What Does the Mean Mean?

Matthew Roughan[a] and Oliver Spatscheck[a]

[a]AT&T Labs – Research,
180 Park Av., Florham Park, NJ, 07932,
{roughan,spatsch}@research.att.com

As the Internet has become a big business, its performance has become an important question. Several companies have started monitoring Internet performance, many ISPs now conduct internal performance measurements, and there are also independent organizations such as RIPE that conduct regular performance monitoring of ISPs. The exact form of measurements may vary, but typically they are based on active probing. A common result presented is an overall statistic combining the measurements over a whole network, and this is frequently used to rate, or rank ISP performance, and is sometimes used as a metric for overall network health. Of course, a single number cannot hope to validly represent all of the information gathered in these types of measurements, but nevertheless, this type of measure is used, and we should ensure that the methodology for compiling the measurements is robust, and appropriate. Most methods that have been used are based on the mean, but it is surprising how diverse the "mean" can be. There are several alternatives (e.g. arithmetic mean versus geometric mean), and each company applying these methods seems to use a different approach. This paper considers five approaches in detail and, explains which of these is the best, and why. In particular, we show that the geometric, and harmonic means, while appealing because of their robustness to outliers, are actually very poor statistics to use when combining Internet measurements, and can cause changes in apparent performance without any real change in the network performance.

1. Introduction

As the Internet has become a big business, its performance has become an important question. Several companies have started monitoring Internet performance, and publishing results. Further many ISPs now conduct internal performance measurements, and there are also organizations such as RIPE that conduct regular performance monitoring. Typically, the measurements use active probing to determine performance, though the exact form of measurement may vary, for instance: TCP connection time, DNS lookup time, ICMP echo time, file download time, application performance, or some purpose designed protocol. In addition to the times measured above, one can, in some cases measure packet losses, but we shall concentrate here on the time based measures.

Each method has advantages, and disadvantages, but these details are not of prime importance in this paper, which is instead concerned with how to compose these measurements together to get overall measurements of performance. That is, given a set of geographically and topologically diverse measurements, how do you combine these to get an overall measure.

Obviously, in combining the results of many measurements there is data loss. Ideally, one would consider the individual measurements separately, to retain this information. However, there is a valid reason for wanting a single metric: it is helpful to assess overall health of a network over time, e.g. to see trends in performance. One can then examine the details to see the causes of certain behavior. Of more dubious merit is the natural desire to use a single number so as to "rate" different ISPs. Many companies now perform this type of rating. We say that it is of dubious merit for many reasons:

- A single number means different things to different people (VPNs care about forwarding across a backbone; web server customer wants good peering, and access links as well).

- A single number can be deceptive. Combining lots of issues (geography, peering, access, back bone, web hosting, ...) into one data point can obscure some details, and highlight others.
- Related to the above point, there is a paradox, referred to as the voting paradox where you can always construct situations where two different voting schemes will report a different result in an election – or in this case, in ranking two Internet Service Providers (ISPs).
- If the result isn't weighted by a representative number of customers in a location, does it mean anything? If you do weight, then the number of customers per location varies by backbone.

However, it seems that companies, and individuals will continue to compose results into a single numbers, despite these problems. Anecdotally, some professionals now have their performance (and thus bonuses) based on their web sites rated performance, as measured by a single company, and a single metric. Hence we cannot ignore these aggregated metrics.

Given that this approach will be used, a reasonable question is "how may we do this with the least loss of information?" The typical approaches are to use a statistical measure, and we have seen three such used in this way: the Arithmetic Mean (AM), the Geometric Mean (GM), and the median. We examine these here, along with the Harmonic Mean (HM) and Trimmed Mean (TM). There are still further alternatives (e.g. the worst case, ...), but we cannot examine the full range of possibilities here and so concentrate on measures of the scale of the distribution. The AM (or average) is the measure with which most people have experience. The stated reasons for measurement companies using alternatives are related to their robustness to outliers.

The principle highlighted here is that the measurements should measure a property that a customer is interested in. For instance, if a customer is interested in the customer latency of a web site, the measurement should be an average weighted by the number of people viewing his page from each region. However, each type of customer will have diverse requirements (some care about worst case, others about average, and weightings would differ by customer), and some may not even know what they need to measure. As an example of the complexity here, the utility of a web site is not linearly proportional to the RTT to the web site, but rather, the RTT's exact value is almost irrelevant as long as it lies below some threshold, and once it crosses that threshold the utility quickly decreases. In this case, the best performance metric might be the number of cases which exceed the threshold. Part of the aim here is to elucidate the properties of the statistics presented, to allow an informed choice when deciding what to use for a particular application.

