# Exploring Spatial Relationships and Identifying Influential Nodes in Cellular Networks

**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 provides 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 its potential application. 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 examine the causality structure in the network using Granger causality to identify key influential indicator base stations within sub-networks. These indicator base stations act as hubs in the wider network and provide additional information about the future states of their neighbors. The penultimate section examines the influential indicator base stations in more detail. Finally, we conclude with a brief discussion of the key points and how we aim to progress this work.**

*Keywords* – **CDR, spatial usage, resource usage, cellular networks, network dynamics**

## 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 increasingly 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 [3-9]. This work provides an 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 the perspective of network infrastructure and 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 explores the influential base stations identified in Section IV. Section VI concludes our work with a brief discussion and a look to future work we aim to carry out.

## II DATA SET

Our data set consists of 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. Where appropriate, both voice calls and SMS are treated as an equivalent data service expressed in bytes and added to cellular data to get the Total Equivalent Data (TED). Voice is encoded in mobile phone networks using adaptive multirate (AMR) codecs. In GSM and wCDMA, a narrowband AMR scheme is used with a typical data rate of 12.2 kbps. A higher quality wideband AMR is used in LTE and offers superior quality at a data rate of 12.5 kbps [10, 11]. Higher and lower data rates are possible, but for this paper a rate of 12.5 kbps will be used in converting voice channels to an equivalent data session. Text messages will be treated as a 200 byte message with 1 second duration. 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 [9] and (ii) airtime [12]. 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 Figure 1). This was particularly challenging (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 considered by [13] but would result in reduced spatial granularity.

Airtime, as defined by [12], essentially quantifies the amount of time a subscriber uses radio and spectrum resources. In the 3G standard a subscriber requests, and is allocated, a radio channel when the subscriber has data to send [14]. 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 [15]. The value of the inactivity timer is configurable by the network operators [14] 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 therefore used as it is more closely related to radio resource usage and less prone to swamping by a small group of voracious subscribers.

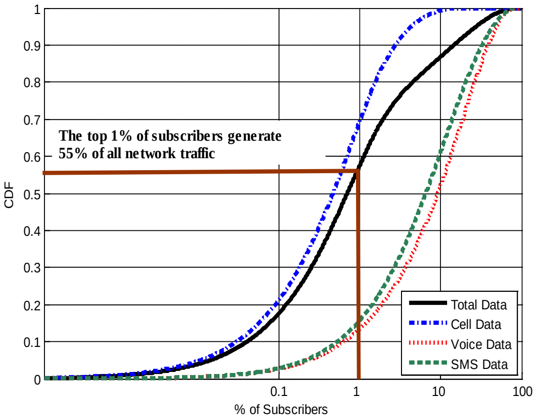

Figure 1: CDF of normalised traffic over the percentage of subscribers (all subscribers). Note that here voice and SMS are treated as an equivalent data service as explained in [8], cell data is the 3G cellular data in bytes and total data is the summation of all three expressed in bytes.

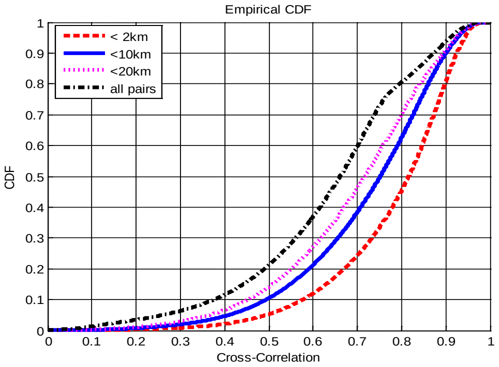

Figure 2: CDF of the cross-correlation between all pairs of base stations and also within certain distance bands based on hourly airtime data over the course of two weeks.

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 [16]. Figure 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 be dependent on the distance between the base stations as shown by the groups in Figure 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 identify base stations that provide the most information about their future sub-network usage, we now focus on 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 improves the of 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 than is possible with only the information in the past of $X_1$ alone [8]. Thus, Granger causality is present in the direction $X_2$ to $X_1$, provided that the inclusion of $X_2$ in the model improves the prediction of $X_1$ by a statistically significant amount. However, this relationship is not necessarily symmetrical and thus '$X_2$ Granger-causes $X_1$' does not imply that '$X_1$ Granger-causes $X_2$' [12].

For example, suppose we have two time series $X_1(t)$ and $X_2(t)$, both having a length of T. As in [17] 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 $\varepsilon_1$ & $\varepsilon_2$ 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 statistically significant. One such method is the F-test; to be considered statistically significant the F-value should be greater than a desired significance threshold ranging from 0 to 1 [17]. The closer the significance threshold is to zero the greater the significance of the result. The Akaike Information Criterion (AIC) [18] was used to estimate the model order.

