Wireless Traffic: The Failure of CBR Modeling

Stefan Karpinski, Elizabeth M. Belding, Kevin C. Almeroth
{*sgk,ebelding,almeroth*}*@cs.ucsb.edu*

Department of Computer Science
University of California, Santa Barbara

***Abstract*— Performance predictions from simulations in wireless networking rarely seem to match the behavior observed once the same technologies are deployed. We believe that one of the major factors hampering researchers' ability to make more reliable forecasts is the inability to generate realistic experimental workloads. To redress this problem, we take a fundamentally new approach to quantifying the realism of wireless traffic models. In this approach, the realism of a model is defined directly in terms of its ability to accurately reproduce the performance characteristics of actual network usage. This direct approach cuts through the Gordian knot of deciding which statistical features of traffic traces are significant. The first major contribution of this work is this new definition of workload realism, together with the analytical and statistical methodology to rigorously assess whether synthetic traffic models meet the definition. The second contribution is the conclusion that commonly used models of wireless traffic distort important metrics for performance evaluation at every layer of the wireless protocol stack. We show by example that this distortion can completely invert the relative performance of protocols. The last and most important contribution is the complete collection of ideas, techniques and analytical tools that will allow the development of more realistic synthetic traffic models in the future.**

I. Introduction

The evaluation of wireless technology requires the generation of workload to test the viability and performance of the new protocol or technique being studied. We believe that lack of realism in traffic workload generation is one of the major limiting factors that prevents simulations and experimental test-bed deployments from accurately predicting the real-world performance of wireless technologies. Today, very little is understood about the impact of different workloads on network performance. Uniform constant bit-rate traffic (CBR) is commonly used to evaluate protocols, but there is no evidence that behavior under such workloads is an accurate predictor of performance under real usage patterns. The inability to experimentally forecast real-world performance is a severe handicap to the entire networking community. It hinders the ability to effectively develop better solutions to the many difficult problems that face emerging wireless technologies.

This paper presents a fundamentally new approach to creating realistic models of workload in networks. Rather than subjectively choosing statistical measures that may or may not actually influence network performance, we define the realism of models directly in terms of their ability to accurately reproduce important metrics. We define a traffic model to be *sufficiently realistic* with respect to a given performance metric, if the model produces metric values that are similar to those observed using the original trace to generate traffic. Using this new definition of sufficient realism, we explore the space of synthetic traffic models, establishing rigorously for the first time that standard but simplistic models of network traffic drastically misrepresent important performance metrics at all levels of the protocol stack.

Our analysis shows that the standard CBR traffic model, and various partially synthetic variations thereof, significantly misrepresent important performance metrics at every level of the network. The delay for application data to traverse end-to-end from the sender to receiver, for example, is the paramount metric for many multimedia applications. We find that the most common traffic model, which uses random end-points, with uniform CBR flows, underestimates average end-to-end delay by more than a factor of 5, on average, and by more than a factor of 11 in 25% of usage scenarios. Network control overhead is overestimated, on average, by a factor of about 2.5, while link control overhead is underestimated by almost a factor of 2. Even for metrics where the average and median misrepresentations are not so extreme, the statistical characteristics of error values often indicate that something is fundamentally unrealistic about the synthetic performance measurements. None of the partially synthetic models manage to accurately represent more than one or two performance characteristics. We use the AODV and OLSR ad hoc routing protocols to show that the relative performance of protocols can be switched when changing from using a CBR traffic model to using real traffic: using random end-point, uniform CBR traffic, AODV appears to induce less link-layer overhead, whereas using real traffic, it in fact induces more.

The first major contribution of this work is the definition of "sufficient realism" together with the analytical and statistical methodology to rigorously test whether synthetic traffic models meet this definition. The second contribution is the conclusion that the most commonly used model of wireless traffic drastically and consistently misrepresents some of the most important metrics for performance evaluation of wireless protocols. Many performance comparisons based on this model may need revisiting. The last and most important contribution is the collection of ideas and analytical tools necessary to create more realistic synthetic traffic models in the future. By applying the methods developed in this work, it will be possible to discover precisely which aspects of network usage affect the realism of performance results. Once this is known, it will become feasible to create models that accurately reproduce those aspects of real wireless usage.

The rest of the paper is organized as follows. In Section II we present motivation and related work. Our experimental and analytical methodology is presented in Section III, while the results of our experiments are explained and analyzed in Section IV. The ramifications of these results are discussed in Section V. Finally, in Section VI, we conclude with how this research may be applied to current wireless studies, and how it points the way to better traffic models for the future.

II. Motivation & Related Work

The history of networking research contains many examples of simplistic models that have proven not only to be inaccurate, but also to drastically skew important characteristics of network behavior. Paxson and Floyd showed that the Poisson packet arrival model, which had been standard for studying wide-area Internet traffic, failed to capture the burstiness and self-similarity of real traffic [1]. The equally common but simplistic Random Way-Point (RWP) mobility model was found to exhibit "density waves" and gradual slowdown of average node speed [2], [3]. In the worst cases, overly simplistic models can switch the relative performance of protocols, thereby invalidating the conclusions drawn from performance comparisons using those models.

The interaction of wireless user and application behavior with the lower layers of the networking stack is characterized by where, when, how much, and to whom data is transmitted. The joint pattern of traffic generation and mobility through time and space completely determines the effect of wireless usage on the lower levels of the network. This is due to the data-agnostic nature of the protocol stack: by design, IP networks treat all data in the same manner [4].[1] The credibility of conclusions derived from simulation or experimental deployment depends crucially on our confidence that the models used to generate traffic and motion in the experiments are sufficiently realistic.

Paxson and Floyd observed in [1] that the interplay between end-point behavior and the network conditions is inherently *closed-loop* in the sense that it is potentially affected by complex feedback. Traffic models typically attempt to preserve the closed-loop behavior of network traffic [5]–[7]. This presents a fundamental difficulty, however, in that it presumes that we know the intent of end-points: what *would* they have done under different conditions? While we can speculate about what an individual node might hypothetically do, we currently do not understand the impact of the full-network traffic pattern upon performance metrics at all—even without trying to account for hypothetical reactions to alternate situations. Especially in the wireless setting, a fuller understanding of total network behavior must be reached before we can sensibly tackle the complexities of multi-level behavioral feedback. Accordingly, in this paper we attempt to provide a first-order approximation of complete network behavior by studying the response of performance metrics to *open-loop* traffic models without multi-level feedback. It is important to realize that while this does not provide a final picture, we currently lack even a first-order understanding of the effect of different workloads on performance. This first-order understanding is an essential initial step.

