbers).

We can model the above measurements using the Binomial Mixture Model (BMM), i.e., the number of observations of a link is a random variable formed by first choosing the class of the link, and then choosing the number of observations of that link based on a Binomial distribution with class dependent observation probability.

3.1 Estimation of parameters: known class

We first consider estimation of parameters assuming we know the class of all links. In this situation, we can consider each class independently as if it followed a single class model. If we have K independent, homogeneous monitors, then the number of times we observe a link in class j will follow a Binomial distribution $B(K, p_j)$ where p_j is the probability that we observe the link with a single monitor. The probability that we observe a class j link k times with K monitors is

$$\text{prob}\{k\} = \binom{K}{k} p_j^k (1-p_j)^{(K-k)}. \tag{3}$$

If we knew that class j (which we denote C_j) has E_j links, then the Maximum Likelihood Estimator (MLE) of p_j given link i is observed k_i times out of K is

$$\hat{p_j} = \frac{\sum_{i \in C_j} k_i}{E_j K}. \tag{4}$$

However, we do not know E_j *a priori*. We only know about links that are observed at least once. Measurement of E_j is equivalent to estimating the number of hidden links!

Ignoring for the moment the class (i.e., the subscript j), in fact what we really observe is the conditional distribution

$$\text{prob}\{k|k>0\} = \binom{K}{k} \frac{p^k (1-p)^{(K-k)}}{1-(1-p)^K}. \tag{5}$$

This is commonly known as a truncated Binomial distribution [4], and estimation of its parameters will lead to estimates of $\text{prob}\{k = 0\}$, and hence an estimate for the number of hidden links.

The MLE for the truncated Binomial distribution is given in [11]. However, there is no simple closed form description of the MLE, but rather the MLE estimator $\hat{p}$ will be the solution to the equation

$$E_{\text{obs}} K p = [1-(1-p)^K] \sum_{i=1}^{E_{\text{obs}}} k_i, \tag{6}$$

where E_{obs} is the number of observed links. In the preceding statistics literature (which dates from as far back as the 50's) some effort went into algorithms and tables to solve this equation without computers. Given modern computing resources it is rather easier to use a simple iterative solution. To find the value of $\hat{p}$ which satisfies this equation we take

$$\begin{aligned} \hat{p}^{(0)} &= \frac{\sum_{i=1}^{E_{\text{obs}}} k_i}{E_{\text{obs}} K}, \\ \hat{p}^{(i+1)} &= \frac{\sum_{i=1}^{E_{\text{obs}}} k_i}{E_{\text{obs}} K} [1-(1-\hat{p}^{(i)})^K]. \end{aligned} \tag{7}$$

We can easily prove $\hat{p}^{(i)}$ converges to a unique fixed point satisfying (6). In practice we found that it converged quickly. The fact that it is a MLE estimate guarantees that it is asymptotically unbiased and efficient. In practice we found that for E as small as 1000 the bias is very small, and the mean-squared error of the estimate is very close to the Cramér-Rao lower bound. Additionally, tests of the errors showed that they were approximately Gaussian. We omit these results to save space, and because they are implicit in following results. An additional implication is that $\hat{E} = E_{\text{obs}}/(1-(1-\hat{p})^K))$ will be a MLE estimator for the total number of links.

3.2 Multi-class observations

In the problem above, we assumed that we knew the classification of the links. We don't *a priori* know this classification, and so we construct an Expectation Maximization (EM) estimation algorithm (see for instance [12]) which estimates both the class, and the class models.

The EM algorithm is an iterative approach that uses two steps: (i) an Expectation step in which we calculate expected values of some "hidden" variable (in this case the class of the links) and (ii) a Maximization stage where we perform a MLE of the system parameters. In more detail define two additional parameters

$$\begin{aligned} w_j &= \text{estimated proportion of links in class } j, \\ c_j^{(i)} &= \text{estimated probability}\{\text{link } i \in \text{ class } j\}. \end{aligned}$$

We start the algorithm by initializing $\hat{p}_j$ and w_j the estimates of the important parameters for our distributions. The choice of initial

Class	p_j	w_j
1	0.010906	0.248714
2	0.140579	0.052389
3	0.345960	0.036864
4	0.557597	0.049963
5	0.758552	0.060776
6	0.917098	0.068741
7	0.998352	0.482553

Table 1: Model parameters for $C = 7$ simulations. The parameters are those found from the EM algorithm applied to the AS-graph data from October 2007.

conditions is not particularly important, though choosing parameters closer to the true parameters will speed convergence. We use the uniform initialization $\hat{p}_j = j/(C+1)$ and $w_j = 1/C$, where C is the number of classes.