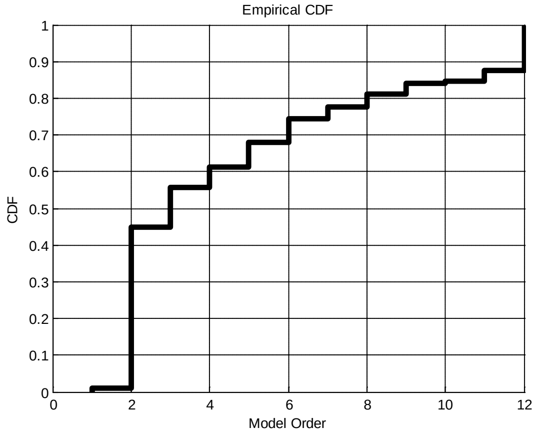

Figure 3: CDF of the model order for each pair of neighbouring base stations using the Akaike Information Criterion with a granularity of one hour.

Using [17] and in a similar fashion to [12] we find the model order using AIC as illustrated in Figure 3. The model order is generally quite low with about 80% of pairings having an order of 8 or less. This suggests that in most casses only a small number of previous samples from causally conneted neighbours are required. For the F-test of significance we set significance threshold level 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 [12]. 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 [12]. This causal graph allows for the exploration and quantification of some causal properties useful in identifying influential nodes [17]. These properties are outlined in the following subsections.

### c) *Causal Density*

Causal density is a global measure of the causal interactivity in a dynamic system. It shows the mean causality over the entire network. A high value of causal density indicates that the constituent parts of the network are coordinated in their activity [17]. It is the average G-causality over all the pairs of base stations examined. Causal density can take on a value between 0 and 1 and gives the average amount of significant Granger causality interactions over the entire network. Granger causality is defined using the causality graph:

$$Causal\ Density = \frac{\sum_{a\in V}\sum_{b\in V-(a)} I[(b,a)\in E]}{\sum_{a\in V}|N_a|}$$

where $N_a$ is the set of neighbours of the base station corresponding to node *a* and *I* is the indicator function [12]. On our network the causal density was found to be 0.38. Unit causal density is a related term used to investigate the local interaction for each base station [17]. Unit causal density quantifies how causally involved a base station is with its surrounding base stations. Unit causal density is the sum of a base stations interactions with its neighbours, normalised by its number of neighbours. A high unit causal density indicates that a node is a causal hub [17].

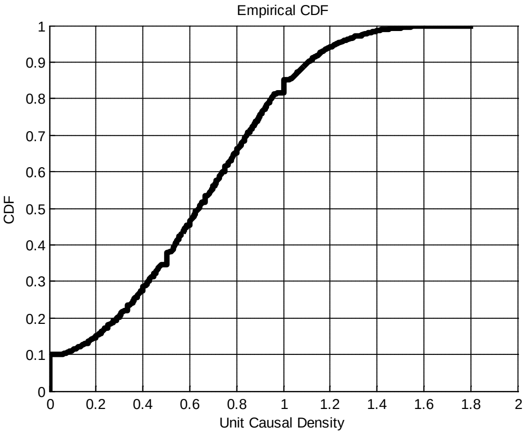

Figure 4: CDF of the unit causal density for each base station on the network.

Figure 4 illustrates the CDF of the unit causal density across the base stations on the network. The median of the causal density was found to be 0.64 with approximately 30% of base stations having a unit causal density of 0.8 or more.

### d) *Causal Flow*

The causal 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 neighbours and which exhibit the results of this influence. Using the causal graph representation, the influence emanating from node *a* is its out-degree (the number of edges going from node *a*). The influence node *a* experiences from its neighbours is given by node *a's* in-degree (the number of edges going into node *a*). Figure 5 illustrates the out and in degree of every node on the network. Note that some nodes have a very strong influence on their surroundings, for example, the top 5% of nodes have an out-degree of 15 or greater.

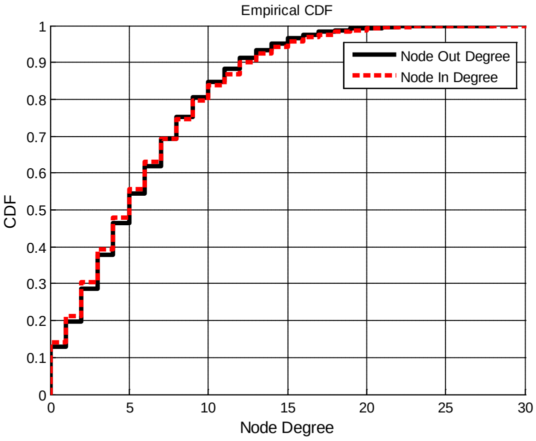

Figure 5: CDF of the out and in degree of every node on the network.

