On Using Crowd-sourced Network Measurements for Performance Prediction

Tova Linder, Pontus Persson, Anton Forsberg, Jakob Danielsson, Niklas Carlsson
Linköping University, Linköping, SE-58183, Sweden

***Abstract*—Geo-location-based bandwidth prediction together with careful download scheduling for mobile clients can be used to minimize download times, reduce energy usage, and improve streaming performance. Although crowd-sourced measurements provide an important prediction tool, little is known about the prediction accuracy and improvements such datasets can provide. In this paper we use a large-scale crowd-sourced dataset from Bredbandskollen, Sweden's primary speedtest service, to evaluate the prediction accuracy and achievable performance improvements with such data. We first present a scalable performance map methodology that allows fast insertion/retrieval of geo-sparse measurements, and use this methodology to characterize the Bredbandskollen usage. Second, we analyze the bandwidth variations and predictability of the download speeds observed within and across different locations, when accounting for various factors. Third, we evaluate the relative performance improvements achievable by users leveraging different subsets of measurements (capturing effects of limited sharing or filtering based on operator, network technology, or both) when predicting opportune locations to perform downloads. Our results are encouraging for both centralized and peer-to-peer performance map solutions. For example, most measurements are done in locations with many measurements and good prediction accuracy, and further improvements are possible through filtering (e.g., based on operator and technology) or limited information sharing.**

I. Introduction

Crowd-sourced measurements and network performance maps summarizing the information from these measurements can be valuable for predicting future download speeds and improving client performance. By summarizing the information from previously observed download speed measurements, these maps allow mobile devices to predict the available bandwidth in different locations and determine opportune times and places to download content.

Bandwidth prediction based on performance maps have been used to minimize download times and reduce the mobile units' energy usage [3], to improve streaming performance [13], [1], [12], [4], and to achieve efficient handovers in multi-homed environments [7], [4]. Careful scheduling of delay-tolerant downloads can also benefit the performance of delay-sensitive (e.g., real-time) applications. For example, scheduling downloads in locations with good bandwidth conditions will result in (relatively) more bandwidth available to delay-sensitive applications in more constrained locations; significantly improving their performance.

While the concept of performance maps have been demonstrated to provide significant performance benefits, many questions remain unanswered. For example, how does the bandwidth variations observed by typical users differ between locations, between operators, and, for a user wanting to use this type of technology, which bandwidth measurements are the best to share among users? In this paper, we use a large crowd-sourced dataset from Bredbandskollen[1] to address these and other open questions. Bredbandskollen is the dominant speedtest service in Sweden. By Feb. 2015 its Android and iOS applications had been used to perform (and collect) roughly 41 million crowd-based download (and respective upload) speed measurements from mobile Internet users.[2] In this paper we focus on the 16 million measurements from mobile (non-WiFi) networks that took place between Jan. 2014 and Feb. 2015, and leverage simultaneously collected meta information such as geographic location and the operator used for each measurement to evaluate the usefulness of crowd-sourced measurements for performance prediction.

The paper makes three primary contributions. First, we characterize the mobile speedtest usage of Bredbandskollen, discuss how the observed usage may impact the service that crowd-sourced performance maps may provide, and develop a scalable methodology to maintain performance map information that simultaneously is both large and sparse. Similar to the mobile network traffic itself, the usage of the service is highly diurnal (with a daily peak-to-valley ratio of 16), suggesting that the measurements may be relatively representative of the performance seen by regular clients. The usage is also highly skewed towards the regions where most people live, with a small fraction of the locations being responsible for the majority of the measurements. For efficient analysis, we split the area of interest into a grid and use a hashmap to perform constant time insertions and lookups. The methodology is motivated by the skew in the locations where the measurements are performed, including the long tail of locations without any measurements (e.g., where nobody lives), and is expected to be applicable to other large-scale performance maps as well.

Second, we analyze the variation (and predictability) of the download speeds observed within and across different locations. Our single-location analysis compare differences in the download speed variations based on factors such as the location granularity, number of measurements per location, operator selection, and the average download speed. We find that there are significant advantages to multi-homing and that locations with more measurements typically see higher average speeds and lower relative bandwidth variations, suggesting that many operators prioritize these regions. Our multi-location analysis extends this analysis by using a hypothesis-based methodology to provide an initial quantification how

[1]Bredbandskollen, http://www.bredbandskollen.se/. We thank Rickard Dahlstrand at the Internet Foundation in Sweden (IIS) for sharing the dataset.

[2]Over the same period, since the start in 2007, 120 million Bredbandskollen speedtests had been performed across both mobile and non-mobile networks.

ISBN 978-3-901882-80-7 © 2016 IFIP

often there are significant download speed differences between neighboring locations that can be leveraged by more advanced techniques for selecting when and where to download large files. We find that in roughly half (44.2%) of the cases, one of two neighboring locations provide significantly (with 95% confidence) better download speed, showing that there can be significant advantages to applications that can select at which of neighboring locations to download content, while traveling along a path (e.g., to/from work).

Third, we present a case-based performance analysis of the relative performance seen by users that use different data sharing policies to determine when to download content along an example path or visiting a sequence of locations. This evaluation is motivated by location-based services with limited access to measurement information, and differences observed across operators and the access technologies used by the users. Using data-driven simulations, we compare the average download speed improvements achievable when some measurement data used for the prediction is missing (e.g., due to limited peer-to-peer sharing rather than through central directory services) and the impact of which measurement information is shared (e.g., if all measurements should be shared, or only those matching a particular operator or network technology). Our results show that there are significant advantages in selectively restricting the information shared, but that these advantage decrease as clients must download during a larger fraction of the locations they visits to complete their downloads. It is also encouraging to see that in all considered cases there are significant performance advantages to use performance maps (regardless of the measurement sharing policy and the sharing level) compared to when clients have no knowledge of prior performance in the different locations.

