Mining Large Scale Cell Phone Data

Jean Bolot
Sprint
Burlingame, California, USA
http://jeanbolot.com/

ABSTRACT

Cell phones are ubiquitous in modern life and the call records collected by network operators are a powerful tool to study the behavior of cell phone users, and how those users use network resources, at previously impossible-to-achieve scales. In this paper we report on results from the analysis of large scale call records data, and more generally of the data generated by mobile users, at a large cellular operator. We consider in particular three kinds of data, namely social network data (who calls whom, how often, etc), location and mobility data (who is where) and spectrum data (who uses how much spectrum in which cell). We describe practical examples of insights derived from mining that data, the impact of the data on areas ranging from marketing to business models or to security, and also consider interesting research challenges ahead.

1. INTRODUCTION

The Internet has become a fundamental component of modern economies, and it provide services, starting with connectivity, that are strategic to companies, governments, families and individual users, and in general to the well functioning of modern life. A growing fraction of those services are accessed by mobile users. Indeed, the size and strategic importance of the mobile Internet, i.e. the Internet as accessed via mobile devices such as laptops or cell phones, is rapidly increasing. Recent reports indicate that the mobile Internet is ramping up in size faster than the "desktop Internet" did in the 80's and 90's; in fact, the estimated total value of the mobile data industry grew by 20% in 2009 - a year of major economic crisis when the global economy decreased by 5% - and mobile data revenues reached $284B [5]. This is now larger than the total PC Internet economy, including Internet content and advertising revenues plus all subscription fees such as monthly dial up and broadband access fees. Furthermore, the number of users of the mobile Internet (measured by the number of users accessing browser-based services on cell phones only) is estimated at between 500 million and 1 billion, almost on par with the total number of PCs connected to the Internet [5, 1]. Thus, cell phones already dominate the Internet, and their importance will continue to grow [4].

A key characteristic of cellular networks and devices is their ability to capture and analyze (at least partial) information on the behavior of mobile users. In particular, operators have routinely captured large scale location data for billing purposes, but also to improve location management or satisfy legal requirements such as E911. More recently they, as well as a number of analytics companies and academic research groups worldwide, have started analyzing a growing variety of data including social network data (who calls whom), location and mobility data (where users are when they call or use services), click-stream data (which sequence of sites users visit, or which sequence of applications and services they use), etc.

In this paper, we report on results from the analysis of such data carried out at a large cellular operator. We consider in particular three kinds of data, namely social network data, location data and spectrum usage data (who uses how much spectrum in which cell). We describe some of the insights derived from mining that data and consider some of the interesting research challenges ahead.

2. SOCIAL NETWORKS

We have analyzed a very large social network gathered from call details records, which reflects the voice and SMS interactions of more than ten million users through hundreds of millions of calls and SMS exchanges. We examined the distributions of the number of phone calls per customer; the total talk minutes per customer; and the distinct number of calling partners per customer. We found that these distributions are skewed, and that they significantly deviate from what would be expected by conventional wisdom, namely power-law and lognormal distributions.

We found instead that our observed distributions (number of calls, of distinct partners, and of total talk time) very closely fit a lesser known but more suitable distribution, namely the Double Pareto LogNormal (DPLN) distribution [6]. We found good fits over time (morning-evening, weekday-weekend) and space (US East Coast-West Coast, urban-suburban).

More importantly, we also found that our graph evolved over time in a way consistent with a generative process based on geometric Brownian motion. Furthermore, this generative process lends itself to a natural and appealing *social wealth* interpretation, and also allows for extrapolations and interpolations. We hope that our success with DPLN spurs further studies involving other datasets and their underlying generative processes. In particular, we hope that our “social wealth” interpretation and analysis will serve as an incentive for social scientists to study the large-scale evolutionary aspects of social characteristics. Indeed, we continue to collect data from our social network for longer-term analysis.

3. LOCATION AND MOBILITY

We have also analyzed call records to understand the mobility patterns of more than a million users over several thousand square miles. We made two contributions to the analysis of mobility patterns of cell phone users. First, using only coarse-grained location information, namely the location of the cell tower associated with a user at the beginning and end of each call, we examined the scaling laws of human mobility, in terms of distance and time. We found that both the distance traveled as well as the duration of calls (on periods) and pauses (off periods) are heavy tailed, in agreement with earlier results (e.g. [3]). However, we found that mobility patterns change during and in between calls, and that patterns are correlated over time, with strength of correlation dependent on activity.

Second, we developed a general technique, using tools from stochastic geometry and Bayesian statistics [7], to refine mobility models as more precise location information becomes available [11]. Thus, we can correct the distributions of distance traveled and direction as coarse location information is augmented by information such as distance to the associated cell tower, signal strength, location of neighboring cells towers, etc. To demonstrate the benefits of our technique, we first showed, using timing measurements from call records, that users are not uniformly distributed in cells. We then showed how that location information impacts the estimated distance distribution and then extended our earlier technique, illustrating the impact of increasingly more precise location information. Our approach is very general and applicable not just to cellular networks, but to other wireless networks such as wireless LANs (WiFi, ...) or adhoc networks.