There have been a significant number of studies of large wireless network deployments [8]–[16]. These analyses have described a wide variety of aspects of wireless network behavior, and provide much insight into the workings of real, deployed wireless networks. These studies present a broad analysis of general system features and trends of specific corporate wireless local-area networks (WLANs) [8], [10], university campus WLANs [11]–[14], [17], [18], and temporary WLANs at conference venues [9], [15], [16]. They also provide a large body of raw data for subsequent analysis and modeling research. Our work provides the methodology for turning this rich foundation of field data into usable, realistic models of workload for a wide variety of networking situations.

The choice of mobility models for mobile wireless simulations can have a drastic impact on important performance metrics [3], [19]–[22]. Moreover, commonly used but simplistic mobility models, such as RWP, exhibit characteristics, including density waves and speed decay, that are categorically dissimilar from any known real-world behavior [3], [20]. In response to this evidence, more realistic mobility models have been proposed [18], [21], [23]. While much of this work focuses on making models that are simply more intuitively appealing [21], [23], some work has begun to capitalized on this newly created wealth of wireless field data, by deriving models from observed usage behavior, rather than intuition alone [10], [18].

In this paper, instead of mobility, we examine an even more fundamental aspect of user behavior in wireless networks: the pattern of traffic generated by users and applications. This aspect of behavior is more fundamental because it applies to all types of wireless networks, not just mobile and ad hoc networks. Moreover, the effect of traffic patterns applies not only to simulations, but also to experimental deployments, which have become the gold standard for wireless protocol evaluation. Experimental deployment sidesteps the issue of accurately modeling the lower layers of the network. Unless traffic and mobility are modeled realistically, however, the experimental results will still be unreliable.

There is a large and diverse body of work on traffic analysis, modeling, and generation [1], [5]–[7], [24], [25]. We are only able to discuss a small, but hopefully representative sampling of this work. Almost all of the traffic generation work has focused on wide-area Internet backbone traffic. The two most prominent traffic generation frameworks are Harpoon and D-ITG. Harpoon [25] uses a traffic trace for self-training, and can subsequently generate synthetic traffic with certain statistical properties based on the original trace. The properties reproduced are the empirical distributions of the following: "file size, inter-connection time, source and

[1]This is violated by some quality of service (QoS) schemes. However, we can simply add QoS metadata—such as traffic classes or urgency flags—to our models of user behavior and the rest of our arguments remain valid. The network is still disinterested in the exact content of the data being transported; only the QoS metadata is relevant.

destination IP ranges, number of active sessions." There is no criterion proposed to determine whether these properties characterize the original traffic adequately—we can only hope that this approximation is good enough. For many purposes, it likely is sufficient; in particular, it is probably appropriate for the intended use in generating traffic for Internet backbone simulations. Wireless networks, however, are particularly sensitive to workload conditions, and sampling form a limited set of empirical distributions does not suffice to reproduce realistic network-wide traffic.

D-ITG [5], [6] generates flows using a simple independent sampling model for packet sizes and inter-packet intervals. The framework contains pre-made models for several common types of Internet traffic. The focus of this project, however, is on providing the infrastructure to generate very large volumes of synthetic Internet-like backbone traffic. No analysis is provided for determining the realism of traffic mixes, or for choosing flow end-points realistically. In wireless networks, these factors are of crucial importance to performance, and cannot be overlooked. Both Harpoon and D-ITG provide excellent traffic generation platforms, but do not provide a systematic framework for understanding or reproducing realistic whole-network workload in the wireless setting.

III. Methodology

The art of simulation lies in knowing which details must to be realistic and which may be abstracted into simpler, approximate models without affecting the accuracy of the results. Clearly, we need not simulate subatomic particle interactions in a wireless network simulator. Instead we use high-level physical models that approximate the real physics well enough that the performance results are the same. Similarly, when modeling network usage, we must require that our models produce the same results as actual user behavior would. This requirement leads us to the following definition:

Definition. A model of user behavior is *sufficiently realistic* if, when compared with actual user behavior, the model, with parameter values extracted from the real data, yields statistically equivalent performance results.

This definition depends on many factors: the type of wireless scenario, the performance metrics under consideration, the actual usage behavior used for comparison, and how strong a notion of statistical equivalence is required. We discuss different measures of error and how to evaluate statistical equivalence in Sections III-D and III-E.

A. Trace Data

Our general methodology is to compare performance metrics in simulations using real traffic patterns from traces to the same metrics in simulations using a variety of synthetic traffic models, including the standard random, uniform CBR model. For our analysis, we use a 24-hour trace recorded in an infrastructured 802.11g wireless LAN with 18 access points, deployed at the 60th Internet Engineering Task Force meeting (IETF60), held in San Diego during August of 2004. The traffic

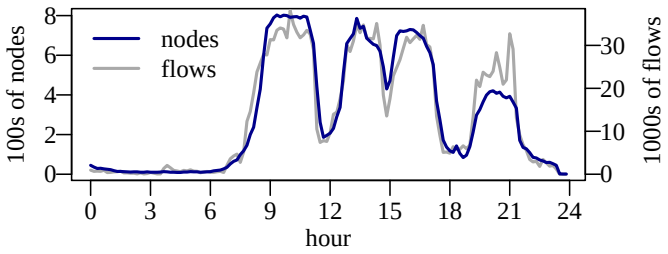

Figure 1: The number of active nodes and flows over time.

trace was captured using `tcpdump` at a single router, through which all wireless traffic for the meeting was routed, including traffic between wireless nodes. The snap length of the capture was 100 bytes, allowing IP, ICMP, UDP and TCP headers to be analyzed. We limit our work to the 24-hour sub-trace recorded on Wednesday, August 4th. This trace contains a broad variety of behaviors and entails a very large volume of traffic: 2.1 million flows, 58 million packets, and 52 billion bytes.

We do not assume or claim that the traffic found at IETF60 is representative of conference settings in general. The observed behaviors are also unlikely to resemble those found in a typical commercial or residential setting. We have chosen this trace, however, because within it can be found behaviors resembling many different types of wireless usage cases. Figure 1 shows the wide variations in the number of active flows and nodes over the course of the trace. In the night and morning hours, the traffic patterns are similar to those one might find in a moderately trafficked business or residential area. During working group sessions, we see highly concentrated, heavy usage patterns. At the zenith of activity, over 800 users, 33 thousand flows, and 1 million packets are seen in a single 10-minute trace segment. At the nadir, a lone node sent only a single 61-byte packet in 10 minutes. All levels of activity between these extremes are represented. Moreover, the mix of traffic types observed changes dramatically over the course of the day, providing a wide representation of possible blends of behavior. This heterogeneity and extreme range of behaviors makes the IETF data set ideal for this evaluation. The variety of activity gives us greater confidence that success or failure of traffic models is not tied to any specific network condition, but is broadly and generally applicable.

