ACM SIGOPS Operating Systems Review (OSR),
Special Issue on Repeatability and Sharing of Experimental Artifacts, January, 2015

Conducting Repeatable Experiments and Fair Comparisons using 802.11n MIMO Networks

Ali Abedi, Andrew Heard, Tim Brecht
School of Computer Science, University of Waterloo
{a2abedi, asheard, brecht}@cs.uwaterloo.ca

ABSTRACT

A commonly used technique for evaluating and comparing the performance of systems using 802.11 (WiFi) networks is to conduct experiments. This approach is appealing and important because it inherently captures critical properties of wireless signal transmission that are difficult to analytically model and simulate. Unfortunately, obtaining consistent and statistically meaningful empirical results using 802.11 networks, even in well-controlled environments, can be quite challenging and time consuming because channel conditions can vary over time.

In this paper, we use 2.4 and 5 GHz 802.11n MIMO networks to study different methodologies that could be used to evaluate and compare the performance of different alternatives used in 802.11 systems (e.g., different systems, configurations or algorithms). We first illustrate that some of the more commonly used methods in existing research are flawed and explain why. We then describe a methodology called *multiple interleaved trials* that, to our knowledge, has not been used for, or studied on, 802.11 networks. We evaluate this methodology and find that it can be used to repeat experiments and to compare the performance of different alternatives. Finally, we discuss other possible applications of this approach for comparative performance evaluations.

1. INTRODUCTION

An important aspect of any type of research is being able to fairly evaluate, compare and draw conclusions regarding the relative merits of multiple competing systems or techniques. In this work, we refer to *alternatives* as solutions or systems that depend on 802.11 networks. Some examples include, comparing different versions of TCP, video streaming techniques, or 802.11 rate adaptation algorithms.

In 802.11 networks, performing fair comparisons of different alternatives can be extremely challenging because channel conditions can vary significantly over time. As a result, analytic models, simulation and channel emulation are enticing because they can be used to ensure that each alternative being compared is exposed to precisely the same channel conditions. Unfortunately, wireless channel models require mathematically modeling the physical properties of signals traveling through space, bouncing off, being absorbed by and passing through walls, ceilings and other objects. Signal propagation is also affected by the materials found in these objects thus requiring different models for different materials. Homes and offices may require different models, because of differences in building materials and furnishings, and it is highly likely that no two homes or offices are identical. More importantly, in order to accurately reflect environments in which WiFi devices are likely to be deployed, such models also need to include models of interference from other WiFi and non-WiFi devices. The use of handheld mobile devices makes obtaining accurate results substantially more difficult, if not intractable.

Therefore, empirical evaluations are highly desirable because they can provide a level of realism and accuracy that is difficult to achieve otherwise. However, empirical evaluations are not a panacea. Because channel conditions can vary over time, conducting repeatable experiments can be very challenging. One of the goals of this paper is to understand the possibilities and limitations of conducting empirical measurements when using increasingly popular 802.11n MIMO networks. We study the efficacy of different methodologies for comparing multiple competing alternatives. We examine both tightly controlled, presumably repeatable channel conditions, and uncontrolled environments with highly variable channel conditions. Our intention is not to discredit past work that use these methodologies, as they may be effective given stable channel conditions. Instead, our goal is to understand these methodologies and their pitfalls.

The contributions of this paper are:

- We examine different existing methodologies for conducting experiments to compare the performance of systems that use 802.11n MIMO networks.
- We show that some commonly used techniques for comparing the performance of different alternatives are flawed, even in highly controlled environments that are free from interference from other WiFi and non-WiFi devices. This could result in misleading conclusions.
- We show that the multiple interleaved trials methodology provides repeatable results and can be used to distinguish differences in performance, even with highly variable channel conditions.

2. BACKGROUND AND RELATED WORK

It is generally understood that interference from WiFi and non-WiFi devices, and movement of individuals in the vicinity of the experiment, can affect channel conditions in 802.11 networks, and thus, repeatability. Nevertheless, experiments are widely used because of the realism they provide.

2.1 Repeatability in Experimental Evaluation

Several studies [1, 2, 3, 4, 5, 6] have considered the issue of repeatability when performing empirical performance evaluations of 802.11 networks. Ganu et al. [1] focus on the re-

Copyright is held by the author(s)

peatability and reproducibility of 802.11 experiments. They report on variations across five runs in a semi-controlled environment using the 802.11b standard. However, they do not provide any statistical analysis of these experiments in order to determine whether or not results could be repeated. In this paper, we present results of several 802.11n MIMO experiments along with a statistical analysis of these results and report on the ability to repeat experiments, as well as fairly compare multiple alternatives.

Burchfield et al. [2] study different factors that affect the repeatability of 802.11 experiments. They show that the outcome of an experiment can change significantly by a simple change in the environment such as interference or swapping the location of the sender and receiver. Therefore, they suggested using wireless channel emulation to solve the repeatability problem. Wireless channel emulators can repeat identical channel conditions across trials [2, 3]. However, they suffer from a lack of realism, as they rely on simplified models of wireless channels [7]. In Section 5.3, we present a methodology that can achieve repeatability even in the presence of external interference.

In an effort to make experiments involving mobility repeatable, robots have been used to carry a wireless node [4, 5, 6]. These robots follow the same path for each experiment to try to minimize the variation caused by following slightly different paths in every run. Rensfelt et al. [6] evaluate the potential effect of using robots, instead of a person carrying a mobile device, on the repeatability of the results. They show that robots can lower the variability of received signal strength indication (RSSI) measurements in 802.15.4 sensor networks. In Section 5.3.3, we study the effect of using a toy train to repeat mobile experiments (in terms of RSSI and throughput) in more widely used 802.11n networks.

2.2 Methodologies for Experiments

To address problems with repeating experiments, researchers suggest avoiding situations where the channel conditions are likely to change rapidly, by conducting experiments "in the middle of the night", "when no one else is around", or in the 5 GHz spectrum to avoid interference [8, 9, 10]. There are three problems with such evaluations. The first problem is that they may not be representative of the conditions under which devices are most often used and, more importantly, the conditions that may prove to be most difficult for different alternatives to handle (i.e., when channel conditions change over time). The second problem is that there is an implicit assumption that running an experiment during the night avoids non-WiFi interference [8, 11]. However, recent studies report significant levels of non-WiFi activity during the night [12, 13, 2]. The third problem with this approach is the assumption of repeatability, which may not be true if not checked. The consequence of these problems is that if external interference is not monitored during an experiment, and repeatability is not established, it may affect the conclusions that can be drawn from these experiments. In essence, differences in performance may be mistakenly attributed to differences between alternatives, when in fact they may be due to differences in channel conditions encountered during the experiments. In this paper, we conduct several experiments to determine whether or not experiments can be repeated in environments with and without interference from other WiFi and non-WiFi devices. We also show (in Section 5.1), as pointed out by others [2], that nighttime channel conditions are not representative of daytime channel conditions.