Our other aim is to make some suggestions about the methodologies used in making measurements. The schemes considered may be hierarchical: it not necessary to use the same statistic to aggregate results from a single path or locality, as that used to aggregate data from different paths or localities. This paper considers shows that when combining measurements it makes sense to use the median, trimmed or geometric mean to combine results from a single path (due to their robustness to outliers), but in combining diverse paths we should use an arithmetic mean (possibly weighted). The arithmetic mean is the least susceptible to distortions that don't really represent performance. That is, with other statistics the apparent performance may change through minor changes that do not actually change the network performance except in details that the particular statistic is sensitive to. The reason the arithmetic mean is least susceptible when combining geographically diverse data is that it performs in the way closest to the average performance seen by a customer, or user, whereas statistics like the geometric mean have little relation to a real customer.

The paper starts by providing some statistical background in Section 2. Then, in Section 3 we consider the effect of aggregating measurements from geographically, and topologically different paths. The noteworthy fact is that the statistics do not perform the same when analyzing disparate distributions as they do in analysing a single distribution. In Section 4 we combine local and global information together and show which combinations have the best performance.

2. Statistical Background

The purpose of this section is to describe the basic statistics (AM, GM, median, TM, and HM) examined here, and illustrate the properties of each, in particular the susceptibility/reliance of each to outliers.

These statistics are well known, and their properties have been previously elucidated in many places, as have the reasons each is typically used. However, in the context of Internet measurements, a number of companies have started using statistics such as the GM due to its robustness to outliers. We argue here that, while this is a valuable aim in any area where the data are inherently heavy-tailed, there are alternatives other than the GM which are also robust, in particular the median and TM, and as we shall see later in this paper there are reasons to avoid the GM here.

Take a set of N data samples X_i all independently drawn from the same statistical distribution. The arithmetic mean (or sample mean) is $\text{AM}(X) = \frac{1}{N}\sum_{i=1}^{N} X_i$, The AM is frequently used to combine data to reduce the impact of errors in measurements. That is, if we wish to measure some quantity X, and the measurements contain errors ε_i, we can obtain a more accurate estimate using the AM. The reason we can do so is the law of large numbers, which says that (under suitable conditions) the AM will converge to the true value X as the number of measurements becomes large. More generally, the Central Limit Theorem (CLT) shows the error in the result becomes normally distributed for large N with known variance. In this application, the AM would be the equivalent of trying to measure some underlying performance parameter of the system, using a number of observations. More generally, the basic task is to find some measure of the central tendency of the measurements.

One of the conditions of the CLT (and law of large numbers) is that the the distribution of the random variables in question has finite variance. Distributions with "heavy-tails" may not satisfy this criteria, but even when they do, the heavy tail results in a *very* slow convergence rate, and so many samples are required [1,2]. A heavy-tailed distribution is one in which there is a significant probability of a large event. A typical example is the Pareto distribution, which has distribution function $F(x) = 1 - (b/x)^{\alpha}$, where α is a shape parameter, and b a scale parameter. The Pareto distribution only has finite variance for $\alpha > 2$, and in fact, the AM of Pareto random variables does not converge at all for $\alpha \leq 1$. Such distributions have been shown to be fundamental to most of traffic modeling [3], and have been suggested to occur in active performance measurement.

For such distributions the AM has poor properties – it may not converge at all, and even when it does it requires very many samples. This is a stated reason for avoiding the use of the AM. An obvious alternative is the median, or 50th percentile of the distribution, that is, the value x for which the distribution function $F(x) = 0.5$. The median is an obvious choice because it depends on the body of the distribution, not the tail.

There may, however, be some concern that the median explicitly fails to capture any of the tail behavior. Some people may wish to have the tail represented in their measure of central tendency, though still avoiding the problems of the AM. This has lead to the choice (in at least one measurement company) of the GM which is given by $\text{GM}(X) = \sqrt[N]{\prod_{i=1}^{N} X_i}$, which can be more easily calculated by taking the exponential of the sum of logs. The normal use of the GM is where the values measured are multiplied together to obtain some larger measure – for instance, in computing the average growth over three years, one would multiply the percentage growth in each of the three years, rather than adding. However, taking the log of the data reduces the length of the tail – for instance, the log of Pareto distributed data has an exponential distribution – and so the GM is less sensitive to outliers in a set of measurements, but it does not discount the tail altogether.

