Fundamental Bounds on the Accuracy of Network Performance Measurements

Matthew Roughan
School of Mathematical Sciences
the University of Adelaide, Adelaide 5005, Australia,
<matthew.roughan@adelaide.edu.au>

ABSTRACT

This paper considers the basic problem of "how accurate can we make Internet performance measurements". The answer is somewhat counter-intuitive in that there are bounds on the accuracy of such measurements, no matter how many probes we can use in a given time interval, and thus arises a type of Heisenberg inequality describing the bounds in our knowledge of the performance of a network. The results stem from the fact that we cannot make independent measurements of a system's performance: all such measures are correlated, and these correlations reduce the efficacy of measurements. The degree of correlation is also strongly dependent on system load. The result has important practical implications that reach beyond the design of Internet measurement experiments, into the design of network protocols.

Categories and Subject Descriptors

C.2.3 [**Computer-Communications Networks**]: Network Operations—*network monitoring*

General Terms

Performance,Measurement

Keywords

Network performance, Internet measurement, load balancing, error estimation, measurement planning.

1. INTRODUCTION

Network performance measurement is a topic of active current research. Internet performance, in particular, has received much attention, and is the topic of several Internet Engineering Task Force RFC's [1, 2, 3, 4, 5], and a major business of several companies (e.g. Matrix NetSystems, Keynote, Niksun, Brix Networks, etc). The basic goal is to improve network performance through monitoring, and this has to a large extent been a success. However, there is a gap in the theoretical underpinnings of performance measurement. In particular, very little has been written about the important problem of quantifying the accuracy of these measurements.

Quantifiable bounds on the accuracy of performance estimates are obviously important for answering questions such as "how long should we measure the network?", or "at what rate should we send measurement probes?". Such questions immediately arise in an network operations context, both in terms of improving long term network performance, and detecting performance problems. However, there are many other places where answering such questions is important. For instance, in the active queue management protocol RED (Random Early Detection) [6], one uses an averaged version of the queue length to obtain a measure of the current network performance. The choice of the time scale at which to average (via an exponentially weighted moving average) is an important parameter of this protocol. Likewise, in many load-sensitive routing schemes (for example, that used in the early ARPANET [7], or such as proposed in [8]), one wishes to obtain measures of the performance of various links in the network to better distribute traffic across the network. Even in TCP, one estimates the Round-Trip Time (RTT) in order that packet timeouts can be adapted to network conditions.

In all of the above, one wishes to perform measurements over short time periods, for instance to detect changes (performance problems) as quickly as possible, or to allow the network to adapt more quickly to traffic changes. On the other hand, there is often a cost involved in the measurement process (in collecting the data, sending active probes, or changing network routing), and so one does not wish to perform such measurements frivolously. One wishes to ensure these measurements have at least the accuracy required for the application in question. Hence, obtaining bounds on accuracy for measurements is an important topic that should receive more research.

This paper presents formulae for such bounds, though we only have analytic results for a somewhat unrealistic model, and simulation results for a more realistic system. The most important finding of this paper is that there are fundamental bounds of the accuracy we may achieve with network performance estimates. These bounds resemble in many ways the Heisenberg uncertainty bounds, and arise, at least in part for the same reasons. Heisenberg's inequality is

$$\Delta x \, \Delta p \geq \frac{h}{2\pi}$$

where Δx and Δp are the unknown errors in position and momentum, respectively. It arises because, when one measures, say the location of a particle, one must bounce a photon on the particle. The impact of the photon changes the momentum of the particle by an unknown amount. One can reduce the energy of the photon to reduce the range of uncertainty in this change in momentum, but only by reducing the photon's frequency, thereby reducing the accuracy of the localization gained through the measurement. An alternative version of Heisenberg's inequality is

$$\Delta T \, \Delta E \geq \frac{h}{2\pi}$$

where ΔT and ΔE are the unknowns in time, and energy of an interaction.

The essential dilemma underlying Heisenberg's uncertainty principle is that our own observations perturb the system, thereby reduc-

SIGMETRICS'05, June 6–10, 2005, Banff, Alberta, Canada.

ing the accuracy of those same observations. In observing a queueing system the same dilemma recurs. In this context, imagine a queueing system with arriving traffic given some fixed (unknown) traffic intensity. We could observe the average delay of packets through this system by sending *probe packets* through the system and measuring their delay. Such probe packets sample from the queue delay distribution, and so we take their sample mean, and use this as a measure of the average network delay, though note that any such estimate contains statistical error (apart from measurement errors). One might assume that to improve the accuracy of one's estimate, we need only increase the rate of probe packets. This has the obviously deleterious impact of reducing performance for all packets, and so is not a particularly practical solution to the problem, but it is often believed that although impractical this solution can provide results of arbitrary accuracy (after one corrects for the additional delays created by the probe packets themselves). However this is not the case. One cannot obtain arbitrary accuracy using probe packets. Further, we are actually worse off in this case than in particle physics, because the series of network performance measurements are strongly correlated, reducing their incremental value in improving our measurements.

As one probes the system in question more rapidly, increasing the load on the system, one increases the amount of correlation in the probe measurements. Correlations in a series of measurements reduce their efficacy in measurement, and so, for high probe rates, one actually reduces the accuracy of the resulting estimate. Thus, we end up in the same situation where accuracy is reduced by the measurements themselves. In this case, our inequality is not quite as simple as Heisenberg's, taking the form

$$\Delta T \, \Delta W^2 \geq f(\rho)$$

where ΔT is the observation time, and ΔW is the error in the estimate of queueing delays, and ρ is the *traffic intensity* a normalized measure of load (the system is stable for $\rho < 1$) at least for the tractable problem considered here. Notice that we refer to the observation time as ΔT because this is the interval within which we cannot localize our estimates of queueing delay any more finely, without a loss of accuracy. The relation above is highly load sensitive, in particular, the average queue length for the system considered above is $\rho/(1-\rho)$, but the function $f(\rho)$ grows like $(1-\rho)^{-4}$, and so accuracy becomes worse for heavily-loaded systems. In fact, the error grows much more quickly than the queue length itself, as the system becomes more heavily loaded. While, not as analytically tractable, we extend these results to more realistic systems through simulation, and find that similar results hold, however, we are in fact worse off, as for an arrival process with Hurst parameter $H \in (0.5, 1)$, the asymptotic relation becomes

$$\Delta T^{2-2H} \, \Delta W^2 \geq g(\rho),$$

which requires us to collect more data to gain the same reduction in estimate accuracy.

Apart from the fundamental bounds on the accuracy of performance estimates, this paper also shows how one can use these bounds to design Internet measurement experiments. Note that, although the scenario describe above applies to active probe measurements, the theory applies both to these, and to other forms of performance measurements, including passive measurements of performance.

The results have implications for Internet protocols, as well as measurements. For instance, the results suggest the TCP's conservative approach to congestion control is appropriate. TCP's congestion control kick's in when packet loss is detected. However, such performance problems occur when a queue is heavily-loaded. The results presented in this paper suggest that estimates of congestion are likely to be highly inaccurate, and so in reality severe congestion could lie behind even a mild congestion indication such as a single packet loss. The results also seem to suggest that TCP Vegas cannot work well in competition with TCP Reno. Its RTT estimation, which is based on relatively few samples, cannot hope to maintain an accurate estimate of queue lengths when those queue lengths can be long, as they would in the presence of competing TCP Reno connections (without additional active queue management). Similarly, load balancing schemes that attempt to adapt quickly to large demands will always have to face the fact that their estimates of delay will be inaccurate, exactly when they desire the most accuracy (when the system is under heavy load). Finally, Internet performance tomography methods that rely on multiple packet measurements (as opposed to methods using multicast packets), are dependent on the correlation structure in the measurements they make, and these results should be directly relevant to such studies.

The paper is organized as follows. In Section 2 we introduce the background to this work, namely the technical basis for Internet performance measurements. In Section 3 we present the fundamental theory behind this work, largely drawn from the simulation literature, though extended to apply to Internet measurements. In particular, we present a new results concerning the accuracy of results drawn from Poisson samples. In Section 4 we illustrate and expand on these results, to both validate them, and make clearer their meaning to a reader. In Section 5 we extend these results in a number of ways, in particular, we note that the simple Markov model from which we can gain analytic results is not adequate for accurate Internet modeling, and we use simulation to study a more realistic Long-Range Dependent (LRD) model. This shows that the Markovian bounds are probably quite conservative in comparison to real network performance measurement bounds. In Section 6 we consider an example, once again to try to make the results more concrete to a reader, and to bring home the practical scale of the measurement problem, whose implications we then consider in Section 7. Finally we conclude the paper in Section 8. The key results of the paper appear in (9), where we demonstrate the impact of Poisson sampling on performance measurement results, and in (17), (19) and (20), where we give the fundamental performance bounds for queue performance.

2. PERFORMANCE MEASUREMENT TECHNOLOGIES

There are many measurements one may collect from a network, for instance, traffic (via SNMP, Netflow, or packet monitors), topology measurements (via traceroute studies, or router configuration collection), and direct performance measurements (via active probes of the network). We will focus here on measurements that provide measures of network *performance*, though note that other supplemental measures may be required (e.g. network topology) to make sense of this data. Many examples of tools to perform such measurements may be found at [9].

