Can They Hear Me Now?: A Case for a Client-assisted Approach to Monitoring Wide-area Wireless Networks

Sayandeep Sen*, Jongwon Yoon*, Joshua Hare*, Justin Ormont, Suman Banerjee
University of Wisconsin-Madison
{sdsen, yoonj, hare, ormont, suman}@cs.wisc.edu

ABSTRACT

We present WiScape, a framework for measuring and understanding the behavior of wide-area wireless networks, e.g., city-wide or nation-wide cellular data networks using active participation from clients. The goal of WiScape is to provide a coarse-grained view of a wide-area wireless landscape that allows operators and users to understand broad performance characteristics of the network. In this approach a centralized controller instructs clients to collect measurement samples over time and space in an opportunistic manner. To limit the overheads of this measurement framework, WiScape partitions the world into zones, contiguous areas with relatively similar user experiences, and partitions time into zone-specific epochs over which network statistics are relatively stable. For each epoch in each zone, WiScape takes a minimalistic view — it attempts to collect a small number of measurement samples to adequately characterize the client experience in that zone and epoch, thereby limiting the bandwidth and energy overheads at client devices. For this effort, we have collected ground truth measurements for up to three different commercial cellular wireless networks across (i) an area of more than 155 square kilometer in and around Madison, WI, in the USA, (ii) a road stretch of more than 240 kilometers between Madison and Chicago, and (iii) locations in New Brunswick and Princeton, New Jersey, USA, for a period of more than 1 year. We justify various design choices of WiScape through this data, demonstrate that WiScape can provide an accurate performance characterization of these networks over a wide area (within 4% error for more than 70% of instances) with a low overhead on the clients, and illustrate multiple applications of this framework through a sustained and ongoing measurement study.

*Primary authors.

IMC'11, November 2–4, 2011, Berlin, Germany.

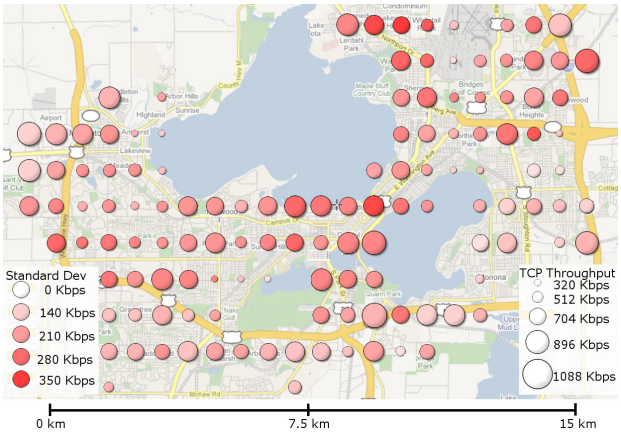


Figure 1: A Snapshot of TCP throughput distribution within our NetB network, covering a 155 sq. kilometer city-wide area. Throughputs are collected based on 1MB downloads, collected using WiScape. Each dot corresponds to a circular area of radius 600 meters.

Categories and Subject Descriptors

C.2.3 [**Network Operations**]: Network Monitoring, Public networks; C.2.1 [**Computer Communication Networks**]: Network Architecture and Design—*Wireless Communication*

General Terms

Documentation, Experimentation, Measurement, Performance, Reliability

Keywords

Cellular Networks, Measurement, Client Assisted

1. INTRODUCTION

The ability to observe an entire network's performance is an important precursor to understanding and predicting its behavior, and in debugging its performance problems. Gathering such detailed observations at a network-scale is challenging for any network, whether wired or wireless.

In wired networks, such as enterprises or ISPs, operators typically deploy multiple monitoring nodes in carefully chosen vantage points within the network to capture and aggregate necessary information [1]. In the context of WLANs,

multiple early efforts emulated these wired approaches by deploying similar monitoring nodes in the wired part of the network [2]. However, wired-only observations fail to capture the impact of *location-specific* RF characteristics. Hence, more recent WLAN monitoring efforts, e.g., Jigsaw [3], chose to deploy numerous wireless sniffers across an entire physical space, e.g., a campus building. While deployment of such a widespread wireless monitoring infrastructure is still feasible in building-wide settings, the logistics of densely deploying and managing such infrastructure is impractical when the wireless networks are significantly larger in scale. In particular, it is virtually impossible to densely deploy wireless sniffers to monitor the performance of a *city-scale or a nation-scale* cellular data network. In this paper, we examine a different solution for monitoring and measuring such large-scale wireless networks, one that leverages the potential assistance of a large number of clients. More specifically, we present *WiScape,* a framework to characterize the *wi*reless land*scape*, using data collected from three commercial cellular data networks over more than one year across large regions: (i) more than 155 square kilometers in and around Madison, WI, in the USA, (ii) a road stretch of 240 kilometers between Madison and Chicago, and (iii) targeted regions in the cities of New Brunswick and Princeton in New Jersey, USA. Figure 1 presents a snapshot of some of our measurement data collected from one of the monitored cellular networks across Madison, WI. The figure partitions the entire area into coarse-grained zones (each zone is a 0.2 sq.km.), with only a sampled subset of zones shown, and the size of the circles represent the average values of the TCP download throughputs (the shade of the circle represent the variance of throughput samples).

In this paper, we also demonstrate, via experimentation, how network operators and users (applications) can benefit from data accumulated by the WiScape framework. For instance, in our experiments we found that a network operator can use WiScape to easily identify significant changes in user experiences within their own network, while an application such as MAR [4] (which uses multiple cellular network interfaces to provide aggregated wireless bandwidth into vehicles) can improve its own performance by up to 41% by leveraging WiScape collected data. Finally, an approach such as WiScape can potentially serve as a *performance watchdog* and can provide a neutral view of different commercial wireless networks over time and space.

Client-assisted monitoring in WiScape

Cellular data networks nationwide are placing increasing emphasis on performance, mobility and wide-area coverage. As these networks attempt to provide ubiquitous connectivity to large geographic areas, operators continue to seek better tools to observe network performance at all locations. Each network operator sends out its RF monitoring trucks during initial deployment of cellular towers and periodically after that to various neighborhoods. Occasionally, if the operator receives a large volume of consumer complaints of network performance from a certain area, they would also conduct additional RF surveys at those specific locations [5]. Each such RF survey is quite labor-intensive.

Furthermore, user complaints are unlikely to capture a vast majority of network performance issues that occur. Users often only complain when a problem is particularly serious and persistent, causing major disruptions to the user over a long period of time.

In WiScape we propose to measure the network's performance, *as perceived by the clients and through the help of clients.* More specifically, in this approach diverse mobile clients measures network properties based on instructions from a central controller to map out the performance across the entire network. Since the clients are naturally mobile, they are perfectly positioned to monitor the network performance from various vantage points. If implemented successfully, this approach can mitigate significant costs that operators might otherwise have to incur in order to collect data of the same level of richness and detail. In addition, such an approach provides us with unique data from the client's point of view, which is not available otherwise.

This high-level idea is actually a fairly common one and different variants of it have been referred to as crowd-sourcing, war-driving, and participatory sensing. Such approaches have been used to collect locations of WiFi APs worldwide, and have been proposed in various types of health-related applications (air pollution, audio noise level, and radiation exposure monitoring across a city), as well as social interactions (detecting presence of friends nearby) [6, 7]. In the wireless setting, there are now ongoing efforts that attempt a similar approach to collect performance data of different cellular networks. Examples include RootWireless [8] a company that distributes mobile phone applications that collect measurements from volunteers to generate coverage maps of cellular operators, the 3gtest application from U Michigan [9] and AT&T's "Mark the Spot" iPhone application that allows iPhone users to record the location of where a phone call was dropped.

While the main idea is relatively simple, the core technical challenge in designing an effective, scalable, and useful system lies in its ability to manage the volume of measurements required and the manner in which measurement tasks can be coordinated across multiple clients. We comment on this issue next.

WiScape approach and usage scenarios: In a client-based monitoring system, if all clients are requested to collect performance measurements all the time, the volume of such measurement traffic could prohibit useful activity in the network. Such an effort could also place a significant burden on the client devices leading to quicker depletion of the limited battery power of these devices. Therefore, the key in designing a client-based monitoring infrastructure is to ensure that the volume of data collected is low, yet is adequate to present the operators and users with a broad understanding of network performance. At the same time, since this approach is able to collect measurements only from specific locations clients are available at any given instant, the number of measurement samples available from any arbitrary location and at any desired time is likely to be quite sparse, often zero, and hence not statistically significant. Therefore, in WiScape *we need to aggregate collected measurements from clients, both in time and space* so that it is statistically significant for observations.

Burdened by above considerations, we partition the world into *zones* (around 0.2 sq. kilometer each) and time in each zone into *epochs* (a few tens of minutes). We define zones such that measurements within each zone have relatively low variance most of the time. We define epochs such that statistics across multiple consecutive epochs of the same zone have

low variance. (Note that epochs may have smaller durations in zones with rapidly changing performance observed by clients.) In other words, each epoch for each zone is the smallest time-space granularity that WiScape attempts to accurately estimate to provide a stable measure. Based on our models, we require around 100 measurement samples to estimate network layer performance of each epoch of a zone, such as throughput, delay, loss, and jitter. We believe for most zones this measurement volume is easy to obtain, especially for zones in dense urban areas with many users. which often require greater attention from network operators.

The nature of data collection in WiScape also dictates the type of its use. Given our intent of collecting a small amount of data, WiScape will miss many of the short-term and transient variations, e.g., as a result of sudden burst of active users arriving in a given location and then disappearing within a few minutes. However, any persistent network behavior (persistent in the order of an epoch, typically tens of minutes) will be captured by our system quite accurately. We show this through multiple examples in Section 4. An interesting such example was WiScape's detection of 4× increase in latencies in a specific zone of two cellular networks in Madison, encompassing the UW-Madison football stadium, for nearly 3 hours on a football Saturday when nearly 80,000 people packed into the stadium for the game.

In the rest of the paper, we explain how we designed the WiScape framework through detailed measurements and statistical analysis of the data and make the following key contributions:

- We establish the feasibility of client-assisted monitoring of wide area wireless networks by carrying out extensive measurements over a duration of more than 1 year, spanning a geographical area of more than 155 sq.km. across multiple cities, a long road stretch of 240 k.m. and across three 3G cellular networks. Traces used for the paper will be made publicly available through CRAWDAD [10].
- We design and implement WiScape — a monitoring system that bins measurements into epochs and zones and collects a relatively small number of measurements per epoch per zone. We establish appropriate parameters for epochs, zones, and the number of measurement samples through detailed data collection, analysis, and experimentation.
- We demonstrate the benefits of WiScape through multiple simple use cases: (i) to quickly detect somewhat persistent changes to network behavior and alert network operators of need to perform detailed investigations of such changes, (ii) to apply WiScape collected data to improve the performance of multi-network applications like MAR and "multi-sim."