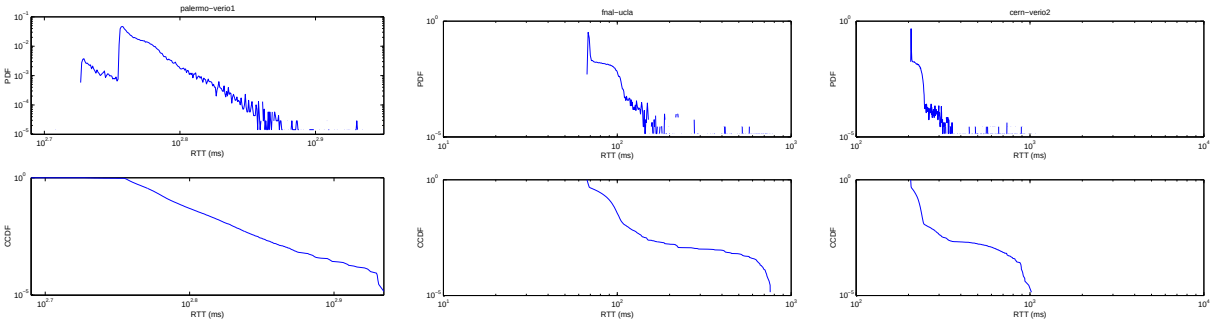
Another alternative that has similar properties to the GM is the $\text{HM}(X) = (\frac{1}{N}\sum_{i=1}^{N}\frac{1}{X_i})^{-1}$, which is more robust to outliers than the GM. The HM has not been used in Internet measurements, but we include it here for comparison.

Finally, the Trimmed mean (TM), which is the AM after the upper and lower α percentiles are removed from the data) is quite robust to outliers, while only omitting a small part of the distribution. We shall remove the upper and lower 5 percentiles (rounded up) here.

2.1. Real RTT measurements

If the distribution of the quantity of interest (RTT, TCP connect time, etc.) were Pareto, or log-normal we would have a natural reason to examine the data on a log scale, and hence the GM is a reasonable metric. The question is whether typical Internet metrics take this form. Some aspects of Internet measurement do – for instance the length of Internet flows exhibit heavy, power-law tails [3].

However, the literature on the distributions of performance metrics seems more limited. There majority of current literature examines correlations between measurements, e.g. [4,5]. In order to fill the gap, we have used data gathered by NIMI (the National Internet Measurement Infrastructure) [6–8] by Vern Paxson and Yin Zhang. The dataset provides the RTTs between a sample of the approximately 50 nodes of NIMI for three days in January 2001.



(a) palermo to verio1 (b) nether to virginia (c) fnal to ucla

Figure 1. Examples of the NIMI RTT data.The graph shows the PDF and CCDF on log-log axis.

Unfortunately, there is no simple distribution that works here – to see why see the examples in Figure 1. The distributions can be multimodal, and have several regimes in the tail. From observing over 100 of these, it is clear that while the distribution tails are considerably heavier than exponential, they could not be described as having a power-law tail (except over some range, in some cases). In hindsight this should be obvious: the time a packet can exist in the Internet is limited by the TTL, and so any distribution of packet delays will be truncated. When one computes the statistics above on these measured distributions one finds that the impact of the tail is minimal – in only 8 of 176 cases was there any marked deviation between the statistics, and even in these case it was not large.

Given the above results, why would anyone consider using anything but the AM? On reason is data errors – outliers not reflecting real network performance. Despite one's best efforts to do experiments without such artifacts, they may appear.

A second reason is that in some cases the measurement is not at the IP layer, but at the TCP, or application layer, which may introduce more complex delays than are seen by simply measuring IP performance. TCP connection times are sometimes used to measure performance, because they can be made to almost any host. However, if the TCP SYN packet is lost, the sender waits for a timeout before retransmitting the packet (and similarly the receiver waits for a timeout before resending the SYN-ACK). The initial timeout is quite large: measurements suggest that 3 seconds is common, and it doubles as each successive packet is lost. Hence a single lost packet can increase the response time by two orders of magnitude (from around 30-50ms to 3 seconds), resulting in a fairly heavy-tailed distribution, albeit one which is not well modeled by a Pareto distribution.

We can simulate this using a RTT that is normally distributed around some mean (50ms in the example presented), but with some probability p that the SYN packet (or its response) is lost, in which case we add 3 seconds to the RTT[1]. Simulations of the statistics showed once again that the median and GM, and HM are quite unaffected by the unusual outlier events, while the AM is strongly affected by p, and also does not converge quickly. The TM is stable until $p = 0.05$ when the probability of a lost packet becomes higher than the trimmed part of the distribution, whereupon it quickly converges to the AM.