There are a number of ways in which one may collect data about network performance:

- **Direct measurements:** it is quite possible for a router, or switch to maintain information about its own performance. For instance, to maintain data on the number of packets or bytes in buffers, or the number of packets dropped for various reasons. Such information can then be collected at regular intervals through a mechanism such as SNMP [10, 11], and in fact there are several MIBS defined for this purpose. In principle, such information could be at an arbitrary level of detail, however, in practice there are limits on the (fast) memory required to store such information, and also on the rate at which it is collected (SNMP is not a particularly efficient mechanism). Hence, such data might be collected typically every five minutes, and despite its potential to be one of the best sources of performance measurements available, it is actually one of the worst. This is primarily a technological issue, however, and the situation might be improved in the future.
- **Passive traffic measurements:** Passive traffic measurements can be used to infer network performance through multiple measurements of the same packet [12, 13]. By measuring the arrival time of a packet (or its acknowledgement) at multiple points, we can

infer the delay between these points. This approach can also provide data of very fine detail, however, it also has limits. Firstly, dedicated packet monitors are typically cheap, but involve non-negligible installation and maintenance cost, and so are not usually installed everywhere. Hence, ones ability to perform such measurements is limited to the locations of packet monitors. This issue might be partially alleviated through the implementation of packet sampling, a technology that allows the router to maintain statistics (such as arrival times) for a sample of the packets that traverse it. Such sampling also reduces problems that arise from the very large volumes of data that would arise from storing data regarding every packet on the network. However, such monitoring requires precisely synchronized clocks[1][2], which, while also technically feasible, requires additional hardware (e.g. GPS receivers) whose installation adds to the difficulty of such measurements, regardless of whether it is performed by packet sampling or a dedicated monitor. Finally, passive monitoring of this type can only infer performance on paths that are used. If no data traffic arise on a path, one cannot infer performance on this path, even though individual components of that path may experience high loads due to traffic from unmonitored paths.

- **Active probes:** The third method used to infer network performance is the well developed approach of active probing, for instance see [15, 16, 1, 2, 3, 4, 5]. In this approach, one deliberately sends *probe* packets into the network with precisely controlled departure times, and measures their arrival times elsewhere in the network. Such probes have the same synchronization issues as passive traffic measurements, and the probe boxes also require installation in a network, but these boxes are not as tightly coupled into the network. For instance, a packet monitor typically requires that an optical splitter be installed in the optical fibre of a network, whereas a packet probe can be installed by any customer of a network. There are also many different types of probes one can mount, with different applications and functional attributes, e.g. ICMP echo response times, TCP SYN/ACK response times, DNS response times, HTTP page downloads, as well as dedicated probe protocols. These factors have lead to active probing being the most widely deployed form of IP network performance measurement methodology.

For a practical comparison of some of the above techniques see [17].

Network performance can mean many things, for instance: reachability, delay, loss, jitter, reordering, and bulk throughput. One can also form more complex functions of these metrics to attain measures such as the subjective performance of an application, e.g. VoIP. However, in this paper we will concentrate on the **delay** aspect of network performance. It is likely that the results herein can be extended to consider loss, and quite possibly other performance metrics such as jitter, however, we concentrate on delay in this paper.

Packet delays are comprised of a number of components:

1. *Packet processing delay* is the delay to perform tasks such as forwarding table lookup, and is very small in modern high speed routers (e.g. $\ll 1$ ms).

2. *Packet transmit time* is the time from starting to send the first bit of a packet onto the wire, until the last bit is finished being transmitted, which is given by the packet size (including framing bits) divided by the link bandwidth. Note this is small for high bandwidth links, e.g. $\sim 4.8\mu$s for a 1500 byte packet on a OC48 (2.5 Gbps) link.

3. *Propagation delay* is the delay a packet experiences on the wire, and is given by the physical fibre (or wire) distance divided by the speed of light in optical fibre (66% of speed in air, 300,000 km per sec). e.g. $\sim$30 ms for a direct East to West Coast transmission in North America.

4. *Queueing delay* is the time spent by a packet in queues, which depends on load, and can be quite large, e.g. 0.2 seconds, even on single OC48 line cards.

The two components that are significant, and therefore of primary interest in most networks are propagation delay, and queueing delay. Propagation delay is determined by network topology and routing, and for the purpose of this paper we shall consider it to be a constant (see [18] for a more realistic view), which is derivable from other network data (the topology and routing information collected elsewhere), or from long-term measurements of the network. Hence, within the context of this paper, the goal is to measure the queueing delays. These delays may be seen as drawn from a random distribution, and the goal of this paper is to estimate parameters of this distribution, in particular the mean (which is one of the most basic and important parameters of the distribution).

In addition to statistical variations in queueing delays, there are measurement errors. We have noted above that one-way delay measurements require precisely synchronized clocks. Without such, one would have errors in measurements. Another example is TCP's measure of RTT (gained from packet acknowledgements), which also contained (in the past) large errors because of a coarse clock resolution (at one point standard implementations of TCP used 500ms clocks). Similarly, any set of performance measurements contains errors and artifacts. Apart from clock errors, these arise from processes that we do not wish to measure, for instance, delays in time-stamping a packet once it is received at a monitor. Such errors are unavoidable (though one can go to significant efforts to reduce them) and so we shall include the fact of measurement errors in our work, though note this is not what we are referring to in talking about estimation accuracy, which refers to the accuracy of estimates of distributional parameters such as the mean.

In the case of active probes, we have a choice in how we can send these probes. A naive approach, sending these at equal spaced intervals can result in problems if, per chance, these aligned with some periodic behaviour in the network in question. There is a more appealing alternative advocated in the RFCs [1, 2, 3, 4, 5], namely Poisson sampling; packets are sent at the epochs of a Poisson process. The elegant Poisson Arrivals See Time Averages (PASTA) theorem [19], guarantees (under mild conditions) that such Poisson sampled packets see the true averaged behaviour of the network, and so we may use these probes to avoid synchronization issues. Poisson streams are actually quite tractable given the new results presented here, regarding Poisson sampled measurements, and so we shall concentrate on Poisson sampled traffic measurements.

The above discussion focusses on packet delays, but the analysis in this paper could equally be applied to measurements of server performance, in which case there are additional measurements available to us, for instance, log files, proxy logs, and client instrumentation logs. In principle, any system that can be modelled as a queueing system is susceptible to this type of analysis, though some approximation may be required to make the analysis tractable.

3. THEORY

The theory we shall apply in this paper derives from theoretical concerns regarding discrete event simulation of queueing systems. Simulation is useful, for instance where the system is not mathematically tractable, or to provide a check on complex mathematical results. However, in the early days of queueing simulation, computing power was orders of magnitude less than it is today, and an ever present issue was how long to run a simulation, or even whether

[1]Precise synchronization is needed for measures of one-way delays. It is not needed to measure loss, or for round-trip times, which can be measured at the same location, and hence using the same clock. However, Internet routing is fundamentally asymmetric, and so there are many scenarios where one cares about the one-way delay.

[2]Note that there are also post-processing techniques intended to remove systematic inaccuracies introduced by clock skew [14].

a meaningful simulation could be run at all. There were many papers, for instance see [20, 21, 22, 23, 24, 25] that considered the issue in detail (drawing on past results found in [26, 27, 28, 29, 30]). This work began with methods to analyse simulation results [20, 21, 22], and using this analysis determine whether the simulation need be continued, but later developed detailed formulae for computing *a priori* how long a simulation need be run, a fact of crucial interest in the decision of whether a set of simulations would be of value [23, 24, 25]. It is this theory that we shall draw directly from here, as there are direct parallels between the requirements of simulation and measurement. In fact, the simulation problem is very similar to the measurement problem, albeit without the technical difficulties in collecting the data that exist in real measurement systems. Hence, we shall follow the formalism of [24] very closely.

3.1 Measurement

The underlying assumption of most performance measurement is that there is a stationary stochastic process, of which we wish to measure some parameter. For instance, we might assume that the network in question is a queueing system in equilibrium, and that our measurements are intended to characterize the performance of this system. We shall define our stochastic process of interest $X(t)$ for $t \geq 0$ to be wide-sense stationary, which means that its mean, variance and auto-covariance are all constant with respect to t for all $t \geq 0$, and can consequently be written $E[X(t)] = \bar{X}$, $\text{Var}\,(X(t)) = \text{Var}\,(X(0))$, and $\text{Cov}(X(t), X(t+s)) = R(s)$, respectively.

In simulation we can guarantee our system is stationary (modulo initialization impacts), but in reality, we cannot assume any networking system is truly stationary. This fact is critical to the problem we consider here. We wish to measure our system over periods where stationarity is a reasonable approximation, i.e. $t \in [0, T]$ such that the conditions of stationarity above are a reasonable approximation. If T is small (say $<$ 1ms), this is very likely to be the case, whereas if T is large (say $>$ 6 hours) it is unlikely to be true (given the daily cycles in Internet traffic). Thus, the shorter the interval over which we conduct our measurements, the better placed we are with respect to our assumption of stationarity. Further, shorter time intervals allow us better localization of performance in time. However, as we shall see, measurements over a shorter time interval will be less accurate. Hence, there is a tradeoff between the accuracy loss through short time interval measurements, and the accuracy loss because the underlying process is not stationary (i.e. its parameters change during the interval of measurement).

It is important to note that, although it is not stated explicitly, many measurement studies *implicitly* make this assumption of stationarity. Without such an assumption, many reported results are meaningless. For instance, consider the result of estimating the delay in a network over a 24 hour cycle, in which the network experiences poor performance during the busy hour. An average measure over the day could quite easily indicate reasonable performance, even though users during the peak hour will experience poor performance.