Another common approach used when conducting empirical performance evaluations is to run experiments multiple times and report the average performance and some notion of the variability of the obtained results. Unfortunately, some papers [11, 14, 9] do not report the standard deviation or confidence intervals, despite running multiple experiments. In Section 5.2, we show that ignoring confidence intervals may lead to incorrect conclusions.

Other papers [15, 16, 17, 18] do report the standard deviation or confidence intervals. Interestingly, in our evaluation, we observe that confidence intervals, while useful, can also be misleading. In Section 5.2, we obtain statistically significant differences (i.e., non-overlapping confidence intervals) for two sets of identical experiments. We show that if an experiment using alternative A is repeated for multiple iterations, followed by multiple iterations using alternative B, changes in channel conditions can result in inaccuracies that are not reflected in the confidence intervals. In Section 5.3 we propose the use of and evaluate a variation of this technique that addresses this issue.

A trace-driven evaluation methodology like T-RATE [19] could be used to collect traces using real experiments and then use those traces to evaluate rate adaptation algorithms using identical channel conditions. Unfortunately, T-RATE is only designed to be used to evaluate rate adaptation algorithms and has only been developed for 802.11g networks.

3. EXPERIMENTAL SETUP

We have created a small test bed for conducting experiments using cubicle-based office space in a university campus building. Our test bed consists of desktop systems containing TP-Link TL-WDN4800 dual-band wireless N PCIe adapter. These cards use the Atheros AR9380 chipset, contain three antennas and support three streams (i.e., a 3x3:3 MIMO configuration). In stationary experiments, we use two desktops, 4 meters apart, with no line of sight. For mobile experiments, we use a laptop configured to use a TP-Link TL-WDN4200 dual-band wireless N USB adapter. This adapter contains an Ralink RT3573 chipset and also supports a 3x3:3 configuration. To maximize repeatability, for some experiments, we use a small electric train on which a laptop is placed. The train is operated at walking speeds.

We conduct experiments using both 2.4 and 5 GHz 802.11n networks. Unless otherwise specified, we choose channels that are not used by any other access points or devices. In our 5 GHz experiments, we enable the optional 40 MHz channel width to increase the available bandwidth. We ensure that there is a minimum of 40 MHz of separation between the channel used by our network and those used by other networks. This helps to avoid channel leakage from adjacent channels, which can cause performance degradation [20]. In our 2.4 GHz experiments we do not use 40 MHz channel widths in order to limit external interference in this spectrum. The optional short guard interval feature is enabled in all of our experiments. To ensure that unknown, or unwanted, interference from other devices is avoided, we continuously monitor all of our experiments using an AirMagnet Spectrum XT [21]. This analyzer is able to detect and classify both WiFi and non-WiFi interference.

On the desktop machines, as well as the laptop, we use Ubuntu 12.04 with Linux kernel version 3.13.0. The ath9k

(Atheros) and rt2800usb (Ralink) device drivers are provided by the backports-3.14-1 package. We use iperf [22] to generate UDP traffic between the sender and receiver at as high a rate as possible, to fully utilize the network infrastructure. The sending device in all of our experiments uses the *Minstrel-HT* 802.11 rate adaptation algorithm, which is the default rate adaptation algorithm used by the Linux Ath9K driver. Although we could have used `tcpdump` to record much of the information reported in this study, we use detailed information obtained directly from the ath9k driver. Previous modifications to the ath9k driver are used, which record highly detailed information for every packet [19]. MAC layer frame aggregation is enabled to increase the efficiency of the 802.11n MAC layer.

4. EXPERIMENTAL METHODOLOGIES

One of our goals in this paper is to better understand the efficacy of existing methodologies that have been used to empirically measure and compare the performance of 802.11 networks. We use the term *trial* to refer to one measurement, typically obtained by running a benchmark or micro-benchmark for some period of time (the length of the trial). An *experiment* can be comprised of multiple trials executing the same benchmark, where the results of the experiment are reported over the multiple trials (e.g., the average of the measurements obtained over the trials).

Because it is well known that channel conditions can vary over time, we are interested in understanding repeatability and degree of variability across multiple experiments. Clearly if the goal of empirical research is to compare multiple alternatives in order to draw conclusions about which is the best choice, it is important that the conditions under which the different alternatives are tested do not change significantly, in order for the comparison to be fair.

The approaches we examine in this work are:

- **Single Trial Experiments**: In this approach, an experiment consists of only a single trial. This is a surprisingly common approach used in existing work [10, 8, 23]. In most cases, multiple wireless environmental setups might be considered (e.g., mobile, stationary, with and without hidden terminals), however, comparisons are made and conclusions are drawn using only a single trial.
- **Multiple Consecutive Trials**: This approach recognizes that possible changes in channel conditions can lead to variability. As a result, multiple trials are used. All trials for the first alternative are run, followed by the second alternative and each of the remaining alternatives.
- **Multiple Interleaved Trials**: This approach requires interleaving each of the alternatives being studied. One trial is conducted using the first alternative, followed as soon as possible by one trial with the second, and so on until each alternative has been run once. When one trial has been conducted using each alternative we say that one *round* has been completed. Rounds are repeated until the appropriate number of trials has been conducted to complete an experiment. If channel conditions are affected at regular intervals, and the intervening period coincides with the length of each trial, it is possible that some alternatives are affected more than others. Therefore, we recommend a random reordering of alternatives for each round. In essence, a randomized block design [24] is constructed where the blocks are intervals of time (rounds) and within each block all alternatives are tested, with a new random ordering of alternatives being generated for each block. In this paper, to make this methodology easier to describe and understand and because we did not find it necessary to randomize trials, we use the same ordering for each round.

Next we evaluate the efficacy of these methodologies for fair and repeatable comparisons of multiple alternatives using throughput and RSSI (RSSI is also a popular metric [1, 2, 6]). When multiple trials are used, we include 95% confidence intervals computed using the Student's t-distribution.

5. EVALUATING REPEATABILITY

The results in Sections 5.1 to 5.3.2 are obtained by running a single experiment where one stationary sending device is communicating at as high a rate as possible to a single stationary receiver for 24 hours. Nothing changes over that time except for possibly the channel conditions. We then divide the results obtained over that entire time period into chunks of time and compare results over those time periods as though they are "different" alternatives. However, because in reality the same alternative is used in all experiments, the results obtained across all experiments should ideally be identical, or close enough so as to be statistically similar. That is, the results should be repeatable. Results that are not similar indicate that there is a flaw with the methodology.

5.1 Single Trial Experiments

Often when experimenters wish to compare two or more alternatives, a single trial of each alternative is used. For example, running alternative A for 60 seconds, followed by alternative B for 60 seconds, and using these results to compare performance differences between A and B.

We collect throughput and RSSI measurements over 24-hours and then divide it into 60-second experiments and compare all consecutive sets of two experiments. Figure 1 shows measured throughput (the top line) and RSSI (the bottom line) for each of these experiments. The x-axis represents time from midnight of one day until midnight 24 hours later. In this and many subsequent graphs, the left y-axis shows throughput and the right y-axis shows RSSI. In practice, an experimenter using this methodology has only one data point for each alternative, rather than the hundreds presented here, hence our focus on only consecutive data points. We include 24 hours worth of data to see how this methodology works at different times of the day (i.e., with different degrees of variability).