The remainder of the paper is organized as follows. Sections II-IV is dedicated to each of our primary contributions. First, Section II presents a high-level characterization of the dataset and the usage of the crowd-sourced Bredbandskollen service. Then, Section III presents analyze the bandwidth variations within and across locations, before Section IV presents a multi-location case study. Finally, related works (Section V) and conclusions (Section VI) are presented.

II. Characterization of Measurement Usage

A. Dataset and limitations

Bredbandskollen is the most popular speedtest service in Sweden. In this study we analyze all speedtest measurements performed by mobile users testing their cellular (non-WiFi) Internet speed via Bredbandskollen's Android or iOS application between Jan. 2014 and Feb. 2015. In total, this dataset include over 16 million measurements. For each measurement, the application measures the upload speed, download speed, and latency. All tests are carried out against the geographically closest Internet eXchange Point (IXP). For each test, the application also records a timestamp, the geographical coordinates where the test was performed, and information such as which network technology[3] and operator was used.

The dataset is highly diverse and includes measurements from 3,184 different phone types, with various iPhone and iPad versions responsible for 38.7% and 21.7% of the measurements, respectively. The majority of the measurements are performed by users from one of the top-four national operators: Telia (32.7%), Tele2 (10.9%), Telenor (10.3%), and Hi3G (9.0%). For operator specific evaluations, we will focus on measurements performed in these four operators' networks.

Limiting ourselves to Jan. 2014 to Feb. 2015 allows us to limit potential effects due to changes in measurement infrastructure and ever improving Internet speeds. We note that the service sometimes appears to be used for diagnostics and the dataset therefore include zero-speed measurements. For our analysis we have limited ourselves to measurements that result in non-zero download speed.

B. Efficient Bandwidth Map Design

We have found a high skew in where the measurements take place, with most measurements being performed in the most populated areas of Sweden. This illustrated in Figure 1, which shows the location of each measurement point. First, it is observed that most measurements take place in the southern parts of Sweden or along the coastal region. This distribution matches well with where people live. For example, 90% of the population live in the southern 1/3 of the country, and most of the people in the northern 2/3 of the country live along the cost. Second, we can see the highest concentration of measurements around the country's three biggest cities: Stockholm, Gothenburg, and Malmo (close to Copenhagen on the map, but on the Swedish side of the Baltic Sea).

To better understand the download speed variations, both within and across locations, we create a bandwidth map that contains information about all measurements, and then use the map to retrieve statistics about each location. Our map divides the world in location buckets ("locations" for short) and associate each measurement with the location in which the measurement took place. Motivated by the many potential measurement locations (e.g., Sweden alone is 450,295 km^2 and divide into more than 11 million 200×200 m^2 squares) and the long tail of locations without measurements, we create a bandwidth map in which only locations with non-zero measurements are stored in a hashmap. To ensure easy lookups, for each measurement, we use a hash key based on the square-coordinate index of each measurement, after translating the x-y index pair into a unique text string. This approach allows constant time insertions and lookups.

In addition to allowing easy access to all measurements within a location, this approach also makes it easy to identify and analyze download speed correlation between neighboring locations. The fast and easy lookup of neighboring locations, is allowed by our choice to use a deterministic method that takes the square-coordinate index of each location as arguments when calculating the hash key for each location. Given knowledge about the square-coordinate index of one location, it is therefore trivial to lookup the measurements of the neighboring locations, which simply is offset by one square-coordinate index (in either x or y direction). In the following, we use this structure to analyze the dataset and provide insights into bandwidth differences and variations within and across locations. Naturally, similar hashmap-based bandwidth maps can easily be implemented for other large-scale crowd-sourced systems.

[3]This includes various 2/2.5G (e.g., GPRS, EDGE), 3G (e.g., UMTS, CDMA, HSPA, HSDPA, HSPAP), and 4G (LTE, LTE Advanced) technologies.

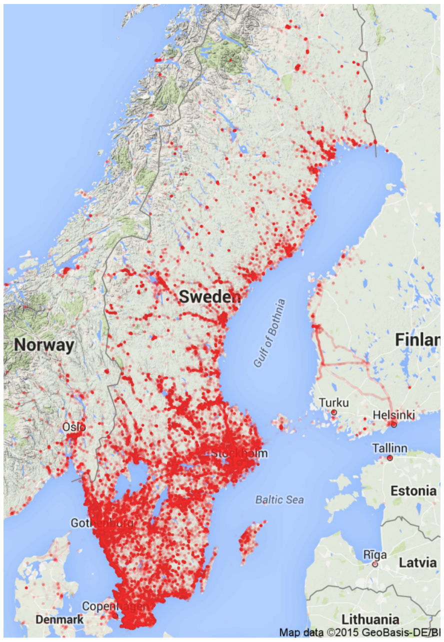
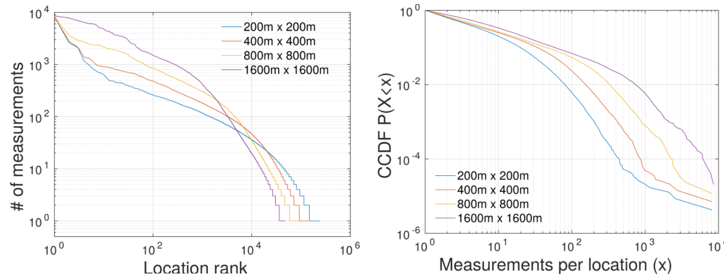
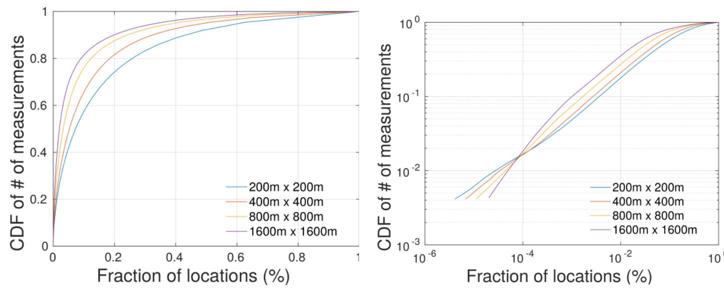


Fig. 1. Geographic measurement overview.



