Characterising Spatial Relationships in Base Station Resource Usage

Emmett Carolan, Seamus C. McLoone and Ronan Farrell
The Callan Institute
National University of Ireland Maynooth (NUIM)
Kildare, Ireland
ecarolan@eeng.nuim.ie, seamus.mcloone@eeng.nuim.ie & ronan.farrell@nuim.ie

Abstract - **This work preliminarily introduces an up to date measurement-driven examination of the spatial characteristics of network resource usage. The data set used is from a large nationwide 3G cellular network comprised of several thousand base stations. Firstly, we discuss our data set and how it can be appropriately used. Next, we examine the spatial correlation between base stations in terms of radio resource usage. We find significant spatial correlation, particularly for proximate base stations. We also examine the causality structure in the network using Granger causality and explore a metric for the identification of key indicator base stations. These indicator base stations act as hubs in the wider network and provide additional information about the future states of their neighbors. Finally, we conclude with a brief discussion of how we wish to build on this work.**

I. INTRODUCTION

In the past two decades mobile phones and devices utilising the mobile phone network have become ubiquitous in modern society. Mobile phone penetration has approached and in some nations exceeded 100% [1]. Cellular networks are undergoing and will continue to experience a large and sustained increase in demand for network resources [2]. As operators move to add capacity, a detailed understanding of the underlying dynamics of resource usage is ever more important. To this end, some recent works have begun to make use of large scale data sets provided by network operators to identify important facets of network usage [3-8]. This work provides a brief examination of spatially significant behavior with regards to resource usage from a network perspective. We aim to investigate i) the spatial correlation of resource usage from a network infrastructure perspective ii) identify key highly connected base stations that provide the most information about their local sub network. These topics are relevant to network providers in the areas of resource planning (hardware/spectrum), management and measurement.

The remainder of this paper is structured as follows: Section II will outline key information about our data set. Section III will focus on examining the spatial correlation between base stations and their radio resource usage. Section IV examines the causal structure present in the network and a way to identify the most influential base stations. Section V concludes our work with a brief discussion and a look to future work we hope to carry out.

II. DATA SET

Our data set is two weeks of nationwide Call Detail Records (CDRs) collected in 2011 from one of the Republic of Ireland's cellular phone networks. The data set includes information on all calls, SMS and cellular data usage of over one million people communicating on a network comprised of over ten thousand base stations. The privacy of individual subscribers is paramount, thus all personal information in the dataset is anonymised and cannot be used to identify individual customers. No information was provided relating to the content of any call, SMS or data session.

III. SPATIAL CORRELATION

In this section we examine how spatially correlated the network usage is. We find that there is a significant amount of spatial correlation present in the network. There are two main metrics used to describe resource usage (i) traffic load in terms of bytes [8] and (ii) airtime [9]. Traffic load in terms of bytes is problematic for our application as on our test network a small number of extremely high data users (mainly USB dongles and to a lesser extent bill pay smartphones – possibly tethered) were heavily skewing the data and masking underlying patterns (see Fig. 1). This was particularly problematic (especially at off-peak times) due to the fine granularity at which we were examining the network i.e. every hour & every fifteen minutes at the base station level. One possible method to mitigate this is outlined in [10] but would result in reduced spatial granularity.

Airtime as defined by [9] essentially quantifies the amount of time a subscriber uses radio and spectrum resources. In the 3G standard [11, 12] a subscriber requests and is allocated a radio channel when the subscriber has data to send. The allocated radio channel is revoked when the subscriber is inactive for a certain period of time defined by the inactivity timer (usually about 10 seconds) [13]. The value of the inactivity timer is configurable by the network operators [11, 12] and a subscriber can move between an active and dormant state multiple times within a single connection session. The airtime is thus defined as the amount of time a subscriber holds onto the radio channel (either in an active or dormant state).

Airtime is thus used as it is more closely related to the radio resource usage and less prone to swamping by a small group of voracious subscribers.