3. Geographic and Topological Aggregation

In the previous section we were estimating some quantity (for instance the RTT between two nodes), where the measurements had some error, or variability around the "true" value. A statistic like the AM obeys limit theorems that cause it to converge to a single value that describes the central tendency of the

[1] A better model is presented in Cardwell, Savage and Anderson [9] but the full details of the results are not needed here.

distribution, and is therefore associated with the underlying quantity we wish to measure. However, we saw that for some distributions – those with heavy-tails – the AM diverged, or did not converge quickly, which provided the motivation for using the alternative statistics.

In this section we consider what happens when you use these statistics to combine estimates from a number of paths. In essence, we are now using the statistic to combine data from different distributions. While one may argue about the weight that should be given to outlier measurements from a single path or locality, individual paths should never be considered outliers – each is distinct and important to incorporate in the results. We do not wish to discount the RTT measurements between two important nodes simply because it is unusually large.

3.1. Simple illustrative simulation

We start with a simple illustration. We use a very simple simulation of a network with no buffering, or forwarding delays, and propagation delays which depend only on the shortest surface path distance between nodes. The typical methodology of a company wishing to measure Internet performance use a set of *monitors* (at geographically dispersed points) to make measurements to some set of *servers* in the network to be measured.

In this simulation we have three monitors (at Atlanta, New York, and San Francisco) and three servers (at Chicago, New York, and San Francisco). The three server points are chosen because they are three of the largest cities in the US (from the perspective of traffic) and each has good peering connectivity, and so are likely places for servers. In particular, all else being equal, one might expect Chicago to be the best choice of the three for a server location, because a server on the east or west coast would be far from the large population centers at the opposite coast, while Chicago is not so far from either. The RTTs between these locations were calculated using a program called `geod` [10] to compute the distance between cities. Note that the RTT between a server and monitor in the same city is some small time Δ.

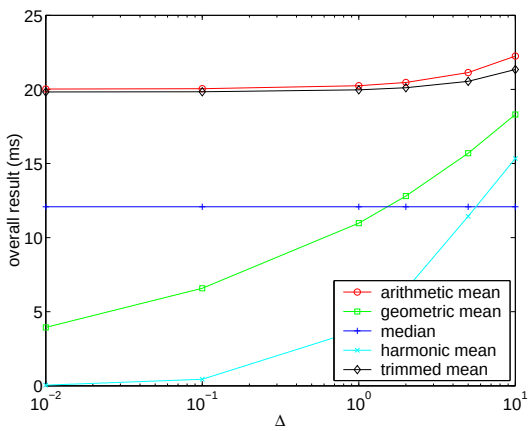
The statistics for each server, and overall and shown in Table 1 for $\Delta = 0.01$ ms, which would not be unrealistic if the server and monitor were on the same LAN segment. The most noteworthy point is that though the GM and HM suggests that Chicago is the worst city (by an order of magnitude) to place a server, the other statistics support the intuition that Chicago is the best place to put a server.

Server	AM	GM	median	HM	TM
Chicago	16.944	14.789	11.478	13.229	11.478
New York	17.836	1.710	12.080	0.030	12.080
San Fran.	25.296	2.426	34.460	0.030	34.460
overall	20.025	3.944	12.080	0.045	19.828

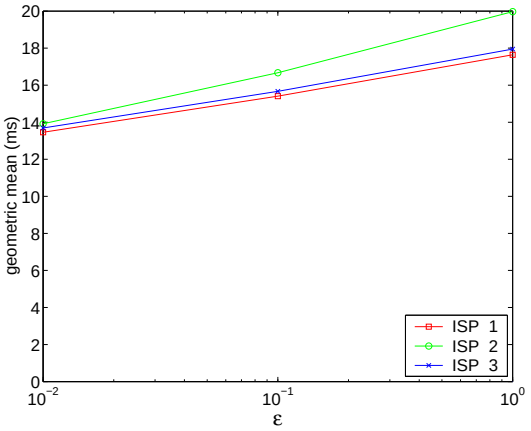
Table 1
Results by server ($\Delta = 0.01$).

Why is there such a difference between the inferences of different statistics. We can discover the reason by examining their behavior as we vary Δ. Figure 2 (a) shows the statistics over all the monitors and servers. The x-axis show the log of Δ. We can see that as Δ varies, the median remains constant, and the AM varies only a little, but the GM varies quite considerably (by more than a factor of 3). Thus we see that the GM is sensitive to the short delay between a monitor and server in the same city. The reason that Chicago fairs so poorly in the GM is simply because there is no monitor in Chicago!

To summarize, the GM is highly sensitive to the smallest measurement, which will typically be between a monitor and server in the same city (and same peer). This is hardly the "overall" measurement that we desired. If this were a purely geographic problem, we might be able to allay concerns in some manner, but consider the situation with respect to peering. If there is a monitor in a city in a single peers network, a comparison between peers would show an unjustified improvement to the network performance for the peer containing the monitor, leading to statistics that favour networks containing more monitors. Adding more monitors is not a suitable method for fixing this problem. The sensitivity still remains, but the measurements will then be complicated enough to obscure the source of the discrepancy.