In the next section, we present some details on our measurement and data collection efforts used designing and evaluating different aspects of WiScape. Subsequently, in Section 3, we present the overall design of WiScape including related validation. In Section 4 we demonstrate some uses of data collected by WiScape, and finally discuss related work and present our conclusions in Sections 5 and 6 respectively.

2. PRELIMINARIES

Our measurement setup consists of a measurement coordinator running on a desktop in our university laboratory, with well provisioned connectivity to the Internet, that periodically requests and collects measurements from different client devices (based on Windows and Linux platforms). In our measurements, we have gathered data from three different cellular networks with nation-wide footprints, referred to as NetA , NetB , and NetC [1]. The data collection process has been ongoing in multiple stages for more than one year now (Table 1) and different clients in our measurement setup had different capabilities and characteristics as discussed next.

Data collection process: While we have collected measurement data for both uplink and downlink, in this paper, we focus on the downlink direction. This is motivated by the observation that most of data traffic is downlink. Our data collection has been done using multiple platforms, some of which are mounted on vehicles (public transit buses in Madison, intercity buses, as well as nodes mounted on personal vehicles), while others are static.

Wide-area: The spatially biggest datasets are labeled *Standalone* and *WiRover* . The *Standalone* dataset was collected using up to five public transit buses in Madison, covering an approximate area of 155 sq. kilometer in this city. These public transit buses typically run from 6am to midnight and each particular bus gets randomly assigned to different routes each day. Even in a single month, this set of buses is able to cover a significant fraction of Madison and its neighboring cities. The *WiRover* data collection process is the newer incarnation of the *Standalone* process, in which all of these bus-mounted nodes now are equipped with two network interfaces (NetB and NetC), and provide free WiFi service to bus passengers using the multi-network setup [13]. In addition to the public transit buses of Madison, we also placed additional nodes on two intercity buses between Madison and Chicago, a distance of more than 240 kilometer. Over time, these buses generated multiple measurement values for each location along this path stretch. We did not evaluate if any bias was introduced by the periodic nature of bus routes on the collected data.

Spot: The vehicular setup cannot provide us with long running contiguous measurements from a specific location. To study cellular network performance over a longer timescale, we selected some indoor locations to continuously collect data for up to 5 months. These included multiple locations in Madison, WI, and Princeton and New Brunswick, NJ. We describe our criteria for selecting the specific locations in Section 3.1. These datasets provide a more detailed and fine-grained view than is possible using with the vehicular collection methods of our *Wide-area* data. We apply these datasets to understand network performance over time for a given static location as will be demonstrated in Section 3.2.1.

Region: This consists of multiple datasets: *Proximate-WI*, *Proximate-NJ*, and *Short segment*. The two *Proximate* datasets were collected in neighborhoods close to the previously selected *Spot* locations. All three datasets consist of targeted measurement data to understand the feasibility of

[1]Since our goal for this paper currently is to explore a measurement framework, and *not* to answer which of these networks perform best or worst in different locations, we did not find it useful to reveal the identities of these nation-wide cellular providers.

Networks	NetA	GSM HSPA [11], Uplink (≤1.2Mbps), Downlink (≤ 7.2Mbps)
	NetB	CDMA2000 1xEV-DO Rev.A [12], Uplink (≤ 1.8Mbps), Downlink (≤ 3.1Mbps)
	NetC	CDMA2000 1xEV-DO Rev.A [12], Uplink (≤ 1.8Mbps), Downlink (≤ 3.1Mbps)
Hardware	Server	3 Desktops with well provisioned wired Internet connection
	Client	3 Laptops with 3 cellular data cards & GPS
Measurement params	Transport protocol (TCP/UDP), Transmission duration (10sec∼5min)	
	Inter packet delay (1msec∼100msec, adaptively varies base on available capacity)	
	Download size (200 and 1200Bytes for UDP, 100Bytes∼2048Bytes for TCP)	
Params logged	Packet sequence number, Receive timestamp, GPS coordinates	

Table 1: Measurement setup details.

Group	Name	Span	Months	Nets	Location
Spot	*Static-WI*	5 locations	5	A, B, C	Madison, WI
	Static-NJ	2 locations	1	B, C	New Brunswick, Princeton, NJ
Region	*Proximate-WI*	Vicinity of the static locations	5	A, B, C	Madison, WI
	Proximate-NJ	Vicinity of the static locations	1	B, C	New Brunswick, Princeton, NJ
	Short segment	20 km road stretch	3	A, B, C	Madison, WI
Wide-area	*WiRover*	155 sq.km. city-wide area and a 240 kilometer road stretch	6	B, C	Madison, WI and Madison to Chicago
	Standalone	155 sq.km. city-wide area	11	B	Madison, WI

Table 2: Different data sets and details of locations. All measurements used TCP and UDP flows, except *Standalone* which used ICMP pings instead of UDP flows.

composing infrequently collected measurement samples from multiple (and potentially diverse) sources for estimating network performance, as will be seen in Section 3.3. These measurements were collected using client devices placed inside personal automobiles and regularly driven over fixed routes.

All of our measurements reported in this paper were collected using laptops or single-board computers equipped with different models of cellular modems (some were USB-based and others were PCMCIA).

Measurements collected: The *Spot* measurements and *Region* measurements collected a specific set of performance metric over three cellular networks, including TCP and UDP throughput, UDP packet loss rate, application level jitter measured in terms of Instantaneous Packet Delay Variation (IPDV) [14], application level RTT, and ICMP-level RTT (NetB only).

Throughput measurements were not conducted while using the *WiRover* system, as they would had affected the network performance experienced by the clients of the WiRover system. Hence, we only collect latency measurements using UDP pings, roughly 12 pings a minute. Details regarding measurement settings for each dataset are summarized in Table 2 and Table 1.

Effect of vehicular mobility on measurements: In our effort to collect measurements from a vast region over sustained durations, we were forced to utilize vehicles traveling at varying speeds. To understand the effects of the vehicular speeds on our data we analyzed the distribution of RTT latency (UDP ping test) as a function of vehicular speed for the zones in our *WiRover* dataset in Figure 2(a).

As can be seen from the plot, there was very limited correlation (correlation coefficient mostly close to zero) between the latency and the vehicle speed. We also plot the CDF of the correlation coefficients which were measured from each zone in Madison and on the path from Madison to Chicago in Figure 2(b). The plot shows that 95% of zones had little correlation (0.16) between the speed of vehicle and latencies observed, for typical vehicle speeds ranging from 0 km/h to 120 km/h. The absence of a correlation between the speeds at which these measurements were collected assures us that our datasets are representative of cellular network performance, which are independent of (typical) vehicle speeds.

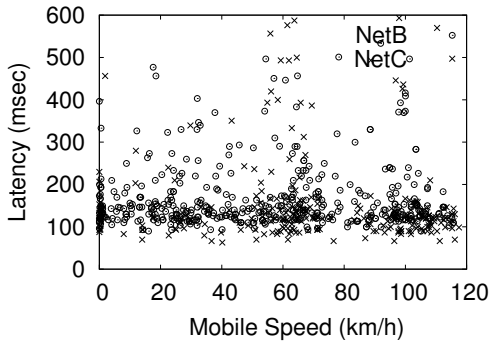
3. DESIGN OF WISCAPE FRAMEWORK

In this section we describe the design of WiScape. Figure 3 summarizes the flow of this section. First, we analyze our *Wide-area* datasets to characterize the performance of cellular networks over a large spatial region to determine if data is aggregatable in space. In Section 3.1 we use these datasets to determine the appropriate size of zones for our measurement framework. In Section 3.2 we use our *Spot* and *Region* datasets comprising of measurements collected at finer time scales to analyze the performance variations of the three cellular networks at fine-grained and coarse-grained time scales at multiple locations. In Section 3.3.1 we determine the number of measurement samples necessary to determine the bandwidth at a zone with certain degree of accuracy. Then in Section 3.2 we determine the frequency with which the measurements should be repeated. Finally, in Section 3.3 we analyze our *Region* dataset to ascertain the feasibility of carrying out client-sourced, coarse grained performance estimation for cellular networks, involving multiple clients.

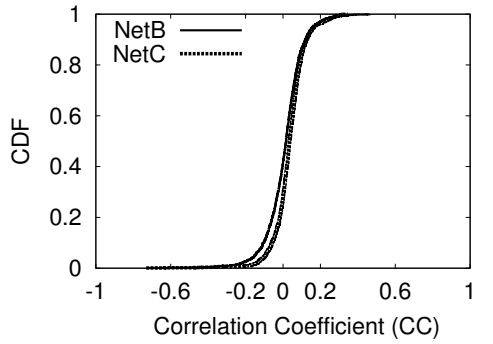
3.1 Aggregation in space (zones)

As it is not feasible to blanket monitor a entire wide-area wireless network we must aggregate data into spatial zones. We desire zone sizes which are small enough to ensure similar performance at all locations inside the zone but big enough to ensure enough measurement samples can be collected for each zone to properly characterize the network's performance. For this purpose we analyze the variation of TCP bandwidth measurements for NetB collected in *Standalone* dataset across city locations by dividing them into circular zones of radius varying from 50 to 750 meters in steps of 100 meters. We have not experimented with other shapes of zones.

In Figure 4 we plot the CDF of relative standard deviation (standard deviation of samples/mean of samples) for all zones, for which we have at least 200 samples per week over the duration of the measurement study. The left most



(a) *WiRover: Vehicle Speed (vs) Network latency*



(b) *WiRover: CDF of Correlation Coefficient*

Figure 2: Latency is weakly correlated with typical vehicular speeds. In Figure 2(a), the latencies are mostly around 120 msec, with no observable trend with increasing speeds. In Figure 2(b), the CDF of correlation coefficient between latency and vehicular speed is less than 0.16 in 95% of zones.

curve corresponds to zone size of 50 meters while the right most curve corresponds to zone size of 750 meters. Furthermore, the relative standard deviation of for 80% of the zones is around 2.5% for zones with radius of 50 meters and 7% for zones with radius of 750 meters[2]. The increase can be explained by the change in terrain conditions across bigger zones. As can be seen from the plot, despite increasing zone radius the relative standard deviation tends to vary only slightly. We pick a zone radius of 250 meters as 80% of the zones with 250 meter radius have relative standard deviation less than 4% and 97% of zones have a relative standard deviation of 8% or lower. The low relative standard deviation, implies that the characteristics of locations inside the zone are mostly similar.