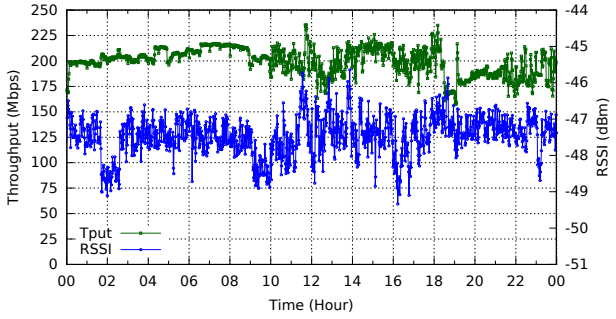

Figure 1: Consecutive Experiment Variability

For this methodology to be sound, each set of consecutive data points needs to be within some tolerance level based on the desired granularity at which differences in alternatives

are to be compared. Figure 2 highlights (by zooming in on) two subsets of the 24 hours. The purpose is to provide a more detailed view of the results at the level of individual experiments and to examine the differences between these two periods of time.

The period of time from 04:00 - 06:15 uses the top x-axis, while the time from 17:00 - 19:15 is plotted using the bottom x-axis. These times were chosen because they include periods of relative stability (04:00-06:15) and variability (17:00-19:15). These subsets could be thought of as roughly corresponding to the middle of the night and some working hours (for the graduate students using this lab and these offices).

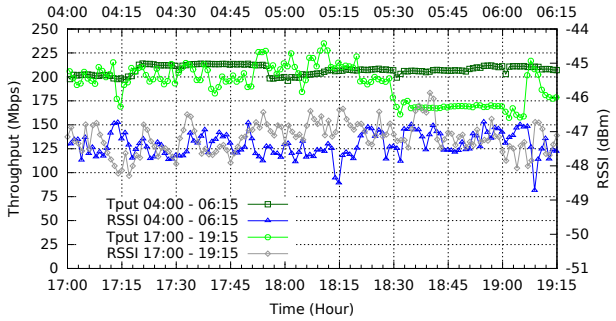

Figure 2: Consecutive Experiment Variability

In Figure 2, throughput is roughly 160 Mbps at 19:06 and then jumps to 200 Mbps for the trial's immediate successor, 19:07. This is an increase of 25% in two consecutive trials whose start times differ by only 60 seconds. From this level of variation, it is evident that a single trial experiment could be misleading if one alternative were to be measured at time 19:06 and another at 19:07. In this case, a difference of 25% might be incorrectly attributed to the change in alternatives, rather than the change in channel conditions. This problem may be exacerbated by adding more alternatives, as channel conditions may change for each of the alternatives. For example, if several alternatives were examined between 18:00 and 8:15, each would be subject to different channel conditions (as can be seen by the changes in RSSI) and obtain different throughputs. While the variation observed between 04:00 and 06:15 in Figure 2 is lower than that observed during 17:00-19:15, there does exist two consecutive trials where a difference of around 10% is observed (at 04:55). This suggests that, for large enough differences between alternatives, the single trial technique may be fine during the middle of the night (or a period of low variability) but rarely during the day, making it an unreliable technique overall. The fluctuations of the throughput during this experiment are mainly due to the movement of people working in nearby cubicles. Additionally, the environments experienced during both time periods were interference-free and the period from 04:00 to 06:15 was entirely stationary (no movement of devices or people), leading to a level of stability that may not be representative of the environments in which devices are most often used.

RSSI measurements in Figure 2 are similarly variable. During the day (17:00-19:15), RSSI in consecutive experiments differs by up to 0.92 dBm, while at night RSSI differs by up to 1.26 dBm. Even though RSSI is not a complete predictor of performance, it is an indication of the variability of channel conditions. Additionally, fluctuation is expected to increase in environments with interference and higher levels of mobility. Judd et al. [25] observed high levels of variability in RSSI in a stationary environment with lots of mobility, such as a lobby or a hallway. However, we observed that fluctuations occur even in fairly stable environments. Interestingly, and also noted by others [26], stronger signals do not necessarily correlate with higher throughput.

GUIDELINE: *In order to draw conclusions using single trial experiments, channel conditions must be extremely stable. As a result, conclusions from these experiments cannot be extended to other environments possessing greater variability (which may be considered more representative of environments in which devices are actually used). More importantly, even if channel conditions are believed to be stable, we do not know of any way to guarantee, or verify, that channel conditions did not change from one trial to the next. Thus, our recommendation is to avoid using this methodology.*

5.2 Multiple Consecutive Trials

As we saw in the previous section, a considerable amount of variability can exist even in interference-free, stationary environments. A possible, and commonly used, solution is to run experiments multiple times and report performance metrics using a mean and some indication of variability across experiments, instead of using a single measure. In practice, when more than a single trial is collected, experimenters often conduct all of the trials for a given alternative, before proceeding to the next. This makes sense, as there is generally some setup time involved in switching between alternatives. For example, changing WiFi cards or a software configuration (e.g., unloading and loading a device driver in order to change the rate adaptation algorithm).

To evaluate this approach, we again divide the same data collected during the 24-hour period into 60-second trials but now combine consecutive trials together to constitute an experiment. While experiments may not be performed in such a continuous fashion, we use this technique as it allows us to fairly and easily utilize the same data when comparing competing methodologies. We compute the throughput and RSSI for different numbers of trials. Recall that during the 24-hour period, the only thing changing is channel conditions and, because we ensure that there is no WiFi or non-WiFi interference, channel conditions *should be* relatively stable (especially compared with more challenging scenarios that include interference). If the experiments are repeatable, we should see no statistical difference between the two consecutive alternatives A and B (i.e., if the 95% confidence intervals overlap, we consider the result to be repeatable). However, if a statistically significant difference is observed (the confidence intervals do not overlap), the measurement technique can potentially lead to erroneous conclusions. A conclusion may be drawn that performance differences are due to different alternatives, rather than changes in channel conditions. Even though, in this case, A and B are identical.

Figure 3 shows the results of applying the multiple consecutive trials technique to the 24-hour data that we used in the previous section. We focus on 4 hours of data (16:00 - 20:00) where higher channel variability was observed (again because we are interested in determining if experiments can be used to evaluate alternatives when channel conditions are variable). The three plots in this figure show the average throughput and RSSI with 95% confidence intervals for both alternatives when using 5, 10, and 15 trials (top, middle and bottom, respectively). We consider 5 trials to be

small but include it because a number of previous studies [9, 17] examine 5 or even fewer trials. As in the previous section, we only compare two consecutive experiments. However, because several trials are used, we test for overlapping confidence intervals to assess repeatability.