Before using the traces, it is necessary to extract application-level behavior from the trace header data. First, we split the trace into individual packet flows. A flow is a series of packets sharing the following five attributes: IP and transport protocols (raw IP, ICMP, TCP, UDP); source and destination IP addresses and TCP/UDP port numbers. Next, the quantity of application-initiated data contained in each packet is calculated. For non-TCP packets, this quantity is simply the size of the transport-layer payload, but for TCP the calculation is more complicated: only new data transfers, explicitly initiated by the application are counted. Data retransmitted by TCP is disregarded, and empty ACKs are ignored. SYN and FIN flags in packets (even empty ones) are counted as a single byte each, since they are explicitly signaled by the application.

Behavior Level	Model	Description
Flow Topology	**T**race	Mapping of flow end-points to wireless nodes taken directly from trace data.
	Sink	One end-point is internal and the other external. The internal node is randomly chosen.
	Uniform	Both end-points are uniformly randomly chosen from all of the wireless nodes.
Flow Behavior	**T**race	Each flow has the actual start time, end time, total data sent, and number of packets from the trace.
	Uniform	All flows have the same duration, volume, packet rate, and data rate as the trace average.
Packet Behavior	**T**race	The sequence of packet sizes and inter-transmission intervals is taken directly from the trace.
	Uniform	Packets sizes are uniform and the inter-packet interval is constant (CBR).

Table I: The three orthogonal levels of traffic behavior, and the traffic models used for each level.

B. Simulations

We use the Qualnet wireless network simulator to perform our experiments. We simulate a stationary multi-hop 802.11b network using the Ad hoc On-demand Distance Vector (AODV) routing protocol, with nodes placed randomly in a square field with sides of 1500 meters. In addition to the active nodes corresponding to trace IPs, equally many passive "infrastructure" nodes are added to each simulation: these nodes initiate no data and simply serve as additional network relays. Our simulations resemble multi-hop mesh networks of the kind that are increasingly studied and deployed for delivery of broadband access in residential, corporate and conference settings. We do not attempt to reproduce the physical environment of the original wireless network, nor do we simulate mobility. The only aspect of the original network's behavior that is reproduced is the total pattern of network-wide traffic.

There are a number of potential objections to this approach. We use single-hop trace data to drive multi-hop simulations; the physical environment, node mobility, handover behavior, and closed-loop dynamics (including TCP feedback) of the original wireless setting are not faithfully reproduced. One must keep in mind, however, that the goal of this research is *not* to understand the conditions of the original network. Rather, we are using the traffic behaviors observed as examples to help us better understand how different types of workload can affect performance metrics. In particular, we aim to understand how real workload compares with common synthetic traffic models. Of course, the reason for such objections is that networking researchers understand that the many aspects of behavior interact with each other in a complex and nearly inextricable manner. However, before we can hope to understand the interaction between workload and other features affecting network behavior, we must study traffic patterns alone, and learn to model them with reasonable accuracy in the absence of additional complicating factors. Accordingly, in this study, we detach application level traffic patterns from the other factors influencing network conditions, and study them in isolation.

The 24-hour trace is split into 144 10-minute segments, each of which serves as the basis for a set of simulations using different traffic models. The traffic models range from a completely realistic trace-driven model, to a standard CBR traffic model. Various partially synthetic intermediate models, described in Section III-C, are simulated to study the impact of different aspects of traffic behavior on network performance. To preserve the fairness of the performance comparison, we keep as many features as possible constant across different traffic models. The traffic generated by each synthetic model preserves as many characteristics from the original trace as possible, within the constraints of the model. Moreover, the following features are preserved across all models: the numbers of wireless nodes, the number of flows, the number of application-initiated data units sent, the total bytes of application data sent, and the average flow duration (and therefore the average data rate).

C. Traffic Models & Performance Metrics

We separate our traffic generation models into three orthogonal and nearly independent levels of behavior:

1) **Flow End-Point Topology**: which nodes communicate with each other, and how frequently; i.e. how flow end-points are mapped onto nodes in the network.
2) **Flow Behavior**: high-level parameters for each flow, including start time, end time, packets sent, bytes sent.
3) **Packet Behavior**: sizes of individual packets, and the intervals between their transmission.

For each level of traffic behavior, we compare several different behavior models. The different levels of models and variants at each level are listed and described in Table I and illustrated with specific examples in Figure 2.

The three levels of behavior are orthogonal, and can be varied almost independently. The exception to their independence is that the trace-based packet behavior model can only be used when the flow behavior model is also trace-based. Once flow behavior is decoupled from the actual trace, there is no natural way to preserve packet behavior. This eliminates three combinations of models and leaves nine viable behaviors, abbreviated by the first letters of their flow topology, flow behavior, and packet behavior models: TTT (fully trace-based), TTU, TUU, STT, STU, SUU (entirely synthetic, sink topology), UTT, UTU, UUU (entirely synthetic, uniform topology).

We have selected nine performance metrics at the application, network, and link layers of the protocol stack. These metrics are commonly used to evaluate the performance of new wireless protocols:

1) **Application:** average end-to-end delay, average jitter, total received throughput.
2) **Network:** AODV control overhead (RREQ/RREP/RERR), RREQs initiated per node, routing queue drop rate.