[2]In Figure 1, some zones have a relative standard deviation greater than 0.3 (mean = 1080 Kbps, dev = 350 Kbs). These zones in Figure 1 correspond to regions with very few samples (less than 200 hundred samples) and hence are not considered while plotting Figure 4.

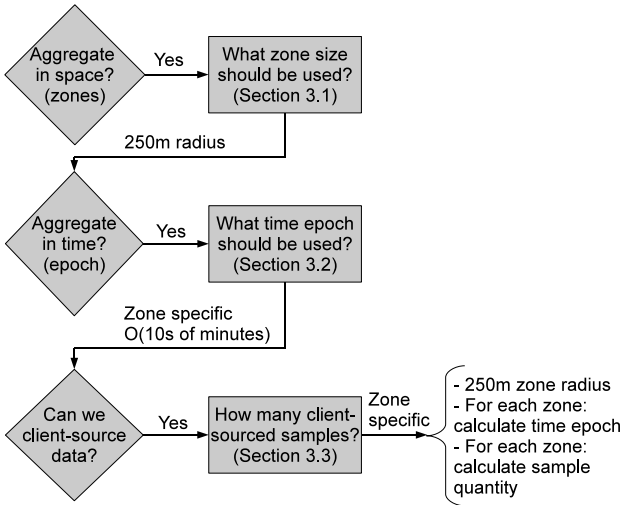


Figure 3: The flow of text in Section 3, describing the design choices made in WiScape.

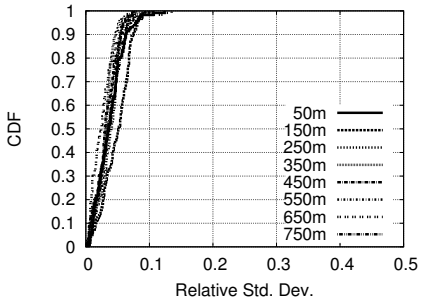


Figure 4: CDF of relative standard deviation of TCP throughput across a cross-section of the city with NetB as a function of increasing zone radius, only zones with more than 200 data-points have been considered.

We find that the TCP throughput does not vary significantly for the cellular network. Specifically, we note that 80% of zones have a relative standard deviation between 2% and 8% regardless of zone size. Moreover, less than 2% of the zones have a standard deviation of 15% or higher. Based on the above observation we selected *representative* zones with overall performance variability for NetB that was between 2% and 8% and zones with TCP throughput variability of the other two networks that was less than 15%. These *representative* zones are used for our *Spot* data collection, as seen in Section 3.2.

We also examined data from WiFi-based networks as reported by others (GoogleWiFi [15], RoofNet [16] and us (MadCity Broadband [17]) in prior work on how throughput measurements for cellular networks might compare to that of such WiFi-based networks. Such prior work report high and sudden variations in achievable throughputs in the WiFi networks, often due to the use of unlicensed spectrum, random access nature, and the characteristic of the spectrum itself. This is contrast to the more coordinated access meth-

ods and the licensed nature of the cellular spectrum that provides some performance stability across epochs as defined above. Hence, epochs in WiFi system are likely more difficult to define than compared to these cellular systems. The low degree of variability in cellular performance is the motivation for exploring the feasibility of estimating cellular data network performance using a small number of measurements.

A closer look: To understand the stability of measurements within individual zones, we use the *Static* and *Proximate* datasets. As noted in Section 2, data for our *Proximate* dataset was collected by driving around in a car within a 250 meter radius from corresponding *Static* dataset locations. The *Proximate* dataset, provides us with network performance measurements from multiple locations in close vicinity of the locations in *Static*. The measurements in *Proximate* dataset are, thus, representative of the kind of measurements we can expect to gather for a given zone from a set of clients in a real deployment of WiScape system. The data for *Proximate* dataset was collected for each zone over a span of 5 months in Madison and 1 month in New Brunswick. In the rest of this section, we present results for a single zone from Madison and one in New Brunswick and omit the results for the remaining five static locations. We examine how the average throughput measured from *Static* subset relates to the throughput measurements from the corresponding *Proximate* measurements.

We present the average and standard deviation for the *Static* and corresponding *Proximate* measurements in Table 3. From the table we note that the client sourced measurements form a reasonable approximation of the expected performance at a given location.

We observe that the average UDP throughput of NetB-WI for the ground truth and the client sourced UDP traces are 876 Kbps and 855 Kbps respectively, where the percentage of error is less than 1%. The observation holds true even in case of representative zones from New Brunswick which has higher degree of performance variation compared to zones in Madison.

The jitter values reported in the *Proximate* dataset are also close to 7 msec for NetA-WI which matches the corresponding *Static* dataset jitter value shown in Table 3. Similarly, the jitter for NetB-WI and NetC-WI are around 3 msec in the *Proximate* dataset which again matches the *Static* jitter value of the two networks at the location, as shown in Table 3. We also have noted the same behavior for NetB-NJ and NetC-NJ whose jitter values are 2.8 msec and 1.6 msec respectively. From the above results we find that measurements collected across multiple locations within a zone are close to each other.

Summary: We choose a radius of 250 meters for zones as 97% of such zones in Madison have low (8%) relative standard deviation for TCP throughput for NetB .

3.2 Aggregating in time (zone-specific epochs)

We analyze data from *Spot* dataset to understand the performance of the three cellular networks over different granularities of time. As noted in Table 2, the *Spot* data was collected at five distinct locations in Madison and two locations in New Jersey, for all three networks, to characterize the performance of the cellular networks at a fine granularity. In particular we study coarse (30 minutes) and fine (10 seconds) time scale variations of different performance parameters such as throughput, loss rate etc. and in Section 3.2.2 we explain the mechanism for calculating the epoch duration for the monitored zones.

3.2.1 Performance at different time granularities

We look at *Spot* data measurements to characterize the performance variability of cellular networks. We present data from two representative locations, one in Madison and another in New Brunswick where the relative standard deviation (standard deviation/average) of any of the parameters (TCP and UDP throughput, Jitter, Loss rate) was less than 0.15, for the entire monitored duration. The highest relative standard deviation of 0.15 was observed for TCP throughput at both locations. We observed similar properties for the other four measured locations in Wisconsin and one other location in New Jersey, but do not present them in this paper for the sake of brevity.

Coarse time scale: We present the average throughput, jitter, and error rates, averaged in 30 minute bins collected in Madison and New Brunswick in Figure 5(a,b,c,d) and 5(e,f,g,h) respectively. As can be seen from Figure 5, for the selected location in Madison, the NetA network on an average offers throughput benefit greater than 50% for both TCP and UDP over the worst performing network. We also find that the variance in throughput across all three networks over the entire duration is less than 0.15 of their long term average. Moreover, all three networks have a packet loss rate less than 1% with a very low variation (Figure 5(d)). We find from Figure 5(c) that the jitter is around 3 msec for NetB and NetC networks while it is around 7 msec for the NetA network.

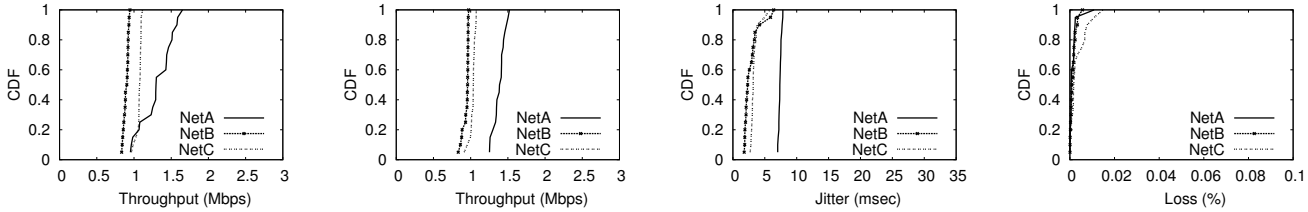
For the location in New Brunswick, looking at Figure 5(e,f), we find that the TCP and UDP throughput for NetB and NetC has higher variability than the location in Madison. Although the overall variation is still lower than 0.15. Akin to the location in Madison, both networks have low jitter (less than 3 msec) and packet loss (less than 1%).

Fine time scale: In Table 4, we present the standard deviation for throughput, jitter, and loss rate calculated for 10 seconds bins and 30 minute bins for all three networks for both locations, to compare and contrast the network characteristics at fine time scales with coarse time scales.As can be seen from the table, the standard deviations over coarse and fine timescales vary significantly. For example, at the location in Madison, the standard deviation of TCP throughput is 211 Kbps at coarse timescales, whereas it is around 377 Kbps at finer timescales, a difference of 159 (377-211) Kbps. Similar observations can be drawn for other metrics across all the networks. This difference in standard deviation is expected as 30 minutes is a large duration of time which can hide large fluctuations in performance. We can make similar observation for the measurements collected at the location in New Brunswick. The high degree of variation at short time scales effectively rules out the use of small and infrequent measurements to estimate performance.

Finally, given the relatively low overall jitter (less than 10 msec) and no losses in the networks, we desist from presenting further jitter and loss performance results for the sake of brevity.

3.2.2 Calculating zone specific epochs

To determine the zone specific epoch duration, we need to determine the granularity of time over which a given met-

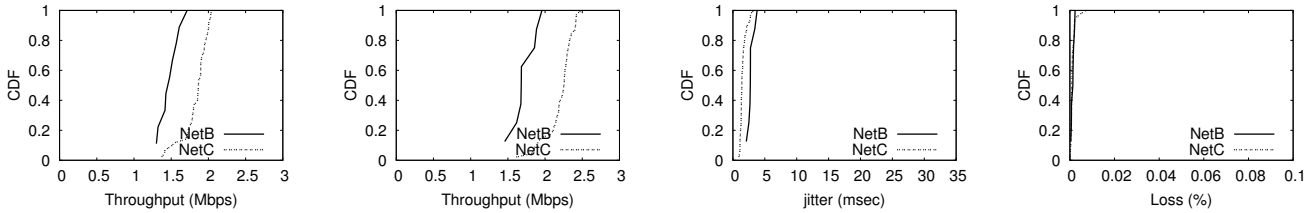


(a) TCP Throughput (in WI)

(b) UDP Throughput (in WI)

(c) Jitter (in WI)

(d) UDP Packet loss (in WI)



(e) TCP Throughput (in NJ)

(f) UDP Throughput (in NJ)

(g) Jitter (in NJ)

(h) UDP Packet loss (in NJ)

Figure 5: CDF of long term (30 min) average data. Plots (a)-(d) correspond to a location in Madison and (e)-(f) correspond to a location in New Brunswick. The variation in throughput across all the networks at two locations is below 15%. Location in New Brunswick shows higher variance in throughput. The average and variation of both jitter and loss are low across all locations.