In Figure 3, when examining throughput, we see that for 5 trials there are two consecutive non-overlapping confidence intervals for experiments at 16:50, 17:50 and 19:05. Increasing the number of trials to 10, we see non-overlapping confidence intervals at 16:00, 16:50, 17:50, 18:20 and 18:35. This may be due to the fact that more trials increases the amount of time between the start of consecutive experiments and, therefore, increases the likelihood that the environment changes between these two experiments. This can lead to problems if experiments happen to be run during these periods of variability. While it may be tempting to assume that an astute experimenter would recognize these periods of variability, in reality, it is unlikely that they have 24 hours of experiments comparing the two or more alternatives; instead, they are likely to have only one data point for each alternative. Unfortunately, we are not aware of any way to determine, before or after the experiments have been run, that the observed differences are due to different alternatives, rather than changes in channel conditions. Recall that in our case we know that all alternatives being examined are identical. These issues become more pronounced when comparing more than two alternatives, as the start times between compared alternatives are further apart. For example, consider 5 alternatives with 15 trials starting at time 18:12. Several of these results have non-overlapping confidence intervals. In the next section, we explain how a simple modification to this technique can overcome these issues.

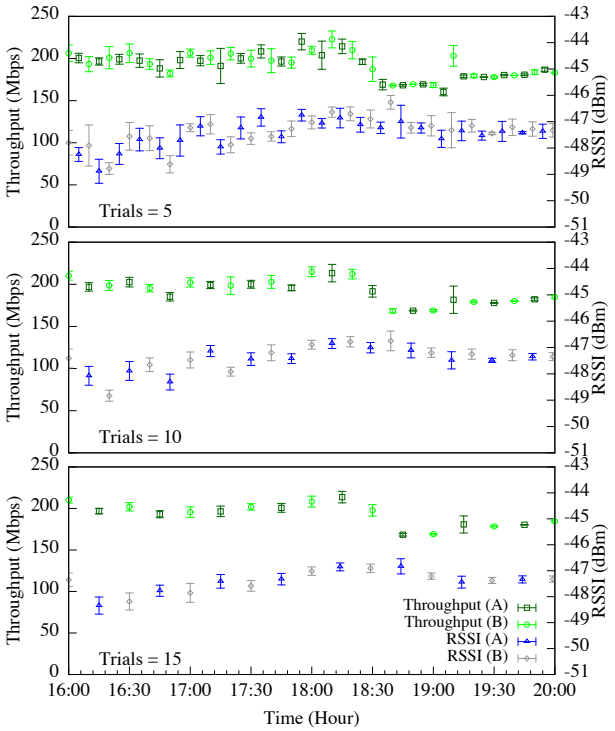

Figure 3: Multiple consecutive trials (2 alternatives)

GUIDELINE: *In a controlled and interference free environment we were unable to reliably repeat experiments using the technique of multiple consecutive trials. Moreover, the confidence intervals obtained using this technique may provide a false sense of rigor and validity when comparing different alternatives. Because channel conditions may change over time and because we do not know of any way to guarantee or verify that they have not changed, our recommendation is to avoid using this methodology.*

5.3 Multiple Interleaved Trials

In the hope of addressing the shortcomings of the multiple consecutive trials technique, we now evaluate the technique we call multiple interleaved trials. To our knowledge, this method has not been explicitly used in previous studies evaluating performance when using 802.11 networks. We are the first to directly study this methodology in the context of WiFi performance analysis.

This perhaps obvious, but important, approach provides the advantage that trials are more closely situated in time, provided each trial does not run for too long (the length of trials is discussed in Section 5.3.5). The intuition is that by interleaving alternatives, they will be exposed to channel conditions that are more similar than when using multiple consecutive trials. This is particularly important as the number of alternatives being compared grows. For example, if two alternatives are compared using multiple consecutive trials, if a microwave-oven (or any device that generates interference) is used during A's trials but not during B's trials, A will likely experience unfairly low throughput. However, with multiple interleaved trials, if the trials are short relative to the length of time the microwave is on, both alternatives are likely to be subjected to the interference. If the trials are sufficiently long, such that one alternative is affected significantly more than the other, it should show up as wider confidence intervals for that alternative.

5.3.1 *Stationary: Two Alternatives*

We begin by using the same 24-hour data used in the previous evaluations in Sections 5.1 and 5.2. This ensures that the same conditions are experienced using all three methodologies. We process the data by using the first 60 seconds of data for the first trial of the first alternative (A) and the next 60 seconds for the first trial of the second alternative (B). This completes the first round or trial. This process is repeated until N trials have been obtained. We then compute the mean throughput, RSSI and confidence intervals from the N trials and plot that data. This process is repeated until all 24 hours of data have been used. We thus produce multiple pairs of data points comparing alternatives A and B over different times of the day. Since A and B are identical, we expect to see no statistical difference in any of the pairs of data points comparing the two alternatives.

Figure 4 shows these results for 5, 10, and 15 trials from a period of the day with the high variability (16:00 - 20:00). In all cases (including times not shown) confidence intervals for A and B overlap. This suggests that experiments using multiple interleaved trials are repeatable for two alternatives for the given data. Increasing the number of trials tightens the confidence intervals. Recall that with multiple consecutive trials, increasing the number of trials decreased the confidence intervals, however the results were not repeatable.

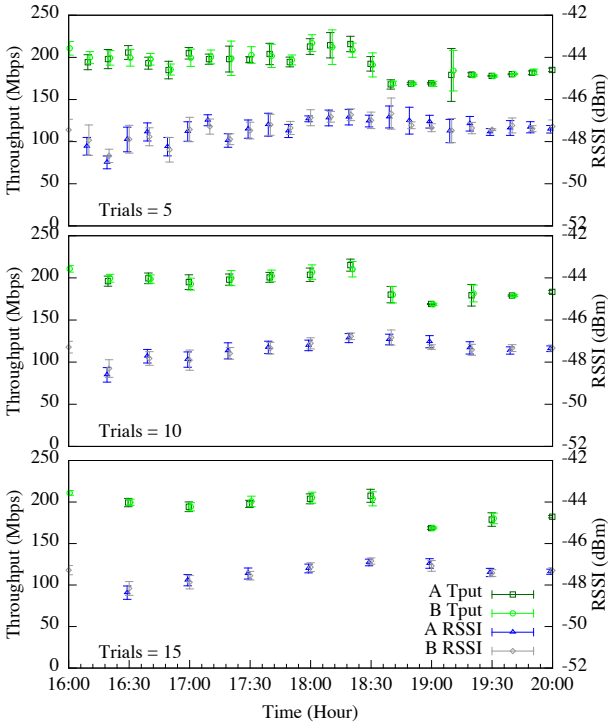

Figure 4: Multiple interleaved trials (2 alternatives)

