Summary Review Documentation for

“Obtaining In-Context Measurements of Cellular Network Performance”

Authors: A. Gember, A. Akella, J. Pang, A. Varshavsky, R. Caceres

Reviewer #1

Summary: This paper consists of two parts: It first makes a case for doing “representative” measurements, i.e., measurements when (or near when) the user is active. The authors find that the latency and other metrics are different when the measurements are done when the device is idle vs when the user is active. In other words, performing background measurements are unlikely to measure the “real” performance of the network.

Second, the paper presents a crowd-sourced measurement platform that will allow researchers to do representative measurements.

Strengths: The motivation for this work is the observation that most crowd-sourced measurement frameworks perform network measurements either at random time intervals, or when the user manually conducts the measurement. In other words, the measurements are not conducted when the user may actually use the network; this implies that the performance seen by the measurement studies will be very different from the performance of the network when the user actually uses the network. To this end, the authors show that there is mean 15ms difference in latency between taking measurements when the phone is idle vs when the phone is active. This difference between idle and active does not stem from signal strength or the length of the idle time. Further, the authors also show the importance of context, while doing measurements. The authors do a thorough job of experimenting under various environmental conditions

The authors then present a measurement platform that allows researchers to perform representative measurements. The idea behind the platform is to only conduct measurements near when the user is active.

Weaknesses: It was unclear if the difference between the “representative” performance measure and the “non-representative” performance measure is big enough to be a problem. Other than this concern, I really liked the paper.

Comments to authors:

- You find an average latency difference of 15ms and a 5-percent difference in loss rate measured when conducting experiments during idle phase and the active phase. However, will these latency and loss rate differences make a difference for user perceived performance (or even the performance that the network operators care about). If this difference is not big enough, then the “non-representative” performance that is measured today can well approximate the “representative” performance measure.

- A simple experiment could be to run some applications under the two scenarios (active vs idle) and see if there is a perceptible difference in application performance. I think this will greatly strengthen the paper.
- The measurement platform allows users to take measurements only when the user is active. One concern that I have is, how long will it take to collect enough samples to do meaningful analysis. For example, let’s say I want to make measurements when the user is at a certain location. However, the users are rarely active when they are at that location, although they pass through that location. This is not uncommon. Your experiments show that that number of instances when the user is active during the day is less than 15. But this means that I need to wait a really long time before I can get enough measurement samples.
- You do not really motivate the problem in terms of who will use this. The work is presented from the operators point of view, but to me this work is most useful for researchers who need to conduct measurements.
- You also need to make a case for why researchers/operators can’t do passive measurements. For example, can researchers get most of the latency measurements done based on the latency actually seen by the user when using an application?

Nits:

- In Figs 16 and 17, may you can change the legend from “Experienced” to “Ground truth”. I don’t know what “experienced” means.
- In general, I think it will be useful to add some meta information to the captions to mark if the results are from the controlled experiments or from the network traces or from the RRC traces

Reviewer #2

Summary: The paper looks into crowdsourcing measurements to gain an understanding of mobile network performance. Specifically, the authors look into when a measurement should be conducted and on what device. An interesting angle to active measurements. However too many variables at play and too small a deployment to make meaningful conclusions.

Strengths: Interesting problem; Large dataset for initial analysis

Weaknesses: Too many variables, not adequate quantification; Small deployment size

Comments to authors: The paper is mostly well written. A new twist to active measurement framework that looks into the context of measurements.

The major complaint I had about the paper is the size of the deployment. Given the many variables at play, I am not really convinced at the end if the objective of the paper will be met if the deployment size is made really large. Some of the experiments involving context are too small scale to make any meaning conclusions. At the end it wasn't also clear what is "representative" about the measurements, when many factors are at play. Wouldn't there be a range of values one would get for a given "query" from the different users still operating in the same context.

The premise of the paper is based on figure.1, where "active range" is rather a derived metric. What you present is probably true, but it's useful to quantify that indeed the way you model the active ranges, matches some ground truth. Why 5 minutes periodicity? In this context, where did the number 17 come from and what is self interference.