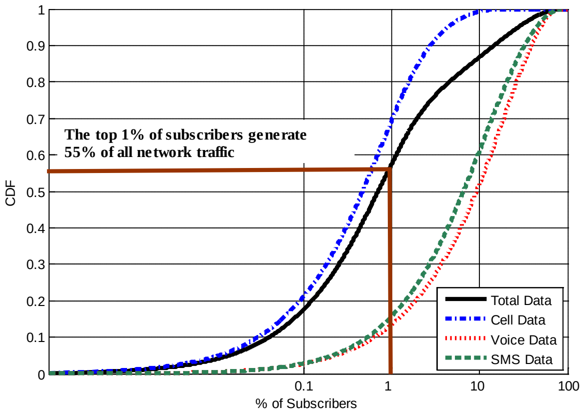

Fig. 1: CDF of normalised traffic over the percentage of subscribers (all subscribers). Note here voice and SMS are treated as an equivalent data service as explained in [8], cell data is the 3G cellular data in bytes and total data is the summation of all three expressed in bytes.

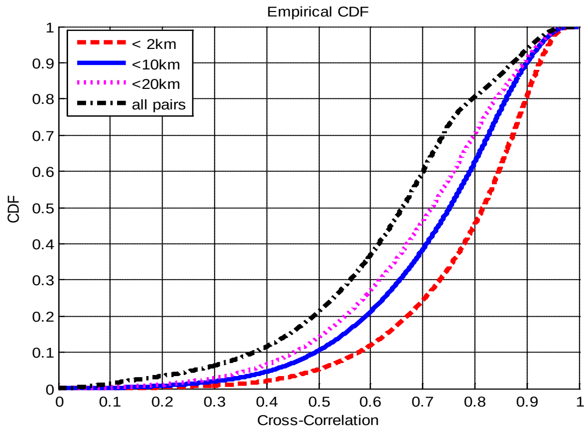

Fig. 2: CDF of the cross-correlation between all pairs of base stations and also within certain distance bands based on hourly airtime data over the course of two weeks.

Using the airtime for each base station we now investigate the extent of the spatial correlation on the network by cross-correlating pairs of base station's time series with one another. Cross-correlation is a widely used statistical method of measuring the similarity (the degree of correlation) between two time series [14]. Fig. 2 shows the cross-correlation calculated at zero lag for all base stations on the network and also for base stations based on certain distance ranges over two weeks of data at a granularity of one hour. Similar results were also obtained for the 15 minute interval but are omitted due to their similarity. The cross-correlation between base stations was found to be quite high with the one hour interval displaying slightly higher values than the 15 minute interval. The median cross-correlation was approximately 0.65 for the one hour interval and 0.5 for the 15 minute interval. 80% of base stations had a cross-correlation greater than or equal to 0.5 for the one hour interval. Cross-correlation was also found to depend on the distance between the base stations as shown by the groups in Fig. 2. For example, the median cross-correlation between cells within 2km of each other was 0.8 falling to 0.7 for all cells within 20km.

IV. CAUSALITY

To find the base stations that furnish us with the most information about their sub-networks future states we turn our attention to causality. The causal relationship between sub networks of base stations can provide extra information for the predication of traffic loads and thus allow for the appropriate allotment of spectrum in advance. Our chosen method for exploring this is Granger causality [8].

a) *Granger Causality*

Granger causality establishes if one time series is useful in forecasting another [8]. One stochastic variable X_2 Granger causes another stochastic variable X_1 if information in the past of X_2 helps predict the future of X_1 with a better accuracy then is possible with only the information in the past of X_1 alone [8]. Thus Granger causality is present in the direction from X_2 to X_1, provided that the inclusion of X_2 in a model improves the prediction of X_1 by a statistically significant amount. This relationship is not necessarily symmetric and thus 'X_2 Granger-causes X_1' does not imply that 'X_1 Granger-causes X_2' [9].