The algorithm then acts as follows:

```
While (not converged) do
  E step
    estimate c_j^(i)
      c_j^(i) ← ŵ_j P{k_i|K, p̂_j}
  M step
    for j=1 to C
      While (not converged) do
        p̂_j ← (Σ_i k_i c_j^(i)) / (K Σ_i c_j^(i)) [1 − (1 − p̂_j)^K]
      end while
      ŵ_j ← Σ_i c_j^(i) / (E_obs (1 − (1 − p̂_j)^K))
    end for
end while
```

where $P\{k_i|K, p_j\}$ is the Binomial distribution $B(K, p_j)$ given in (3), k_i is the number of observations of link i, and E_{obs} is the total number of links observed.

Convergence occurs when the total change in the estimates $\hat{p}_j$ falls below $\epsilon = 10^{-6}$. The EM algorithm is in general guaranteed to converge, and in our example we find it converges (for this case) reasonably quickly (results below).

We also perform a hard classification of observed links for use in assessing the relationship between link class and policies. We select the class with the highest likelihood, i.e., the class of link i is $\text{argmax}_j c_j^{(i)}$. The estimated number of observed links in each class is defined to be $\hat{E}_j = E^j_{\text{obs}}/(1-(1-\hat{p}_j)^K)$, where E^j_{obs} is the number of links observed in each class.

3.3 Performance

The first test of the above algorithm is its performance when the model (e.g., the value of C) is correct. We use a set of realistic parameters derived from the AS-graph data (for October 2007) and shown in Table 1. We will use a total number of observed links $E = 50,000$, which lies within the range of observations over the time interval of our data. We simulate the observations of the links 5000 times using the BMM with $C = 7$. Results for estimates of $\hat{E}$ are shown in Figure 1 in the form $100\hat{E}/E$ so that we can see the relative errors as a percentage. We can see that bias is very small and that errors are of the order of $\pm 1\%$, and approximately Gaussian. The hard classification of the links was correct for 94% of links (across all 5000 simulations, and 50,000 links).

4. MODEL SELECTION

When we consider real data, there is an additional problem: C is unknown. From the perspective of simplicity we can easily argue

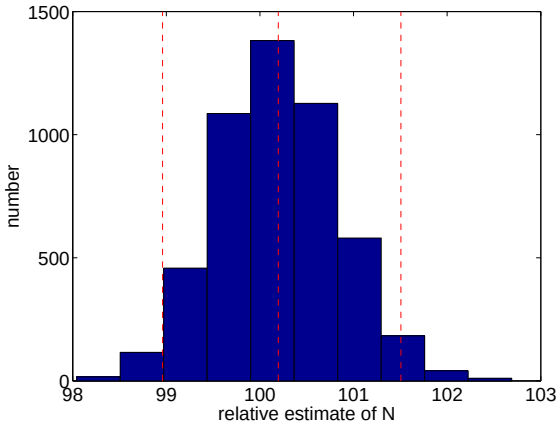


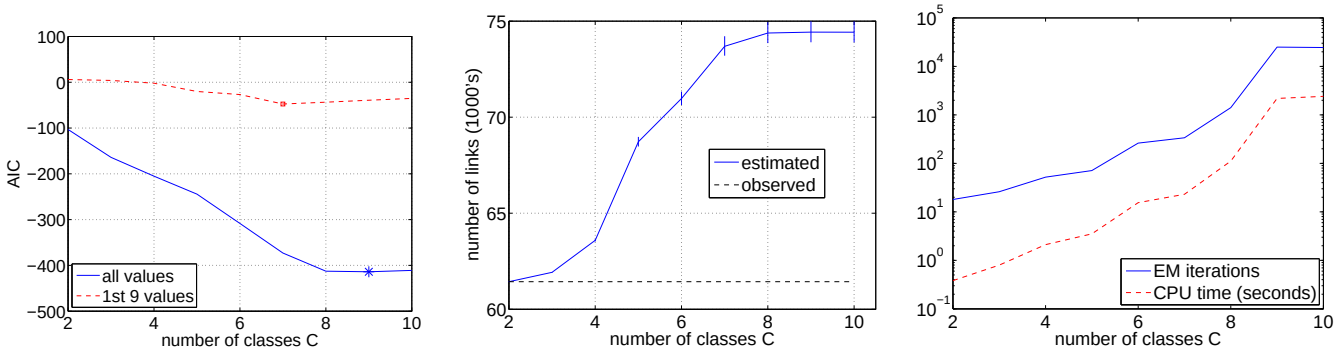
Figure 1: Results of 5000 simulations of EM estimation algorithm with 7 classes (as shown in Table 1). Vertical lines show the 2.5th and 97.5th percentile, and mean.