(a) Rank plot (b) CCDF

Fig. 2. Distribution of the number of measurements per (non-empty) location.



(a) Lin-lin scale (b) Log-log scale

Fig. 3. Relative concentration of measurements across non-empty locations.

C. Location and Measurement Concentration

To understand the skew in the locations where the measurements take place, we calculate the number of measurements associated with each location bucket, when using one of four different bucket sizes: 200×200 m^2, 400×400 m^2, 800×800 m^2, and $1,600 \times 1,600$ m^2. Following common practice, we use rank plots, Cumulative Distribution Function (CDF) plots, and Complementary CDF (CCDF) plots to capture the concentration of events; in this case the locality of measurements.

Figures 2(a) shows a rank plot of the number of measurements per location, where location "ranks" are sorted from the locations with most to the fewest number of measurements. Figure 2(b) shows a CCDF of the fraction of locations with more measurements than a given sample threshold N, as a function of N. Both these plots focus on the locations with the most measurements. We note that depending on location-bucket granularity there are only between 4 (less than 0.003% of the non-empty 200×200 locations) and 200 locations (approximately 0.8% of the $1,600 \times 1,600$ locations) with more than 1,000 measurements. Regardless of granularity, there are however more than 1,000 locations with more than 100 measurements (corresponding to between 0.8% and 7% of the non-empty locations, depending on bucket granularity).

The small fraction of locations with many measurements matches the intuition that there is a high skew in the locations where the majority of measurements take place. The concentration in measurements are captured by Figures 3(a) and 3(b), which show CDFs of cumulative fraction of total number of measurements associated with the locations that makes up the fraction (x) with most measurements. While there are differences in the absolute concentration, in general, between 10-20% of the locations (with non-zero number of measurements) are responsible for between 80-90% of all measurements. This suggests similar skews as with Pareto principle. For much of our later analysis we will therefore focus on the location buckets with the most measurements. In particular, we will typically focus on the top-15% of locations. With this choice, we end up using a threshold N of 15, 25, 30 (for the first three granularities, respectively), each set being responsible for 70%, 70%, and 80% of the measurements, respectively. To ensure that we always have more than 1,000 sample points, we use $N = 20$ for the $1,600 \times 1,600$ granularity case; resulting in 18% of the locations and 90% of the measurements being captured.

D. Time-of-Day Analysis

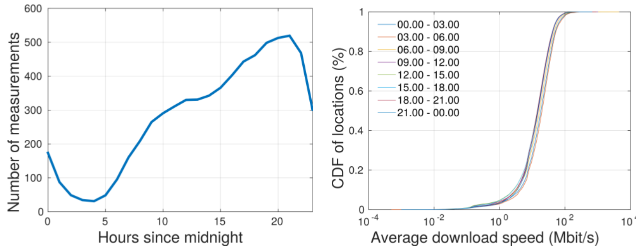
To better understand any biases in the dataset and the service usage, we next take a closer look at the hourly usage pattern (Figure 4(a)) and the "download speed profiles" for different times of the day (Figure 4(b)), where a download speed profile consists of a CDF of the download speeds observed across all measurements associated with that profile. We use three-hour time buckets to distinguish profiles.

We observe a significant diurnal pattern (Figure 4(a)) in the number of measurements per hour, as function of time, with a peak-to-valley ratio of 16 (525 to 33). The diurnal pattern matches well with the expectation of when the networks are most in use. While we have observed some non-negligible differences in the average and median download speeds for the different times of day, with the biggest differences in median being between 3:00-6:00 (20.7 Mbit/s) and 18:00-21:00 (18.1 Mbit/s), the download speed profiles (Figure 4(b)) have in general very similar characteristics, regardless of the time of day. Motivated by the relatively small differences and the lack of measurements during nighttime in most locations, for the remainder of this paper, we do not differentiate between measurements within a single locations based on time-of-day.

III. Bandwidth Variation Analysis

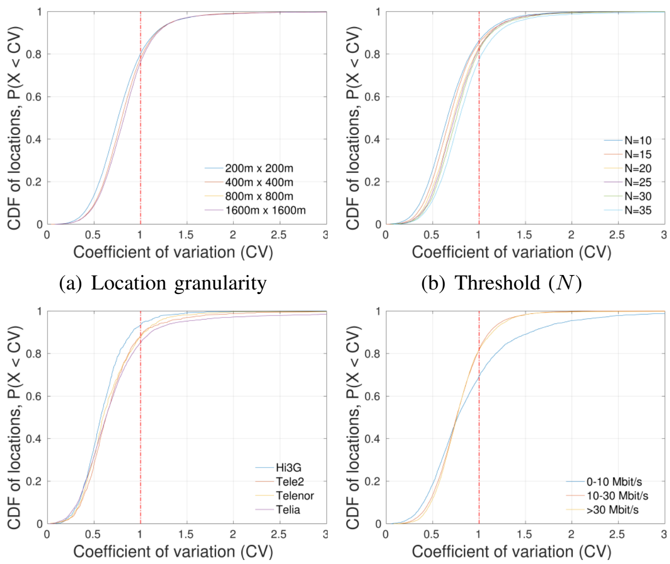
A. Single-location Variation Analysis

To better understand if and how past measurements can be used to predict the network conditions in a location, we first consider the relative variation observed in the measurements



(a) Diurnal usage (b) Download speeds

Fig. 4. Time-of-day analysis.