(a) Sensitivity analysis



(b) Comparison of three ISPs with servers in NY

Figure 2. Two examples.

3.2. A second simple example

As a second simple example, we consider a comparison between two ISPs, the first a backbone provider, and the second a customer of the first, with only a single PoP, and a poor connection to the backbone. However, the customer (based in NY for the purpose of this discussion) has a monitor agent on the same LAN segment as its server, while the tier 1 ISP has its servers in a hosting center one (short) step from the backbone (note monitors are located on the backbone with negligible additional delay). We shall use the same monitoring points as before (with the addition of the second monitor in NY for ISP 2), and consider a comparison of servers in NY city.

We will simulate this as above, using geographic distance to compute the delays between servers and monitors, but with an extra RTT delay Δ associated with transfer from the tier 1 ISP to its hosting center, a RTT delay δ to ISP 2 from ISP 1, and a RTT delay ε between ISP 2's monitor and server. Figure 3 shows the topology and the network delay times.

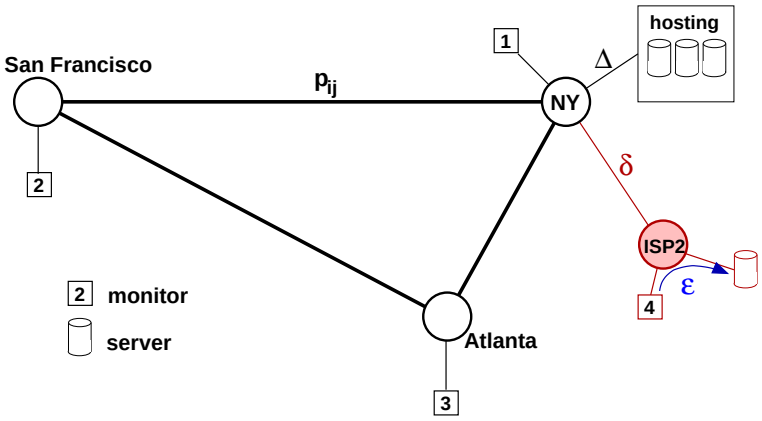


Figure 3. Example 2: two ISPs. The customer ISP (ISP 2) is has the single shaded spur node in NY.

The aggregate statistics for each server are also shown in Table 2, given $\Delta = 0.1$, $\delta = 5$, $\varepsilon = 0.01$. The large backbone ISP has a clear advantage in the AM, TM and median measurements because of the extra 5 ms delay on ISP 2's link to the backbone. However, when we consider the HM and GM, ISP 2's performance appears to be substantially better. Clearly, the small delay ε between the monitor and server on ISP 2's network is dominating the performance measure.

ISP	AM	GM	median	HM	TM
1	14.724	4.007	8.640	0.388	8.640
2	17.127	2.509	11.040	0.040	11.040

Table 2
Aggregate statistic for example 2 (given in milliseconds).

3.3. A more realistic simulation

It might be easy to dismiss the above arguments by saying that a typical monitoring company has many more monitoring points, and so the problem goes away. To illustrate that this is not the case, we provide a more realistic example. In this example, we use the Keynote 25 (25 monitoring points based in the USA) as our monitoring points. To make the example even more realistic we do not use simple geometric distances to estimate delays, but rather we use a set of real measurements between the Keynote agents themselves to estimate the propagation delay between agents (via the min RTT measurement). It is assumed that each server will be located a short delay ε from the corresponding agent i. We could conduct a number of experiments on this basis, but the most illustrative is to do a comparison of servers based in different ISPs, but in the same city (we use real ISPs for this purpose but anonymize them because otherwise these results might be subject to misinterpretation – they are not real measurements of the ISP performance). We chose NY city as a good illustrative example because several ISPs have Keynote agents there (but any other city would serve).

We cannot measure directly the time ε, and so we vary this to see the effect on the measurements. The value of ε has almost no effect on the overall AM and no effect whatsoever on the median, but has a rather strong effect on the GM. Figure 2 (b) shows the GM for three Tier 1 ISPs with Keynote agents in NY, with respect to the value of ε. The GM varies so much that, given different values for ε in each ISP, one might reasonably achieve any ranking of the three ISPs.

One issue raised above is that often raw measurements are rounded to some degree, in some cases because the clocks involved have finite precision. The above shows that such rounding in the data can actually impact the results disproportionately. Particularly if the measurements are made on different platforms which have different accuracies, as often occurs in distributed measurement infrastructures where nodes are not all deployed simultaneously, or on the same type of hardware.