The main metric in evaluating the framework was TCP throughput. It would have been interested to see other metrics like latency, loss etc

Reviewer #3

Summary: The authors use traces from 20,000 real users to demonstrate that measurements of 3G performance may not be representative of the performance users see when actually using their device. They then build a crowdsourcing measurement platform that allows network administrators to trigger measurements shortly after periods of use—but only if conditions have not changed substantially—and demonstrate that, for their study of 12 users, it reports substantially different performance than a randomly sampled study would have.

Strengths: This is an ideal IMC paper: it uses real data to motivate a measurement problem, develops a new measurement methodology, implements it, and demonstrates good performance. The paper is well written, and clearly motivates the need for measurements targeted at so-called "representative" usage. The authors built a real measurement platform, and demonstrate that it accurately detects changes in context that may lead to substantial differences in measured performance.

Comments to authors: I enjoyed reading this paper. My biggest complaint is that the results are not surprising: i.e., that performance metrics are sensitive to changes in context. That said, clearly the state of the art does not take it into account, and your platform does. My second concern is that your evaluation only uses 12 users, and does not have a "ground truth." While you show that the measurements vary substantially between context-aware and random, and that context-aware seems to be closer to "experienced," it does not actually show the performance during periods of activity, which it would seem you should have been able to collect by instrumenting the actual apps used by your volunteers. Said another way, how do we know "experienced" is actually a good proxy for the ground truth? 12 users also pales in comparison to 20,000, obviously, and it's hard to know conclusively whether the results will bear out when scaled up to a provider-wide measurement.

Stepping back a bit, though, I wonder a bit about the higher-level motivation for this work. What, precisely, does an operator do with the data they gather? How much data is enough to act upon? Will there be enough comparable points to draw conclusions? This isn't a problem with the work, *per se*, but does make me wonder about the impact it will have.

Reviewer #4

Summary: This paper shows that mobile network performance is influenced by many factors including context (location of the mobile (hand, pocket, etc), mobility vs stationary, etc) and user interaction with the phone. The authors argue that measuring performance should be done when the user is actively using the phone, in a context similar to active use but in a way that does not harm user experience.

This paper has two parts: the first part gives strong evidence that one must measure performance when the user is active or near-active and in the same context as active use. The second part designs a measurement system and tests it on a small user deployment.

Strengths: The observation that performance varies with context and phone usage is valid, and is well documented in the paper. While it is not exactly surprising, it does show that naive performance measurements can get it really wrong. In a few cases, the authors go beyond just observing the results, they try to understand the root causes for these differences, and offer interesting insights. The results shown in the first part of the paper are compelling.

Weaknesses: First of all, the paper misses a clear problem statement: why do we care about "representative" performance, and not the "real" performance? Is it impossible to measure actual, real performance? Why?

The second part of the paper is not as solid as the first part. First, the authors do not discuss the different choices we have to build a measurement system to test performance. Their system design seems to make some assumptions that constrain its design, but these assumptions are not spelled out. In other words, I have no clue why the system is designed as it is—and this really needs to come out of your paper.

I don't understand why one wouldn't just profile the performance of applications the user is engaging with - I mean, your system is checking interface stats to see if the user is actively using the network; couldn't you just pull connection stats from the kernel to decide how well the network is working? Say we just analyze user connections to facebook.com (or Google or any other popular site) to measure their throughput and RTT (one can imagine better performance metrics, of course). The performance differential between different users accessing some Facebook server/CDN at the same time will be (mostly) due to the differences in performance in the last hop, i.e. in the cellular network. Why is this obvious approach not suitable?

The paper is quite wordy - most of the section have a preview part that is quickly repeated in the body of the sections, and then again summarized at the end. This feels like you are trying to fill up space. It's better to have a shorter, clearer paper than a 14 page paper with lots of repetitions.

Writing could be sharpened in many places, many phrases are windy and difficult to follow.

Comments to authors: This can be a really good paper - and the first part is quite thorough and nice. I like the fact that you try to understand the root cause for the results observed.

However, the choices in the second part and the end-goal itself are not well motivated, and this reduces the strength of the paper.

I am confused as to why the goal of this paper is what it is and not the more natural "real" performance. I admit there may be big hurdles in measuring and using "real performance" numbers, but the paper should at least tell us what these are.

Reviewer #5

Summary: This paper proposes a framework for the measurement of the user experience on cellular networks that takes into account the actual context. The authors argue that to better quantify the user experience measurements need to be taken within the same context as the actual use and propose a framework that is able to schedule measurements to capture that particular context. The authors implement their solution and test it.