However, with an assumption of approximate stationarity, we now have to choose the optimal point for T. This is our point of departure from simulation design. In simulation design the limiting factor (for T) is the computation required to simulate the system in question for this interval of time. Given the large increases in computing power available now, it is rare for this to be the limiting factor in a simulation. However, in our case, there is a practical limit on T imposed by the nature of stationarity in the *real* system under observation. Furthermore, if our goal is to detect changes in the system (for instance performance problems), then temporal localization of performance estimates is important, because this directly impacts the time to detect these changes.

The first approach (for example see [31, 32]) to designing measurement experiments is to apply a simple form of the Central Limit Theorem (CLT). For instance, take a set of independent identically distributed measurements X_i for $i = 1, \ldots, N$, with mean $\bar{X}$ and variance $\sigma_X^2 < \infty$, and sample mean

$$\hat{X}_N = \frac{1}{N} \sum_{i=1}^{N} X_i, \tag{1}$$

then the sample mean converges (in distribution) as

$$\sqrt{N} \left(\hat{X}_N - \bar{X} \right) \rightarrow N(0, \sigma_X^2), \tag{2}$$

where $N(0, \sigma^2)$ denotes a normal distribution with zero mean, and variance σ^2. The theorem shows that the sample mean converges to the true mean, and that the variance of the sample mean about the true mean decreases in proportion to N. From this, one can compute (at least approximately) how many data points are required to achieve a given accuracy. There is no dependence on the time of these samples, and so, one could increase N either by extending T, the measurement interval, or by increasing the rate at which measurements are made.

The CLT (as described above) only applies to independent measurements X_i. What happens if the measurements are correlated in some fashion? In this case, we can replace the above results with the following CLT[3]. Given X_i, which are drawn from a stationary process with mean $\bar{X}$, and auto-covariance $R(s)$, then the sample mean converges as follows

$$\sqrt{N} \left(\hat{X}_N - \bar{X} \right) \rightarrow N(0, s_{\hat{X}}^2). \tag{3}$$

Note that this is identical to the simple CLT results, except that σ_X^2 has been replaced with $s_{\hat{X}}^2$ which is referred to (see [24]) as the *asymptotic variance* of the process X, which is defined by the above relationships to be

$$s_{\hat{X}}^2 \equiv \lim_{N \rightarrow \infty} N \,\text{Var} \left(\hat{X}_N \right) \tag{4}$$

One may compute the asymptotic variance using the following relationship (from [21]),

$$s_{\hat{X}}^2 = \sigma_X^2 + 2 \sum_{i=1}^{\infty} R(i), \tag{5}$$

where $R(i) = E[X_j X_{j+i}] - E[X_j]^2$ is the auto-covariance function. Notice that the correlations increase the asymptotic variance $s_{\hat{X}}^2$ dramatically compared to σ_X^2, with an equivalent reduction in the accuracy of estimates.

In the context of a continuous time process $X(t)$ being sampled at times t_i for $i = 1, \ldots, N$, to give $X_i = X(t_i)$, we can see that faster sampling does not grant the same improvements as sampling over a longer interval. For instance, consider uniform sampling times $t_i = i\Delta t$ for a process with exponential auto-covariance function $R(t) = \exp(-\beta t)$, then

$$\begin{aligned} s_{\hat{X}}^2(\Delta t) &= \sigma^2 + 2 \sum_{i=1}^{\infty} R(i\Delta t) \\ &= \sigma^2 + 2 \sum_{i=1}^{\infty} \exp(-\beta \Delta t)^i \\ &= \sigma^2 + 2 \frac{\exp(-\beta \Delta t)}{1 - \exp(-\beta \Delta t)}. \end{aligned} \tag{6}$$

Given a fixed time interval of measurement T, then the number of samples available will be approximately $N \simeq T/\Delta t$, and so the variance of the estimator will be

$$\text{Var} \left(\hat{X}_{T/\Delta t} \right) = \frac{\Delta t \sigma^2}{T} + \frac{2 \Delta t e^{-\beta \Delta t}}{T(1 - e^{-\beta \Delta t})}. \tag{7}$$

[3] Under conditions discussed in Section 5.

which converges to $\frac{2}{\beta T}$ as $\Delta t \to 0$. Hence we see that we cannot achieve arbitrary accuracy by more rapid sampling, as we can by increasing T, or if the correlation function $R(s) = \sigma^2\delta(s)$ (where $\delta(\cdot)$ is the Dirac-delta function). Notice that the above could also have been derived directly from the continuous version of the above results, i.e.

$$\lim_{T\to\infty} T \text{ Var}\left(\hat{X}_T\right) \quad = \quad 2\int_0^\infty R(u)du, \tag{8}$$

because, quite obviously, samples from a continuous process cannot give us more information than continuous observations of that process. However, given that most Internet measurements are a discrete-time in nature, we shall primarily consider measurements that are formed via discrete samples from a continuous-time process. Hence, in the following we shall assume that we have measurements $X_i = X(t_i)$, for some set of sample points t_i.

Note that PASTA implies that Poisson samples (i.e. where the t_i are given by a Poisson process) will see the true time-averaged behaviour of the system[4], so Poisson sampling is valid in this context, but note that $N(T)$ will be a Poisson random variable. In asymptotic results, however, we shall substitute its expected value $E[N(T)] = \lambda T$. In Appendix A, Theorem A.2 we prove the following result (given Poisson sampled measurements with rate λ),

$$\lim_{N\to\infty} N \text{ Var}\left(\hat{X}_N\right) = \sigma_X^2 + 2\lambda\int_0^\infty R(u)du, \tag{9}$$

where the integral is finite.

As noted above, $N \sim \lambda T$ (for large N), so

$$\lim_{T\to\infty} T \text{ Var}\left(\hat{X}_N\right) \to \frac{\sigma_X^2}{\lambda} + 2\int_0^\infty R(u)du. \tag{10}$$

The result makes a great deal of sense. For very high sampling rates, i.e. $\lambda \to \infty$, this tends to the continuous measurement result (8), whereas, if we reduce the sampling rate λ, while extending the measurement time T, so that N remains constant, and then taking the limit as $N \to \infty$, we get

$$\lim_{N\to\infty}\lim_{\lambda\to 0} N \text{ Var}\left(\hat{X}_N\right) \to \sigma_X^2, \tag{11}$$

which is to be expected, because samples will be at least several multiples of the *correlation scale* apart, so that the correlations will be negligible, and so the simple CLT result applies.

Note that if one *thins* Poisson measurements, for instance by dropping measurements, independently with probability $1-p$, then the resulting sampling process is still a Poisson process (with rate reduced by a factor of p), and the results above still hold, though with new arrival rate λp, and of course, one must on average observe the system for time $T = N/(\lambda p)$ to obtain N measurement.

Note that the above results are asymptotic approximations for large N. For the purposes of this paper, this suffices. We primarily consider fundamental limits here, and these occur for large numbers measurements where the above limits are quite accurate. However, if one were concerned with making more accurate estimates for small N, one could apply the Gauss-Markov theorem to obtain the Best Linear Unbiased Estimator (BLUE) for $\bar{X}$. The BLUE is not a practical estimator here, however, because it requires pre-knowledge of the auto-covariance of the process (which we see later is dependent on load), and so we will not consider this directly here. Note also that [23] presented more efficient estimators than those considered here, but once again these are impractical, though in this case because they require direct measurement of the length of the busy period, which we do not have access to in most Internet measurements, and would not, in any case help with non-regenerative systems such as those exhibiting LRD.

Notice also that in all of the above results, we assume that the measurements themselves contain no errors. In fact, it is easy to incorporate independent errors into the results, as the variance of the sum of independent random variables is the sum of the variances thereof. Hence, unbiased errors with variance σ_E^2 will add σ_E^2 to the asymptotic variance.

3.2 Queueing delays

The importance of the above results lies in the fact that packet delays are not independent. Theoretical consideration of the way queueing occurs leads to correlations (see below), and measurements of packet delays have shown correlations in practice [33]. Hence, when considering the accuracy of mean packet delay estimates, one should incorporate correlations into the model. We can do this in an analytic fashion in some simple queues, leading to simple results describing the accuracy of queueing delay estimates.

Another way to understand these results is to consider that many queueing systems are regenerative when the system is empty. That is, the ends of busy periods form *renewal* points. In order to measure properties of such a queue, one should observe the system over several of these renewals, i.e. for several busy periods. Note though, that the length of the busy period grows as load increases, leading to longer observations times. Morse [26] describes this same phenomena in terms of *relaxation times*. In this terminology, when the queue experiences higher loads, it not only has a longer (average) queue length (it grows as $(1-\rho)^{-1}$), but also larger *excursions* around this queue length, which grow as $(1-\rho)^{-2}$. This can also been seen as an increase in the correlation scale of the process, which is directly related to the length of the busy period. Hence there is a requirement that measurements be further apart to compensate. However, the formulation above, in terms of the auto-covariance function of the waiting times allows us to make analytic statements about the variance of the estimators, given N measurements.