3.4. Reversal

In this section we present one more illustration of the issues above. However, in this case we use real measurements directly, to illustrate the point. To do this we reverse the problem: normally we measure from the monitor to the server, but let us temporarily reverse the roles of these devices and measure from the server to the monitor. We could place servers in all of the most interesting positions, and gain an understanding of the tradeoffs in positioning a server in this way, thereby optimizing our performance under some metric. We shall use ping (ICMP echo) to make the measurements (we are interested here in how to put measurements together, not in the potential limitations of pings). We shall do pings from two "server" sites to all of Keynote's North American monitors. Then we can compute the statistics of the measurements for each "server" site.

meas. point	AM	GM	median	HM	TM
backbone 1	52.48	41.26	56.85	24.11	52.50
backbone 2	62.60	42.46	64.00	16.92	60.38

Table 3
Results by server (in ms).

The results are shown in Table 3. The results support our assertion: though in this case the GM reports the same ranking as the AM, TM and median, the two GMs are far closer than the other metrics, and the HM does report the opposite ranking.

3.5. Geographic weighting

As an aside, in many cases it would make more sense to weight the various statistics used. For instance, if a statistic is intended to represent an average of the customers view of the data, then it would make sense to weight each value by the number of customers based in that region. It is easy to weight an AM. Weighting a GM is also possible in the log domain, but it is not clear that this retains the meaning. Weighting of a median is a little more difficult, but possible. The technique is to give each measurement a weight, and after sorting the data compute the cumulative sum of the weights until one reaches the half-way point. Rigorous weighting methods for the other two methods are not known to the authors.

In actuality, weighting in a meaningful way is hard when the group performing the measurements do not have data such as customers by location (for instance if the measurements are made by an independent measurement company) and further in measurements between providers the actual customer base would be different. Even weighting by population can be hard because it is not always clear from external data which parts of the population are served at which point in the network (particularly in regional PoPs). Weighting can also lead to strange effects, such as a change in the overall performance metric because of a change in the geographic distribution of customers rather than an actual change in network performance. For these reasons weighting does not seem to be commonly used, but the ability to do it easily, and meaningfully in the AM, and median is an advantage.

4. Combined Geographic and Statistical Modeling

The previous examples assume performance is dominated by propagation delay, and has no random delays such as might be seen from queueing of packets in buffers, or due to random application layer delays. In reality, there are going to be both geographic effects on our measurements, and random components in the individual measurements. Hence in general we must consider the combination of these effects on our total measure.

We may use one statistic to combine measurements from the same path (is essence to estimate the underlying delay from measurements with noise) and then use a second statistic to combine the data from different paths into one result. We refer to the former as the local method, and the latter as the global method. In this section we examine what happens when you use different combinations of global and local method, and make a number of suggestions regarding which combinations are most useful. We base the results on a series of simulations, but in order to make the simulations as realistic as possible we have used the data obtained using NIMI (discussed earlier) to populate the simulation model. NIMI randomly samples paths between nodes, and so we do not have a completely connected graph of data. In fact the largest completely connect clique (found using the max clique solver at `http://rtm.science.unitn.it/intertools/clique/`) has 9 nodes (fnal, gatech, sandia, sony, ucla, uky, umass, utokyo, and verio2). This clique has a wide variety of nodes, including North American, and International cases, and so seems a reasonable set of points.

In this simulation we generate N measurements between each of the nodes above based on the empirical Cumulative Density Function (CDF) for the RTT measurements from NIMI. We can simulate samples from these distributions, in addition to computing the statistics theoretical value directly from the CDF. We compare each of the five statistics described above as both the local, and global method. Note that in only a few cases, such as the AM of the AM do the statistics commute, or are they associative.

We use four performance measures to assess these combinations. In each case the results are based on 400 simulations each with for $N = 60$ measurements along each of 45 paths.

relative bias: To measure the deviation from the true value we look at the relative bias $E\left[X - \bar{X}\right]/\bar{X}$, where $\bar{X}$ is the true value of the performance measure based directly on the empirical CDFs of the data set. We consider the relative measure because each of the statistics in question can take a different value. The results are shown in Figure 4 (a). The results show that the bias is generally small (below 1% in most cases). The only cases where it might be considered significant are when the HM is used as the global in conjunction with the TM or median as the local statistic.

relative RMSE: the relative Root Mean Square Error (RMSE) defined by rRMSE $= \sqrt{E\left[(X - \bar{X})^2\right]}/\bar{X}$. The rRMSE is an alternative measure for how close the result comes to the true value given a finite sample

of data. The results in Figure 4 (b) are larger when the HM or median is used for the globally.
Sensitivity to large outliers: We deliberately replace one randomly chosen data point from each simulation with an outlier (of 100 seconds), and assess impact using the rRMSE defined above. Figure 4 (c) shows these results, and clearly the poorest performer is using the AM for both global and local statistic, but poor results occur for all cases with the AM as the global statistic, or the HM as the local statistic.
Sensitivity to small outliers: Alternatively we use a small outlier (1.0e-6). Figure 4 (d) shows these results, and there are two very poor performers: the HM local with the GM or the HM as global statistic.