Strengths: Potentially interesting framework that could allow a network operator to schedule measurements across devices

Weaknesses: I am not convinced about the motivation of the work - it seems that the authors are "inventing" a problem to solve (more below).

A lot of the measurements presented about the impact of the device orientation are well known (performance depends on where the device is placed).

Comments to authors: The entire premise of the paper is that network operators need to be able to quantify the performance of the cellular service within the user context. Measurements that could be collected at random times and random locations are assumed to not be useful given that they do not really correspond with what the user is experiencing. As such, the authors propose to use an agent on the phone that is able to first monitor when the user is actually interacting with the device, and perform a follow on measurement only if the original context is unchanged.

The solution proposed based on this premise is rather reasonable and some results are presented for downlink throughput.

The main weakness of the work, in my opinion, is that of its motivation. I think the authors are essentially inventing a problem to solve.

1. Say, that these measurements are for the network operator for network dimensioning or provisioning. If that is the goal, then it does not really matter whether the performance measured is the one experienced. They need to know that with the current deployment "any" use would be acceptable. The performance you measure is going to be under the same conditions but any user could as well use the service in any other location, and knowing the locations with poor performance (whether experienced now or not) is not relevant in my mind. 2. If the measurements are for the user itself, then taking a measurement 20s later within the same context and for TCP download does not capture the performance experienced by the user anyway. Given that you install an agent on the phone to detect

activity and context, then why not measure the actual performance for the application that was actually used.

The one element of the solution that is interesting is the ability for the operator to schedule performance tests across a number of devices with specific criteria, but that is a very small piece of the overall solution and not really a research challenge the way I see it.

The results presented by the authors on the impact of device position on the performance are not novel in their own right. A lot of the indoor localization work, just to name one specific area, have demonstrated this impact before.

Lastly, the argument that passive measurements "in-network" are not able to capture context may be true, but they still correspond to performance experienced. It seems that the only thing missing is whether the device was on and the user was interacting with it. Given that specific applications cannot be used without active user interaction, then one could imagine solutions where passive measurements could be mined to identify those measurements that could reflect actual use if that would be of interest.

As a result, I am not convinced about the motivation of the work or its actual usefulness. I think the authors need to rethink whether the problem addressed is indeed a problem to start with.

Reponse from the Authors

Many concerns raised by reviewers stem from the ambiguity of the term "representative", prompting us to instead use the term "in-context" to refer to the type of measurements we intended to capture. In-context measurements are cognizant of the conditions (e.g., location, time, user activity, etc.) under which the measurements occur. In particular, there is a frequent need to obtain a measure of the network performance users experience solely when they are interacting with their device. Hence, "whether or not a user is actively using their device," along with the physical context of the device during this interaction, is an important condition to consider. We use this new terminology throughout the paper.

Another common concern was our focus on conducting active measurements near the times a device is actively used, as opposed to (passively) quantifying the "real performance" of application traffic. Our paper focuses on active in-context measurements because of the difficulty in eliminating the confounding factors of application variability [17] and our desire to only capture how the cellular network performs. However, our observations on how in-context measurements must be conducted can be used to filter passive measurements. We have clarified our stance on active versus passive measurement in Sections 2.2, 3.5, and 4.

Specific Comments:

Significance of performance differences [A]: We explain in more detail in Section 3.2.2 the importance of quantifying even small performance differences.

Definition of active range [B]: We added additional details on our definition of active range to Section 3.2.1. Prior work identified that changes in radio resource state occur after 17 seconds.

System design [D]: In Section 4, we added justification for our design choices and described how the system could be modified to leverage passive measurements.

Enough measurement samples [A,C]: We performed additional analysis using the network dataset to show there are sufficient measurement opportunities (Section 4.1.1).

Range of values [B]: We clarified in Section 4.1.1 that we compared CDFs of throughput for the three measurement types and observed similar ranges for ground truth and in-context measurements.

Choice of ground truth [C]: We clarified in Section 4.1.1 that we do not use performance from actual application traffic as ground truth because of variability between applications and the influence of cross-traffic from multiple applications.

Other metrics in evaluation [B]: We added to Section 4.1.1 that we choose to measure throughput because it inherently captures several metrics, including latency and loss.