	NetA-WI		**NetB-WI**		**NetC-WI**		**NetB-NJ**		**NetC-NJ**	
	Static	Proximate	Static	Proximate	Static	Proximate	Static	Proximate	Static	Proximate
TCP (Kbps)	1242 (196)	1266 (180)	845 (63)	827 (82)	1067 (61)	1005 (78)	1494 (222)	1549 (196)	1850 (201)	1869 (159)
UDP (Kbps)	1241 (101)	1257 (135)	867 (67)	855 (89)	1017 (62)	962 (72)	1690 (290)	1748 (248)	2204 (221)	2245 (166)
Jitter (msec)	7.4 (0.4)	8.5 (0.6)	3 (1.6)	5.4 (1.6)	3.4 (1.2)	5.6 (2.4)	2.8 (1.5)	2.8 (0.9)	1.6 (0.9)	1.5 (0.6)
Loss (%)	∼0	∼0	∼0	∼0	∼0	∼0	∼0	∼0	∼0	∼0

Table 3: Table showing the closeness average and standard deviation (in parentheses) of different nearby locations (*Proximate* dataset) from the same zones for each network.

ric is stable. A metric should be estimated for each epoch independently. We use the Allan deviation measure [18] to determine the epoch for which the metric is stable. The Allan deviation is used to calculate the frequency stability of a variable and is defined as the square root of the Allan variance. Allan variance is then defined as the variance between two measurement values formed by the average of the squared differences between successive values of a regularly measured quantity. The sampling period of the measurement also forms a parameter which determines the time granularity at which the Allan deviation is measured. The difference from standard deviation arises from the usage of immediate measurement values to calculate the difference terms, instead, of using the long term mean.

It is mathematically expressed as,

$$\sigma_y(\tau_0) = \sqrt{\frac{\sum_{i=1}^{N-1}(T_{i+1}-T_i)^2}{2(N-1)}}$$

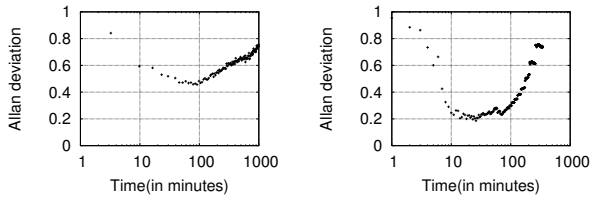
Where, T_i are the averaged measurement values collected at time instance i and N is the total number of available measurement values. A low Allan deviation implies that the current values do not differ much from the previous values. In contrast, large Allan deviation would signify that the coherence of the measured metric is changing.

We present the Allan deviation of UDP throughput at the two zones for the NetB network using the *Proximate* dataset in Figure 6 as an example. In the figure, the x-axis of the plot represents the periodic burst duration while the y-axis represents the corresponding Allan deviation. We find that, for the zone in Madison, Allan deviation becomes the lowest around a time duration of about 75 minutes. This value is higher (mostly greater than 0.5) at both smaller and larger values. For the zone in New Brunswick we find that Allan deviation is lowest around 15 minutes. We pick this minimum value of the Allan deviation is the epoch duration for the corresponding zone. Epochs for other metrics can similarly be determined using the above method.

In WiScape, we collect measurements from clients to get stable estimates in each epoch for a zone, re-starting this process as we move from one epoch to the next. Hence, for the representative zone from Madison, the measurement

	NetA-WI		NetB-WI		NetC-WI		NetB-NJ		NetC-NJ	
	Long (30m)	Short (10s)	Long (30m)	Short (10s)	Long (30m)	Short (10s)	Long (30m)	Short (10s)	Long (30m)	Short (10s)
TCP (Kbps)	211	370	33	102	36	96	126	408	167	414
UDP (Kbps)	77	241	39	82	38	94	153	429	182	365
Jitter (msec)	0.2	0.7	1.3	2.1	0.7	1.6	0.5	1.6	0.5	1.0
Loss (%)	∼0	∼0	∼0	∼0	∼0	∼0	∼0	∼0	∼0	∼0

Table 4: Table showing the standard deviation of long term (30 min) and short term (10 sec) data for each network. The standard deviation of short term data is significantly higher than that of long term data.



(a) Location in WI (b) Location in NJ

Figure 6: Allan deviation for UDP throughput measurements at a given zone for NetB using *Proximate subset* traces. For the measured data, the Allan deviation is lowest around 75 minutes, which corresponds to the epoch of the zone.

process repeats every 75 minutes, while for the zone in New Brunswick it repeats every 15 minutes.

Summary: When aggregated at finer time scales (tens of seconds), the network metrics vary significantly more, than when aggregated at a coarser time scale (tens of minutes). Hence, we use the minimum value of the Allan deviation in each zone to determine the epoch of that zone. This value is estimated regularly for each zone.

3.3 Composability of client sourced measurements

We use client sourcing to collect measurements from different client devices, leading to estimation of network properties for each epoch in each zone. Composability of measurements collected from diverse sources would be feasible only when they are similar (to a certain) extent to one another. In our work, we have only used laptop or single-board computer (SBC) based hardware, each equipped with different cellular modems. This section shows that composability across this class of clients is, indeed, possible. However, composability of measurements from a mobile phone and a laptop equipped with a USB modem may not always work well. This is because a mobile phone, among its other characteristics, has a more constrained radio front-end and antenna system, than a USB modem. Potentially data collected from such devices with different capabilities need to go through a normalization or scaling process. We have not addressed such types of composition in this work. Instead we suggest that we group devices into broad categories — mobile phones, laptops or SBCs with USB or PCMCIA modems, etc., and perform client-assisted monitoring for each individual category separately. Given our experimentation was performed using laptops and SBCs equipped with cellular modems (as this was the platform used in our wide-area data collection efforts for various practical and logistical reasons), we demonstrate that composability within its category. Future work would require us to re-create some of these results with the mobile phone category as well as as examining techniques for normalization across categories, a significant effort unto itself.

To demonstrate the closeness of client sourced samples to stationary data, we evaluate a) if the probability distribution of the measurements collected at the *same location* (same GPS coordinates) by *different clients* at *different times* within the time epoch are statistically similar to the overall long-term distribution at that location and b) if the probability distribution of the measurement samples collected by *different clients* at *different locations* (within a bounded distance) during the *same time* epoch are statistically similar to the overall long-term distribution at that location. While (a) measures the temporal variability, (b) measures the spatial variability of the measurement samples inside a zone.

We measure the similarity of two probability distribution functions, using the symmetric Normalized Kullback-Leibler Divergence (NKLD) between the data from the *Static* dataset and the *Proximate* dataset for a given location. The symmetric NKLD is a measure of the dissimilarity between two distributions.

The Kullback Liebler divergence (KLD) quantifies the relative entropy between two probability distributions which are generated from a common event. The KLD is zero for two identical probability distributions. To rectify the asymmetric nature of the metric we use a symmetric and normalized version of the metric as used in [19]. The normalized symmetric Kullback Leibler metric,

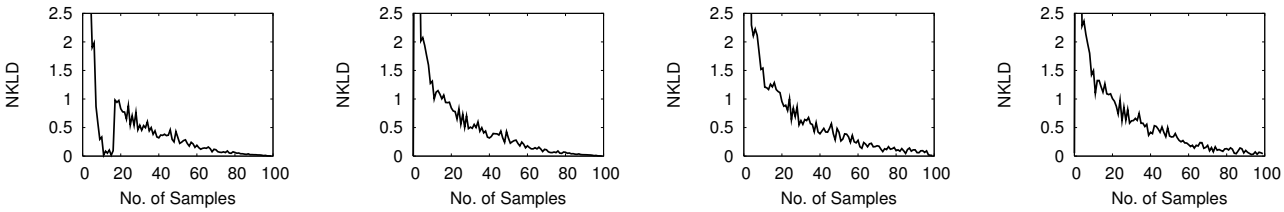
$$NKLD(p(x), q(x)) = \frac{1}{2}\left(\frac{D(p(x)||q(x))}{H(p(x))} + \frac{D(q(x)||p(x))}{H(q(x))}\right)$$

where, $p(x)$ and $q(x)$ are the two probability distributions based on a common set χ.
$H(p(x)) = \sum_{x\in\chi} p(x)log(1/p(x))$ is the entropy of the random variable x, with probability distribution $p(x)$, and,

$$D(p(x)||q(x)) = \sum_{x\in\chi} p(x)|log\frac{p(x)}{q(x)}|$$

is the Kullback-Leibler divergence. A small value of NKLD would signify that the two distributions are "close". For our experiments, we take an NKLD value of 0.1 and lower to signify that the distribution of measurements are similar. We plot the KLD distributions for UDP throughput for the NetB network in Figure 7.

Temporal variability of samples: We randomly select two measurement traces of two clients of progressively in-



(a) NKLD for samples collected at same location (GPS co-ordinates) at different times in WI.

(b) NKLD for samples collected at different locations (within a zone) at the same time in WI.

(c) NKLD for samples collected at same location (GPS co-ordinates) at different times in NJ.

(d) NKLD for samples collected at different locations (within a zone) at the same time in NJ.

Figure 7: Plot of NKLD for UDP throughput (a) and (c) shows that samples collected at temporally different instances at same location are highly similar, (b) and (d) shows data collected at spatially different locations which are in the same zone are highly similar. Plots (a) and (b) corresponds to location in Madison, Wisconsin, while (c) and (d) corresponds to location in New Brunswick, New Jersey.

creasing time durations with the same GPS coordinates and calculate the divergence of this distribution with the overall distribution consisting of all measurements, this process is repeated across 100 iterations and the average of the NKLD is calculated. We plot the results in Figure 7(a) and Figure 7(c). We find that for the location in Madison, by the time we have accumulated 50 to 60 samples, the NKLD goes down to a 0.1 signifying that the two distributions are similar to each other. For the location in New Brunswick, we find that the NKLD goes below 0.1 once we have accumulated 80 to 90 samples. Furthermore, the two distributions become similar once we have gathered around 120 samples.

We need higher number of samples in New Brunswick, due to the greater degree of variation of its performance compared to the network in Madison.