There are many results describing the transient behaviour of simple queueing systems, and thence their auto-covariance functions. We will present the most simple here for the purpose of exposition as even these may become fairly complex, but note that many generalizations are possible. The system we consider is the M/M/1 queue, i.e. a queue with a Poisson arrival process (of rate λ) of packets whose service times are exponential (with mean $1/\mu$). We denote the traffic intensity $\rho = \lambda/\mu$ and note that the queue is stable in the sense that it is positive recurrent for $\rho \leq 1$, but that the expected length of the busy period is infinite for $\rho = 1$ (because the distribution of the length of the busy period has a heavy-tail in the critical transition between stability and instability), and so we only consider queues with $\rho < 1$. The M/M/1 queue is very well studied, with many text book results, e.g. see [34]. For instance, the mean and variance of the number of packets in the system are

$$\bar{N} = \frac{\rho}{1-\rho}, \quad \text{Var}\,(N) = \frac{\rho}{(1-\rho)^2}.$$

The mean and variance of the number of packets in the buffer (not counting the packet in service) are

$$\bar{B} = \frac{\rho^2}{1-\rho}, \quad \text{Var}\,(B) = \frac{\rho^2}{(1-\rho)^2}.$$

The mean and variance of the waiting time are

$$\bar{W} = \frac{1}{\mu}\,E[N]\,, \quad \text{Var}\,(W) = \frac{1}{\mu^2}\,(\,E[N] + \text{ Var}\,(N))\,.$$

and the mean time spent in the system (by a packet) is

$$\bar{T} = \frac{\frac{1}{\mu}}{1-\rho}.$$

[4]There are obvious implications for Poisson vs deterministic sampling, when estimating spectra, or correlation functions, but these do not concern us here.

Note that $\bar{W}$ may be what we are interested in measuring, but that $\bar{T}$ is what is actually measured by an active probe, and includes both the queueing delay, and the packet transmission and processing times.

The M/M/1 auto-covariance results are not as simple, because of their dependence on the transient behaviour of the M/M/1 queue (often represented in terms of Bessel functions). However, one can find the auto-covariance of this queue in [26, 20], where the auto-covariance for the number of packets in the system is given as

$$R(s) = \frac{\lambda\mu(\mu-\lambda)}{\pi}\int_0^{2\pi}\sin^2\theta\frac{e^{-ws}}{w^3}d\theta, \tag{12}$$

where

$$w = \lambda + \mu - 2\sqrt{\lambda\mu}\cos\theta, \tag{13}$$

Morse [26] gives the integral of the auto-covariance $R(s)$ over s to be

$$\int_0^\infty R(u)du = \frac{\lambda\mu(\lambda+\mu)}{(\mu-\lambda)^4}, \tag{14}$$

which we can simplify by dividing numerator and denominator by μ^4 to get

$$s^2_{\hat{N}} = 2\int_0^\infty R(u)du = \frac{2\rho(1+\rho)}{(1-\rho)^4}\frac{1}{\mu}, \tag{15}$$

which is the asymptotic variance for the estimates of the number of packets in the system. See also [24, (22)], where the additional factor of $1/\mu$ in the above result comes from the fact that in [24] time is scaled so that $1/\mu = 1$. Given such a scaling, the observation time is given in units of number of (average) service times, whereas in (15) observation time is given in absolute units (e.g. seconds). We shall consider here estimates of two quantities, the mean queueing delay and the mean number of packets in the buffer, denoted by $\bar{W}$ and $\bar{B}$, respectively. The latter is not suitable for measurement by active packet probes, but can be measured using other methods such as statistics collected by a router. The auto-covariance is different for different quantities such as $\bar{W}$ and $\bar{B}$. Examples are presented in [23, 24] where the following results appear

$$\begin{array}{rcl} s^2_{\hat{B}} & = & \frac{2\rho^2(1+4\rho-4\rho^2+\rho^3)}{(1-\rho)^4}\frac{1}{\mu} \quad \simeq \quad \frac{4\rho^2}{(1-\rho)^4}\frac{1}{\mu}, \\ s^2_{\hat{W}} & = & \frac{\rho(2+5\rho-4\rho^2+\rho^3)}{(1-\rho)^4}\frac{1}{\mu} \quad \simeq \quad \frac{4\rho}{(1-\rho)^4}\frac{1}{\mu}, \end{array} \tag{16}$$

where once again the factor of $1/\mu$ arises because of the different time scale from [24].

The above results assume continuous measurement of the system delays, which we cannot do here. Imagine that we can measure delay for all departing packets, then we would in fact have a Poisson sampling with rate λ (the arrival rate of packets to the system). In many cases we might sample from the set of packets traversing the system, and hence sample at rate $\lambda_s = p\lambda$. Similarly, when using Poisson probes, we have a probe rate λ_s, though in this case the total traffic rate to the system would be the sum of probes and real traffic. Given such samples from the arriving traffic, equation (9) gives

$$\begin{array}{rcl} s^2_{\hat{B}}(p,\lambda,\mu) & \simeq & \frac{\rho^2}{(1-\rho)^2} + p\frac{4\rho^3}{(1-\rho)^4} \\ s^2_{\hat{W}}(p,\lambda,\mu) & \simeq & \frac{1}{\mu^2}\frac{\rho(2-\rho)}{(1-\rho)^2} + p\frac{4\rho^2}{(1-\rho)^4}, \end{array} \tag{17}$$

where the asymptotic variance is with respect to the number of measurements N, not the time interval T, which would involve dividing the above formulae by $\lambda_s = p\lambda$, with the best case bound (with respect to T) occurring for $p = 1$, i.e. where every packet is sampled.

3.3 Errors

In Whitt [24] the author considers both relative and absolute errors, as either may be more important in a particular context. In the Internet measurements under consideration there is a significant constant component (the propagation delay), which we assume here we can measure with minimal errors over the long term. Hence, if relative errors were considered, they should be relative to the queueing delay plus the propagation delay. It is therefore more reasonable to consider absolute errors here, given that the propagation delay may vary for different sets of measurements, independently from the queueing delay.

Given that there is enough data for the asymptotic normal distribution to be a reasonable approximation over the body of the data, we can use Gaussian confidence intervals to estimate the potential errors in an estimate. For instance, if we wish to assess the $(1-\beta)$th percent confidence intervals for the estimate of the waiting time (notionally the region which we believe the true value to fall, $(1-\beta)$% of the time, given the estimate) then these are

$$\hat{W} \pm z_{\beta/2}\sqrt{\frac{s^2_{\hat{W}}(p,\lambda,\mu)}{T}} \tag{18}$$

where $z_{\beta/2}$ are such that i.e. $p(-z_{\beta/2} < N(0,1) < z_{\beta/2}) = 1-\beta$. For example, $z_{2.5} = 1.96$, so the 95th percent confidence intervals would be $\hat{W} \pm 1.96\sqrt{s^2_{\hat{W}}/T}$.

Writing this another way, we consider the time interval over which the measurements were conducted to be the unknown in time (the most accurate extent to which we can localize our estimate) ΔT, and the $(1-\beta)$ confidence interval for $\hat{W}$ to be the unknown error in that quantity $\Delta W = z_{\beta/2}\sqrt{s^2_{\hat{W}}/\Delta T}$. Then we can write

$$\Delta T\,\Delta W^2 \;\geq\; z^2_{\beta/2}\; s^2_{\hat{W}}(p,\lambda,\mu), \tag{19}$$

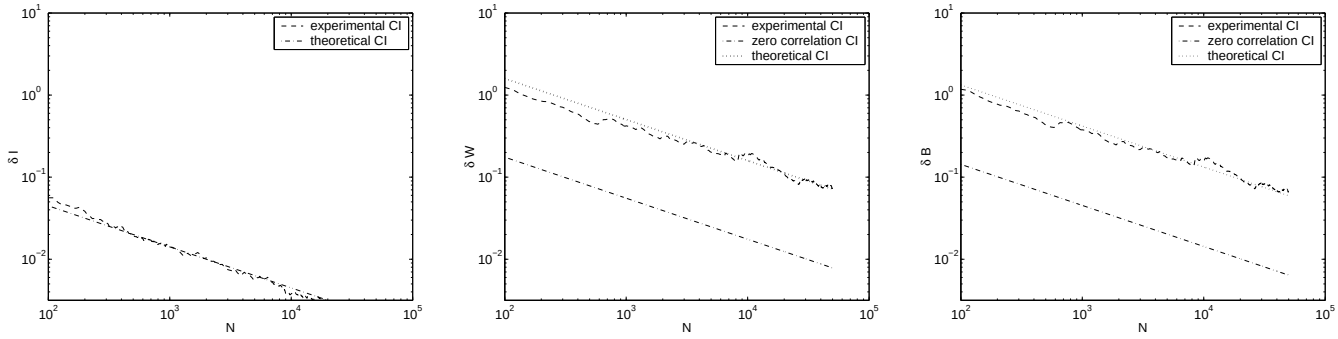
and likewise

$$\Delta T\,\Delta B^2 \;\geq\; z^2_{\beta/2}\; s^2_{\hat{B}}(p,\lambda,\mu). \tag{20}$$

4. NUMERICAL RESULTS

The work on simulations has long been known. However, its implications, in terms of bounds to the accuracy of network performance estimates has not been fully explored (to the author's knowledge). In our first results we consider fundamental bounds where we assume that we have precise measurements of the quantity of interest for all packets. These measurements could be obtained through careful passive monitoring of all traffic entering and exiting the system in question. There is no additional information available to improve these results regardless of how cleverly one collects data, and these measurements do not distort the system under observation.

We shall present some examples of these results through a comparison of the theoretical results with simulation experiments. The initial simulations are of the M/M/1 queue, and we perform these simulations for 100,000 arrivals, discarding the first half of the simulation to reduce bias from the non-equilibrium initialization (we initialize the system as empty). For each set of parameters we run 30 simulations, and use the ensemble to estimate the variance of estimators (for varying N), and thence the 95th percent confidence intervals for the data. Figure 1 presents log-log graphs for the magnitude of the confidence intervals bounds $1.96s_{\hat{X}}$ (both theoretical as derived above, and as derived from the simulations) for the inter-arrival time distribution, the waiting times, and the number of packets in the buffer with respect to the number of measurements N. Note that the error bounds from simulation and theory are close in each case. The error bounds for estimating the inter-arrival times are orders of magnitude lower than those for the queue, because the inter-arrival time variance is smaller, and there are no correlations in these measurements. For another comparison, Figure 1 (b) and (c) show the error bounds that would result from using the system variance σ_X, which does not incorporate correlations, rather than the asymptotic variance $s_{\hat{X}}$. One can see that this results in larger errors than for the inter-arrival times, but would underestimate the true errors by order of magnitude, because it does not include the impact of correlations in the measurements.