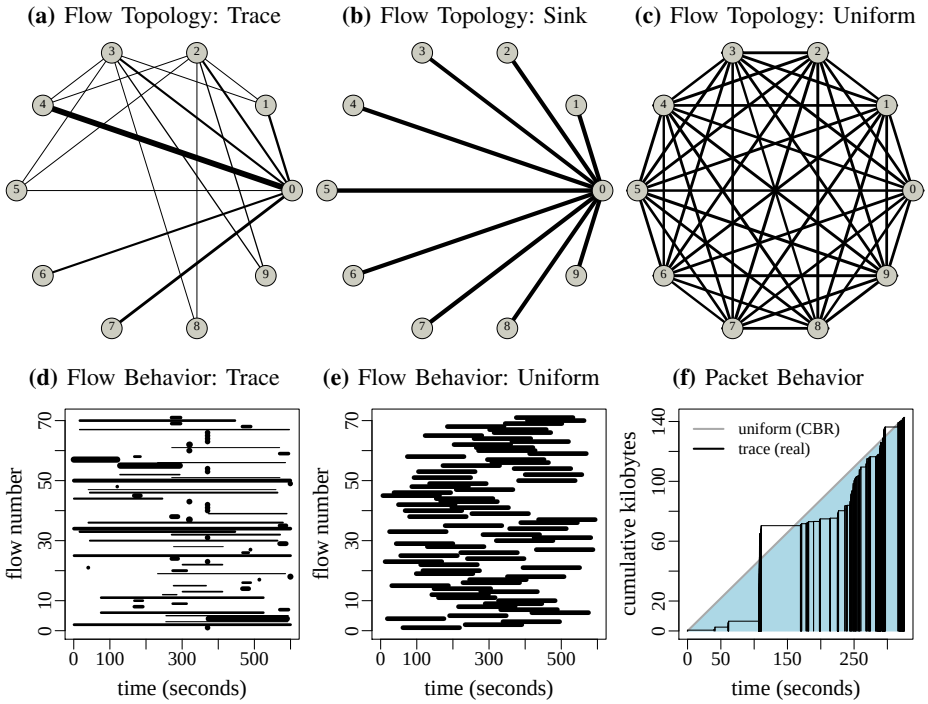

Figure 2: Examples illustrating the different traffic models for the three levels of behavior. Figures 2a, 2b and 2c, show example flow topologies. The width of each line is proportional to the logarithm of the number of flows between the nodes it connects. In each graph, node zero is the gateway to the Internet. Uniform and trace flow behavior examples are plotted in Figures 2d and 2e. The time axis indicates when the various flows start and end; the width of each flow line is proportional to the logarithm of its data rate. Flow numbers are assigned arbitrarily. Figure 2f compares packet behavior for the uniform model (i.e. CBR), with the trace of an actual flow. In the uniform model, the cumulative data sent increases smoothly over time (gray diagonal line). In the actual packet trace, the data transmissions are variable both in size and in inter-transmission interval, leading to a "lumpy" cumulative data history (black step-function).

3) **Link (MAC):** control overhead (RTS/CTS/ACK), packet retransmission rate, retransmission failure rate.

For each of the 144 10-minute trace segments, we have run simulations using each of the nine traffic models, for a total of 1,296 simulations.

D. Measures of Error

The simulations described in Section III-C provide us with the raw data to compare performance metrics for synthetic traffic models with those for real traffic traces. To assess the realism of these models, however, we need a measure of how inaccurate the synthetic performance values are when compared to the real values. Let x be the value of a performance metric using real traffic, and y the value of the same metric using an alternate traffic model, M. Some common measures of error are the difference $[y-x]$, the ratio $[y/x]$, and the standard error $[(y-x)/x]$. These are all reasonable measures of error; but which is most appropriate for assessing the realism of performance metrics? Instead of picking one arbitrarily, we will first consider the properties that an ideal error function should have, and then use those properties to determine the best measure of error. Moreover, we will show that the unique measure of error that exhibits these ideal properties is the *log-ratio* of metric values:

$$\log(y/x) = \log(y) - \log(x). \tag{1}$$

In this discussion, $E(x, y)$ is a generic error function applied to the synthetic value, y, with respect to the real value, x.

The first property that an error function should have is *insensitivity to common factors*. That is, if both values are scaled by the same constant, the error should be unaffected:

$$\forall\, x, y, c : E(xc, yc) = E(x, y) \tag{2}$$

There are three major motivations for this requirement:

1) Changing units should not affect error values.
2) Error values for "large" and "small" scenarios should be directly comparable. Scenarios with large x values will naturally have larger raw differences between x and y. This requirement allows scenarios of different scales to be compared fairly and without bias.
3) Changing between metrics that differ by a known constant for each scenario should not affect error values.

The last point is best illustrated by an example. Consider two closely related performance metrics: average throughput, t, and total bytes received, r. Suppose that there are f flows in a given scenario with average duration, d. Since $t = r/fd$, the metrics t and r contain the same information—they differ only by a known constant in each simulation scenario. Equation 2 ensures that the errors of these metrics are the same:

$$E(t^{\mathrm{TTT}}, t^{\mathsf{M}}) = E\left(\frac{r^{\mathrm{TTT}}}{fd}, \frac{r^{\mathsf{M}}}{fd}\right) = E(r^{\mathrm{TTT}}, r^{\mathsf{M}}). \tag{3}$$

The difference measure does not satisfy Equation 2, but the ratio, standard error, and log-ratio error measures all do.

The second property that an ideal error function should have is *additivity of compounded errors*. If two independent causes of error each induce some factor of misrepresentation, then the combined error should be the sum of the errors caused by each factor separately:

$$\forall\, x, c_1, c_2 : E(x, xc_1c_2) = E(x, xc_1) + E(x, xc_2). \tag{4}$$

This property allows us to compare error values meaningfully across different traffic models. For example, if flow topology and packet behavior affected some performance metric independently with no interaction effects, we would expect that

$$E(x^{\text{TTT}}, x^{\text{UTU}}) \approx E(x^{\text{TTT}}, x^{\text{TTU}}) + E(x^{\text{TTT}}, x^{\text{UTT}}). \quad (5)$$

If these two values differ significantly, there must be some interaction between the two levels of behavior that introduces more error than can be explained by each separately. Without the property of additivity given in Equation 4, such a comparison would not be possible or meaningful.

Additivity of compounded errors also implies two desirable properties that are easily derived from Equation 4. It forces the error of an accurate representation to be zero: $E(x, x) = 0$. It also forces underestimation and overestimation to be treated symmetrically. The error of underestimating by some factor is opposite but equal to overestimating by the same factor:

$$E(x, x/c) = -E(x, xc). \quad (6)$$

It is easily verified that the difference, ratio, and standard error measures do not satisfy Equation 4, and the difference, as noted, does not satisfy Equation 2. The log-ratio is the only metric presented that satisfies both conditions. Moreover, it can be proved that $\log(y/x)$ is the *only* differentiable function that satisfies both (up to a constant). In the Appendix, we present a proof of this claim. Throughout the rest of the paper, we use the log-ratio to measure the error of performance metrics.

E. Tests of Statistical Equivalence

In this section, we consider the values of performance metrics as random variables, drawn from unknown distributions. We present three tests for the statistical equivalence of the metric values induced by synthetic and real traffic. Let M be a traffic model as before and let X be a performance metric. Let X_k^{M} be a random variable representing the value of X in the k^{th} scenario using the traffic model M. If the distribution of X_k^{M} is the same as that of X_k^{TTT}, then both the median and mean values of the log-ratio $R_k^{\mathsf{M}} = \log(X_k^{\mathsf{M}}/X_k^{\text{TTT}})$ should be zero. The first two tests check the plausibility of precisely these hypotheses. The third test separates small, medium, and large simulation scenarios, and test their means separately to catch any size-dependent performance bias.

The Median Test. If M induces realistic performance, then the median of each log-ratio variable, $R_k^{\mathsf{M}} = \log(X_k^{\mathsf{M}}/X_k^{\text{TTT}})$, should be zero. The k^{th} indicator variable is defined as

$$I_k^{\mathsf{M}} = \begin{cases} 0 & \text{if } R_k^{\mathsf{M}} < 0, \\ 1 & \text{if } R_k^{\mathsf{M}} \geq 0. \end{cases} \quad (7)$$

If the median value of R_k^{M} is truly zero, then I_k^{M} is a Bernoulli variable with probability parameter $p = \frac{1}{2}$. The variables I_k^{M} are all independent since they come from separate simulations, and cannot affect each other's outcomes. Therefore, the sum $S_n^{\mathsf{M}} = \sum_{k=1}^{n} I_k^{\mathsf{M}}$ should follow a binomial distribution of n trials with $p = \frac{1}{2}$. The median test applies the exact binomial cumulative distribution function (CDF) for n and p to the observed value of S_n^{M}, yielding a p-value: the probability that such an extreme value would occur by chance under the hypothesis that the median of each R_k^{M} is zero.