in favor of smaller C, even if it creates small errors in the model fit to the data, and so we need to trade off model accuracy with the quality of the fit to the data. This tradeoff is often captured through information criterion, e.g., the Akaike Information Criterion (AIC). In the context of normally distributed errors, it is defined by AIC $= n[\ln(2\pi \text{RSS}/n) + 1] + 2P$, where RSS is the Residual Sum of Squared errors, n is the number of data points and P is the number of model parameters. The minimum value of the AIC can be used to select the model that best satisfies the trade off between model simplicity and accuracy. We compare the estimated and empirical values of the truncated BMM's distribution, i.e., prob$\{k|k > 0\}$, so $n = K$ and $P = 2C$. We also calculate a second version of the AIC, recognizing that the critical values (for estimating the number of hidden links) of the BMM distribution are those corresponding to class 1 and 2, and hence we calculated a RSS for the first 9 elements of the distribution. Figure 2 (a) shows the two AICs. They take their minima for $C = 9$ and $C = 7$. Figure 2 (b) shows the estimated number of links with respect to C. Note that it varies insignificantly (with respect to the 95th percentile confidence intervals shown on the plot) for $C = 7{-}10$. Hence in tests (e.g., in Section 3.3) we have used $C = 7$ because of the reduced computational cost (see Figure 2 (c)).

Figure 3 shows the observed distribution of link observations, and the estimated BMM for $C = 7$. We can see that the fit is quite satisfactory. Note that the 0 term of the histogram does not appear in the observations because this data point is censored by our very lack of observations. The BMM extrapolates this value estimating the number of hidden links. The parameters of this distribution are shown in Table 1.

5. HOW BIG IS THE AS-GRAPH?

We now have the required results to answer the question of interest: how big is the Internet? More precisely, "how many links are there in the AS-graph?" We use the above algorithm choosing C based on the AIC test. Figure 4 shows the number of observed and inferred links since January 2004, along with the number of usable monitors. We can see that the number of monitors has not changed much, but that there is clear growth in both the observed, and estimated number of links. The trend is approximately linear, as shown by the linear trend fitted to the data. The trend avoids some of the potential problems with variance of individual estimates (careful examination of the largest deviations in early 2005 suggests that



(a) AIC values; the marker indicates the minimum value.

(b) Estimated number of links.

(c) Number of iterations, and the CPU times (in seconds).

Figure 2: AS graph estimates for October, 2007 with respect to the number of classes C.

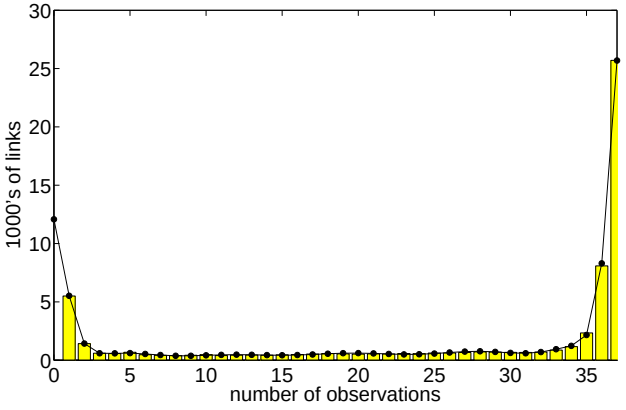


Figure 3: The fitted distribution for $C = 7$. Bars show the number of links with respect to how often each is observed, and the curve shows the estimated distribution.