Spatial variability of samples: We randomly select locations which are 50-250 meters apart from each other and simultaneously start UDP downloads using two clients at both locations for a duration of 2 minutes. The test is repeated at 10 different locations. We plot the divergence of the distribution of throughput values collected from two locations in Figure 7(b) for Madison and Figure 7(d) for New Brunswick. We find that with 80 and 100 measurements in Madison and New Brunswick respectively the NKLD is less than 0.1. This signifies that by the time we have accumulated around 100 samples at two locations the distribution of such samples becomes similar to one another in both the representative locations.

Based on above results, we conclude that client sourced measurements can be used as a estimator of the ground truth for a zone.

3.3.1 Example: client sourced throughput estimation

We intend to determine the minimal amount of measurements necessary to estimate the network's performance at a given location with a certain degree of accuracy. In this section, we use throughput estimation as an example. We note that similar methods can be used for client-sourced estimation of other metrics such as jitter, loss and latencies etc..

A lot of research has focused on estimating the available network bandwidth for wired as well as WiFi based networks [20, 21]. In contrast, few studies have concentrated on characterizing the available bandwidth for the cellular networks. Availability of an accurate and efficient estimation algorithm is vital for client-assisted monitoring.

Network-Location	UDP	TCP
NetA-WI	90	60
NetB-WI	60	40
NetC-WI	40	40
NetB-NJ	120	120
NetC-NJ	70	50

Table 5: Table showing the number of back-to-back measurement packets to be sent to estimate TCP/UDP throughput within an accuracy of 97% of the expected value.

We experimented with two such bandwidth measurement tools: Pathload and WBest [20, 21]. To estimate the accuracy of these tools we take the average of UDP throughput measured over 100 seconds for 10 iterations as the *ground truth* at that location. We then define relative error as $E = \frac{X - G_{UDP}}{G_{UDP}} \times 100\%$, where X is the result from available bandwidth measurement tools (i.e., Pathload or WBest) and G_{UDP} is the ground truth UDP throughput. In our evaluations we found that neither of the two tools give an accurate approximation. WBest consistently under-estimates the actual bandwidth by up to 70% while Pathload underestimates up to 40%. Similar benchmarking results are also reported in [22]. Hence, we carry out simple UDP downloads over a duration of time to measure the network performance. In the rest of this section, we determine how many such samples should be sent to fairly accurately ($\sim$97%) estimate the network throughput at a specific location. We intend to diagnose the reason behind the estimation inaccuracies for the two bandwidth measurement tools as part of our future work.

How many packets necessary? We revisit our TCP and UDP throughput measurements from our *Proximate* datasets to determine the minimum number of packets to be collected for attaining a maximum accuracy in estimating the expected performance of a zone.

We select a given number of client collected packets and calculate their average. We then compare it with the ground truth throughput at that instant (calculated as mentioned above). We repeat this process 100 times for a given packet size. We present the number of packets necessary to attain an accuracy of 97% in Table 5. We find that for the

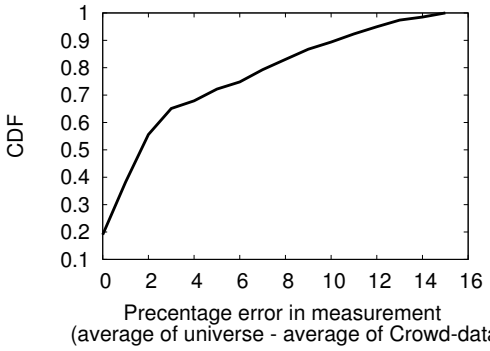


Figure 8: Error of WiScape measurements in comparison to our extensive measurements collected.

zone in Madison, we need 90 TCP packet measurements to obtain an accuracy within 97% of the expected measurement for NetA . From the same table, we can estimate the expected TCP throughput which is within 97% of the expected throughput by collecting as many as 40 back-to-back measurement packets for NetC for both TCP and UDP. The number of packets are marginally higher for NetA as compared to NetB and NetC as the network performance varies more for NetA clients (Figure 5, Table 4).

For the zone in New Brunswick, we find that we need 120 packets for estimating the TCP and UDP performance of NetB network. Whereas, for the NetC network we need to send only 70 UDP and 50 TCP packets back-to-back for a estimation accuracy of 97%. With an expected cellular data-rate in hundreds of Kbps, a client can thus, finish a measurement in less than a second.

Summary: *We validate that network performance estimation using a small number of measurements collected by different clients inside a zone is indeed feasible.* Specifically, we find that for the monitored zone, the distribution of the observed metric becomes almost similar to that of any other client present in the same zone (or from the same client at an earlier time epoch) we have accumulated more than 80 packets.

3.4 Putting it all together

We envision a simple user agent in each client device, e.g., as part of the software in the mobile phones or bundled with drivers of cellular NICs. A measurement coordinator, deployed by the operator or by third-party users, will manage the entire measurement process. Each cellular device periodically reports its coarse-grained zone (based on associated cellular tower) to the measurement coordinator[3]. Based on this zone information, our measurement coordinator periodically provides each mobile device with a measurement task list.

When a mobile device performs a task, it is required to collect more precise zone information at which the task is initiated as well as completed. If the mobile phone has a built-in GPS receiver, it is possible for it to obtain zone information quite easily. However, alternate techniques to obtain zone information include triangulation and fingerprinting based techniques, using the cellular, WiFi, or Bluetooth interfaces [23, 24, 25].

[3]Current cellular systems already collect such zone information from all mobile devices in order to correctly and quickly route calls and traffic to them.

The rate of refreshing the measurements for each zone would depend on the coherence period of that zone as determined by looking at the Allan deviation.

For a given zone, once in every coherence time-period, the measurement coordinator will provide a measurement task to each active mobile client with a probability, chosen such that the number of measurement samples collected over each iteration is sufficient for estimating accurate statistics, as determined by the NKLD algorithm. Once the selected clients report their measurements, the server checks if the measured statistic has changed substantially from its previous update (say by more than twice the standard deviation). In such a situation the server would update its record for the zone with the new value.

Validation: To analyze the accuracy of our WiScape framework, we partitioned our *Standalone* dataset which consists of around 400 zones with 200 or more samples, into two subsets (*Client sourced data* and *Ground truth*). For each zone, we assume that the entire *Ground truth* set provides our expected value (consisting of up to 125,000 packets for various zones). Figure 8 shows the CDF of error in estimation of TCP throughput for the WiScape data from the *Client sourced* dataset and the *Ground truth* data. As can be seen from the plot, WiScape data has less than 4% error in estimating the TCP throughput for more than 70% of the zones. The maximum error in performance measurements is around 15%, which indicates that WiScape is able to determine the necessary measurement parameters for each zone and provide a fairly accurate performance estimate.

Discussion: We note that there is an important trade off between the volume of measurements collected, the ensuing accuracy, and the energy and monetary costs incurred. Our design in WiScape defines one specific design choice in this multi-dimensional space. Many other alternatives are certainly possible and would make for interesting exploration in the future.

4. APPLICATIONS OF WISCAPE

In Section 4.1 we demonstrate how client-assisted monitoring of networks can help discover zones with highly variable network performance. Variability in network performance can be an indicator of possible network problems. Hence, client-assisted monitoring can help network operators short-list zones which need further detailed diagnosis. Finally, in Section 4.2 we characterize the potential performance enhancement for two applications when using coarse grained measurements. Both applications use more than one cellular network.

4.1 Helping operators

To ensure that the network performance at the deployed regions is above a certain quality, the cellular service providers carry out periodic drive-by tests to validate the performance of their network. This involves using a car equipped with network measurement equipment, and then carrying out network performance tests at specific locations. However, such tests are labor intensive and hence not scalable for wide area wireless networks. Client-assisted monitoring can help network operators in this regards by pin-pointing zones with performance characteristics significantly different than neighboring zones.

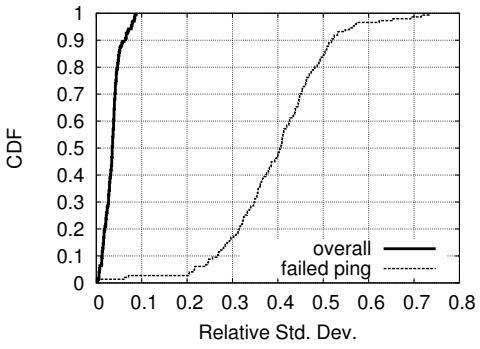


Figure 9: CDF of Relative standard deviation (standard deviation/average) of TCP throughput for all zones (with 250 meter radius) and those with more than 20 days with at least one ping failure.

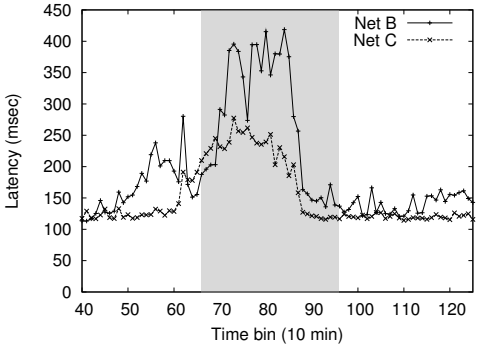


Figure 10: Network latency (averaged over 10 minutes) during a football game. The scheduled time of the game is covered by the shaded region.

Identifying locations with variable performance

Let us assume the network operator intends to determine potential locations with highly variable throughput (say relative standard deviation greater than 20%). This information would be difficult to deduce from a relatively low number of client sourced measurements because of the fact that the accuracy of client sourcing depends on low variability in network performance. We note that while small throughput tests conducted infrequently every tens of minutes might miss out on zones with highly variable performance, other infrequently calculated metrics may be used to detect such variability. To highlight such a metric, we revisit our *Standalone* dataset. As mentioned in Section 2, we present data for ICMP ping tests in our *Standalone* dataset. From this dataset, we first determine zones, with radius 250 meters, that have multiple ping test failures. In Figure 9 we present the CDF of relative standard deviation of all the zones with more than 200 measurements and those zones with at least one failed ping tests every day, for a period of 20 consecutive days or more. As can be seen from the plot, zones with 20 or more consecutive days with at least one ping failure have a very high variation in their relative deviation of TCP throughput. For example, 65% of the links have a relative deviation of the order of 40%. We also find that zones with back-to-back ping failures constitute 97% of the zones with relative standard deviation above 20%. This is in contrast with the majority of other zones which have less than 1% relative standard deviation.