The Mean Test. We use Lyapunov's generalization of the Central Limit Theorem (LCLT) to test the hypothesis that the mean of each R_k^{M} is zero. We present the theorem and its application to the series R_k^{M} in the Appendix. Here we simply present the resulting test statistic and its usage. If the mean of each R_k^{M} is zero, then LCLT implies that the test statistic

$$\hat{Z}_n^{\mathsf{M}} = \frac{\sum_{k=1}^{n} R_k^{\mathsf{M}}}{\sqrt{\sum_{k=1}^{n} (R_k^{\mathsf{M}})^2}} \quad (8)$$

converges to a standard normal distribution for large values of n, where n is the number of simulated scenarios. In this case, $n = 144$, which is fairly large by traditional statistical standards. The p-value of the mean test is given by applying the standard normal CDF to the test statistic, $\hat{Z}_n^{\mathsf{M}}$.

The 3-Mean Test. Performance behavior in scenarios with a large number of nodes or flows is often very different from behavior in small scenarios. In some cases, R_k^{M} is skewed positively for one group, but negatively for another. In such cases, R_k^{M} can pass the median and mean tests even though behavior in each case is unrealistic. To catch such situations, we split the simulations into three groups by the number of flows: the lower, middle, and upper thirds. The 3-mean test simply applies the mean test to each of these groups separately and uses the minimum p-value of the three. This reduces the power of the test, since each group is smaller, but can catch cases where the error has size-dependent biases that cancel out on average.

IV. Results

Our simulation results are summarized in Figure 3. Each subfigure shows a single performance metric. The distribution of errors for each traffic model is visualized with a box-and-whisker plot. These plots allow immediate assessment of realism: a good traffic model should have log-ratio values that are tightly clustered around the center, with a small, evenly balanced box. Additionally, the mean and median markers should be close to the center. Complementing the visual display of summary statistics, Figure 3 also lists three p-values to the right of each box-and-whisker plot. These are, in order, the p-values for the median, mean and 3-mean tests described in Section III-E. Each test catches a different type of unrealistic statistical behavior.

The UTU model, for example, which uses real flow behavior but synthetic flow topology and packet behavior, does a very good job of accurately reproducing realistic average end-to-end delay (Figure 3b): the median and mean are both close to zero, and the p-values are all greater than 0.05. The standard uniform CBR model (UUU), on the other hand, underestimates average end-to-end delay by between a factor of 2 and 11 half of the time, and by more than a factor of 11 a quarter of the time. Its p-values for delay are all less than 0.005.

Figure 3a shows that the STU traffic model passes the mean test but not the median or 3-mean tests. This result indicates

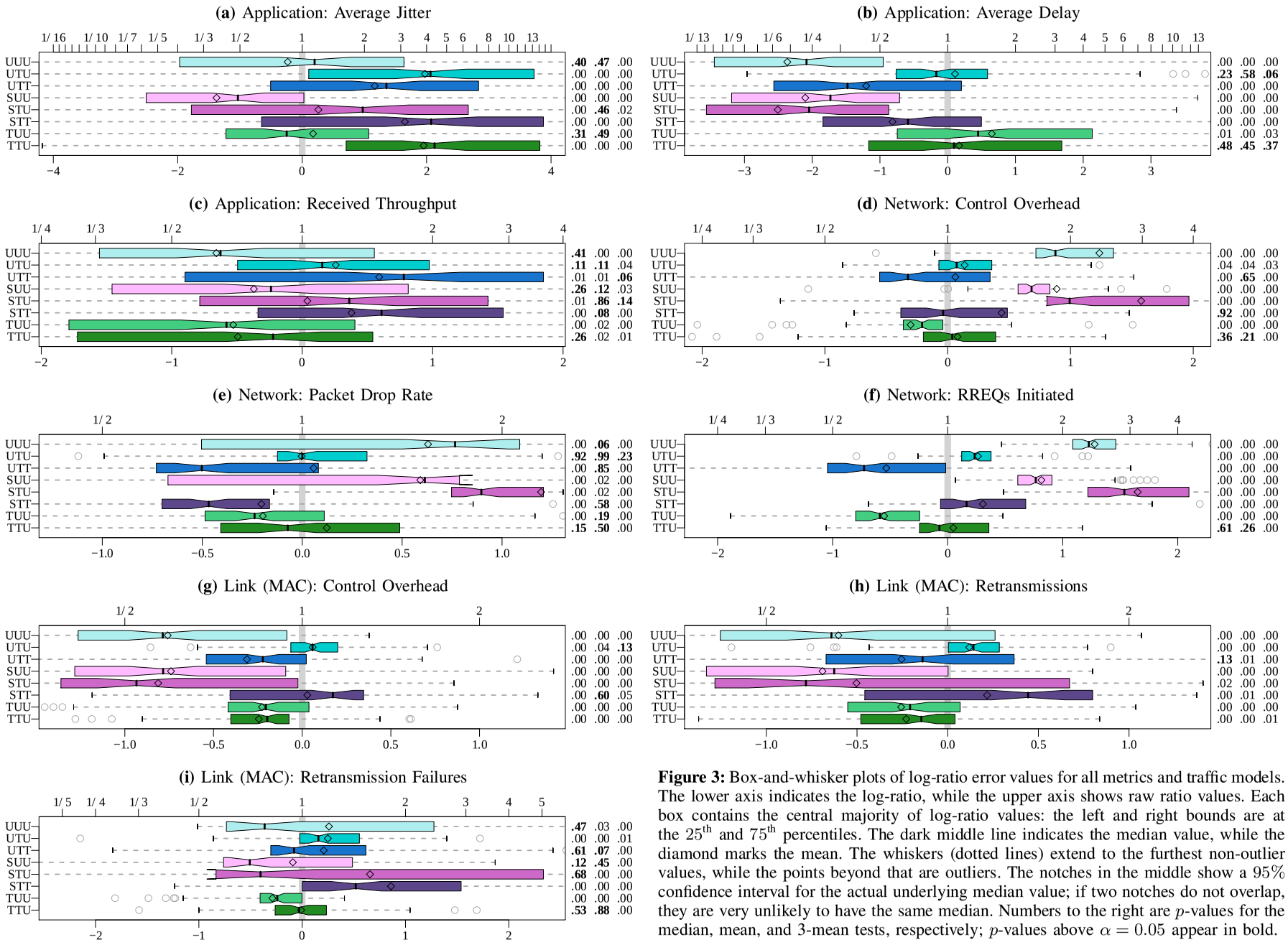

Figure 3: Box-and-whisker plots of log-ratio error values for all metrics and traffic models. The lower axis indicates the log-ratio, while the upper axis shows raw ratio values. Each box contains the central majority of log-ratio values: the left and right bounds are at the 25^{th} and 75^{th} percentiles. The dark middle line indicates the median value, while the diamond marks the mean. The whiskers (dotted lines) extend to the furthest non-outlier values, while the points beyond that are outliers. The notches in the middle show a 95% confidence interval for the actual underlying median value; if two notches do not overlap, they are very unlikely to have the same median. Numbers to the right are p-values for the median, mean, and 3-mean tests, respectively; p-values above $\alpha = 0.05$ appear in bold.