We attempt to bring the general features of the graphs out by setting a threshold (at 0.01) and saying a method is bad if it exceeds this threshold in one of the performance measurements. The actual value of this threshold is arbitrary, and the number of measurements will determine the exact pattern of the result, but this is simply an attempt to summarize the previous four graphs which should be referred to for the true perspective. Figure 4 (e) shows the summary: (paler) green for good, and (darker) red for bad.

We can further rule out using the GM for the global operation due to the work of Section 3. Hence the most useful approaches involve using the TM, GM or median for the local operation, and the AM or TM for the global. The ability to weight the AM in an meaningful way makes it more attractive for the global operation. Thus the best approach is to use the AM for the global operation, and the TM, GM or median for the local operation.

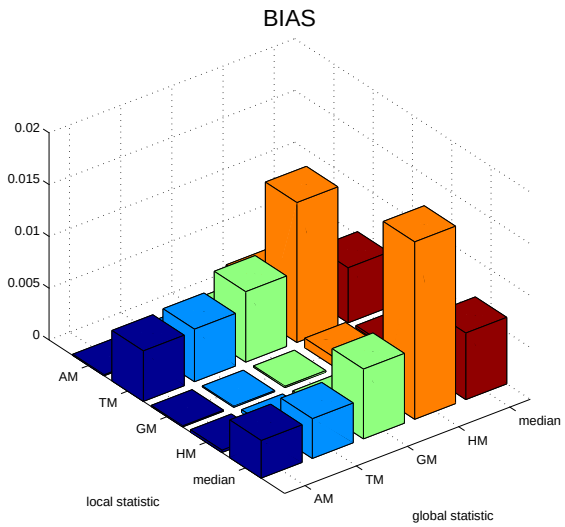
5. Conclusion

In conclusion, we have found that blind use of statistics such as the GM can be dangerous, and produce misleading results. After consideration of five statistics (the arithmetic, trimmed, geometric and harmonic mean, and the median) we found that it made sense to use different statistics for local aggregation (across a single path, or locality) and global aggregation (across all paths or localities). The preferable global operation was the arithmetic mean or simple average, and the preferable local operations were the geometric or trimmed means, and the median. In the future it might be interesting to consider other statistics. For instance the AM and HM are L^p norms (and the GM can be phrased in this form) and so we might be able to consider the best statistic over this whole class of statistics.

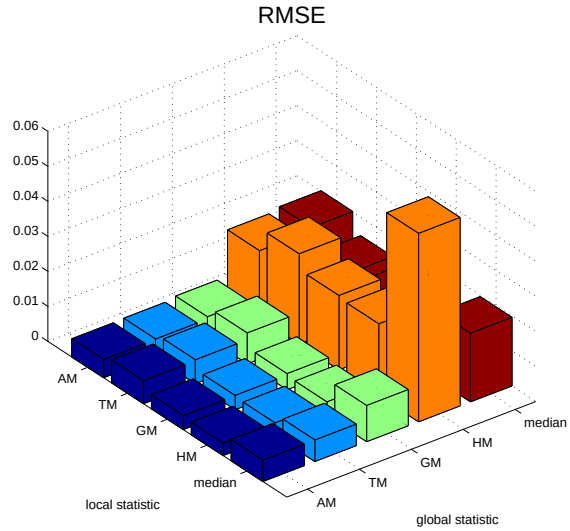
Acknowledgments: We would like to thank Vern Paxson and Yin Zhang for the use of the data collected by them using NIMI, and we would further like to thank Yin for some useful comments.