(a) Inter-arrival times. (b) Waiting times. (c) Packets in the buffer.

Figure 1: Error bound measurements for the M/M/1 queue with $\rho = 0.8$.

Figure 1 shows results for one value of traffic intensity, $\rho = 0.8$. In Figure 2 we compare the empirical error bounds (derived from simulation) over a range of parameters, by varying the arrival rate of the system to alter ρ. Note that, as we vary the arrival rate, we change the variance of the inter-arrival times, and so this has a minor impact on our estimates of the inter-arrival times. However, in contrast, the arrival rate has a strong impact on the errors for the waiting time and packets in the buffer measurements. This impact arises because increasing the load on the system increases the correlations in the measurements, and makes them less accurate. In extreme cases (e.g. $\rho > 0.9$) the results do not even converge to the asymptotic results until a significant number of measurements have been made (of the order of 10000 arrivals).

4.1 Active probing bounds

The bounds above assumed that we had perfect measurements of the system in question. Of course, in reality, most measurements are based on some kind of sampling. Where such sampling involves passive measurements, either uniform, or Poisson, the previously reported results apply. However, if one applies active probing, then the active probes themselves impact the performance of the system in question. One could also imagine dropping arrivals to help control some system's performance through active queue management schemes. Although both of these activities may be detrimental to the system in question, a natural question, is "what are the fundamental bounds on estimator accuracy, whatever one does?"

Consider a system with arrival rate λ_A to which we add probes at rate λ_S, then the total arrival rate of traffic will be $\lambda = \lambda_A + \lambda_S$. In most measurements, it is desirable for probes to have the same service time distribution as normal requests/packets, because if they may be otherwise distinguished, it can lead to gaming of the measurement traffic to artificially improve performance results by giving preferential treatment to probe packets. Hence we take $\mu_S = \mu_A = \mu$, and hence $\rho = \rho_A + \rho_S$. In this scenario, $p = \lambda_S/\lambda$, and therefore $\rho = \rho_A/(1-p)$. The result, from (17) is shown in Figure 3 (a) which shows the impact of sampling rate p. The three solid curves show the empirical results, and the three dashed curves show the theoretical results. Note that the value of N is the number of sampled measurements, which decreases as p decreases, so for small p we would not have as many sample measurements on which to draw, in any one simulation. If the critical issue were the time of the measurements, the curves for small p would be displaced to the right, to lie much closer together. This is an indication that sampling hurts less in these situations where there are strong correlations. We loose relatively little information by dropping some of the closer together, and hence more highly correlated measurements.

Our objective is to determine how much better we could do by sending more probe packets. Intuitively, the previous argument should suggest to a reader that fast probing does not necessarily help, particularly given that it increases the load on the system in question. However, we can formally assess this by minimizing $s^2_{\hat{W}}(p, \lambda, \mu)$, over all possible values of p for which the system remains stable. We can rewrite $s^2_{\hat{W}}$ as a function of p, λ_A, and μ_A. The result, derived from (17), is a rational function of polynomials (in p), which we can minimize (in our case numerically), to get the results charted in Figure 3 (b) and (c). Figure (b) shows the optimal active probe rates λ_S, normalized by average service time with respect to ρ_A. Note that the optimal rate for probe packets is small for low ρ_A, because in this case the asymptotic variance is small in any case, and more probing is unnecessary, whereas, the optimal probe rate for high loads is of a necessity low because the load induced by the probes themselves worsens the performance of the measurements by increasing the load, and thence the asymptotic variance.

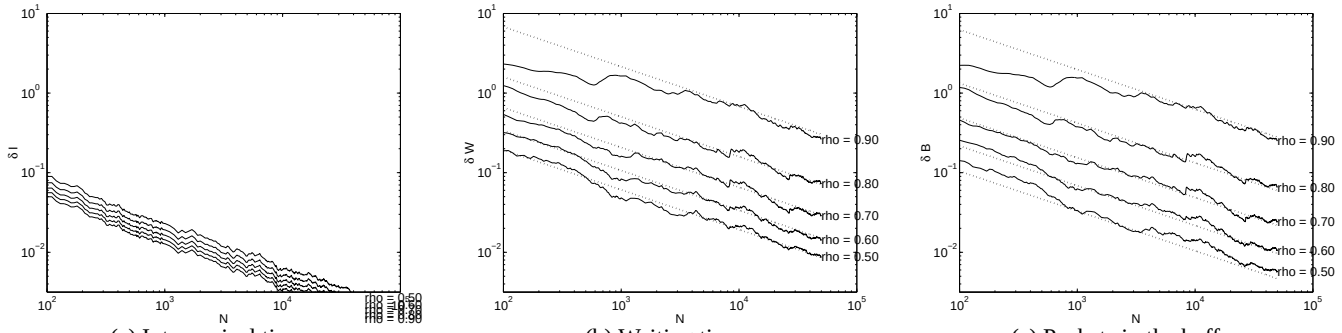
Figure 3 (c) shows the resulting square root of the asymptotic variances for the optimal active probe rate, and the square root of the asymptotic variance given passive measurements of all packets. Notice that the optimal active probe measurements have a substantially higher error than the passive measurements, in fact worse by more than a factor of 2 over a wide range of values of ρ_A.

This substantial performance reduction for the optimal active probing fits our analogy to Heisenberg's uncertainty principle rather nicely. Probing more rapidly increase the system load, and creates more variance in the results than the additional measurements can reduce, principally through inducing a longer correlation scale on the measurements. There are no fundamental barriers to converting these results into analytic formulae, though the resulting formulae would be rather complex, and we have preferred here to notate this limit by $f(\rho_A)$, and illustrate the result with the graphs.

5. EXTENSIONS

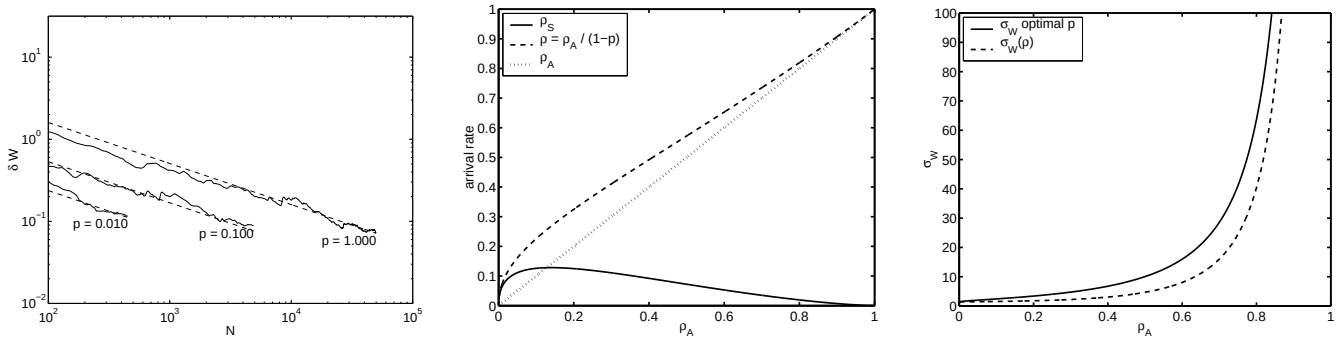
It is not the contention of this paper that the M/M/1 model is a good model for Internet systems — it clearly is not. Despite the inaccuracy of the M/M/1 model, we can learn a great deal from this simple case. Most importantly, that there are bounds in the accuracy of a finite set of measurements, and these bounds are both larger than one might expect (given the variance of the queueing process) and highly dependent of the load on the queue. These qualitative facts seem to be true in all of the extensions observed by this author (see below). Further, although they may not be particularly accurate, it is to be hoped that the above results might be useful for "back of the envelope" type calculations for estimating measurement rates and durations for some Internet studies. However, it would be desirable to have more accurate results.

Unfortunately, it is not practical to estimate the covariances of a real trace, and thereby estimate the asymptotic variance of an estimator, because the covariance estimates themselves will contain errors with magnitude larger than the asymptotic variance we wish to estimate. Hence, given the current results, we need to start with a model. Within this limitation, there are many possible models, some of which are more appropriate for Internet systems. Many already



(a) Inter-arrival times. (b) Waiting times. (c) Packets in the buffer.

Figure 2: Error bound measurements for the M/M/1 queue for various values of ρ.



(a) Waiting time errors, given different sampling rates..

(b) The optimal active probing rates, given base load ρ_A on the system in question.

(c) The square root of the asymptotic variance for the optimal active probe rate as compared with the complete information asymptotic variance..

Figure 3: Results of sampled probing.

have known results developed in the literature. For a start, although we only consider the M/M/1 queue above, results already exist for M/G/1 systems. For the M/G/1 queue (with finite variance service times) [24]

$$\bar{B} = \frac{\rho^2(c_S^2+1)}{2(1-\rho)}, \quad s_{\hat{B}}^2 \simeq \frac{\rho^2(c_S^2+1)^3}{2(1-\rho)^4}. \tag{21}$$

where $c_s^2 = (m_2 - m_1^2)/m_1^2$ is the squared coefficient of variation of the service times (m_i denotes the ith moment of the service time distribution). e.g. for the exponential distribution, $c_S = 1$, for the deterministic distribution $c_S = 0$. Note that the average queue length for this system is $E[Q(0)] = \frac{\rho^2(c_S^2+1)}{2(1-\rho)}$, so these results are similar in nature to those above, though c_S has a large impact on the results, as one might expect. Similarly results exist for arbitrary Markov processes [25], networks of queues [24], and queues with multiple types of customers [24], as well as Reflecting Brownian Motion (RBM), which can be used to model other queueing systems [24].