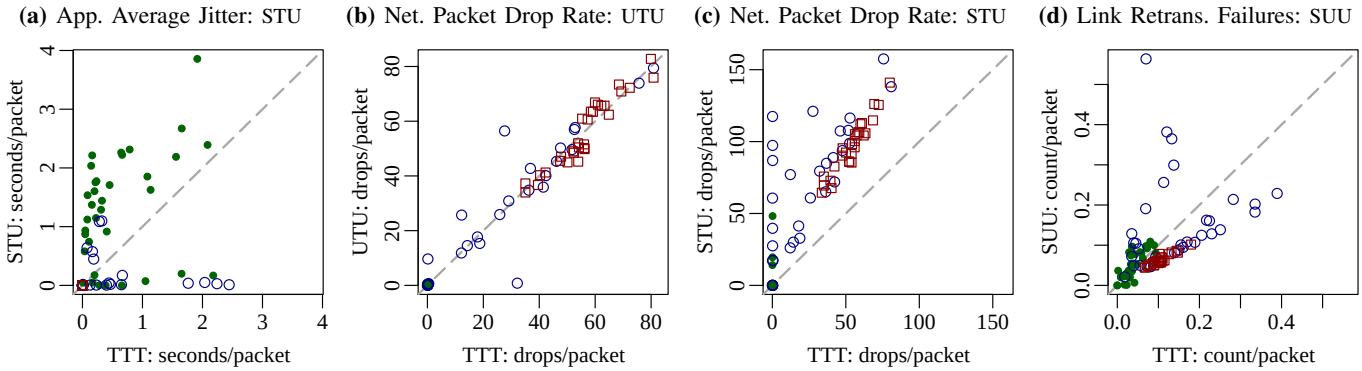

Figure 4: Scatter plots of various metrics for an alternate traffic model (STU, UTU, SUU) versus real trace traffic (TTT). Each data point represents a pair of matched simulations: the x-coordinate is the metric value for trace traffic, the y-coordinate is the metric value for the synthetic model. Scenarios are plotted according to the amount of flows in their trace segment: the bottom third with filled dots; the middle third with circles; the top third with squares.

that this model tends to overestimate jitter, but in some subset of cases it significantly underestimates instead, yielding a mean near zero, but a skewed median. This analysis is verified when we examine a scatter plot of metric values for STU versus TTT traffic models. In Figure 4a we can see that for the majority of simulations, the STU value for jitter exceeds the TTT value. In a significant minority, however, the STU model yields very small jitter values, while the TTT model gives large values. Even though the average error is near zero, this is not a model that reproduces realistic jitter.

The UTU traffic model accurately reproduces packet dropping behavior in routing queues. Figure 4b shows a scatterplot for such a realistic model: the data points are well clustered symmetrically around the diagonal line. We conclude that the primary influence on the packet drop rate is flow behavior. However, if we use a sink topology model, as in Figure 4c, the drop rate becomes inflated. The sink topology introduces an unrealistic routing bottleneck in the network, causing excessive queue overflows for all sizes of scenarios. The uniform topology model does not exhibit this bottleneck, and thereby avoids producing this unrealistic performance artifact. This demonstrates that while the uniform topology model is generally less realistic than the sink model, for certain metrics, their relative quality is the opposite.

Figure 4d illustrates a case where the 3-mean test catches unrealistic behavior that the median and mean tests do not catch. From Figure 3i we can see that the mean and median error values for the SUU model are fairly small. The scatterplot, however, shows that this model significantly underestimates the failure rate for large scenarios with many flows (squares), while overestimating the rate in smaller scenarios.

There are few, if any, positive conclusions that can be drawn from these results. The primary message is that these synthetic traffic models, especially the standard uniform CBR model, consistently misrepresent the most important performance metrics. The traffic model that performs the best overall is the UTU model, which uses real flow behavior with uniform flow topology and uniform packet behavior. This model, however, still fails statistical tests of realism for all but two of the metrics considered. Further development of traffic models is needed before it becomes possible to generate traffic in simulations or experimental deployments such that a single set of experiments can realistically evaluate all aspects of network performance.

V. Discussion

What are the ramifications of these results? The discovery that the most commonly used traffic model for wireless networks drastically misrepresents important performance metrics may shed some light on the lack of trust in results from wireless simulations. It is now well established that network usage behavior—both mobility, and, with this research, traffic patterns—have an impact on network performance that cannot be ignored. Even experimental deployments cannot avoid the need for more realistic traffic workload models. While using a real, physical network successfully sidesteps simulation problems below the application layer, without realistic traffic models, reliable, meaningful performance predictions remain beyond our reach.

A. Relative Performance Comparisons

Performance evaluations are primarily used to compare new protocols with existing ones. It remains possible that while misrepresenting important metrics, synthetic traffic models preserve the relative performance of protocols. In order to test this hypothesis, we have run further simulations, revisiting a classic comparison of ad hoc routing protocols: AODV *vs.* OLSR. We focus on the trace segments with 75 or fewer active nodes, since published performance comparisons of ad hoc routing protocols have typically not used more nodes than this. For each scenario, we have run simulations with 10 different seed values using the AODV and OLSR routing protocols and the UUU and TTT traffic models, allowing us to compare relative performance results when switching traffic models.

The key result of our comparison is that, for certain performance metrics, **the relative performance of AODV and OLSR is switched, simply by changing traffic models.**

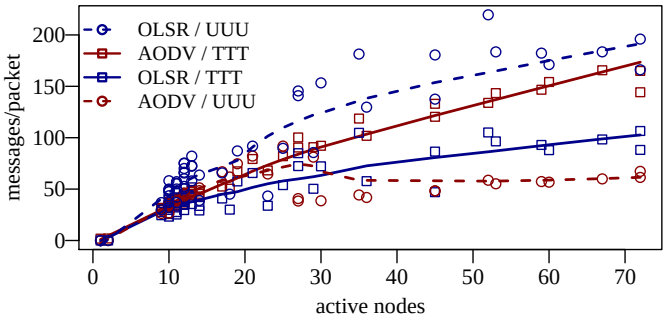

Figure 5: Comparison of MAC control overhead for AODV and OLSR using both the UUU and TTT traffic models. The fit lines show locally weighted, smoothed performance trends.