Formally, suppose we have two time series $X_1(t)$ & $X_2(t)$ both having a length T. As in [11] we can describe the two time series using a bivariate autoregressive model:

$$X_1(t) = \sum_{i=1}^{p} A_{11,i}X_1(t-1) + \sum_{i=1}^{p} A_{12,i}X_2(t-1) + \varepsilon_1(t).$$

$$X_2(t) = \sum_{i=1}^{p} A_{21,i}X_1(t-1) + \sum_{i=1}^{p} A_{22,i}X_2(t-1) + \varepsilon_2(t).$$

where $p < T$ is the model order i.e. the maximum number of lagged observations of X_2 used to predict the current value of X_1 or vice versa at time t. The matrix A contains the model coefficients while ε_1 & ε_2 are the residuals of the autoregressive model. X_2 granger causes X_1 if all the coefficients of A_{12} are non-zero i.e. if the residuals are reduced by the inclusion of the second time series in the model. In practice a threshold is set to determine if the relationship is significant. One such method is the F-test - to be considered statistically significant the F-statistic should be greater than some desired significance threshold ranging from 0 to 1 [11]. The closer the significance threshold is to zero the stricter the test. [11] provides two different methods of determining the model order, the Akaike Information Criterion (AIC) [12] and the Bayesian Information Criterion (BIC) [13].

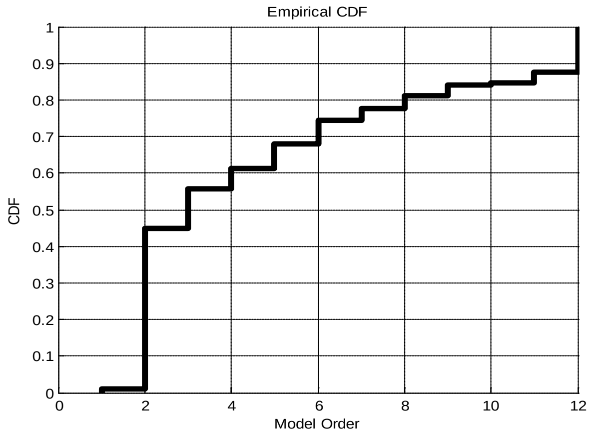

Fig: 3: CDF of the model order for each pair of neighboring base stations using the Akaike Information Criterion with a granularity of one hour.

Using [11] and in a similar fashion to [9] we find the model order using the Akaike Information Criterion as illustrated in Fig: 3. The model order is generally quite low with about 80% of pairings having an order of 8 or less. For the F-test of significance we set the critical value to 0.05. The causality is tested for every pair of neighboring base stations in both directions. On this network 38% of base stations pairs were found to have a statistically significant causal relationship in at least one direction at a granularity of one hour.

b) *Identifying Influential Base Stations*

To examine the network as a whole we create a causality graph using the pair-wise causal relationships [14]. The resulting graph of Granger causality interactions is a directed graph G = (V, E) where V is the set of vertices, E is the set of edges. Thus, each base station becomes a node on the graph and there is an edge from node a to b (i.e. (a,b) ϵ E) if there is a significant Granger causality interaction between them and they are neighbors in terms of coverage area [9].

The graph representation allows us to examine which base stations are the influencers and which are the influenced i.e. which base stations have a causal influence on their neighbors and which exhibit the results of this influence. A metric known as causal flow is used to quantify this relationship. The causal flow of a base station is the difference between the causal relationships it exerts on its neighbors and the causal relationships it experiences itself from its neighbors. In terms of the causality graph the causal flow of a node is the difference between its out-degree (number of edges emanating from the node) and in-degree (number of edges pointing to the node).

A node with a positive causal flow can be viewed as a source or an influential node while a node with a negative causal flow can be viewed as a sink or an influenced node. Fig. 4 illustrates the CDF of the causal flows of each base station on the network. We see that about 20% of base stations on the network have a large causal flow in either a positive (≥2) or negative direction (≤ -2).