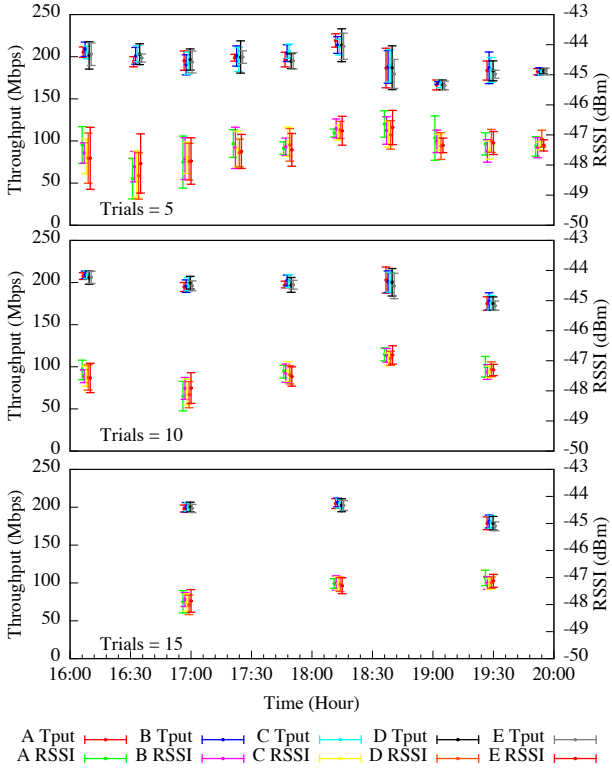

Figure 5: Multiple interleaved trials (5 alternatives)

5.3.2 *Stationary: Five Alternatives*

To this point we have focused on comparing two alternatives. However, it is often desirable to compare multiple competing alternatives. Comparing more alternatives makes the evaluation more challenging because each round takes longer to complete and trials from the same alternative are farther apart in time. Consequently, trials may experience more disparate channel conditions. The different trials for each alternative can be separated in time either because of more alternatives, as is the case in this section, or perhaps because of setup time required to switch between alternatives. In Section 5.3.5 we also consider the interval between trials by examining trials of different lengths.

We again use the same 24-hour data and multiple interleaved trials to compare five alternatives. Data is processed in the same way as for two alternatives, except there are now five alternatives (so each round lasts 5 minutes). The results are shown in Figure 5. Each group of points, illustrated by different colors, represents the comparison of 5 alternatives. This graph shows that the 95% confidence intervals for the mean throughput and RSSI for all five alternatives for 5, 10 and 15 trials overlap. The overlapping confidence intervals indicate that each alternative was subjected to roughly similar channel conditions, and demonstrates that for this environment, results are repeatable for five alternatives using multiple interleaved trials. For these experiments, increasing the number of trials decreases the size of the confidence intervals, allowing the study of finer differences between multiple alternatives. However, we expect that diminishing returns would come into effect as the number of trials grows.

GUIDELINE: *The technique of using multiple interleaved trials increases the probability that the compared alternatives will be subject to similar channel conditions. This leads to repeatable experiments and should permit fair comparisons of different alternatives. Because multiple trials are being conducted, the size of the confidence intervals is a reliable indicator of the variability in channel conditions across different trials, and also provides a gauge as to whether or not differences in the observed means are statistically significant.*

5.3.3 *Mobile: No Interference*

Previous experiments were conducted using stationary devices with no interference. Since mobility increases the variability of wireless channels, we evaluate the efficacy of multiple interleaved trials in mobile environments where a WiFi receiver is moved by a motorized toy train or by a person.

We set up a toy train track in a roughly oval shape that allows for a fairly wide range of received signal strengths due to varying degrees of path obstruction and distances from the Access Point (AP). At the starting point, there is very little blocking the line of sight between the laptop and the AP. At the farthest point, there are several cubicle walls consisting of metal, wood and fabric in the way. Although, the distance between the closest and furthest points on the track are roughly only 5 meters, path loss due to cubicle walls affect the 5 GHz signal propagation sufficiently to achieve a 15 dBm range of signal strength variation.

We place our laptop on the train and aim to maximize repeatability by starting the train from the same position in each trial. Each trial consists of two laps around the track at approximately walking speed. We find that throughput experiments are relatively repeatable in this environment. Figure 6 (top) shows the throughput (bottom cluster) and RSSI (top cluster) measured for each second. Each trial lasts 60 seconds and we plot the throughput obtained during each second of all 20 trials. We add wider dark lines indicating the maximum and minimum values obtained across all runs

to better indicate the range of values obtained. As expected, each trial is different, however, all trials follow similar trends with throughput increasing near the AP (near the 0, 30, and 60 second marks) and decreasing as it gets further away.

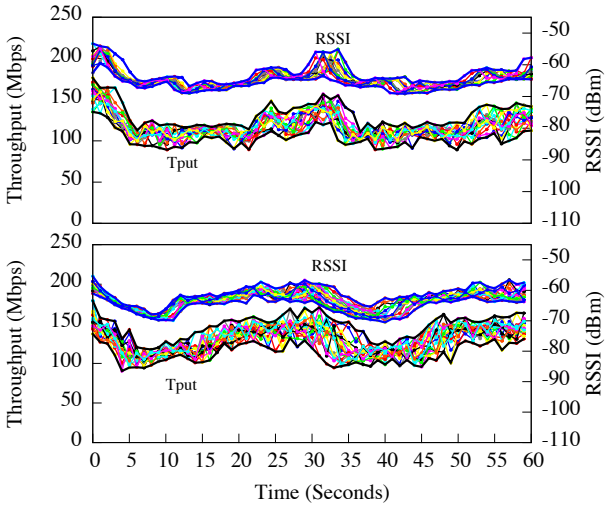

Figure 6: Mobile: toy train (top) and walking (bottom)

We now repeat the same experiments, with a person carrying the laptop. Markers on the floor are used to follow roughly the same path as that of the train. Holding the laptop at waist height, we measure throughput for 60 seconds; again making two passes around the track for each trial. Figure 6 (bottom) shows that the throughput (bottom cluster) and RSSI (top cluster) trends are fairly consistent across trials, although less consistent than those obtained with the train. This is expected as walking a path is naturally less accurate due to the difficulty of maintaining consistent speeds, positioning of the arms and body, and due to the body's impact on the received signals.

We should not (and do not) compare performance results obtained from the train and walking trials. The height of the laptop and the line of sight have changed between the two experiments. Instead, we use interleaved trials to analyze the collected data as though two alternatives are being studied in each of the toy train and walking experiments.

Figure 7 shows the mean throughput achieved using the multiple interleaved trials technique for different numbers of trials. The bars with 3 trials use data from the first 6 runs (two alternatives are interleaved), 5 trials - the first 10, and 10 trials - all 20 runs. These graphs are presented to examine differences in the variability and repeatability of results obtained during the train and walking experiments. Again, if the experiments are repeatable, we expect to see no statistically significant difference between these two alternatives (labeled A and B). While the confidence intervals are overlapping in all cases, indicating repeatability, the train experiments exhibit tighter confidence intervals than the walking experiments. This is an important finding as walking experiments are often used for mobile WiFi performance evaluation. If small differences between alternatives are to be measured, the use of a robot (or train) may help reduce variability and increase the likelihood of obtaining statistically significant differences. On the other hand, because increased variability may be due to line of sight interference from the person carrying the device, one should be careful not to assume that results obtained with a train or robot can be applied to situations where people are walking and carrying devices.