Figure 5 shows some of the results from this experiment. The graph plots MAC control overhead against the number of nodes in each simulation scenario. Each data point shows the average overhead for the simulations of that scenario using ten seed values. With the standard UUU traffic model, OLSR consistently has much higher overhead, across the board, than AODV. When real traffic is used, however, OLSR performs significantly better than AODV. A similar switch occurs for the number of MAC retransmissions. This experiment demonstrates that the use of unrealistic traffic models can change the relative performance of different protocols, as well as skew absolute performance measurements.

B. Generality of Results

The most significant limitation on the generality of this analysis is that it is based entirely on a single data set from IETF60—albeit a large and varied one. It is possible that traffic in this trace happens to produce network performance that is unusually dissimilar to standard traffic models. This data set, however, represents a highly heterogeneous collection of network usage behaviors, from slow and steady off-peak usage, to extremely heavy peak usage: over 800 users, 33 thousand flows, and 1 million packets in a single 10-minute trace segment. Despite the broad variety of behaviors, the results are consistent: in all types of usage scenarios, simplistic traffic models, like uniform CBR, systematically skew important performance measurements at all levels of the network. While the precise results for other data sets might differ, it is very unlikely that CBR traffic models will happen to accurately reproduce realistic performance in other experiments. This paper provides strong evidence that better traffic models are needed for performance evaluations.

C. Towards Realistic Models of Wireless Workload

What would better traffic models look like? How can we create them? One possible approach is to use actual traffic traces as we have done. This approach is unsatisfactory, however, because it provides the experimenter with almost no control over experiments. Synthetic models have parameters, which can be tweaked as necessary—adjusting, for example, the number of active nodes in a simulation, without affecting other parameters. Traffic traces, on the other hand, must be used without significant alteration if they are to actually provide the desired realism. The "messiness" of the performance comparison from trace data in Figure 5 illustrates why using traces directly is not ideal: each data point differs not only in the number of nodes shown on the x-axis, but also in other dimensions, such as the number of flows and packets, and the average flow duration. The result is a highly noisy comparison, affected by many unseen parameters. Only by applying a local smoothing algorithm are trends somewhat elucidated.

Instead of using trace data directly, it should be possible to configure a synthetic traffic model based on observations from a real data set, and then run side-by-side simulations using the synthetic model and the real data, producing statistically equivalent performance results. This is precisely what our definition of sufficient realism entails. The work in this paper provides the tools to measure how close to this ideal a model is and in what areas it needs improvement. Without this feedback, any improvements in realism are purely guesswork. Our breakdown of traffic behavior into three orthogonal levels also allows the problem to be approached in smaller pieces, rather than being solved all at once.

The next step towards better traffic models, is to investigate which aspects of real traces may be altered without detrimentally affecting the resulting performance metrics. For example, to test whether a complex time-series model of packet behavior is necessary, we randomize the order of the packet sizes and/or inter-packet intervals and compare performance using these randomized traces against performance using the original traces. If the performance is unchanged, we can conclude that no complex time-series model of packet behavior is necessary: sampling the packet sizes and inter-packet intervals from empirical distributions is sufficiently realistic. If, on the other hand, the performance characteristics are altered by shuffling packets, then some time-series model of packet behavior is needed. By partially randomizing the packet order in specific ways, the exact limits of realism necessary can be found. A similar approach will allow the development of realistic models for the other levels of network usage behavior.

VI. Conclusions

This research rigorously quantifies the impact of a variety of synthetic traffic models on performance metrics that wireless researchers use to evaluate new technologies and protocols. The first step in this assessment process was to formally define what it means for a network usage model to be sufficiently realistic. In essence, a model is considered realistic if it produces performance results that are statistically equivalent to those produced by real usage. A well-defined, objective measure of realism for traffic models has not previously existed. Evaluations of realism have formerly relied on essentially arbitrary statistical measures of similarity to real traffic, which may or may not affect the performance metrics that researchers care about. The definition of sufficient realism leads us to our general experimental approach: we use differential analysis comparing performance metrics derived from real traffic with

those derived from synthetic traffic models. The theoretical contributions of this analysis are:

1) An in-depth analysis of the desirable mathematical properties of a measure of error for performance metrics.
2) Proof that the unique measure of error that satisfies these properties is the log-ratio of metric values.
3) Three rigorous tests of statistical equivalence between synthetic and real performance results.

These analytical tools allow the evaluation of realism over a collection of drastically different usage scenarios. Evaluation over a heterogeneous collection of scenarios is essential to establishing the credibility of usage models. Moreover, these theoretical results are equally applicable to other types of usage models—for example, mobility.

On the practical side, this paper gives crucial insight into why most researchers do not trust simulation results: with the traffic models commonly used, the results are unlikely to reflect real performance. Simultaneously, it indicates that the same problem will also affect experiments using physical deployment of test networks, unless those networks are subjected to real workloads. The only way to address this fundamental lack of realism is to develop usage models that reproduce important performance metrics more accurately. Our theoretical results provide the tools necessary to do this. The development of better traffic models should begin with real traces, and proceed by incremental changes, checked by differential analysis. First, alter a small aspect of the trace, simulate, then compare. If the realism of the results is unaffected, the traffic feature altered was inessential. Otherwise, it is a feature of behavior that must be captured in a realistic traffic model. This approach will allow the precise mapping of which aspects of traffic patterns have an impact on performance, and which ones can be safely abstracted away.

APPENDIX

Theorem. The unique differentiable function satisfying Eqs. 2 & 4 is $E(1,e)\ln(y/x)$. *Proof:* Let $f(z) = E(1,z)$. Eq. 2 gives $E(x,y) = E(1,y/x) = f(y/x)$. Eq. 4 gives: $f(z) = f(z/w) + f(w)$. Differentiation by z yields $f'(z) = w^{-1}f'(z/w)$. In particular, if we choose $w = z$, we get $f'(z) = z^{-1}f'(1)$. Integration by z gives: $f(z) = f'(1)\int z^{-1}dz + c = f'(1)\ln(z) + c$. By Eq. 4, $f(1) = 0$, so $c = 0$. Thus $f(e) = f'(1)\ln(e) = f'(1)$. We conclude that $f(z) = f(e)\ln(z)$, so $E(x,y) = f(y/x) = f(e)\ln(y/x) = E(1,e)\ln(y/x)$, as desired. ∎

Theorem. [*Lyapunov's Central Limit Theorem*] Let $\{R_k\}_{k=1}^{\infty}$ be a series of independent variables with $\langle R_k\rangle = 0$. Let $s_n^2 = \sum_{k=1}^{n}\left\langle\left|R_k^2\right|\right\rangle$ and $r_n^3 = \sum_{k=1}^{n}\left\langle\left|R_k^3\right|\right\rangle$. For each n, let Z_n be the normalized mean of $\{R_k\}_{k=1}^{n}$: $Z_n = \sum_{k=1}^{n} R_k/s_n$. If $\lim_{n\to\infty} r_n/s_n = 0$, then $\lim_{n\to\infty} Z_n \sim \mathcal{N}(0,1)$ (the standard normal distribution). (See [26] page 229.)