(a) Location granularity (b) Threshold (N)

(c) Operator (d) Average download speed

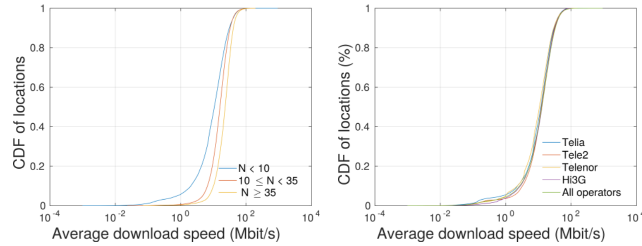
Fig. 5. Cumulative Distribution Function (CDF) of the Coefficient of Variation (CV), across large number of sample locations.

within each location. The relative variation is typically measured by the coefficient of variation (CV), sometimes referred to as the relative standard deviation, defined as the ratio $\frac{\sigma_i}{\mu_i}$ of the standard deviation (σ_i) and average (μ_i), and estimated as $\frac{s_i}{\overline{x}_i}$, using their respective sample measures s_i and $\overline{x}_i$.

Figure 5 shows CDFs of the coefficient of variation (CV) of the download speeds, as calculated across all locations i satisfying different criteria. In particular, we show curves to illustrate the impact of using different (a) location granularities, (b) thresholds for the minimum number of measurements per location, (c) operators, and (d) locations for which the average download speed falls into different speed ranges. When interpreting these results, we note that typically locations with a CV less than one are considered as low-variance locations, and locations with CV greater than one are considered to have high variance. To put this in context, we note that exponential distributions have CV=1, Erlang distributions have CV<1, and hyper-exponential distributions have CV>1.

Based on these figures, we make four observations. First, and perhaps most importantly, the majority of locations are low-variance locations. For example, for 12 of the 17 curves more than 80% of the locations are considered low-variance. Only for the low-average-speed case (69%) is there less than 75% low-variance locations. This is interesting as low-variance locations provide better prediction opportunities.

Second, the results are relatively insensitive to the location granularity (Figure 5(a)) and the minimum number of measurements per location (Figure 5(b)). In both plots, both the distributions and cross-over points are very similar for all curves. The insensitiveness with regards to granularity is encouraging, as the use of coarser granularity provides greater opportunities to aggregate information from many measurements.



(a) Number of measurements (b) Operators

Fig. 6. CDF download speed profiles.

Third, while there are some smaller variations across the top-four operators (Figure 5(c)), for all four operators more than 85% of the locations have low-variance. Hi3G have the highest fraction (93%), low-variance locations, whereas Telia has the lowest fraction (85%). The somewhat lower fraction for Telia may be due to Telia covering more locations, including some less covered regions where the mobile Internet conditions are not as good as within the cities. This conjecture is supported by the fourth and final observation, that the fraction of high-variance locations is much higher for low-bandwidth connections. The higher variance for low-bandwidth conditions, is exemplified by the 0-10 MBit/s curve in Figure 5(d). Here, 69% are consider low-variance locations, compared to 82% of the 10-30 MBit/s and 30+ Mbit/s locations.

To further validate our conjecture, we compare the CDFs of the (average) download speeds observed for locations with different number of measurements. Figure 6(a) shows the download speed profiles for three intervals of sample thresholds (some of which are smaller than the thresholds used to determine which locations to include in our analysis). Consistent with our conjecture, the download speeds are significantly lower (note the logarithmic axis scale) for the locations with few measurements (e.g., $N < 10$ curve) than those with many (e.g., $N \geq 35$ curve). Having said that, we have not observed any significant differences in the overall download speed profiles of the operators (Figure 6(b)).

B. Pairwise Head-to-Head Comparison

In cases when one operator provides better download speeds in a location, knowing the "winner" may allow multi-homed users to switch to the best operators in each location [4], [7] and multipath-TCP users may be able to better utilize the differences in speeds across parallel connections [16]. In contrast, if there typically is no clear "winner", solutions that aggregate all measurements (across operators) may allow for added accuracy in locations with otherwise few measurements.

To better understand how often there are statistical differences observed between two operators, we use hypothesis testing. For this analysis we identify a large number of pairwise sample sets, perform hypothesis testing on each such pair, and report the fraction of pairs for which the test is rejected. This methodology allows us to calculate and compare the fraction of locations in which the download speed difference between two operators are statistically significant.

TABLE I. PERCENT REJECTED PAIRWISE T-TEST WHEN COMPARING OPERATORS' AVERAGE DOWNLOAD SPEEDS.