Identifying locations for additional provisioning

Coarse grained estimates can also help network operators determine places where additional resources might be needed to satisfy periodic surge in demands. For example, Figure 10, shows the network latency of two cellular networks near a football stadium (80,000 seating capacity) during a football game. The shaded region in the plot represents the scheduled time of the football game. As can be seen from the plot, for the duration of the game the average ping latencies go up from 113msec to 418msec, an increase of the order of 3.7X for NetB. As the duration is in order of 100s of minutes, infrequent periodic monitoring can detect the above event and help operators take corrective measures.

4.2 Improving client performance

To show the potential benefits of a WiScape like system for clients, in Section 4.2.1 we show that for a large number of zones the performance of one cellular network is persistently better than other networks over large duration of time, and hence observable using infrequent measurements. In Section 4.2.2 we explain how such information can be utilized by clients with multiple cellular connections to choose the best network of operation for each zone.

4.2.1 Persistent network dominance

We intend to understand if the relative performance characteristics of different cellular networks are persistent over large periods of time (for each zone). For this purpose, we define *persistent network dominance* as follows: when the lower 5 percentile of the best network's metric is better than the upper 95 percentile of other networks in a given zone, we say the zone is *persistently dominated* by the best network. The fact that the lowest 5 percentile of performance of the dominant network is better than the 95 percentile of the other networks implies that the dominance is persistent over time and hence observable using infrequent measurements made by a WiScape like system. In Figure 11, we present the percentage of zones with a persistently dominant network, in terms of RTT latency collected from the *WiRover* dataset, as a function of the zone size. As we see, persistent network dominance is observed in 85% of the zones and across different zone sizes. The consistently better performance of one network at a given zone can be explained by observing that the network performance is dependent on the base-station location, technology, and traffic load on the base-station; a combination which would be expected to vary across different network operators.

We use measurement data from our *Short segment* dataset to further investigate the presence of persistent network dominance. The measurements were collected with our vehicle driving across this stretch of roadway regularly for a period of 5 months, at average speeds of 55 km/h. We show a part (10 km) of the road stretch in Figure 12. Each circle corresponds to a zone of 250 meters radius and the shade on the circle corresponds to the network which performs best in that zone.

We plot average TCP throughput performance of NetA, NetB, and NetC networks for each zone over entire experiment duration in Figure 13. In conformance with our observations of persistent dominance in terms of latencies, we

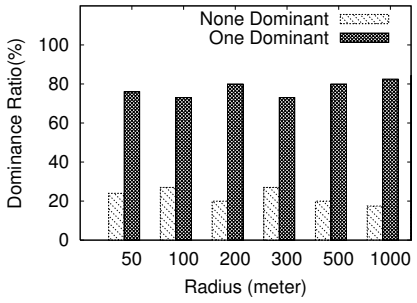


Figure 11: Most of the zones are persistently dominated in terms of network latency, by either NetB or NetC regardless of the zone size.

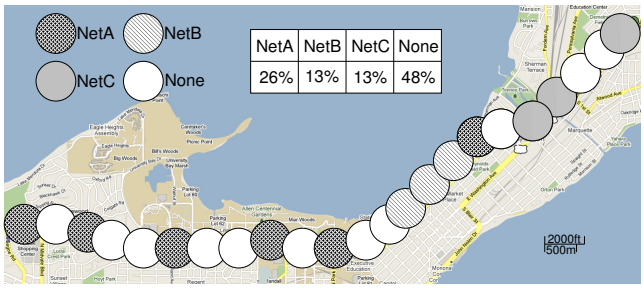


Figure 12: Map depicting the Short segment dataset and the dominant network for each zone. Over all zones we observe that 52% of zones have a dominant network.

find that for a significant number of zones a specific network offers better performance on an average than the other two networks. For example, at zone 20 (as marked in the x-axis) the performance difference between the best network gives 42% higher throughput than the next best network over the entire measurement set. Similarly, the performance at zone 4 of the best network is almost 30% higher than others. We also find that multiple zones exist where none of the networks give clear performance advantage for the entire set of measurements. We identify zones where the lower 5 percentile of the best performing network is better than the upper 95 percentile of other two networks. The inset table in Figure 12 shows the number of zones where one network dominates other networks. From the table we note that there are 52% of zones where one network gives better performance than other consistently over the measurement period. We color the zones in Figure 12 based on which network dominates it. A white color indicates a lack of a persistently dominant network.

4.2.2 *Application performance improvement*

We present two application scenarios which can benefit from approximate network quality estimates for a specific location. The first is a client equipped with a mobile phone that has two or more SIM cards and hence can connect to any *one* of two or more alternate cellular networks at a given point in time. We call this the *multi-sim* application. Such phones are cheaply available in the market today, e.g., Samsung Fizz [26] and Jinpeng S3288 [27], and are gaining in popularity in developing countries like India and China. In

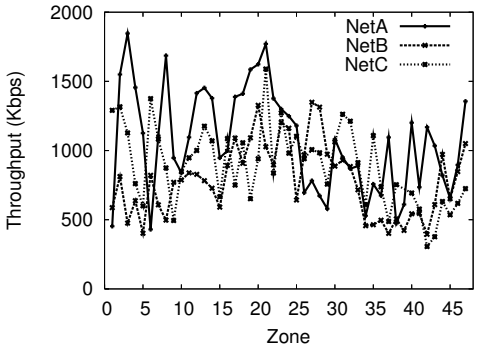


Figure 13: TCP performance on each zone across 20 km stretch of road for three monitored networks.

absence of any knowledge of which network gives the best performance at the current location, the clients with such phones would be forced either to select a network in a random fashion or to carry out measurements to ascertain the network quality for all the networks.

The second application that can benefit from location specific information is a MAR [4] — a multi-network Internet gateway mounted on a vehicle that can aggregate bandwidth from all networks that it simultaneously connects to. The scheduler in MAR stripes traffic flows to different network interfaces. While authors in [4] suggest using location specific network performance information to further optimize performance by intelligently mapping data requests to interfaces based on locality of operation, we highlight the benefits of such a scheme over a simple multi-interface striping algorithm where all currently active requests from different clients are mapped onto different cellular networks in a round robin fashion.

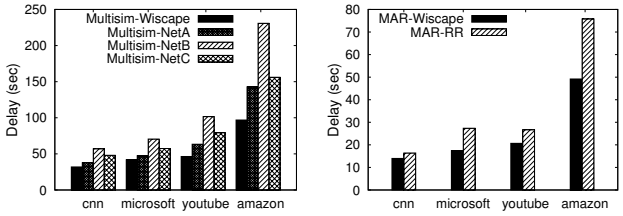
To illustrate the benefits of coarse throughput estimates for the above two applications, we consider the following experiment scenario. A client (either a MAR gateway or a Multi-sim phone) places back-to-back requests for a set of pages from the Internet while driving on a road stretch depicted in Figure 12. In our experiments, the client requested pages from a webserver hosting a pool of 1000 wep pages with sizes between 2.8 KBytes and 3.2 MBytes, generated using SURGE [28]. Akin to [4] we also experiment with popular Web sites by downloading wepages to a depth of 1 from their starting page. For our experiments we run the car on the same road-segment (Table 2) multiple times during the experiment. We compare performance between a system where data is requested in a round robin fashion on each network. The other system with a monitoring agent uses the GPS to determine the location of the vehicle and based on zone information selects the best network to minimize download latency.

Multi-sim Improvements: We present the results in terms of HTTP latency averaged over ten runs in Table 6. As can be seen from the Table 6, we can decrease the HTTP latency by 30% by selecting best performed interface at a given location. We show the HTTP latency for well known Web pages in Figure 14(a). As can be seen from the plot, our scheme gives the maximum improvement for amazon.com webpage (32% improvement) and minimum improvement for microsoft.com webpage (13% improvement).

MAR performance improvements: Here we compare the download latency for the two schemes. We measured

	Avg.(in sec)	Std.(in sec)
WiScape	87.66	8.33
NetA	124.26	14.90
NetB	158.55	33.69
NetC	145.46	14.89
MAR-WiScape	25.72	3.48
MAR-RR	36.8	6.44

Table 6: Average latency and standard deviation for downloading 1000 HTTP files. We can improve HTTP latency by 30% using Multi-sim-WiScape.



(a) Multi-sim improvement (b) MAR improvement

Figure 14: Multi-sim and MAR latency improvements with well known Web pages over round-robin (MAR).

the HTTP latency by running our car with a MAR client with 3 interfaces on a 2.4 Km segment of road (from zone 10 to 15 from Figure 13). We measure the performance of MAR system using network performance information and while mapping client requests to interfaces in a throughput-weighted Round Robin fashion (MAR-RR). As can be seen from Table 6 (last two columns) by using the information provided by WiScape we can decrease the HTTP latency by 32% compared to MAR-RR.

We have also experimented with well known Web pages as described above. We present results in Figure 14(b). As can be seen from the plot, using a locality aware scheme can improve performance by 37% over a naive round robin scheme.

We note that the applications can also estimate network performance through explicit measurements of its own, which would result in a steady measurement overhead on all the network interfaces. Besides, applications like MAR would also need to stop all client traffic thus potentially hampering performance. In contrast a client-sourcing base approach would gather this data ahead of time and can simply make it available to potential clients, at a low overhead. We would also like to note, that we did not account for multiple system level issues, such as energy efficiency, time to switch between links, or presence of client think time into account while calculating the performance. Accounting for such issues might lead to changes in achievable benefits that we present. However, we believe that independent of the specific metric being optimized (energy or completion time) information about link performance can always be leveraged for better performance.

5. RELATED WORK

We compare and contrast our contributions in WiScape with prior work on two separate fronts.

Prior monitoring research: With the rapid growth of cellular based Internet connectivity, cellular providers and third-party developers have started developing client-based techniques to learn about the properties of these networks. They include the AT&T's "Mark the Spot" [29] application, 3gtest [9], and applications by Root Wireless [8]. Unlike these applications, WiScape focuses on a measurement methodology for client-sourcing that systematically reduces the number of measurements required across time periods and zones based on data already collected while ensuring that collected data is statistically useful.

Other recent work has conducted detailed measurements of specific 3G cellular networks to understand their performance for both static and mobile environments [30, 31]. Lie *et al.* presents the characterization of PHY and MAC layer of 3G network and its impact on TCP performance [30]. Akin to [30] where the authors find the DRC (dependent on SINR) to vary significantly over large time scales, we also found a high variation in RSSI over the period of a day. Similar to [30], we did not find any correlation (0.03) between the expected application level TCP throughput and RSSI. In light of the above observation we discarded RSSI statistics from further consideration.