To apply the LCLT to the series $R_k^{\mathsf{M}} = \log(X_k^{\mathsf{M}}/X_k^{\mathrm{TTT}})$, we must show that under the null hypothesis, the assumptions of the theorem are satisfied by this series. First, the null hypothesis, implies that $\langle R_k\rangle = 0$. Variables from separate simulations are independent since they cannot influence each other's values. Formally, $\Pr(X_k|X_j) = \Pr(X_k)$. Therefore the log-ratios are also independent for different k. The last requirement is that $\lim_{n\to\infty} r_n/s_n = 0$. To verify this, we use the estimators $\hat{s}_n^2 = \sum\left|R_k^2\right|$, and $\hat{r}_n^3 = \sum\left|R_k^3\right|$. When $\hat{r}_n/\hat{s}_n$ are plotted on a log-log scatter plot, with n increasing up to the number of simulations, they asymptotically approach a downwardly sloped line as n grows. Thus $\lim_{n\to\infty}\ln(\hat{r}_n/\hat{s}_n)/\ln(n) = c < 0$. This implies that $\lim_{n\to\infty}\ln(\hat{r}_n/\hat{s}_n) = -\infty$, and therefore $\lim_{n\to\infty} r_n/s_n = \lim_{n\to\infty}\hat{r}_n/\hat{s}_n = 0$. This test for the convergence of r_n/s_n is applied to each model and metric pair.

REFERENCES

[1] 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, June 1995.

[2] E. Royer, P. Melliar-Smith, and L. Moser, "An analysis of the optimum node density for ad hoc mobile networks," in *IEEE International Conference on Communications*, Helsinki, Finland, June 2001, pp. 857–861.

[3] J. Yoon, M. Liu, and B. Noble, "Random waypoint considered harmful," in *IEEE Infocom*, San Francisco, CA, USA, April 2003, pp. 1312–1321.

[4] D. Clark, "The design philosophy of the DARPA Internet Protocols," in *ACM Sigcomm*, 1988.

[5] S. Avallone, D. Emma, A. Pescap, and G. Ventre, "A distributed multiplatform architecture for traffic generation," in *International Symposium on Performance Evaluation of Computer and Telecommunication Systems*, San Jose, CA, USA, July 2004.

[6] ——, "High performance internet traffic generators," *The Journal of Supercomputing*, vol. 35, no. 1, pp. 5–26, January 2006.

[7] F. Hernández-Campos, "Generation and validation of empirically-derived TCP application workloads," Ph.D. dissertation, Univ. of North Carolina, Chapel Hill, 2006.

[8] D. Tang and M. Baker, "Analysis of a metropolitan-area wireless network," in *ACM MobiCom*, Seattle WA, August 1999.

[9] A. Balachandran, G. Voelker, P. Bahl, and V. Rangan, "Characterizing user behavior and network performance in a public wireless LAN," in *ACM Sigmetrics*, Marina Del Ray CA, June 2002.

[10] M. Balazinska and P. Castro, "Characterizing mobility and network usage in a corporate wireless local-area network," in *ACM MobiSys*, San Francisco, CA, USA, May 2003, pp. 303–316.

[11] D. Kotz and K. Essien, "Analysis of a campus-wide wireless network," in *ACM MobiCom*, September 2002.

[12] T. Henderson, D. Kotz, and I. Abyzov, "The changing usage of a mature campus-wide wireless network," in *ACM MobiCom*, September 2004.

[13] D. Schwab and R. Bunt, "Characterizing the use of a campus wireless network," in *IEEE Infocom*, March 2004.

[14] F. Chinchilla, M. Lindsey, and M. Papadopouli, "Analysis of wireless information locality and association patterns in a campus," in *IEEE Infocom*, March 2004.

[15] A. Jardosh, K. Ramachandran, K. Almeroth, and E. Belding-Royer, "Understanding congestion in IEEE 802.11b wireless networks," in *ACM/USENIX Internet Measurement Conference*, Berkeley, CA, USA, October 2005, pp. 279–292.

[16] ——, "Understanding link-layer behavior in highly congested IEEE 802.11b wireless networks," in *ACM Sigcomm EWIND*, Philadelphia, PA, USA, August 2005.

[17] D. Tang and M. Baker, "Analysis of a local-area wireless network," in *ACM MobiCom*, Boston MA, 2000.

[18] C. Tuduce and T. Gross, "A mobility model based on WLAN traces and its validation," in *IEEE Infocom*, Zürich, Switzerland, March 2005, pp. 664–674.

[19] T. Camp, J. Boleng, and V. Davies, "A survey of mobility models for ad hoc network research," *Wireless Communications and Mobile Computing*, vol. 2, pp. 483–502, September 2002.

[20] J. Yoon, M. Liu, and B. Noble, "Sound mobility models," in *ACM MobiCom*, San Diego, CA, USA, September 2003, pp. 205–216.

[21] A. Jardosh, E. Belding-Royer, K. Almeroth, and S. Suri, "Towards realistic mobility models for mobile ad hoc networks," in *ACM MobiCom*, San Diego, CA, USA, September 2003, pp. 217–229.

[22] Q. Zheng, X. Hong, and S. Ray, "Recent advances in mobility modeling for mobile ad hoc network research," in *ACM Southeast Conference*, Huntsville, AL, USA, April 2004, pp. 70–75.

[23] A. Jardosh, E. Belding-Royer, K. Almeroth, and S. Suri, "Real-world environment models for mobile network evaluation," *IEEE Journal on Selected Areas in Communications*, March 2005.

[24] V. Paxson, "End-to-end routing behavior in the internet," in *ACM Sigcomm*, Palo Alto, CA, USA, 1996, pp. 25–38.

[25] J. Sommers and P. Barford, "Self-configuring network traffic generation," in *ACM/USENIX Internet Measurement Conference*, Taormina, Sicily, Italy, October 2004, pp. 68–81.

[26] W. Feller, *An Introduction to Probability Theory and Its Applications*, 3rd ed. New York, NY, USA: Wiley, 1968, vol. 1.