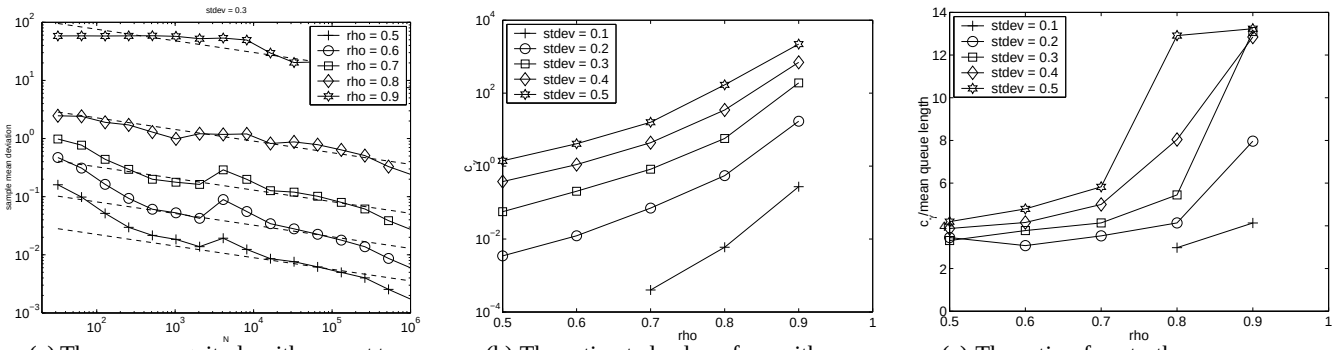
Despite the power of the results above, it is highly noteworthy that the systems for which we have valid approximations do not include systems which exhibit infinite variance, or Long-Range Dependence (LRD). It is now widely accepted that packet network traffic is *self similar* over a wide range of timescales, and exhibits LRD [35, 36]. Further, heavy-tailed distributions (often exhibiting infinite variance) are the norm. The above queueing scenarios were all Short-Range Dependent (SRD), and so their accuracy should be questioned. Unfortunately, even the more complex form of the CLT does not hold in these cases — for a start, for a LRD system the integral $\int_0^T R(u)du$ does not converge as $T \to \infty$. However, it seems quite likely that one can extend these results using the generalized CLT to allow accurate characterization of errors for these systems.

In more detail, for a discrete time, stationary process with auto-covariance function $R(k)$, LRD refers (under our definition) to the property that $\sum_{k=0}^{\infty} R(k)$ diverges [37]. As such, we obviously cannot apply the standard CLT. Likewise, the CLT depends on finite second moments for the distributions involved. Fortunately, an alternative theorem does exist for LRD data (the generalized CLT)[5]. It states that a stationary LRD process X_{H}, with a slow, power-law decrease in the auto-covariance function for large lags (i.e. $R(k) \sim c_\gamma |k|^{2H-2}$ as $k \to \infty$, $H \in (0.5, 1)$.) has, as $N \to \infty$

$$\mathrm{Var}\left(\hat{X}_{\mathrm{H}}\right) \to \frac{c_\gamma N^{2H-2}}{H(2H-1)}. \tag{22}$$

In contrast to the standard CLT, the variance decreases much more slowly with N. In fact the rate of decrease in the variance is now a function of H (typically referred to as the Hurst parameter). Figure 4 (a) shows the errors in the estimate $\bar{X}$ as a function of the number of data N on a log-log plot. Note that the case $H = 0.5$ corresponds to white Gaussian noise where the data is uncorrelated and therefore (17) applies. As H increases (corresponding to increasing LRD), the variance decreases more slowly, until in the extreme case $H \to 1$ the variance would not decrease no matter how much data we collect. Note that similar results have been seen in the context of simulation of systems involving heavy-tailed distributions [38, 39]. In these cases, the authors noted much slower convergence of estimates such as the sample mean of the queue length. The correspond-

[5] A version of the CLT also exists for heavy-tailed random variables, but we do not consider this here, as for heavy-tailed distributions the deviation from the standard CLT arises through individual events that are many times larger than the average, rather than correlations in the data. It seems likely that such distributions play a smaller part in performance measurement than in traffic measurements.



(a) The error magnitude with respect to the number of measurements N given traffic arrivals with standard deviation 0.3 and mean rate given by ρ.

(b) The estimated value of c_γ with respect to the mean and standard deviation of the traffic, denoted ρ and stdev respectively..

(c) The ratio of c_γ to the mean queue length.

Figure 4: Error bound measurements for a queue with LRD inputs. The traffic arrival process is FGN with $H = 0.8$ for a server with unit capacity.

ing Heisenberg like relationship is of the form

$$\Delta T^{2-2H}\,\Delta W^2 \;\geq\; z^2_{\beta/2}\,\frac{c_\gamma}{H(2H-1)}. \tag{23}$$

Some asymptotic results exist that suggest that given LRD inputs to a queue, the correlation structure of the queueing process will also be LRD (which stands to reason), with the same parameter H. The limiting factor in the analysis here is that we do not not have closed forms for the correlation function of queues driven by LRD input traffic. The magnitude of the correlations, c_γ will vary dependent on the system load, as well as other parameters of the input arrival process, such as the Hurst parameter H and the processes variance, but we do not have a closed form for $c_\gamma(\rho, \sigma^2_X, H)$.

Given the lack of theoretical results to provide an analytically form of the asymptotic variance of a queue with LRD inputs, we use simulation to derive such. In this case we simulate a queueing process with input traffic given by discrete-time samples from a Fractional Gaussian Noise (FGN) process. We generate approximate FGN sequences using the spectral synthesis method used also in [40], and then vary the average rate, and variance or this arrival process by simple additive, and multiplicative factors. We perform 100 simulations for each parameter value, each of 2^{21} time intervals, again dropping the first half of the measurements to avoid initial transients in the system.

Figure 4 (a) shows the empirical error bounds that result from this simulation (the solid curves, with noted markers). The figure shows, for a process with fixed standard deviation (stdev$= 0.3$), and Hurst parameter ($H = 0.8$, which is a fairly typical value for traffic data), the impact of varying ρ. We can see in this case, that the insight from the M/M/1 queue applies, only more strongly. That is, for finite length data we have fundamental limits on the accuracy with which we can know the average queueing delay, and these limits decrease more slowly for LRD dependent data.

Figure 4 (a) also shows lines (with slope $1-H$) fitted to the empirical data as a method for estimating c_γ. We perform this fitting over a wider range of parameters, and display the results in Figure 4 (b). Clearly, the parameter c_γ is highly dependent of the parameters of the arrival process. However, interestingly, the ratio of this parameter to the average queue length, shown in Figure 4 (c), appears less sensitive to the input process parameters than for the M/M/1 queue.

6. AN EXAMPLE

In this section we attempt to provide some additional insight by considering a concrete example. Consider, for instance, a queue being fed by 1500 byte packets (1500 is the MTU for Ethernet packets, and so a common packet size) which arrive in a Poisson stream. Given such arrivals, we can treat this as a M/D/1 queue, whose asymptotic variance is one eighth that of the M/M/1 queue as a result of (21), and the fact that $c_S^2 = 0$ for deterministic service times. Given (17), we can see that the asymptotic variance with respect to the measurement time T is

$$s^2_{\hat{W}}(p,\lambda,\mu) \simeq \frac{1}{p\mu^3}\frac{\rho(2-\rho)}{2(1-\rho)^2} + \frac{\rho}{2(1-\rho)^4}\frac{1}{\mu}, \tag{24}$$

where we note that $\lambda_s = p\lambda$, and the measurement interval is long enough that the approximations hold. Given a desired error ϵ the duration of measurements required to achieve this will be

$$T = \frac{1.96^2 s^2_{\hat{W}}(p,\lambda,\mu)}{\epsilon^2}. \tag{25}$$

Given a fast link, e.g. an OC48 (2.48 Gbps) link, for which packet transmission times are 4.8 μs, and the link can carry roughly 200,000 packets per second, we can see that the first term in $s^2_{\hat{W}}(p,\lambda,\mu)$ is very small, and so the asymptotic variance is almost independent of sampling rate. However, also note that the measurement interval must be long enough that $T = \lambda N$, is a reasonable approximation, which would not be the case for $N < 10$. Hence there is a second bound imposed on the measurement time by this restriction.

Figure 5 shows such a set of results, for a reasonable range of values of ρ. Notice, as we would expect from (24) the asymptotic variance bounds are nearly independent of p, except for very small sampling rates p and low ρ. In fact, these bounds suggests that one is much better off sampling at a very low rate. However, there is a point at which the time required to collect enough samples becomes the primary limitation, for instance, when $p = 10^{-9}$ we can see that we need to observe the system for around 10^6 seconds (for low traffic ρ) in order to observe around 10 packets. Given this tradeoff, and the plotted results, one might deduce that a sampling rate of $p = 10^{-9}$ was not unreasonable. We could easily make the sampling rate entirely time dependent, so we sample x packets per second regardless of the traffic rate ρ. Given the results insensitivity to p, the a sample rate of a packet every 1000 seconds would be adequate to produce results as accurate as the asymptotic variance would allow.