Similar studies have also been conducted on outdoor WiFi mesh networks [32, 15]. Again, such prior works are primarily measurement studies and do not focus on our focus of a client-sourced measurement framework with goals of minimal data collection from diverse clients.

Related applications: The novelty of our work is that we collected long term city-scale data and built the WiScape framework which harnesses the performance measurement of 3G network to maximize the performance of multi-network applications, e.g., MAR [4]. Many other vehicular networking systems have been designed and deployed in recent years, each with different target applications. Examples include VanLAN [33, 34], a WiFi based Internet service into vehicles [33, 34], PluriBus [35, 36] a WiFi, 3G, and WiMAX based system with similar goals but with different algorithms, DieselNet [37, 38] that mostly focused on delay tolerant networking and opportunistic Internet services. We believe that many of these systems could potentially leverage client-based data collected by WiScape to better optimize their data striping algorithms (analogous to our design of improvements to MAR).

6. CONCLUSION

In this paper we presented the design for a client-assisted network monitoring system. Through extensive measurement over a period of more than one year, in and around Madison and small parts of New Jersey, we have validated the possibility of carrying out client assisted network monitoring. With experimentation we show how client-assisted network monitoring can help cellular network users and operators. We believe this work is merely a starting point in larger scale measurements and network monitoring, spanning multiple cities, state, or across the whole country.

Cellular data traffic volume is set to increase dramatically in near future, placing enormous load on the infrastructure. Moving forward, we intend to expand the spatial and temporal reach of our client-assisted cellular data network monitoring, with the goal of understanding the effects of increased cellular networks on performance.

We hope to organically grow our efforts in the months and years to come. Specifically, we intend to extend our study

to bigger cities, where high number of users, would present a more challenging monitoring problem.

To deploy our ideas developed in WiScape, we plan to integrate our proposed sampling techniques into a publicly available cellular network based measurement and monitoring tool called Network Test [39], available for both the Android and iPhone platforms currently. We believe this would further enrich our understanding of the client-sourced network measurement process.

7. ACKNOWLEDGEMENTS

We would like to thank Shan-Hsiang Shen, Lance Hartung, Hoewook Chung for their help in collecting measurements. We thank the Madison Metro Transit and the Van Galder Bus Company, for letting us use their buses to collect measurements. Sayandeep Sen, Jongwon Yoon, Joshua Hare, Justin Ormont, and Suman Banerjee have been supported in part by the US National Science Foundation through awards CNS- 1059306, CNS-0855201, CNS-0747177, CNS-0916955, CNS-1040648, and CNS-1064944.

8. REFERENCES

[1] C. Fraleigh, S. Moon, B. Lyles, C. Cotton, M. Khan, D. Moll, R. Rockell, T. Seely, and C. Diot. Packet-level traffic measurements from the sprint ip backbone. In *IEEE Network*, 2003.

[2] D. Kotz and Essian K. Analysis of a campus-wide wireless network. In *MobiCom*, 2002.

[3] Yu-Chung Cheng, John Bellardo, Péter Benkö, Alex C. Snoeren, Geoffrey M. Voelker, and Stefan Savage. Jigsaw: solving the puzzle of enterprise 802.11 analysis. In *SIGCOMM*, 2006.

[4] P. Rodriguez, R. Chakravorty, J. Chesterfield, I. Pratt, and S. Banerjee. Mar: A commuter router infrastructure for the mobile internet. In *MobiSYS*, 2004.

[5] B. Cankaya, V. Watcon, and Metters D. Wireless cell site finder and customer service system. In *United States Patent 7236767*, 2007.

[6] Aman Kansal, Michel Goraczko, and Feng Zhao. Building a sensor network of mobile phones. In *IPSN*, 2007.

[7] Emiliano Miluzzo, Nicholas D. Lane, Kristóf Fodor, Ronald Peterson, Hong Lu, Mirco Musolesi, Shane B. Eisenman, Xiao Zheng, and Andrew T. Campbell. Sensing meets mobile social networks: the design, implementation and evaluation of the cenceme application. In *SenSys*, 2008.

[8] Root wireless inc. http://www.rootwireless.com.

[9] Smart phone 3g test. http://www.eecs.umich.edu/3gtest/.

[10] Community resource for archiving wireless data at dartmouth. http://crawdad.cs.dartmouth.edu/.

[11] High-speed packet access. http://www.3gamericas.org.

[12] N. Bhushan, C. Lott, P. Black, Attar R., Y. Jou, M. Fan, Ghosh D., and Au J. Cdma2000 1xev-do revision a: A physical layer and mac layer overview. In *IEEE Comm. Magazine, 44(2)*, 2006.

[13] Wirover. http://host.madison.com/wsj/news/local/article_e1a67156-dea9-11df-9332-001cc4c002e0.html.

[14] C. Demichelis and P. Chimento. Ip packet delay variation metric for ip performance metrics (ippm). RFC 3393, IETF.

[15] Mikhail Afanasyev, Tsuwei Chen, Geoffrey M. Voelker, and Alex C. Snoeren. Analysis of a mixed-use urban wifi network: when metropolitan becomes neapolitan. In *IMC '08*.

[16] John Bicket, Daniel Aguayo, Sanjit Biswas, and Robert Morris. Architecture and evaluation of an unplanned 802.11b mesh network. In *ACM Mobicom*, 2005.

[17] Vladimir Brik, Shravan Rayanchu, Sharad Saha, Sayandeep Sen, Vivek Shrivastava, and Suman Banerjee. A measurement study of a commercial-grade urban wifi mesh. In *IMC*, 2008.

[18] D.W. Allan. Time and frequency characterization ,estimation and prediction of precision clocks and oscillators. In *IEEE Transactions*, 1987.

[19] V. Shrivastava, D. Agrawal, A. Mishra, S. Banerjee, and T. Nadeem. Understanding the limitations of transmit power control for indoor wlans. In *IMC*, 2007.

[20] Manish Jain and Constantinos Dovrolis. Pathload: A measurement tool for end-to-end available bandwidth. In *PAM Workshop*, 2002.

[21] Mingzhe Li, Mark Claypool, and Robert E. Kinicki. Wbest: A bandwidth estimation tool for ieee 802.11 wireless networks. In *LCN*, pages 374–381. IEEE, 2008.

[22] D. Koutsonikolas and Y. Charlie Hu. On the feasibility of bandwidth estimation in 1x evdo networks. In *MICNET*, 2009.

[23] P. Bahl and V. N. Padmanabhan. Enhancements to the radar user location and tracking system. In *INFOCOM*, 2000.

[24] Moustafa Youssef and Ashok Agrawala. The horus wlan location determination system. In *MobiSys*, 2005.

[25] A. Haeberlen, Flannery E., Ladd A., A. Rudys, D. Wallach, and L. Kavrakt. Practical robust localization over large-scale 802.11 wireless networks. In *MobiCom*, 2004.

[26] Samsung fizz phone. http://www.gsmarena.com/samsung_c5212-2709.php.

[27] Jinpeng s4100 mobile phone. http://www.chinatronic.com/products.php/S4100.

[28] Paul Barford and Mark Crovella. Generating representative web workloads for network and server performance evaluation. In *ACM SigMetrics*, 1998.

[29] At&t news release. http://www.att.com/gen/press-room?pid=4800&cdvn=news&newsarticleid=30336&mapcode=.

[30] Xin Liu, Ashwin Sridharan, Sridhar Machiraju, Mukund Seshadri, and Hui Zang. Experiences in a 3g network: interplay between the wireless channel and applications. In *MobiCom*, 2008.

[31] Wee Lum Tan, Fung Lam, and Wing Cheong Lau. An empirical study on the capacity and performance of 3g networks. *IEEE Transactions on Mobile Computing*, 7(6):737–750, 2008.

[32] J. Robinson and E.W. Knightly. A performance study of deployment factors in wireless mesh networks. 2007.

[33] Aruna Balasubramanian, Ratul Mahajan, Arun Venkataramani, Brian Neil Levine, and John Zahorjan. Interactive wifi connectivity for moving vehicles. In *SIGCOMM*, 2008.

[34] Ratul Mahajan, John Zahorjan, and Brian Zill. Understanding wifi-based connectivity from moving vehicles. In *IMC*, 2007.

[35] Mahajan, Ratul and Padhye, Jitendra and Agarwal, Sharad and Zil, Brian. E pluribus unum: High performance connectivity on buses. Technical Report MSR-TR-2008-147, Microsoft Research TechReport, 2008.

[36] Aruna Balasubramanian, Ratul Mahajan, and Arun Venkataramani. Augmenting mobile 3g using wifi. In *MobiSys*, 2010.

[37] Aruna Balasubramanian, Brian Neil Levine, and Arun Venkataramani. Enhancing interactive web applications in hybrid networks. In *MobiCom*, 2008.

[38] Xiaolan Zhang, Jim Kurose, Brian Neil Levine, Don Towsley, and Honggang Zhang. Study of a bus-based disruption-tolerant network: mobility modeling and impact on routing. In *MobiCom*, 2007.

[39] Network test. http://networktest.org/.

Summary Review Documentation for

"Can They Hear Me Now?: A Case for a Client-Assisted Approach to Monitoring Wide-Area Wireless Networks"

Authors: S. Sen, J. Yoon, J. Hare, J. Ormont, S. Banerjee

Reviewer #1

Strengths: Lengthy explanation of the choices of the parameters (size of zones, epochs and number of samples).
Large set of measurements over space, time and wireless networks.
Evaluate the benefits of a multi-network strategy for data connectivity.

Weaknesses: The framework idea is expected, so a key contribution is the choice of parameters. However since it is mainly based on data collected in once city, it is not clear how reusable the parameters for over environments like countryside or dense urban areas like NYC.

Comments to Authors: It is great to see such attention spent on choosing the right parameters for your system. I do not have many comments as the paper was well written and the idea was clear.

Since you collect data from buses, I was wondering if it would not introduce some bias in your data set. For instance, data in one location often collected around the same time.

Aggregation in space:

It seems that you have done your study by aggregating over time instead of keeping time constant.
How would the standard deviation look like if you would compare different zones during the same time of the day?
It would be nice to get similar results from other environments like Manhattan or the countryside.
Did you evaluate the incremental benefits of having zones of different sizes or different shapes (e.g., clustering)?
RTT is an important metric for wireless environments. Why is it not part of the zone size evaluation?

I find your last section on multi-Sim, MAR, network dominance interesting. You might want to develop that aspects in future papers.

Reviewer #2

Strengths: The paper has rich data: 3 cellular network performance measured by clients at several different locations. The use of client-side measurements for network monitoring makes a lot of sense.