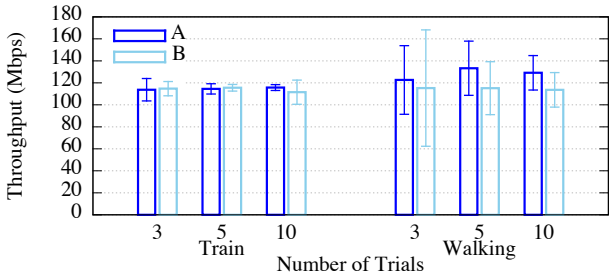

Figure 7: Mobile: no interference

GUIDELINE: *Human controlled mobile experiments introduce significant variability, making it desirable to use automated tools such as an electric train or robot. The multiple interleaved trials technique may permit the use of experiments that involve people walking with mobile devices when automation is not feasible. However, it may be difficult to discern small differences between alternatives.*

5.3.4 Mobile: Person Walking with Interference

To further test the reliability of results obtained using multiple interleaved trials, we conduct a few experiments using a 2.4 GHz network where WiFi and non-WiFi interference are present. We first use a channel that overlaps with multiple other uncontrolled APs (called the "Busy Channel") and then a channel with neighboring, but non-overlapping, channels (called the "Unused Channel" because it is unused except for our experiments). We refrain from referring to this second channel as interference-free due to the potential for channel leakage effects described in Section 5.1 and [20]. Instead, these two experiments can be thought of as examining scenarios with significant amounts of WiFi interference and with limited WiFi interference. Both channels are subject to non-WiFi interference. We follow the same procedures as the walking experiments, described in Section 5.3.3, using twenty 60-second trials. However, in this case we started much closer to the AP and walked to a point significantly farther from the AP. This was done in order to cover a much wider range of signal strengths.

We plot raw throughput and RSSI for each trial in Figure 8 and find that while there is significant variation across trials for both channels, the general trends are quite similar across the different trials. Figure 9 shows the average throughput with 95% confidence intervals. When considering two alternatives we see that the confidence intervals overlap in all cases. This suggests that multiple interleaved trials can be used to conduct repeatable experiments, even in a challenging 2.4 GHz environment with significant interference. Note that the confidence intervals for these experiments are much smaller than the walking experiments shown in Figure 7. This is indicative of the variable and unpredictable nature of these types of experiments. While it might be tempting to conclude that existing techniques could be used in this situation, we again emphasize that we know of no way to guarantee or verify that channel conditions do not change during the experiment in a way that creates unfair conditions for one or more alternatives.

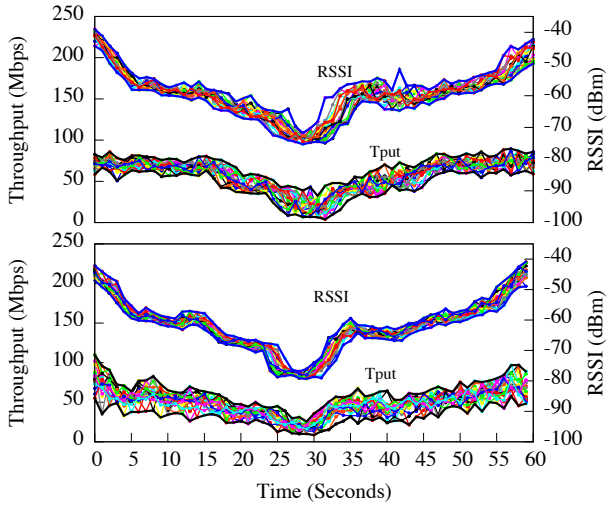

Figure 8: Mobile: unused (top) and busy (bottom) channels

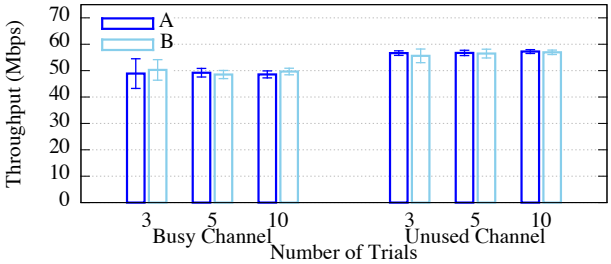

Figure 9: Mobile: 2.4 GHz, busy and unused channel

5.3.5 *Trial Durations*

In addition to ensuring that enough trials are run to obtain tight confidence intervals, an interesting consideration is the length of time required to conduct each trial. One issue is how long trials should be and another is whether or not trials of different lengths can be compared (e.g., when measuring data transfer time) While these are issues that we hope to investigate more deeply in future research, we now provide a brief examination. To study these issues we use the same 24-hour data used previously in Sections 5.1 and 5.2, and compare results obtained using trials of different lengths.

Figure 10 shows the results obtained by considering three alternatives (A, B and C) with trial lengths of 15, 60 and 240 seconds, respectively. In this case each round is 315 seconds. The top, middle and bottom graphs show, 5, 10 and 15 trials respectively. To more easily compare the three alternatives, we have zoomed in on a range of time where variability was relatively high (in this case 12:00 - 20:00).

These results show that even when comparing alternatives that use different durations (15, 60, 240 seconds), all sets of experiments comparing these alternatives have overlapping confidence intervals. This demonstrates repeatability for this technique, for this set of data. We also note that with only 5 trials, the confidence intervals for alternative A (15 seconds) may be slightly larger in some cases than those obtained for the alternatives with longer trial durations. However, as the number of trials increases, those differences seem to diminish. In Section 6, we use alternatives that each require different amounts of time to run in order to test if this methodology can distinguish between alternatives where differences are expected.

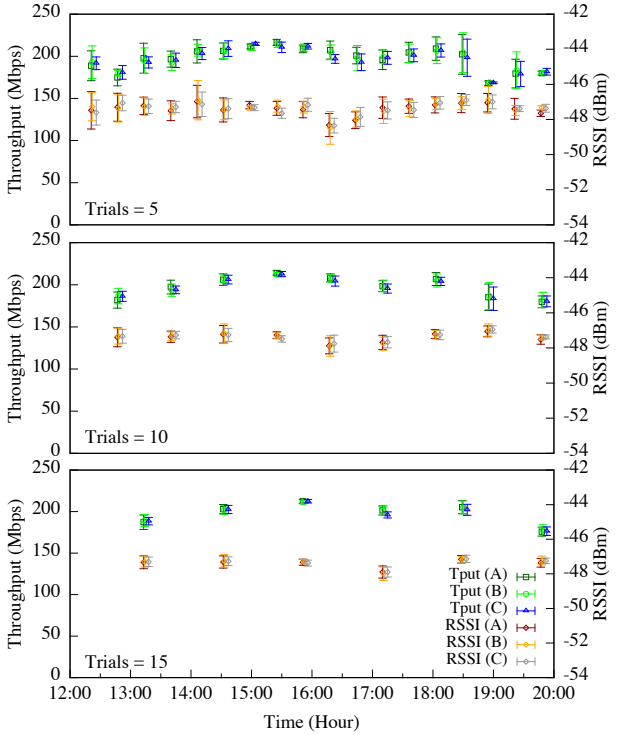

Figure 10: Different trial lengths 15 s (A), 60 s (B), 240 s (C)

6. DISTINGUISHING DIFFERENCES

Thus far all experiments have compared two or more alternatives, each of the same configuration and setup, which allows us to establish repeatability. However, it is often desirable to perform experiments involving comparisons of different alternatives. Generally, we are interested in how a given system, configuration, or algorithm performs relative to other alternatives.