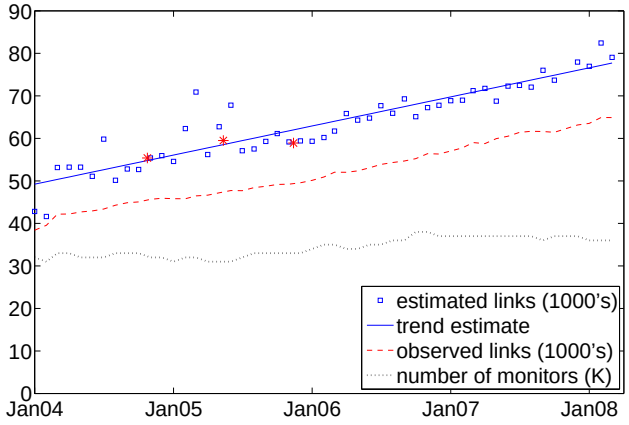


Figure 4: The trend in the numbers of links. The asterisks show previous estimates from [6, 9, 13].

Paper	dataset label	date	$\hat{E}$
Zhang *et al.* [13]	Updates(1M)	2004-10-24	55,388
He *et al.* [9]	All	2005-05-12	59,500
Mühlbauer *et al.* [6]	N/A	2005-11-13	58,903

Table 2: Past estimates of links in the AS-graph.

these are model selection errors). The trend has a growth rate of 18.7 links per day, as compared to the growth in observed links of 16 per day. On a larger scale our results give a yearly growth rate of around 1000 hidden links.

Several prior studies have attempted to estimate total numbers of links: 3 such are shown in Table 2. They have used additional sources of data: e.g., Internet Routing Registries (IRRs), and Looking Glasses, as well as additional route monitors (1000 such in [6]). IRRs in particular introduce the possibility of "false positives", and so they may have overestimated of the number of links. Figure 4 shows each of these estimates by and asterisk '*'. We can see that the first and third are very close to our own estimate, while the second is very close to the linear trend. While not absolute proof, the correlation between the results of different approaches suggests that all are producing reasonably accurate views. For instance, our results suggest that He *et al.* [9] have eliminated the vast majority of false positive links. Good news!

5.1 Do classes have meaning?

Until now, we have treated classes as an abstract division of the links. Our classes were motivated by p2p vs c-p links, but were constructed without any reference to any model of inter-AS policies. The classes incorporate a range of factors including geographic and topological bias, but the prime motivation stemmed from previous work showing that peer-2-peer links are harder to see that customer-provider links. A natural question is "To what extent do our classes reflect actual link policies?"

Dimitropoulos *et al* [14] provides a recent classification of links into types: p2p, c-p, and s2s. We cross-match their types to the links in our data. Table 3 shows the breakdown of each type of link into classes. We can immediately see that class 1 links are largely made up of p2p links, with a few p2p links appearing in class 2. The more easily observed classes are dominated by c-p and s2s links in roughly similar proportions.

These results reflect two findings. Firstly, a significant determining factor of link class is the role for which a link is used, so our classes do have meaning. The class is not a 100% classifier of the link policy, though. There are some p2p links in the easily observed classes, presumably because their place in the network topology makes them easily viewed, though some may be classification errors, either in our algorithm, or in that of Dimitropoulos *et al* [14]. Secondly, the results clearly show that p2p links are much harder to observe, supporting intuition generated by previous studies.

Class	p2p	c-p	s2s
1	74.25	2.32	2.33
2	14.50	4.27	6.98
3	2.66	4.01	5.81
4	1.75	6.31	0.78
5	1.43	8.92	3.49
6	1.09	5.04	5.81
7	4.33	69.12	74.81
observed links	4830	48760	258
estimated links	15990	57400	300

Table 3: % of classes by link policies. Note that almost 90% of p2p links fall into class 1 or 2, whereas well over 90% of c-p and s2s links fall in classes 3-7.

6. HOW MANY MONITORS?

The most obvious question, following the above analysis, is how many monitors should we have? Under the above conditions, we can answer this question. Let us seek to guarantee that we will observe a given link from class 1 with probability $1 - q_j$. How many monitors would we need?

We focus here on class 1, because the observation probability of the other classes is so much higher for any non-trivial number of monitors. For small probability q_1 of missing class one links, the probability of missing any of the other links will be so much lower it will be negligible. The probability of observing a class 1 link, with K monitors, under the above model is simply $1 - (1 - p_1)^K$, where p_1 is the class 1 link measurement probability from a single monitor, which we have shown to be around 0.01. So we want the minimal value of K such that $q_1 \leq (1-p_1)^K$. Rearranging we get $K = \lceil \ln(q_1)/\ln(1-p_1) \rceil$. For $q_1 = 0.05$, 0.01 and 0.001, we get $K = 299, 459$ and $K = 684$, respectively.