Weaknesses: The paper mostly focuses on aggregation, but aggregation would miss interesting temporal and spatial dynamics.

Comments to Authors: I enjoy reading your paper. The goal of understanding the effectiveness of using client-side measurement to quantify the wide-area wireless network performance makes a lot of sense. Understanding the level of aggregation both in time and in space is useful. On the other hand, using aggregation to find the time interval and zones to smooth out the data is simple, and the paper should shorten this part, and spend more time on the more interesting and less obvious part on the applications. Moreover, aggregation filters out interesting variation across time and space, which is equally interesting to the aggregation results if not more and is more challenging technically than aggregation. I would like to see some discussion on this part.

The authors promise to publish their datasets, which would be very useful to the community.

Reviewer #3

Strengths: - detailed measurement study which is well thought out
- empirically informed design
- potential impact of the data.

Weaknesses: - tedious to read, though its not this particular paper's fault
- not clear how actionable the results are except for long term network planning and fault detection.

Comments to Authors: My main comment is on how this data can be acted on. For example, it is not clear if a user observed poor performance, how this data can help him improve or even diagnose his performance problems. While it is nice to have a sense of the network performance at long time scales, it is not clear how useful it is to the end user.

Can we isolate uplink and downlink performance? It would be useful since cellular links are notoriously asymmetric.

How are measurements impacted by client parameters such as battery levels etc? I guess since you are using laptops this is not a concern, but with smart phones battery dictates uplink and downlink behavior which would bias your measurements.

Reviewer #4

Strengths: Extensive measurements have been collected, from a variety of situations. The framework has some practical applications. The authors intend to share their measurements.

Weaknesses: The active measurement approach seems less ideal than a passive one for gathering the information for network providers. The presentation needs quite a bit of improvement. In

particular, it is often difficult to follow which of the variety of data sets is being used. The paper does not consider smart phones, and in general does not consider the various practical issues for deploying this in practice.

Comments to Authors: A more compelling motivation is needed as to why network operators would want to use this framework. I agree that these operators would like more information on where problems are occurring in their network, for the various reasons that you describe in the paper. What I do not agree with is the need to collect (only) active measurements to get that information. Most of the information shown in this paper could be gathered with passive measurements, which have numerous benefits: no overhead on the wireless infrastructure, they are measurements of actual user traffic, they would not require instrumenting the user devices, etc. Only in cases where the user could not connect to the network might active measurements be useful.

I agree that users could take advantage of this framework, e.g., to compare the performance of multiple network providers in a given location. However, that seems to have limited applicability, at least for cell phone users, the majority of whom will have a contract with one provider (even though some will have multiple SIM cards, as stated in your paper).

Since access to the passive measurements may not be available to you, the active measurements that you have collected could fill in. However, the paper currently does not touch on the issue of the benefits of passive measurements for the provider, which is why my rating is lower than it might otherwise have been.

Minor Comments

Section 2
- In the discussion of the WiRover data set, the throughput tests could affect other users in all of the wireless environments, not just the tests done on buses.
- Please clarify where the various tests were run between (i.e., where is the server(s) that is used in each test located?)

Section 3.1: In "A closer look", please explain the intuition behind "driving around in a car within a 250 meter radius"; i.e., people may sit at a location that is within a short distance of an access point, but not too many will drive around in a circle to stay in range.

Section 3.2: It would help to have some more insight on what test duration will provide a reasonable throughput value with low enough variance for the tasks that a user may care about; e.g., someone trying to decide between using provider A and provider B is not going to run two 30 minute tests; they might run two 30 second tests, if that is sufficiently long to make a confident decision
Section 3.2.2: A time duration of 75 minutes seems rather long for a throughput test that is supposed to have "low overhead"

Section 3.3: In the discussion of UDP downloads (or earlier), clarify how many clients you have at your disposal

Section 3.3.1:
- Can't a provider monitor available bandwidth at each access point, rather than run active tests? (i.e., is there any technical reason they could not do this, even if they are not doing so today?)
- While UDP downloads will obtain useful information, it seems like a heavyweight way to do so. why (generally speaking) can providers not obtain this information passively?
- In "Summary", why would there be any doubt that one could use clients under your direct control to estimate network performance?

Section 3.4: clarify what is meant by "sufficient collect accurate statistics"

Section 4: in the caption of Figure 9, shouldn't "200 meter radius" be "250 meter radius"?

Section 4.1
- Clarify why throughput tests "cannot detect zones with highly variable performance"; it seems quite plausible, depending on how you conduct the tests and measure the results
- Note that passive measurements by the provider could also determine this information.

Section 4.2.1: Clarify which is the "best network's metric"

Section 4.2.2
- While there are certainly users who use multiple SIM cards, it seems like it will be a minority of users, due primarily to cost; thus, I do not find the arguments here particularly compelling
- SURGE uses a synthetic workload, does it not? Are you testing against a Web server running these synthetic pages? (If so, please clarify)

Section 5: if you consider my comments on passive measurements, then there will obviously be a bunch of new material to consider here

Reviewer #5

Strengths: Real data, geographically dispersed, over a year. Data submitted to CRAWDAD.

Weaknesses: Most of the hard systems problems were not confronted, even the stated challenge of scale. It is not clear how typical their data is with respect to line-of-sight problems.

Comments to Authors: There are really two parts to this paper: the measurement collection system and the results of the measurements themselves. The measurement collection system seems to ignore all of the hard parts of the problem - including those brought up and stated by the authors. The measurements themselves may be of independent use for others, particularly because they are being made publicly available via CRAWDAD.

For the system, you state in the intro that the core technical challenges are related to scaling, but no where in your system do you *explicitly* consider scaling. Yes, you spend a lot of time trying to figure out how often to measure something (my concerns for that are below) but you never relate it back to the amount of bandwidth or power consumed at the client (as implied by both the abstract and intro) or how many simultaneous clients your system can support or what infrastructure would be required to monitor, for example, a nationwide cellular deployment. Your measurements have fewer than 20 clients in total and no mention was made as to how many of these were working in parallel. Further, it reads

as if all of your measurements were done with laptops instead of handheld devices, so no inference about the power consumption of WiScape can be made either. The lack of these points makes the assertion that client-side monitoring is viable unsupported.

Further, the big challenge that you ignore is deployment: how do you get your measurement infrastructure (be it an app or what have you) deployed on to enough client phones to get decent coverage? In my opinion, this is the hardest part of a client-side measurement system and it's not at all mentioned here.

The fact that WiScape has apparently not been tested on an actual portable handset seems to be a big strike against the viability of this system.

Now, ignoring all of that, there are still problems. Most of the interference in cellular networks is caused by line of sight issues. How typical are your data sources with respect to line of sight? Intuitively, a bus on a road outside does not have significant line of sight issues. Your data description of your spot nodes was not sufficient to determine what line-of-sight issues they had. So, it is not clear how your empirical zone and epoch derivations would change if the client was indoors, for example.

That said, the graphs in Figure 5 seemed very interesting - if this paper is the first to present this level of data (from different regions, over time), please say so.

In Section 3.1, is the assumption that zones are square? If yes, please say so. Any reason not to consider other shapes?

The data in Fig 4. seems to contradict Fig 1. and Fig.9 ; why are all of the relative standard deviations so small in Fig 4 (<0.1), whereas in Fig 1, they seem to be much higher (350 Kbs std dev / 1080 Kbs mean = 0.3 relative std dev) and Fig 9 shows relative stddevs > 0.7: it seems unintuitive that the stddev should be so small in Fig 4 (independent of zone size) when lots of the other Figures show much higher variation.

In Section 3.2, the implicit assumption here is "stable is better", where I am not convinced. Intuitively, operators are not as interested in the average performance as much as the "trouble spots", so by optimizing for inter-epoch stability, are you not just averaging away what the operator wants to know?

The conclusion did not add much to the paper: were there any lessons learned? Are the plans for a WiScape iOS or Android app? What are your concrete next steps?

Response from the Authors

We thank the reviewers for their constructive comments that helped improve the paper. We have fixed the text to address some of the reviewer concerns regarding clarity of description.

Some questions regarding interpolability of our observations in bigger cities, while using various cellular phones, in presence of severe line-of-sight issues etc., can be answered only by increasing the scale of our study to a broader geographical location and by involving more people in the process. We intend to do so in future and have expanded the conclusions section to describe our concrete next steps to this end.

In particular we understand that design, implementation and deployment of client-assisted wide-area wireless network monitoring systems would involve addressing multiple hard challenges, a few of which are described by Reviewer#5. However, designing and deploying the system is not the main focus this paper. Our main contribution is, first, to analyze measurement data for three cellular service providers over the span of an entire city, and at various other locations for a period of two years. Second, we identified characteristics of wide-area cellular networks which help estimate performance efficiently and highlighted how such minimal coarse grained client collected measurements can be leveraged by some applications. We believe that establishing the effectiveness and utility of lightweight coarse grained monitoring is an important first step for designing and implementing a scalable client-assisted monitoring system.

That said, as mentioned in Section 6, we are actively working towards building such a measurement platform for both vehicular communication systems as well as for cellphone users.

We also agree with Reviewer#4 that passive measurement is a potential lightweight alternative for understanding network performance. We are presently in the process of augmenting our measurement mechanisms with such passive estimation techniques. In future we intend to publish our findings on the same.

Finally, we have shown how client assisted in Section 4 can help network service providers and users of multiple cellular cards. We agree with Reviewer#3's concern on the utility of client-assisted measurements for individual users. Client-assisted monitoring in the context of cellular networks is a new technique and we intend explore other aspects of cellular network monitoring that client assisted monitoring can help with.

Specific comments:
Reviewer#3: Test duration of 75 minutes… The measurements for a zone need to be retaken every 75 minutes, the test themselves involve sending 40-120 packets (as noted in Table 5). We have made this clearer in the text.

Reviewer#5: In Section 3.2, ... here is "stable is better", ... We meant "stable is better" in the sense that, we can potentially take less number of measurements to converge to the correct estimate if the network is stable.

Reviewer#5: The data in Fig 4. seems to contradict ... In Figure 1, the zones have a radius of 800 meters, and include locations with less than 200 measurements, hence some zones show high standard deviation. In Figure 9 the higher relative standard deviation is for zones with multiple ping failures, which we took as an indicator of high performance variance. The high standard deviation for such zones, thus, when compared to the small relative standard deviation for the aggregation of all zones (also plotted in same graph) proves our point.

Reviewer#5: Line of sight issue ... All our static measurements were taken in indoor locations and hence were not in line-of-sight from the base station, we have clarified it in the text.