The problem is that the observations times are long. The correlations in the results imply observation times of the order of one million seconds (10 to 12 days). Such observation times are clearly unrealistic given the limitations on stationarity in these measurements, and the desire to rapidly detect changes. Notice that the critical component of the asymptotic variance contains a factor of $1/\mu$, so the situation scales with link speed, meaning that for different link speeds, similar observation times are required. Also note, that while the M/D/1 queue is not realistic, this is almost the easiest possible

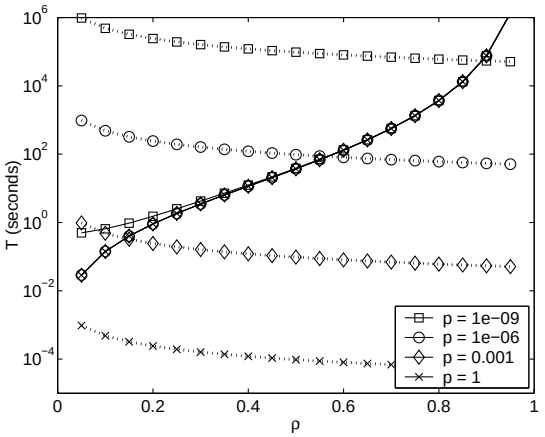


Figure 5: Estimated measurement times to reach 1ms accuracy on the M/M/1 queue, with sampling rate p, traffic intensity ρ and mean service time 4.8 μs. The solid lines are the error bounds predicted by the sample variance, while the dotted lines are the corresponding bound implied by $T = \lambda N$ for N at least 10.

case, and any changes made (for instance by introducing LRD into the arrival process, or increased variance in the service times) would make the situation worse.

We might improve the situation by loosening the accuracy bounds, to say ± 100 ms. This decreases the time requirements by a factor of 10,000, bringing measurement times down to the order of 100 seconds. However, note that this hardly produces results of the accuracy most measurement studies wish to obtain.

The important intuition that one needs to take away is that, if the queue is run at moderate to high loads, the measurement intervals required to estimate queueing delay with any reasonable accuracy are very long. Too long, in fact, for reasonable measures to be made.

7. DISCUSSION

We have seen two important results above

- finite (large) bounds on the accuracy of performance measurements, and
- these bounds are quite load sensitive.

Despite quite varied models, we see this two features in all of the models studied, and their origin in the length of busy periods, and hence the correlations in the process under study, suggest that we are likely to see such impacts in many other systems. Other papers seeking to provide quantifiable bounds to measurement accuracy such as [31, 32], have also reached similar conclusions, though not as dramatic, given that these papers did not explicitly include the impact of correlations.

The results have implications for a range of applications. Obviously, these results impact the design of measurement experiments. Primarily, it seems less important to collect fine grained data than one might naively guess. Correlations in the data reduce the utility of the additional measurements. However, this is not the only way in which these results might impact networks. There are many other places where such measurements are used: for instance in TCP RTT estimation algorithm, active queue management, and load-balancing algorithms. Considering the latter example, in particular, we see that the above results are particularly undesirables.

Load-balancing algorithms make measurements of current network loads and performance, and use these to assign load in such a way so as to balance it across network elements, thereby utilizing resources more efficiently, and improving performance. However, we have now seen that measures of performance will degrade in accuracy at exactly that time when a load-balancing algorithm is needed most (when a system is heavily loaded). Hence, we may expect to seen some instability in such algorithms, and in fact, early load-balancing routing algorithms used in the early ARPANET did indeed display unstable oscillations [7, 41].

7.1 How can we do better

The results above are bad. In particular, they suggest that a measurement system designed to rapidly detect link performance problems would have great difficulties, however, such systems are critical to high reliability networks, where problems must be rapidly detected and diagnosed.

There are a number of approaches we can use to help mitigate this problem. Firstly, often significant problems are not single link problems. A single link failure is not a big issue in most large backbones as they have redundant paths, and excess capacity for carrying rerouted traffic from a link failure. Important problems, such as might arise from a network wide DoS attack, or a network misconfiguration often impact multiple links. Where there are multiple link measurements available, one is back to the simulation case where we may effectively use an ensemble average rather than a time average. To repeat, with less technical jargon, we may get better results by using multiple measurements from different locations.

Secondly, if we had precise traffic monitoring and modeling, we may assess performance through traffic measurements directly. For instance, as noted above, it is much easier to measure the arrival rate to the M/M/1 queue, than to measure the queue itself. From such measurements we may directly estimate the queue. The limitation here is that adequate measurement infrastructure is not widely enough deployed in most large networks, and is unlikely to be so deployed in the near future, and the current Internet traffic models still need considerable practical verification for such networks.

Thirdly, we can continue to run such networks at low loads. Most large backbones are currently run at low levels of utilization. Under such loads, we can expect considerably more accurate performance estimates.

Finally, one method for reducing the correlations in data is to examine differences. These contain weaker correlations, and hence suffer less from the above impacts. Hence, the difference process would be more appropriate when one was performing tasks like anomaly detection which must be fairly rapid.

8. CONCLUSION

The results presented in this paper are important both from a practical, and a theoretical point of view. From the theoretical point of view, we have seen that there are fundamental bounds to our knowledge of Internet performance. In some respects these bounds are similar to Heisenberg's uncertainty bounds in physics. The results generalize results from simulation theory to consider Poisson sampled measurements such as are obtainable in Internet measurements.

Of equal importance are the practical results of this paper that even for a conservative model (the M/D/1 queue), we are not in a good situation with respect to measurement accuracy. Measurement intervals must be long to achieve accurate measurements, and this has impacts on many facets of network design, for instance the design of load balancing mechanisms, or congestion control protocols.

There are many interesting possibilities for continuing this work. For instance, the models above should be verified with real Internet data. Further, the implications of these results should be investigated in a range of applications such as load-balancing applications to determine the extent to which this is a problem for these applications.

Finally, note that we have concentrated here on average network delay measurements. These results seem to also apply to a range of other network performance measurements, for instance server performance measurement, and a range of other statistics of those measurements, such as the median, percentiles, and even transforms such as the Fourier transform of the data. Further, the results should be extendable to other types of measurements, such as loss, and perhaps even packet reordering metrics.

9. REFERENCES

[1] V. Paxson, G. Almes, J. Mahdavi, and M. Mathis, "Framework for IP performance metrics." IETF, IP Performance Metrics, Request for Comments: 2330, 1998.

[2] J. Mahdavi and V. Paxson, "IPPM metrics for measuring connectivity." IETF, IP Performance Metrics, Request for Comments: 2678, 1999.

[3] G. Almes, S. Kalidindi, and M. Zekauskas, "A one-way delay metric for IPPM." IETF, IP Performance Metrics, Request for Comments: 2679, 1999.

[4] G. Almes, S. Kalidindi, and M. Zekauskas, "A one-way packet loss metric for IPPM." IETF, IP Performance Metrics, Request for Comments: 2680, 1999.

[5] G. Almes, S. Kalidindi, and M. Zekauskas, "A round-trip delay metric for IPPM." IETF, IP Performance Metrics, Request for Comments: 2681, 1999.

[6] S.Floyd and V.Jacobson, "Random early detection gateways for congestion avoidance," *IEEE/ACM Transactions on Networking*, vol. 1, pp. 397–413, August 1993.

[7] J. M.McQuillan, G. Falk, and I. Richer, "A review of the development and performance of the ARPANET routing algorithm," *IEEE Transactions on Communication*, vol. COM-26, pp. 60–74, December 1978.

[8] A. Basu, A. Lin, and S. Ramanathan, "Routing using potentials: A dynamic traffic-aware routing algorithm," in *Proc. ACM SIGCOMM*, (Karlsruhe, Germany), 2003.

[9] "CAIDA performance measurement tools taxonomy." `http://www.caida.org/tools/taxonomy/performance.xml`.

[10] J.D.Case, M.Fedor, M.L.Schoffstall, and C.Davin, "Simple Network Management protocol (SNMP)." IEFT, Request for Comments: 1157, 1990.

[11] D. R. Mauro and K. J. Schmidt, *Essential SNMP*. O'Reilly, 2001.

[12] N. Brownlee, "Packet matching for NeTraMet distributions." `http://www2.auckland.ac.nz/net/Internet/rtfm/meetings/47-`, March 2000.

[13] Y. Zhang, L. Breslau, V. Paxson, and S. Shenker, "On the characteristics and origins of Internet flow rates," in *ACM SIGCOMM*, (Pittsburgh, Pennsylvania, USA), August 2002.

[14] S. B. Moon, P. Skelly, and D. Towsley, "Estimation and removal of clock skew from network delay measurements," Tech. Rep. 98-43, Department of Computer Science, University of Massachusetts at Amherst, 1998.

[15] V. Paxson, *Measurements and Analysis of End-to-End Internet Dynamics*. PhD thesis, U.C. Berkeley, 1997. ftp://ftp.ee.lbl.gov/papers/vp-thesis/dis.ps.gz.

[16] V. Paxson, J. Mahdavi, A. Adams, and M. Mathis, "An architecture for large-scale Internet measurement," *IEEE Communications Magazine*, 1998.

[17] P. Barford and J. Sommers, "Comparison of probe-based and router-based methods for measuring packet loss." in submission, (see `http://www.cs.wisc.edu/~pb/publications.html`).

[18] Y. Zhang, N. Duffield, V. Paxson, and S. Shenker, "On the constancy of Internet path properties," in *ACM SIGCOMM Internet Measurement Workshop (IMW '2001)*, (San Francisco, California, USA), November 2001.