7. CONCLUSION

This paper provides us with a way of estimating the number of hidden links in an AS graph. The technique performs well against the best data we have for validation, and allows us to estimate the trend in the size of the Internet — according to our data it grows at 18.7 links per day. We use the model to derive the number of (well placed) monitors that would be needed to see a complete AS-graph with high-probability. We estimate that 700 route monitors would see at least 99.9% of links. The results also support previous studies, and their intuition that peer-2-peer links are much harder to see than customer-provider links.

There are problems we still wish to explore here, for instance, the underlying assumption of our approach that the stratified model captures the dependencies and heterogeneity of our measurements is only approximate, and we wish to further improve the model. Moreover, it is possible that some entire class of links could be missing from current measurements, and we need to investigate this further. True validation of the results over the entire Internet is impractical (for exactly the reasons that make this technique worthwhile), but it is possible that the methodology could be validated on a smaller segment of the Internet, for which precise ground-truth data was available.

Finally, this type of technique could be extended to a large number of Internet measurement problems that have a similar character. For instance, in anomaly detection, we might see our anomalies (say a worm, or DoS attack) through some set of monitors. We could use capture-recapture based techniques to estimate the number of anomalies we are missing.

Acknowledgement

Olaf Maennel was supported in this work by ARC grants DP0557066 at the University of Adelaide. The data used in this paper was derived from the Oregon RouteViews project. We would also like to gratefully acknowledge useful conversations with Randy Bush regarding this work.

8. REFERENCES

[1] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power-law relationships of the Internet topology," in *ACM SIGCOMM*, (Boston, MA, USA), 1999.

[2] J. C. Doyle, D. L. Alderson, L. Li, S. Low, M. Roughan, S. Shalunov, R. Tanaka, and W. Willinger, "The "robust yet fragile" nature of the Internet," *Proceedings of the National Academy of Sciences of the USA*, vol. 102, pp. 14497–502, October 2005.

[3] A. Lakhina, J. Byers, M. Crovella, and P. Xie, "Sampling biases in IP topology measurements," in *Proc. IEEE Infocom*, April 2003.

[4] S. E. Fienberg, "The multiple recapture census for closed populatiosn and incomplete 2^k contingency tables.," *Biometrika*, vol. 59, no. 3, pp. 591–603, 1972.

[5] International Working Group for Disease Monitoring and Forecasting, "Capture-recapture and multiple-record systems estimation i: History and theoretical development," *American Journal of Epidemiology*, vol. 142, no. 10, pp. 1047–1057, 1995.

[6] W. Mühlbauer, A. Feldmann, M. R. O. Maennel, and S. Uhlig, "Building an AS-topology model that captures route diversity," in *ACM SIGCOMM*, (Pisa, Italy), 2006.

[7] "University of Oregon Route Views Archive Project." `www.routeviews.org`.

[8] "Ripe NCC: routing information service raw data." `http://www.ripe.net/projects/ris/`.

[9] Y. He, G. Siganos, M. Faloutsos, and S. V. Krishnamurthy, "A systematic framework for unearthing the missing links: Measurements and impact," in *USENIX/SIGCOMM NSDI*, (Cambridge, MA, USA), April 2007.

[10] R. Bush, J. Hiebert, O. Maennel, M. Roughan, and S. Uhlig, "Testing the reachability of (new) address space," in *INM'07: Proceedings of the 2007 SIGCOMM workshop on Internet network management*, (New York, NY, USA), pp. 236–241, ACM, 2007.

[11] P. F. Rider, "Truncated binomial and negative binomial distributions," *Journal of the American Statistical Association*, vol. 50, pp. 877–883, Sept. 1955.

[12] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, 2001.

[13] B. Zhang, R. Liu, D. Massey, and L. Zhang, "Collecting the Internet AS-level topology," *ACM SIGCOMM Computer Communication Review (CCR) special issue on Internet Vital Statistics*, January 2005.

[14] X. Dimitropoulos, D. Krioukov, M. Fomenkov, B. Huffaker, Y. Hyun, kc claffy, and G. Riley, "AS relationships: Inference and validation," *ACM SIGCOMM Computer Communication Review (CCR)*, vol. 37, no. 1, pp. 29–40, 2007.