	Telia	Tele2	Telenor	Hi3G
Telia	–	$\frac{639}{1278}$=50.0%	$\frac{434}{953}$=45.5%	$\frac{327}{701}$=46.7%
Tele2	$\frac{639}{1278}$=50.0%	–	$\frac{28}{75}$=37.3%	$\frac{40}{63}$=63.5%
Telenor	$\frac{434}{953}$=45.5%	$\frac{28}{75}$=37.3%	–	$\frac{21}{49}$=42.9%
Hi3G	$\frac{327}{701}$=46.7%	$\frac{40}{63}$=73.5%	$\frac{21}{49}$=42.9%	–
All	$\frac{857}{5908}$=14.5%	$\frac{304}{1180}$=25.8%	$\frac{258}{988}$=26.1%	$\frac{202}{805}$=25.1%

For each pair of sample sets, we apply Welch's t-test [15] to calculate the t-statistics for the null hypothesis that the means of those two sets are equal (i.e., $\mu_1 = \mu_2$), with the alternative hypothesis that the means differ (i.e., $\mu_1 \neq \mu_2$). Assuming normally distributed samples in each sample set, but allowing for different variance in each set, the t-score is calculated as

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \quad (1)$$

where $\overline{x}_1$ and $\overline{x}_2$ are the means of the two sample sets, s_1 and s_2 are the standard deviation of two sets, and n_1 and n_2 are the number of measurements in each set. For each pair, we then compare the t-statistics with the significance threshold $t_{\alpha,df}$, where the degrees of freedom is calculated as

$$df = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}, \quad (2)$$

and α allow us to control the significance level. When $t > t_{\alpha,df}$, we reject the null hypothesis (that the means are equal) in favor of the alternative hypothesis (that the means differ). This test is then repeated for all pairs, and we report the fraction of tests that are rejected.

Consider now the potential download speed differences between the top-four operators in our dataset. Table I shows the percentage of cases in which the null hypothesis (that there are no differences in the mean download speeds) are rejected with 95% confidence. Here, we have only conducted pairwise tests when both operators have at least $N \geq 10$ sample points in a location. In addition to comparing the top-four operators, we also include a comparison against a default case when "all" measurements from that location are used as an aggregate, including measurements from any operator.

First, note that the fraction of rejected null hypothesis tests consistently is much smaller for the "all" row (14.5-26.1%) than for any of the other pairwise comparisons (all above 37%). This suggests that the aggregate measurements in a location often can be used as a good estimate of what users of well-used operators in that location experiences. This is in part due to many locations being dominated by a big player, but may also partially reflect the competitive nature of the telecom industry driving operators toward trying to provide their users with at least equally good service as their competitors in a particular location. Yet, in the cases with sufficient measurements from multiple operators, we note that there often is a statistically significant winner when comparing two operators. For example, among the six unique pairwise comparisons we reject between 37.3-73.5% null hypothesis, suggesting that we have 95% confidence that the average download speeds of the operators in these locations differ.

These differences suggest that there may be significant benefits to using adaptive multi-homing. Although there typically is not a consistently dominating player that always have better performance, there often is a clear winner in individual regions. The flexibility to switch to the best operator in each region can therefore significantly improve the average download speeds. Others have drawn similar conclusions [5]. Without a consistent winner there are also advantages to multi-path TCP solutions, which can adapt the bandwidth share across parallel connections, over separate operators, for example [16].

C. Comparing Neighbor Locations

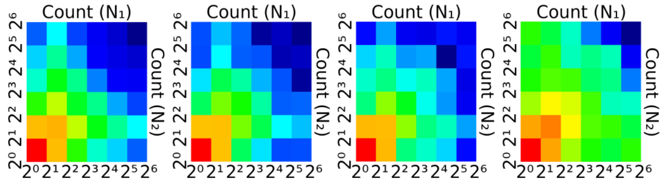
Geographic download speed differences allow mobile clients to select opportune times and locations to download contents, such as to improve download speeds, save energy, and improve conditions for delay sensitive applications. For example, a mobile client moving between two neighboring locations can easily use a bandwidth performance map to decide in which of the two locations to download the content. We next analyze how frequently there are statistically significant differences between the average download speeds observed between neighboring locations.

First, let us again apply our pairwise methodology (Section III-B) to compare the measurements associated with two neighboring locations. For much of this analysis we discuss results in which we vary one factor at a time, starting with a default case in which we use a granularity of 200×200, a 95% confidence level ($\alpha = 0.05$), and require at least $N = 20$ measurements in each of the two locations. In this default case we reject the null hypothesis (that there are no differences in the means) in 44.2% of the location pairs. This shows that in slightly less than half of the cases one of the locations would be the preferred location to perform a download.

Interestingly, the probability that the user will have a clear winner location increases slightly with coarser granularities. For example, compared to our default case with granularity 200×200 (44.2%), the percentage rejected null hypotheses increases to 47.5% with granularity 400×400, to 53.9% with 800×800, and 57.7% with $1,600 \times 1,600$. This finding suggests that coarser granularity maps can help users make better download decisions than fine granularity maps.

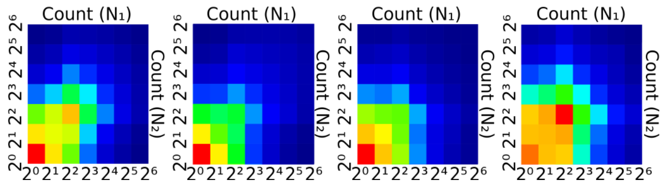
With limited variations within locations (Figure 5(a)), much of the observed increase in the number of neighboring locations with statistical differences is likely due to the improved prediction allowed by the additional measurements associated with larger location buckets. This hypothesis is also supported by the results when looking closer at the impact of the threshold value N. For example, relative to our default case with $N = 20$ (44.2%), the fraction of rejected tests increase to 47.0% with $N = 35$, and go down to 38.9% with $N = 10$.

The impact of the number of measurements are further illustrated in Figure 7, which shows the percentage of pairwise t-tests that were rejected for pairwise tests with different sample counts (N_1, N_2) for the two locations. Here, heatmap bucket boundaries (squares) are defined by the (N_1, N_2) counts on the x- and y-axis. Across all granularities, the biggest fraction of rejects (where one location see significantly better download speeds than the neighbor location) occurs in the heatmap buckets (squares) with most measurements (outermost blue regions). While most pairs fall into heatmap buckets with few measurements (lower left corner) for all granularities,



(a) 200×200 (b) 400×400 (c) 800×800 (d) $1,600 \times 1,600$

Fig. 7. Fraction (not) rejected pairwise t-tests comparing the download speeds in neighboring locations. Here, red is zero rejected and blue is 100% rejected.