In order to study the granularity at which statistically significant differences can be determined between alternatives, we conduct experiments using the interleaved trials methodology. In this example, we vary the amount of the data transferred from a sender to a receiver and measure the time required to complete the transfer. Relative to the baseline data size of 200 MB, we use transfer sizes of 100%, 110%, 120%, 130% and 140%, whose transfer times should ideally, differ by exactly these factors (relative to the baseline case). We chose 5 alternatives in order to determine the degree to which the interleaved trials methodology can be used to distinguish between these different alternatives. While we perform 20 interleaved trials of each size, in both a noisy 2.4 GHz network and an interference-free 5 GHz network, we examine the data obtained after 5, 10, 15 and 20 trials. This is done in order to study the influence that the number of trials may have on the variability of the results and therefore the confidence intervals.

Figure 11 (top), shows the transfer times in the 2.4 GHz network for each of the 20 trials. As can be seen in this graph, the transfer times differ widely between trials. At 2.4 GHz, transfer times are noticeably longer near the beginning of our experiment. This reinforces the need to use the multiple interleaved trials technique; if consecutive trials were used then transfer times for data sizes measured early in the experiment could have appeared unfairly long.

However, using the multiple interleaved trials technique, all alternatives (i.e., data sizes) are subjected to the unfavorable environment. Using the 5 GHz network, transfer times were fairly consistent for each data size, which is expected due to the lack of interference.

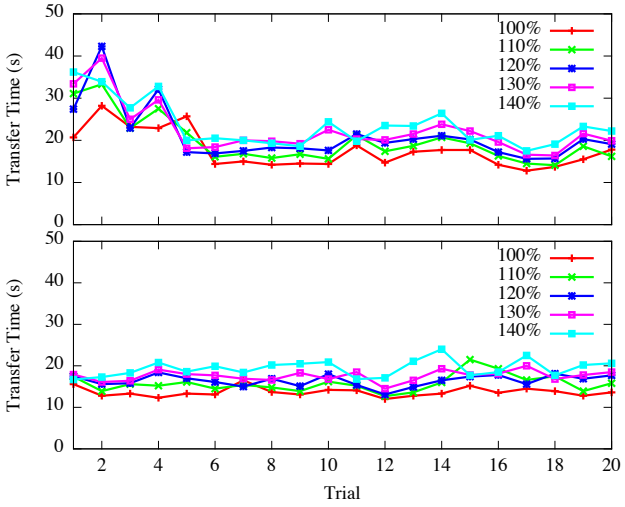

Figure 11: Transfer times: 2.4 (top) and 5 GHz (bottom)

Figure 12 plots the average transfer times and 95% confidence intervals computed after 5, 10, 15 and 20 trials. As might be expected, when using the 2.4 GHz network with more dynamic channel conditions (e.g., due to WiFi and non-WiFi interference), small numbers of trials results in relatively large confidence intervals. This may make it difficult to draw conclusions about differences in performance between alternatives unless those differences are substantial.

Confidence intervals are noticeably wider when using the 2.4 GHz network, compared to the 5 GHz network. As a result, for the 2.4 GHz experiment, only differences as large as 40% are considered statistically significant. However, for the 5 GHz experiment, differences as small as 10% can be distinguished. Generally, a rule of thumb cannot be established regarding the granularity of differences that can be detected, as it depends on the level of variability in the experiment. Note that is not a limitation of the technique but rather the reflection of natural variation of experimental results. The multiple interleaved trials technique naturally and inherently captures variability across trials.

GUIDELINE: *When using the multiple interleaved trials methodology in some environments, a small number of trials may not be sufficient to allow one to determine statistically significant differences between multiple alternatives. It is critical to conduct multiple trials and to compute and report confidence intervals.*

---

7. DISCUSSION

It is important to understand that the results we have obtained are unique to our environment and the times at which the experiments were run. These results should **not** be used to draw conclusions about the granularity of differences that can be distinguished in other environments. We acknowledge the difficulties involved in conducting multiple interleaved trials. It may be time consuming to manually make configuration changes between each trial. This stresses the benefits of automation or scripting of experiments.

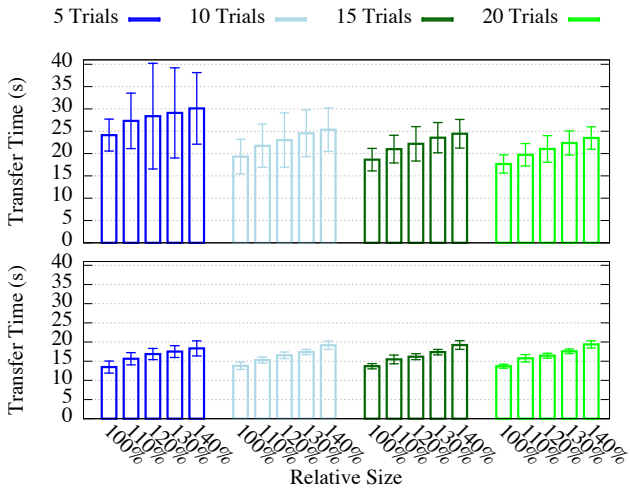

Figure 12: Transfer times: 2.4 (top) and 5 GHz (bottom)

In an ideal world, all trials would be conducted at once to achieve perfect temporal correlation and identical channel conditions between trials. One approach, inspired by Judd *et al.* [27], would be to use splitters/combiners to capture identical signals at multiple receivers. Unfortunately, this has the significant drawback of reducing the received signal strength by $1/N$, where N is the number of splitters. Therefore, this approach may be of limited practicality and we leave its evaluation for future work.

8. USE IN OTHER CONTEXTS

We have found that the idea of using multiple interleaved trials predates our use in at least one other field. In an n-of-1 clinical trial, one individual is observed as different treatments (perhaps including a placebo) are administered over time with data being collected and analyzed to determine the best treatment. This technique has applications in individualized medicine [28].

We believe that multiple interleaved trials are well-suited to, but to our knowledge have not been used, when evaluating the performance of computer systems or networks when the system under test *can not be guaranteed* to be subject to identical conditions across experiments. Some examples might include: experiments on other types of wireless networks, "live" systems like wide area networks or web services, or when using cloud computing environments where CPU, memory and network performance can all vary significantly [29] over time. We plan to investigate some of these use cases in future work.

