Identifying Value in Crowdsourced Wireless Signal Measurements

Zhijing Li, Ana Nika, Xinyi Zhang, Yanzi Zhu, Yuanshun Yao, Ben Y. Zhao, Haitao Zheng
UC Santa Barbara
{zhijing, anika, xyzhang, yanzi, yao, ravenben, htzheng}@cs.ucsb.edu

ABSTRACT

While crowdsourcing is an attractive approach to collect large-scale wireless measurements, understanding the quality and variance of the resulting data is difficult. Our work analyzes the quality of crowdsourced cellular signal measurements in the context of basestation localization, using large international public datasets (419M signal measurements and ∼1M cells) and corresponding ground truth values. Performing localization using raw received signal strength (RSS) data produces poor results and very high variance. Applying supervised learning improves results moderately, but variance remains high. Instead, we propose *feature clustering*, a novel application of unsupervised learning to detect hidden correlation between measurement instances, their features, and localization accuracy. Our results identify RSS standard deviation and RSS-weighted dispersion mean as key features that correlate with highly predictive measurement samples for both sparse and dense measurements respectively. Finally, we show how optimizing crowdsourcing measurements for these two features dramatically improves localization accuracy and reduces variance.

1. INTRODUCTION

As wireless networks continue to grow in size and coverage, network monitoring and management is becoming an increasingly costly and resource intensive task [10]. While it used to be a standard practice to measure wireless performance by covering an area with vehicles and specialized equipment, that is simply impractical today. Instead, companies and research firms are turning to crowdsourcing as a cheap and scalable way to perform network measurements at scale [1].

But just how reliable are these user-contributed measurements? There are obvious reasons to doubt the accuracy and the consistency of user-contributed wireless network measurements. First, unlike specialized measurement tools deployed by network providers, user-contributed measurements tend to be generated using commodity equipment with less accuracy. Second, users are often less tech-savvy, and more likely to introduce errors during operation or through user contexts (*e.g.* driving, phone in pocket). Third, crowdsourced measurements are constrained by the mobility patterns of contributing users. Therefore, measurements will follow user mobility, and are likely uneven in coverage.

With this in mind, it is critical for network providers to understand the value and limitations in crowdsourced network measurements. While crowdsourced measurements can be used for a number of management functions (*e.g.* network performance and coverage measurements [13, 7, 6], transmitter localization and radio map construction [12, 8, 2, 3], spectrum anomaly detection [11]), they are generally not amenable to quantitative analyses, because of the dearth of both measurement data and ground truth datasets.

In this work, we are taking a data-driven, quantitative approach to answering some of these questions, by focusing on the specific application of *basestation localization*. Basestation or transmitter localization is a basic operation in wireless network management, and critical to providers interested in locating misbehaving transmitters or mapping out potential holes in basestation coverage. Besides, nowadays many mobile applications rely on cell tower triangulation to determine user position [4] for lower energy consumption than GPS. However, the public sources of cell tower location are incomplete and inaccurate [2, 5]. Like other management applications, basestation localization uses received signal strength (RSS) measurements gathered by mobile devices. Unlike other applications, analyzing localization performance is tractable today, given the availability of both crowdsourced RSS datasets and ground-truth data on basestation locations.

We are interested in answering several critical questions about user-contributed signal measurements. *First*, how accurately can we locate wireless basestations using RSS measurements and known algorithms, and does accuracy correlate strongly with intuitive properties such as number or density of measurements? *Second*, can machine learning classifiers help improve location accuracy? *Third*, can we develop techniques to identify features or properties of highly accurate measurement instances, and use them to build techniques that produce more accurate results?

Our study uses several large public datasets of crowdsourced RSS measurements gathered by user smartphone apps around the globe through the OpenCellID [2] and OpenBMap [3] projects. They are unique for two reasons: they provide raw signal measurements (compared to aggregate coverage maps), and include ground truth of real basestation locations. In total, we analyze ∼1M cells and 419M signal measurements. Using the ground truth data and existing localization algorithms, we first quantify the predictive quality of crowdsourced data, *i.e.* how accurately can each measurement instance predict the basestation location? We then try to identify and improve the poor localization results by applying supervised learning. Finally, we try to identify key properties of measurement instances that correlate well with localization accuracy, by taking a novel application of unsupervised learning technique we call feature clustering.

2. CONCLUSION

We summarize our findings as follows:

- We apply seven popular basestation localization algorithms to our ground truth datasets, and find that localization results have very high variance across a number of factors, including algorithms, datasets, and scenarios. In addition, there is a significant variance in error even across cell instances in the same dataset.
- We apply ML classifiers to improve localization accuracy. While

overall accuracy is higher, error variance remains high, and our attempts to find key impactful features produce no clear results.

- We then take a novel application of unsupervised learning to identify hidden correlations in the data, which we call *feature clustering*. We define a distance metric between measurement instances based on similarity of their values in key features. Clustering the entire dataset based on pairwise distances produces key clusters that correlate features with localization accuracy of data inside them. From this, we identify RSS standard deviation and RSS-weighted dispersion mean as independent features that identify highly predictive data instances for sparse and dense measurement datasets.
- Finally, we develop an adaptive crowdsourcing technique using these two features. Applying this technique produces dramatic improvements in both increased localization accuracy and reduced variance. We also show that our results could generalize across datasets and geographic regions.

For detailed information, please refer to [9].

3. REFERENCES

[1] http://www.cnet.com/news/verizon-t-mobile-att-sprint-who-is-the-fastest-carrier-in-the-nation/.

[2] http://www.opencellid.org/.

[3] http://openbmap.org/.

[4] http://www.skyhookwireless.com/.

[5] http://www.antennasearch.com/.

[6] ACHTZEHN, A., ET AL. Crowdrem: Harnessing the power of the mobile crowd for flexible wireless network monitoring. In *Proc. of HotMobile* (2015).

[7] GEMBER, A., ET AL. Obtaining in-context measurements of cellular network performance. In *Proc. of IMC* (2012).

[8] LI, L., ET AL. Experiencing and handling the diversity in data density and environmental locality in an indoor positioning service. In *Proc. of MobiCom* (2014).

[9] LI, Z., ET AL. Identifying value in crowdsourced wireless signal measurements. In *WWW* (2017).

[10] LITTMAN, L., AND REVARE, B. New times, new methods: Upgrading spectrum enforcement. *Silicon Flatirons Center* (2014).

[11] PFAMMATTER, D., GIUSTINIANO, D., AND LENDERS, V. A software-defined sensor architecture for large-scale wideband spectrum monitoring. In *IPSN* (2015).

[12] RAI, A., CHINTALAPUDI, K. K., PADMANABHAN, V. N., AND SEN, R. Zee: zero-effort crowdsourcing for indoor localization. In *Proc. of MobiCom* (2012).

[13] SEN, S., ET AL. Can they hear me now?: a case for a client-assisted approach to monitoring wide-area wireless networks. In *IMC* (2011).