(a) 200×200 (b) 400×400 (c) 800×800 (d) $1,600 \times 1,600$

Fig. 8. Number of neighbor pairs in each heatmap bucket. Here, red indicate the bucket with most pairs and blue the buckets with fewest neighbor pairs.

there are in general more location-pairs per high-count bucket when the granularity is coarser. To put the relative skews in perspective, Figure 8 shows the relative number of observed location pairs falling into each heatmap bucket.

We have also found that rejected pairs often are associated with regions with higher download speeds. In all considered cases, the average download speed across the locations with the rejected hypotheses are consistently higher than across the non-rejected pairs for the same case. For example, for the 200×200 case the two averages are 30.2 Mbps and 27.0 Mbps, respectively. As expected, the fraction of rejected tests increases when only requiring 90% confidence (e.g., 51.7% with $\alpha = 0.1$) and decrease when requiring 99% confidence (e.g., 30.9% with $\alpha = 0.01$).

IV. Multi-location Use Case Study

Crowd-sourced measurements allow performance maps to scale and have the advantage of a rich database. However, with many contributing users, measurements will typically be performed using different phone types, across different operator networks, and using different transfer technologies. This raises many questions regarding which information is best shared and distributed among users.

In addition, both centralized directory services and peer-to-peer approaches are possible to share measurement information. While peer-to-peer exchange policies can help avoid single-point of failure problems and naturally provide localized data sharing, peer-to-peer approaches typically limits the measurements that clients have access to. Such limitations can impact the accuracy of the prediction.

This section investigates the performance impact that limitation in the amount of data (e.g., due to limited coverage by peer-to-peer systems) and the type of information that is shared (e.g., by careful filtering based on which operator and network technology was used for each measurement) may have on the performance optimizations performed by a client using the performance maps information. Motivated by one of the most common use cases for network performance maps, for this analysis we consider the average achieved download speed when using the available information to predict opportune times and places to download (and upload) data.

A. *Trace-driven Methodology*

For this analysis we focus on 200,000 measurements that took place within a $20 \times 20\text{km}^2$ area centered in central Stockholm between Nov. 2014 and Feb. 2015. Stockholm is the capital, the highest populated city in Sweden, and also the area with most measurement points (e.g., Figure 1).

We then simulate the performance of a user moving across N locations, with each location defined as a unique 1km^2 rectangle. For each simulation, we randomly select one measurement from each of N location buckets along the path. A sequence of N such random sample measurements represents a sample path.[4] Given a sample path, we then evaluate the performance seen by different policies, where each policy use the remaining measurements in each bucket to predict the best locations to perform the download.

In our evaluation, we assume that a client always downloads when in the k locations with the highest expected (predicted) download speed of the N locations. Given download speed estimations for each location, this approach maximizes the client's expected average download speed over k locations. Furthermore, under this assumption, the client performance of different policies only differs by the information used for the prediction. To simulate different client behaviors and data sharing policies we filter the data used for each simulation. For example, to simulate a client that only has knowledge about 50% of the measurements we filters out 50% of the measurements before making the prediction. Similarly, to simulate a policy that only uses (or shares) information about a particular operator, we only use the data associated with that operator for the prediction.

Note that the k selected locations may differ based on the information that the client has available when making the download speed estimates for each location. In this paper we use five different filtering policies.

- *Full sharing:* Users share and use all available data.
- *Same operator:* For each location, users only use measurements made over the same operator.
- *Same technology:* For each location, users only use information about measurements made over the same network technology (3G or 4G).
- *Restricted sharing:* For each location, users only use information that simultaneously satisfy both the "same operator" and "same technology" policy.
- *Random sharing:* For each location, users only use information about p% randomly selected measurements.

We also compare two baseline policies. First, we include results for a "no sharing" policy that does not use any information at all, but simply picks k (of N) locations at random. This policy provides an example of the performance seen by a user not using any past knowledge for the scheduling. Second, we use an "oracle" policy that uses information about

[4]This independence assumption is motivated by work by Yao et al. [18], which shows that there typically are no correlations between a series of measurements performed along a specific trip, even when performed back-to-back, and time-based moving averages therefore are not suitable for prediction in mobile networks (in contrast to in static environments).

the speeds that the user actually would see in each location, and hence always "guesses" the best k locations for each particular sample path. Clearly, this provides a lower bound that is not achievable in practice, except for the special case when clients download in all N locations (and no scheduling is needed).

When comparing policies, we simply calculate the average download speed of each policy as as the average sample speed across the k locations selected for that policy. Naturally, higher download speeds are better here, as it provides better energy saving opportunities, for example. Each reported value is calculated as the average value over 50 simulations.

B. Simulation Results

Let us first consider the average download speed along a sample path of $N = 11$ locations, starting in a suburb (Hässelby) and ending in downtown (Östermalm). Figure 9 shows the average download speed as a function of the number of download locations k ($1 \leq k \leq N$) along the path, when using the "full sharing" and "oracle" policy. For all k, the "no sharing" policy (not shown) achieves on average the download speeds of the right-most point $k = N$. In this example, the "full sharing" policy achieves noticeable improvements over the "no sharing" policy, but these improvements decrease with increasing k. The much higher speeds of the "oracle" policy indicates that there is much room for improvements.