9. CONCLUSIONS

Experiments have long been used and will remain an important tool for evaluating the performance of 802.11 networks. Because channel conditions vary over time, the difficulty lies in obtaining repeatable and fair comparisons when evaluating competing alternatives. One approach used in the literature is to control for, and essentially eliminate, variability in channel conditions. However, such evaluations are not representative of more variable channel conditions, under which devices are likely to be used.

In this paper, we study the degree to which 802.11n MIMO network experiments can be repeated, with particular emphasis on methodologies for comparing competing systems,

configurations, or algorithms. We find that using existing methodologies, we were not able to reliably obtain repeatable results, even under controlled conditions where there is no WiFi or non-WiFi interference. As a result, we propose the use of and evaluate the multiple interleaved trials methodology, that permits conducting experiments under variable channel conditions. The keys with this approach are that 1) it ensures that all alternatives are subject to similar channel conditions, and is therefore fair, and 2) it can be used to easily and explicitly measure the variability of the results. Using this technique, we are able to obtain repeatable results and distinguish differences among competing alternatives in several challenging scenarios, including mobile environments with both WiFi and non-WiFi interference. We recommend that multiple interleaved trials be considered when conducting 802.11n MIMO experiments and stress that, regardless of the methodology used, it is critical for researchers to understand, quantify and report on the variability of their results.

10. ACKNOWLEDGMENTS

Tim Brecht is partially supported by a Discovery Grant and an Accelerator Supplement from the Natural Sciences and Engineering Research Council (NSERC). Ali Abedi and Andrew Heard are partially supported by scholarships from BlackBerry and NSERC, respectively. The authors thank Xiaoyi (Eric) Liu for his work on the ath9k driver, and Augusto Born de Oliveira, Jean-Christophe Petkovich, Tyler Szepesi and the anonymous reviewers for their feedback.

11. REFERENCES

[1] S. Ganu, H. Kremo, R. Howard, and I. Seskar, "Addressing repeatability in wireless experiments using orbit testbed," in *Tridentcom*, 2005.

[2] R. Burchfield, E. Nourbakhsh, J. Dix, K. Sahu, S. Venkatesan, and R. Prakash, "RF in the jungle: Effect of environment assumptions on wireless experiment repeatability," in *ICC*, 2009.

[3] G. Judd and P. Steenkiste, "Using emulation to understand and improve wireless networks and applications," in *NSDI*, 2005.

[4] D. Johnson and A. Lysko, "Comparison of MANET routing protocols using a scaled indoor wireless grid," *Mobile Networks and Applications*, 2008.

[5] P. De, A. Raniwala, R. Krishnan, K. Tatavarthi, J. Modi, N. A. Syed, S. Sharma, and T. Chiueh, "Mint-m: An autonomous mobile wireless experimentation platform," in *MobiSys*, 2006.

[6] O. Rensfelt, F. Hermans, P. Gunningberg, and L.-A. Larzon, "Repeatable experiments with mobile nodes in a relocatable WSN testbed," in *DCOSSW*, 2010.

[7] B. Blywis, M. Güneş, F. Juraschek, and J. H. Schiller, "Trends, advances, and challenges in testbed-based wireless mesh network research," *Mobile Networks and Applications*, 2010.

[8] I. Pefkianakis, Y. Hu, S. H. Wong, H. Yang, and S. Lu, "MIMO rate adaptation in 802.11n wireless networks," in *MobiCom*, 2010.

[9] L. Deek, E. Garcia-Villegas, E. Belding, S.-J. Lee, and K. Almeroth, "Joint rate and channel width adaptation for 802.11 MIMO wireless networks," in *SECON*, 2013.

[10] K. Nikitopoulos, J. Zhou, B. Congdon, and K. Jamieson, "Geosphere: Consistently turning MIMO capacity into throughput," in *SIGCOMM*, 2014.

[11] S. H. Y. Wong, H. Yang, S. Lu, and V. Bharghavan, "Robust rate adaptation for 802.11 wireless networks," in *MobiCom*, 2006.

[12] Aniket, N. Carlsson, C. Williamson, and M. Arlitt, "Ambient interference effects in WiFi networks," in *NETWORKING*, 2010.

[13] S. Rayanchu, A. Patro, and S. Banerjee, "Airshark: detecting non-WiFi RF devices using commodity WiFi hardware," in *IMC*, 2011.

[14] S. Lakshmanan, S. Sanadhya, and R. Sivakumar, "On link rate adaptation in 802.11n WLANs," in *INFOCOM*, 2011.

[15] C.-Y. Li, C. Peng, S. Lu, and X. Wang, "Energy-based rate adaptation for 802.11n," in *Mobicom*, 2012.

[16] X. Tie, A. Seetharam, A. Venkataramani, D. Ganesan, and D. L. Goeckel, "Anticipatory wireless bitrate control for blocks," in *CoNEXT*, 2011.

[17] H. Rahul, F. Edalat, D. Katabi, and C. G. Sodini, "Frequency-aware rate adaptation and mac protocols," in *Mobicom*, 2009.

[18] D. Gupta, D. Wu, P. Mohapatra, and C.-N. Chuah, "Experimental comparison of bandwidth estimation tools for wireless mesh networks," in *INFOCOM*, 2009.

[19] A. Abedi and T. Brecht, "T-RATE: A framework for the trace-driven evaluation of 802.11 rate adaptation algorithms," in *MASCOTS*, 2014.

[20] L. Deek, E. Garcia-Villegas, E. Belding, S.-J. Lee, and K. Almeroth, "The impact of channel bonding on 802.11n network management," in *CoNEXT*, 2011.

[21] AirMagnet, "Fluke networks." http://www.flukenetworks.com/enterprise-network/wireless-network/AirMedic.

[22] IPerf. http://sourceforge.net/projects/iperf/.

[23] W.-L. Shen, Y.-C. Tung, K.-C. Lee, K. C.-J. Lin, S. Gollakota, D. Katabi, and M.-S. Chen, "Rate adaptation for 802.11 multiuser MIMO networks," in *Mobicom*, 2012.

[24] D. Montgomery, *Design and Analysis of Experiments*. Wiley, 2012.

[25] G. Judd, X. Wang, and P. Steenkiste, "Efficient channel-aware rate adaptation in dynamic environments," in *MobiSys*, 2008.

[26] D. Halperin, W. Hu, A. Sheth, and D. Wetherall, "Predictable 802.11 packet delivery from wireless channel measurements," in *SIGCOMM*, 2010.

[27] G. Judd and P. Steenkiste, "A simple mechanism for capturing and replaying wireless channels," in *ACM SIGCOMM workshop on Experimental approaches to wireless network design and analysis*, 2005.

[28] E. Lillie, B. Patay, J. Diamant, B. Issell, E. Topol, and N. Schork, "The n-of-1 clinical trial: the ultimate strategy for individualizing medicine?," *Personalized Medicine*, vol. 8, no. 2, pp. 161–173, 2011.

[29] J. Schad, J. Dittrich, and J.-A. Quiané-Ruiz, "Runtime measurements in the cloud: observing, analyzing, and reducing variance," *Proceedings of the VLDB Endowment*, vol. 3, no. 1-2, pp. 460–471, 2010.