[19] R. Wolff, "Poisson arrivals see time averages," *Opns. Res.*, vol. 30, pp. 223–231, 1982.

[20] G. S. Fishman and P. J. Kiviat, "The analysis of simulation-generated time series," *Management Science*, vol. 13, pp. 525–557, Mar 1967.

[21] G. S. Fishman, "Estimating sample size in computing simulation experiments," *Management Science*, vol. 18, pp. 21–38, Sep 1971.

[22] G. S. Fishman, "Statistical analysis for queueing simulations," *Management Science*, vol. 20, pp. 363–369, Nov 1973.

[23] A. M. Law, "Efficient estimators for simulated queueing systems," *Management Science*, vol. 22, pp. 30–41, 1975.

[24] W. Whitt, "Planning queueing simulations," *Management Science*, vol. 35, no. 11, pp. 1341–1366, 1989.

[25] W. Whitt, "Asymptotic formulas for Markov processes with applications to simulation," *Operations Research*, vol. 40, no. 2, 1992.

[26] P. M. Morse, "Stochastic properties of waiting lines," *Journal of the Operations Research Society of America*, vol. 3, no. 3, pp. 255–261, 1955.

[27] J. Abate and W. Whitt, "Transient behaviour of the M/M/1 queue via Laplace transforms," *Advances in Applied Probability*, vol. 20, pp. 145–178, 1988.

[28] J. Abate and W. Whitt, "The correlation functions of RBM and M/M/1," *Stochastic Models*, vol. 4, no. 2, pp. 315–359, 1988.

[29] J. Abate and W. Whitt, "Transient behavior of the M/M/1 queue: Starting at the origin," *Queueing Systems: Theory and Applications*, vol. 2, no. 1, pp. 41–65, 1987.

[30] J. Abate and W. Whitt, "Transient behavior of the M/G/1 workload process," *Operations Research*, vol. 42, no. 4, pp. 750–764, 1994.

[31] J. A. Schormans and T. Timotijevic, "Evaluating the accuracy of active measurement of delay and loss in packet networks," in *6th IFIP/IEEE Conf. on Management of Multimedia Networks and Services (MMNS)*, pp. 409–421, Sept. 2003.

[32] T.Timotijevic, C.M.Leung, and J.Schormans, "Accuracy of measurements techniques supporting QoS in packet-based intranet and extranet VPNs," *IEE Proc.-Commin*, vol. 151, pp. 89–94, Feb. 2004.

[33] Andren, Hilding, and Veitch, "Understanding end-to-end internet traffic dynamics," in *Proceeding Globecom'98*, (Sydney, Australia), pp. 1118–1122, 1998.

[34] L. Kleinrock, *Queueing Systems*, vol. II: Computer Applications. John Wiley and Sons, 1975.

[35] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of Ethernet traffic (extended version)," *IEEE/ACM Transactions on Networking*, vol. 2, pp. 1–15, Feb 1994.

[36] V. Paxson and S. Floyd, "Wide-area traffic: The failure of Poisson modeling," *IEEE/ACM Transactions on Networking*, vol. 3, no. 3, pp. 226–244, 1995.

[37] J. Beran, *Statistics for Long-Memory Processes*. Chapman and Hall, New York, 1994.

[38] A. Erramilli, J. Gordon, and W. Willinger, "Applications of fractals in engineering for realistic traffic processes," in *Proceedings of the 14th International Teletraffic Congress - ITC 14* (J. Labetoulle and J. W. Roberts, eds.), vol. 1a, pp. 35–44, Elsevier, Amsterdam, 1994.

[39] M. E. Crovella and L. Lipsky, "Long-lasting transient conditions in simulations with heavy-tailed workloads," in *Proceedings of the 1997 Winter Simulation Conference*

[40] M. Roughan and D. Veitch, "Measuring long-range dependence under changing traffic conditions," in *IEEE INFOCOM'99*, (NY, NY), IEEE Computer Society Press, Los Alamitos, California, March 1999.

[41] J. M.McQuillan, I. Richer, and E. C. Rosen, "A new routing algorithm for the ARPANET," *IEEE Transactions on Communication*, vol. COM-28, pp. 711–719, May 1980.

APPENDIX

A. PROOFS

LEMMA A.1. *For a symmetric function $f(\cdot)$ the following holds*

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} f(i-j) = f(0) + 2\sum_{k=1}^{N} f(k)\left(1-\frac{k}{N}\right),$$

and as $N \to \infty$

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} f(i-j) \to f(0) + 2\sum_{k=1}^{\infty} f(k).$$

where the latter sum converges.

PROOF.

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} f(i-j) &= \frac{1}{N}\sum_{i=1}^{N} f(0) + 2\frac{1}{N^2}\sum_{i=2}^{N}\sum_{j=1}^{i-1} f(i-j) \\
&= f(0) + 2\frac{1}{N}\sum_{i=2}^{N}\sum_{k=1}^{i-1} f(k) \\
&= f(0) + 2\frac{1}{N}\sum_{k=1}^{N-1} f(k)\sum_{i=k+1}^{N} 1 \\
&= f(0) + 2\frac{1}{N}\sum_{k=1}^{N-1} f(k)(N-k) \\
&= f(0) + 2\sum_{k=1}^{N-1} f(k)\left(1-\frac{k}{N}\right) \\
&\to f(0) + 2\sum_{k=1}^{\infty} f(k).
\end{aligned}$$

where the sum converges. □

THEOREM A.1. *Take a wide-sense stationary process $X(t)$, with mean μ_X, variance σ_X^2, and auto-covariance function $R(s) = E[X(t)X(t+s)] - \mu^2$. Form the discrete-time series X_i by sampling at time points $t_i = i$, then the sample mean $\hat{X}_N = 1/N\sum_{i=1}^{N} X_i$ has asymptotic variance*

$$\lim_{N\to\infty} N\,\mathit{Var}\left(\hat{X}_N\right) = \sigma_X^2 + 2\sum_{i=1}^{\infty} R(i),$$

where the sum is finite.

PROOF.

$$\begin{aligned}
N\,E\left[\hat{X}_N^2\right] &= \frac{1}{N}E\left[\left(\sum_{i=1}^{N} X_i\right)^2\right] \\
&= \frac{1}{N}E\left[\sum_{i=1}^{N}\sum_{j=1}^{N} X_i X_j\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} E[X_i X_j] \\
&= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[R(i-j)+\mu^2\right] \\
&= N\mu_X^2 + \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} R(i-j)
\end{aligned}$$

Noting that $R(0) = \sigma_X^2$, and the variance $\mathrm{Var}\left(\hat{X}_N^2\right) = E\left[\hat{X}_N^2\right] - E\left[\hat{X}_N\right]^2$, where the expected value of the sample mean $E\left[\hat{X}_N\right] = \mu$, we can apply Lemma A.1 to get

$$\lim_{N\to\infty} N\,\mathrm{Var}\left(\hat{X}_N\right) = \sigma_X^2 + 2\sum_{i=1}^{\infty} R(i),$$

where the sum is finite. □

THEOREM A.2. *Take a wide-sense stationary process $X(t)$, with mean μ_X, variance σ_X^2, and auto-covariance function $R(s)$. Form the discrete-time series X_i by sampling at time points t_i, drawn from a Poisson process with rate λ, then the sample mean $\hat{X}_N = 1/N\sum_{i=1}^{N} X_i$ has asymptotic variance*

$$\lim_{N\to\infty} N\,\mathit{Var}\left(\hat{X}_N\right) = \sigma_X^2 + 2\lambda\int_0^{\infty} R(u)du,$$

where the integral is finite.

PROOF. The mean μ_X of the process can be extracted much as above, so for simplicity we prove the result for a mean zero process $\mu_X = 0$, and leave the generalization to the reader. Now imagine that our sample times t_i are a Poisson process with rate λ, then, as above

$$\begin{aligned}
N\,E\left[\hat{X}_N^2\right] &= \frac{1}{N}\sum_{i=0}^{N}\sum_{j=0}^{N} E[X_i X_j], \\
&\to f(0) + 2\sum_{k=1}^{\infty} f(k),
\end{aligned}$$

where $f(k) = E[X_{i+k}X_i]$. Note that (for $i > j$) the time interval between t_i and t_j is a sum of $i-j$ exponential random variables, which is an Erlang distribution, with density function

$$\begin{aligned}
p(t)dt &= \text{prob}\left\{t_i - t_j \in [t, t+dt)\right\} \\
&= \lambda\frac{(\lambda t)^{i-j-1}}{(i-j-1)!}e^{-\lambda t}dt.
\end{aligned}$$

The expected value $E[X_i X_j]$ can be expanded (using the theorem of total probability) to give

$$\begin{aligned}
E[X_i X_j] &= \int_0^{\infty} E[X_i X_j | t_i - t_j = u]\, p(u)\, du, \\
&= \int_0^{\infty} R(u)p(u)\, du,
\end{aligned}$$

where the integral is finite. This leads to the following

$$\begin{aligned}
N\,E\left[\hat{X}_N^2\right] &\to E\left[X_0^2\right] + 2\sum_{k=1}^{\infty} f(k) \\
&= E\left[X_0^2\right] + 2\lambda\sum_{k=1}^{\infty}\int_0^{\infty} R(u)\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-\lambda u}\, du \\
&= E\left[X_0^2\right] + 2\lambda\int_0^{\infty} R(u)\sum_{k=1}^{\infty}\frac{(\lambda t)^{k-1}}{(k-1)!}e^{-\lambda u}\, du \\
&= \sigma_X^2 + 2\lambda\int_0^{\infty} R(u)\, du,
\end{aligned}$$

where the integral is finite. □