We next take a closer look at the "random sharing" policy and the impact of the percentage shared. Figure 10 shows the relative download speed difference between the "random policy" and the "oracle" as a percentage, for different levels of sharing p. With our choice to normalize download speeds relative to the download speeds observed with the "oracle" policy, the "full sharing" curve in Figure 10 is simply equal to the percent difference between the two curves in Figure 9. Note that "full sharing" consistently performs the best (closest to zero), and the "random sharing" policy with the least amount of sharing ($p = 1\%$) and $k = 1$ results in the worst relative download performance (compared to "oracle").

Finally, Figure 11 compares the relative performance (download speed increase compared to the "oracle" policy) for each sharing policy. Here, we have simulated 20 different path scenarios; each scenario consisting of 11 randomly selected location squares, from the entire Stockholm area.

We note that all three selective policies ("same operator", "same technology" and "restricted sharing") outperform "full sharing" when the client downloads in only a small fraction of the locations (e.g., $k = 1$ and $k = 4$ shown in Figures 11(a) and 11(b), respectively). For these conditions the "restricted sharing" policy performs the best. As the client uses more locations (e.g., $k = 8$ shown in Figure 11(c)), the benefits become smaller and "full sharing" in fact has the best median performance (red line). This shows that the type of information that is used for the prediction is very important when being selective (small k values) but decrease with increasing k, where it instead may be important to ensure that all locations have a sufficient number of "reasonable" measurements. In all cases, all policies significantly outperforms the "no sharing" policy, highlighting the value of using careful scheduling based on performance maps.

V. Related Work

Prior work have shown that there is very limited correlation between neighboring locations [18], that bandwidth prediction is more successful if location is taken into account [2], and that for a given location, past bandwidth measurements give a good prediction of the experienced bandwidth [17]. Motivated by these and similar observations, researchers have proposed the use of network performance maps.

Network performance maps have been shown useful in many scenarios [17], [8], [19], [3]. For example, performance maps based on commuter traces have been used to reduce the average download times of delay-tolerant downloads, effectively reducing the energy usage of the mobile devices [3], and to achieve smoother video streaming in mobile environments [13], [1], [12], [4], [5]. As an example, Riiser et al. [12] show that bandwidth prediction together with careful quality adaption can help reduce the number of playback interruptions of HTTP-based Adaptive Streaming (HAS), compared to when not using prediction.

In the context of vehicular networks, both crowd-sourced [11] and personal [10] bandwidth maps have been shown to provide good predictions. While mobile devices typically experience worse performance than stationary on the same network [9], history-based prediction can improve download speeds also in high-speed scenarios [17]. Good bandwidth prediction has also been shown to help improve handover selection in multi-homed environments [7], [4].

Transport layer information such as round-trip times (RTTs) can further improve prediction maps [8]. Our dataset does not contain RTT information, TCP retransmissions, threshold, window sizes, etc. Instead, similar to most prior works, we focus only on application layer measurements. In this case, download (and upload) speeds from (to) a server.

While speedtest data similar to that used here have been used in other studies, we are not aware of any study that use such data to evaluate the value of crowd-based bandwidth performance maps. Perhaps closest to ours is the work by Sommers and Barford [14], in which they use similar data to compare the latencies in WiFi and mobile networks in different regions. In contrast to their work, we focus on the value of using crowd-based measurements for performance prediction.

VI. Conclusions

Using a large dataset from Bredbandskollen, this paper evaluates the prediction accuracy and achievable performance improvements that large-scale crowd-sourced datasets may allow when download speed predictions based on these datasets are combined with careful download scheduling for mobile clients. Working with a large and sparse dataset, we first present a scalable performance map methodology, which uses a hashmap-based structure to perform constant time insertion/retrieval of geo-sparse measurement information. Using this methodology, we then characterize the speedtest usage of Bredbandskollen, observing a usage representative of the bandwidth usage itself, including highly diurnal usage pattern and most measurements being in highly populated regions.

We then extend the analysis to answer questions regarding how the bandwidth variations observed by typical users

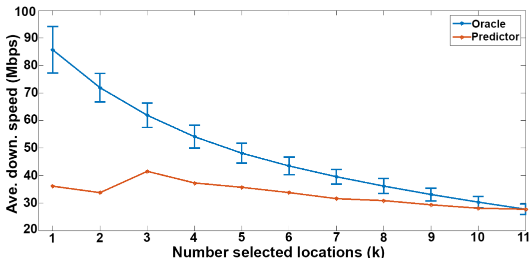


Fig. 9. Average download speeds when using “full sharing” and the “oracle” policy.

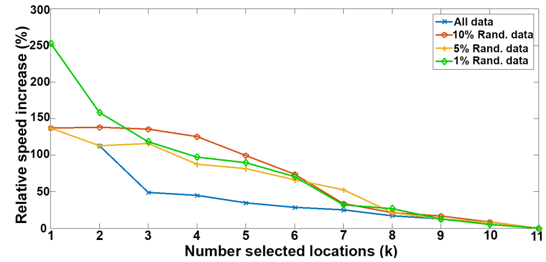
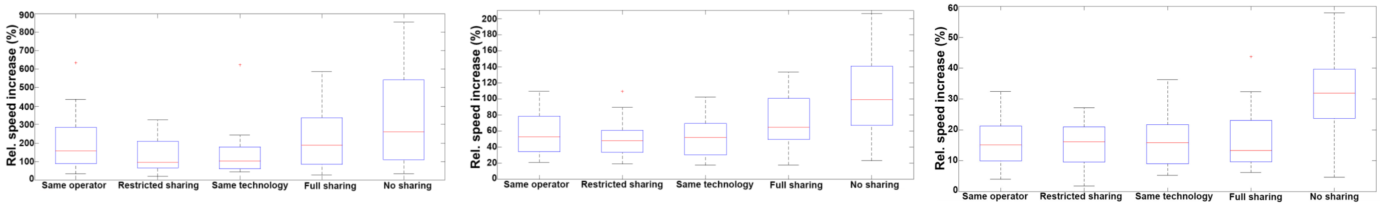


Fig. 10. Impact of fraction of measurements shared with the “random sharing” policy. Download speeds calculated relative to the “oracle” policy.