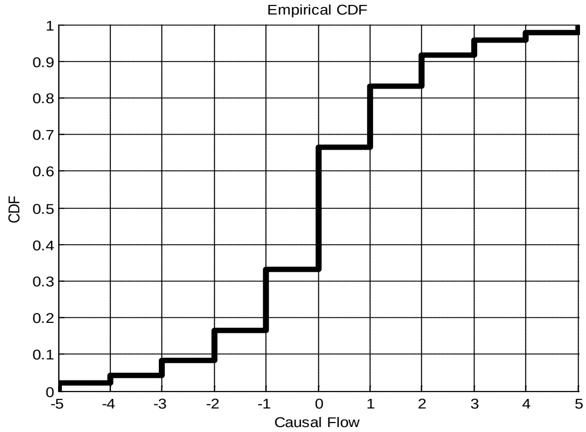

Fig. 4: CDF of causal flow

V. Conclusion & Future Work

This work was a preliminary exploration of the spatial characteristics of network usage and some important methods for identifying influental nodes of interest. A significant amount of spatial correlation was found for base stations in close proximity, dropping off as the seperation distance increases. Also a statistically significant causal structure was found in the network between 38% of neighbouring base stations. A metric for qualifing interesting base stations that act as either sources (influencers) or sinks on the network was also examined. In future work we aim to explore causal paths throughout the network and compare these with various forms of spatial data to look for any intersting trends (do paths follow transportaion networks, streets etc.). We also aim to further explore the properties of causal sources and sinks and identify the drivers of their behaviour. Anther interesting area of exploration is using the Granger causality relationships between base stations to inform models of spectrum usage in subnetworks. This would have applications in local spectrum allocation etc.

This work was supported through the Science Foundation Ireland Centre for Telecommunications Research (SFI-CE-I1853).

References

[1] Y. F. Chuang, "Pull-and-suck effects in Taiwan mobile phone subscribers switching intentions," *Telecommunications Policy*, vol. 35, pp. 128-140, 2011.

[2] Cisco, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2011–2016," http://www.cisco.com/en/US/solutions/collateral/ns

341/ns525/ns537/ns705/ns827/white_paper_c11-520862.html2012.

[3] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das, "Understanding traffic dynamics in cellular data networks," in *INFOCOM*, 2011, pp. 882-890.

[4] D. Willkomm, S. Machiraju, J. Bolot, and A. Wolisz, "Primary users in cellular networks: A large-scale measurement study," 2008, pp. 1-11.

[5] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wang, "Characterizing Geospatial Dynamics of Application Usage in a 3G Cellular Data Network," 2012.

[6] E. Carolan, S. McLoone, S. McLoone, and R. Farrell, "Analysing Ireland's Interurban Communication Network using Call Data Records," presented at the ISSC, NUI Maynooth, 2012.

[7] R. Farrell, E. Carolan, S. McLoone, C., and S. McLoone, F., "Towards a Quantitative Model of Mobile Phone Usage Ireland – a Preliminary Study," presented at the ISSC, NUI Maynooth, Ireland, 2012.

[8] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: Journal of the Econometric Society*, pp. 424-438, 1969.

[9] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das, "Understanding spatial relationships in resource usage in cellular data networks," 2012, pp. 244-249.

[10] J. Doyle, "Estimating Movement From Mbile Telepphony Data," 2013.

[11] A. K. Seth, "A MATLAB toolbox for Granger causal connectivity analysis," *Journal of neuroscience methods*, vol. 186, pp. 262-273, 2010.

[12] H. Akaike, "A new look at the statistical model identification," *Automatic Control, IEEE Transactions on*, vol. 19, pp. 716-723, 1974.

[13] G. Schwarz, "Estimating the dimension of a model," *The annals of statistics*, vol. 6, pp. 461-464, 1978.

[14] Y. Kim, R. Balani, H. Zhao, and M. B. Srivastava, "Granger causality analysis on ip traffic and circuit-level energy monitoring," in *Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building*, 2010, pp. 43-48.