4. SPECTRUM USAGE

Most existing studies of spectrum usage have been performed by actively sensing the energy levels in specific RF bands including cellular bands. Our approach has been to provide a unique, complementary analysis of cellular primary usage by analyzing a dataset collected inside a cellular network. One of the key aspects of our dataset, compared to others examined in related spectrum analysis, is its scale - it consists of data collected over three weeks at hundreds of base stations. We dissected this data along different dimensions to characterize and model primary usage as well as understand its temporal and spatial variations. Our analysis revealed several results that are relevant if Dynamic Spectrum Access (DSA) approaches are to be deployed for cellular frequency bands. For example, we found that call durations show significant deviations from the often-used exponential distribution. Though this can complicate the modeling of primary usage, we found that a random walk process, which does not use call durations, can be used for modeling the aggregate cell capacity. Another novel result we found is that spatial spectrum usage is highly non-uniform, especially during periods of high load, with clusters of sectors whose intra-cluster usage patterns are correlated.

We also considered the more fundamental problem of whether or not spectrum sensing is actually a viable approach to estimate when and how much secondary users can take advantage of available capacity. Indeed, sensing mechanisms that estimate the occupancy of wireless spectrum play a cricital role in enabling non-interfering secondary usage. The problem of designing such mechanisms is, therefore, crucial to the success of approaches based on Dynamic Spectrum Access. We developed key insights into this problem by empirically investigating the design of sensing mechanisms applied to check the availability of excess capacity in CDMA voice networks. We focussed on power-based sensing mechanisms since they are arguably the easiest and the most cost-effective.

We made three main contributions [9]. First, we found that accurate single sensor spectrum sensing is essentially unachievable, i.e. power at a single sensor is too noisy to help us accurately estimate unused capacity. However, we also found that there are well-defined signatures of call arrival and termination events. Using these signatures, we showed that we can derive lower bound estimates of unused capacity that are both useful (non-zero) and conservative (never exceed the true value). Finally, we used a combination of measurement data and analysis to deduce that multiple sensors are likely to be quite effective in eliminating the inaccuracies of single-sensor estimates.

5. FUTURE RESEARCH: BUSINESS MODELS

The capture and availability of large scale cell phone data has enabled, and will continue to enable, a wide range of new services. For example, in the case of location and mobility data, the capture and availability of such data has enabled the development of many location-based or location-aware services, and indeed an rapidly increasing number of such services is now available, ranging from navigation to location-aware advertising, friend finder, etc, and many more are announced or launched on a daily basis. However, this location data, since it enables new services and new economic activities, is seen as economically valuable. This raises the question then of how valuable it is, and how to quantify that value. This is precisely the goal of our recent research.

Using insights from cell phone data, we have developed an analytic framework, namely models and the techniques to solve them, to help quantify the economics of location information [2]. Our aim has been to derive models which can be used as decision making tools for entities interested in or involved in the location data economics chain, such as mobile operators or providers of location aware services (mobile advertising, etc). We considered in particular the fundamental problem of quantifying the value of different granularities of location information, for example how much more valuable is it to know the GPS location of a mobile user compared to only knowing the access point, or the cell tower, that the user is associated with. We have used our approach to derive insights into what is arguably the quintessential location-based service, namely proximity-based advertising.

To our knowledge, our work is the first one to present and analyze economic models which can help understand the economic value generated by mobile users with location based services, for different granularities of location information in wireless networks. We believe that the work provides an important first step towards a general analysis of not just the data itself, but also of the business models enabled by large scale cell phone data.

6. REFERENCES

[1] T. Ahonen, *Mobile as the 7th Mass Media*, London, UK: futuretext, 2008.

[2] F. Baccelli, J. Bolot, “Modeling the economic value of location and preference data of mobile users”, submitted for publication, Nov. 2009.

[3] M. Gonzalez, C. Hidalgo, and A. Barabasi, “Understanding individual human mobility patterns”, *Nature*, vol. 453, pp. 479–482, 2008.

[4] S. Keshav, “Why cell phones will dominate the future Internet, *Computer Communications Review*, vol. 35, no. 2, April 2005.

[5] M. Meeker, *The Mobile Internet Report*, Chichester: Morgan Stanley, Dec. 2009.

[6] M. Seshadri, S. Machiraju, A. Sridharan, J. Bolot, J. Leskovec, C. Faloutsos, “Mobile call graphs: Beyond power-law and lognormal distributions”, *Proc. ACM KDD Conference on Knowledge Discovery and Data Mining*, Las Vegas, Aug 2008.

[7] D. Stoyan, W. Kendall, and J. Mecke, *Stochastic Geometry and its Applications*, Wiley, 1995.

[8] D. Willkomm, S. Machiraju, J. Bolot, A. Wolisz, “Primary users in cellular networks: A large-scale measurement study”, *Proc. IEEE Symp. Dynamic Spectrum Access Networks (Dyspan)*, Chicago, IL, Oct. 2008.

[9] D. Willkomm, S. Machiraju, A. Wolisz, “The problem of spectrum sensing in cellular networks”, submitted for putblication, Mar 2010.

[10] H. Zang, J. Bolot, “Mining call and mobility data to improve paging efficiency in cellular networks”, *Proc. ACM Mobicom ’07*, Montreal, Canada, Sept. 2007.

[11] H. Zang, F. Baccelli, J. Bolot, “Bayesian inference for localization in cellular networks”, *Proc. IEEE Infocom 2010*, San Diego, CA, Apr. 2010.