REFERENCES

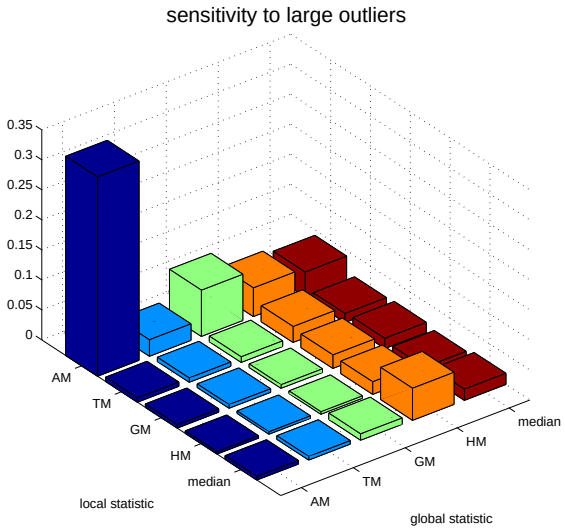
1. A. Erramilli, J. Gordon, and W. Willinger, "Applications of fractals in engineering for realistic traffic processes," in *Proceedings of the ITC-14*, vol. 1a, pp. 35–44, Elsevier, Amsterdam, 1994.
2. M. E. Crovella and L. Lipsky, "Long-lasting transient conditions in simulations with heavy-tailed workloads," in *Proceedings of the 1997 Winter Simulation Conference*, 1997.
3. W. Willinger, V. Paxson, and M. S. Taqqu, "Self-similarity and heavy tails: Structural modeling of network traffic," in *A Practical Guide to Heavy Tails: Statistical Techniques and Applications* (R. Adler, R. Feldman, and M. S. Taqqu, eds.), pp. 27–53, Birkhauser, Boston, 1998.
4. J. Andren, M. Hilding, and D. Veitch, "Understanding end-to-end Internet traffic dynamics," in *IEEE GLOBECOM'98*, (Sydney, Australia), 1998.
5. Q. Li and D.L.Mills, "On the long-range dependence of packet round-trip delays in Internet," in *Proc. IEEE International Conference on Communications*, pp. 1185–1191, 1998.
6. V. Paxson, A. Adams, and M. Mathis, "Experiences with NIMI," in *in Proceedings of the Workshop on Passive and Active Measurement*, 2000.
7. V. Paxson, J. Mahdavi, A. Adams, and M. Mathis, "An architecture for large-scale Internet measurement," *IEEE Communications Magazine*, 1998.
8. Y. Zhang, N. Duffield, V. Paxson, and S. Shenker, "On the constancy of Internet path properties," in *ACM SIGCOMM Internet Measurement Workshop*, 2001.
9. N. Cardwell, S. Savage, and T. Anderson, "Modeling TCP latency," in *INFOCOM 2000*.
10. P. Thomas, "Spheroidal geodesics, reference systems and local geometry," Tech. Rep. S-138, U.S. Naval Oceanographic Office, 1970.



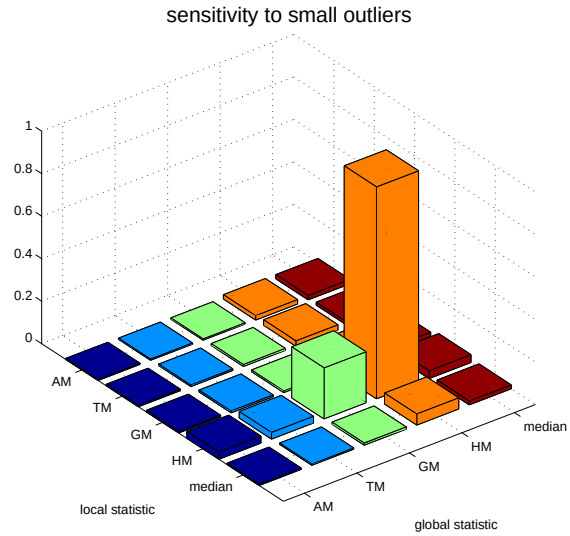
(a) Relative bias.



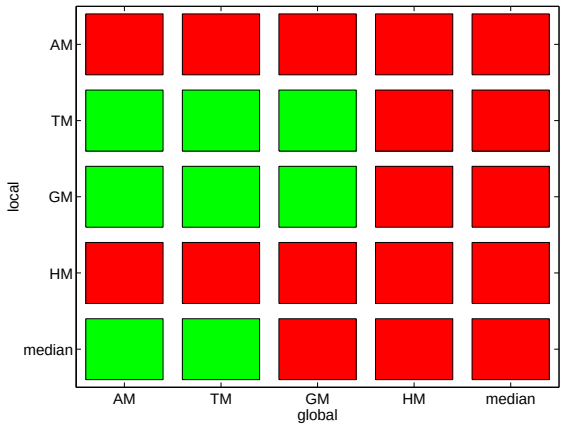
(b) Relative RMSE.



(c) Sensitivity to large outliers.



(d) Sensitivity to small outliers.



(e) A summary of the methods performance: (paler) green for good, and (darker) red for bad.

Figure 4. Performance results for combinations of global and local statistics.