(a) $k = 1$, $N = 11$ (b) $k = 4$, $N = 11$ (c) $k = 8$, $N = 11$

Fig. 11. Boxplot showing relative performance compared to the “oracle” policy, for each of the other sharing policies.

differ between location and operator, for example. For this analysis, we analyze the bandwidth variation and predictability of the download speeds observed within and across different locations, when accounting for factors such as the location granularity, number of measurements per location, operator selection, and the average download speed. Using hypothesis testing we show that there often are significant download speed differences that can be predicted between neighboring locations, that these differences are most significant in the locations with most measurements, and that larger location buckets therefore are beneficial. Finally, we use a data-driven performance study of a geo-smart download scheduler to evaluate the relative performance of users using different subsets of the measurements when predicting opportune locations to perform downloads. This allows us to capture effects of limited sharing or filtering based on operators, network technology, or both. Our results are encouraging for both centralized and peer-to-peer network performance map solutions, and shows that the high skew in measurement locations allows us to achieve additional improvements through filtering (e.g., based on operator and network technology) or reduce overhead through limited information sharing.

Motivated by the performance improvements that we show are possible with the help of crowd-source measurements, we believe that crowd-sourced performance maps provides a valuable tool for mobile clients. Future work includes the design and evaluation of geo-smart schedulers for a richer set of application domains and for multi-homed devices. There are also many interesting open challenges with deploying crowd-based systems in general [6].

REFERENCES

[1] I. Curcio, V. Vadakital, and M. Hannuksela. Geopredictive real-time media delivery in mobile environment. In *Proc. MoVid*, 2010.

[2] P. Deshpande, X. Hou, and S. R. Das. Performance comparison of 3G and metro-scale WiFi for vehicular network access. In *Proc. ACM IMC*, 2010.

[3] A. G. Estévez and N. Carlsson. Geo-location-aware emulations for performance evaluation of mobile applications. In *Proc. IEEE/IFIP WONS*, 2014.

[4] K. Evensen, A. Petlund, H. Riiser, P. Vigmostad, D. Kaspar, C. Griwodz, and P. Halvorsen. Mobile video streaming using location-based network prediction and transparent handover. In *Proc. NOSSDAV*, 2011.

[5] K. R. Evensen, D. Kaspar, C. Griwodz, P. Halvorsen, A. F. Hansen, and P. E. Engelstad. Improving the performance of quality-adaptive video streaming over multiple heterogeneous access networks. In *Proc. ACM MMsys*, 2011.

[6] A. Faggiani, E. Gregori, L. Lenzini, V. Luconi, and A. Vecchio. Smartphone-based crowdsourcing for network monitoring: Opportunities, challenges, and a case study. *IEEE Communications Magazine*, 52:106–113, Jan. 2014.

[7] S. Herborn, H. Petander, and M. Ott. Predictive context aware mobility handling. In *Proc. ICT*, June 2008.

[8] K. Hojgaard-Hansen, T. K. Madsen, and H.-P. Schwefel. Reducing communication overhead by scheduling tcp transfers on mobile devices using wireless network performance maps. In *Proc. EW*, 2012.

[9] K. Jang, M. Han, S. Cho, H.-K. Ryu, J. Lee, Y. Lee, , and S. B. Moon. 3G and 3.5G wireless network performance measured from moving cars and high-speed trains. In *Proc. ACM MICNET*, 2009.

[10] G. Murtaza, A. Reinhardt, M. Hassan, and S. S. Kanhere. Creating personal bandwidth maps using opportunistic throughput measurements. In *Proc. IEEE ICC*, June 2014.

[11] T. Pogel and L. Wolf. Prediction of 3G network characteristics for adaptive vehicular connectivity maps (poster). In *Proc. IEEE VNC*, Nov. 2012.

[12] H. Riiser, T. Endestad, P. Vigmostad, C. Griwodz, and P. Halvorsen. Video streaming using a location-based bandwidth lookup service for bitrate planning. *ACM Trans. Multimedia Comput. Commun. Appl.*, 8:24:1–24:19, Aug. 2012.

[13] V. Singh, J. Ott, and I. D. D. Curcio. Predictive buffering for streaming video in 3g networks. In *Proc. IEEE WoWMoM*, June 2012.

[14] J. Sommers and P. Barford. Cell vs. WiFi: On the performance of metro area mobile connections. In *Proc. ACM IMC*, 2012.

[15] B. L. Welch. The generalization of student’s problem when several different population variances are involved. *Biometrika*, 34:28—-35, 1947.

[16] D. Wischik, C. Raiciu, A. Greenhalgh, and M. Handley. Design, implementation and evaluation of congestion control for multipath tcp. In *Proc. NSDI*, 2011.

[17] J. Yao, S. Kanhere, and M. Hassan. Improving QoS in high-speed mobility using bandwidth maps. *IEEE Transactions on Mobile Computing*, 11:603–617, Apr. 2012.

[18] J. Yao, S. S. Kanhere, and M. Hassan. An empirical study of bandwidth predictability in mobile computing. In *Proc. ACM WiNTECH*, 2008.

[19] J. Yao, S. S. Kanhere, and M. Hassan. Geo-intelligent traffic scheduling for multi-homed on-board networks. In *Proc. ACM MobiArch*, 2009.