To get a more holistic view of the influence of a node while taking into account the influence it experiences, a metric known as the causal flow is employed. The causal flow of a node (base station) is the difference between the causal interaction it exerts on its neighbours and the causal interaction its neighbours, in turn, exert on it. Thus, on the causality graph, the causal flow is the difference between the node's out degree and its in-degree. Nodes with positive causal flows are causal sources while nodes with negative causal flows are causal sinks. The more positive or negative the flow is, the stronger the source or sink is respectively.

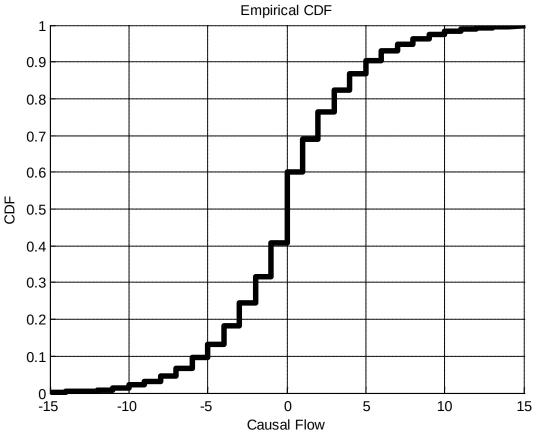

Figure 6: CDF of the causal flow of each base station on the network.

Figure 6 shows the CDF of the causal flow for each base station on the network. The information presented in Figure 6 can be used to identify causal sources and sinks in the network. For example, 10% of base stations are causal sources with causal flows greater than or equal to five. Conversely, 10% of base stations are causal sinks with flows less than or equal to negative five. The strong sources and sinks identified in Figure 6 will be further examined in the following section.

## V EXAMINING SOURCES AND SINKS

In the previous section base stations that exert/experience influence on/from their neighbours were identified. These base stations were known as sources and sinks respectively. In this section we examine these sources and sinks and compare them with each other and the general network to see if they have any special properties that stand out.

### a) *Sources, Sinks & Usage*

Figure 7 shows the CDF of each base station's total equivalent data (see section II Data Set) usage grouped by their causal flow. The three grouping are strong sources (top 10% of base stations ranked by causal flow), all base stations and strong sinks (bottom 10% of base stations ranked by causal flow). It is readily apparent that the strong sources experience much higher usage than the other two groups. For example, the median total equivalent data usage of a strong source base station is approximately 4.5 times that of the median for all base stations on the network.

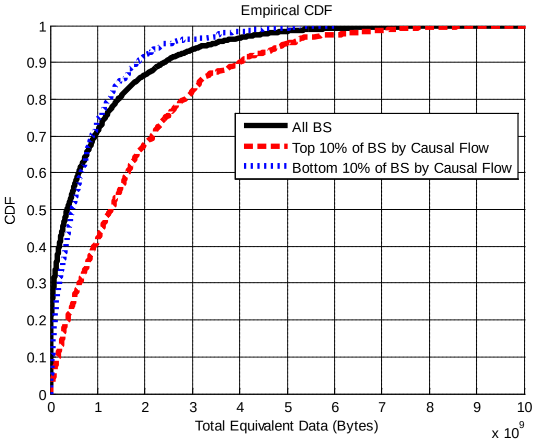

Figure 7: CDF of the Total Equivalent Data used per base station ranked by their Causal Flow. The top 10% represent strong sources while the bottom 10% represent strong sinks.

### b) *Sources, Sinks & Connections*

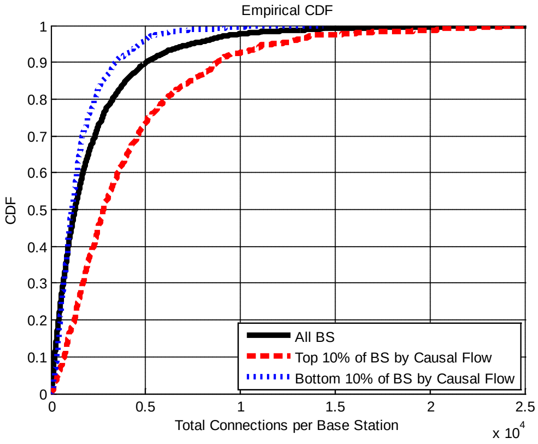

Figure 8: CDF of the total number of connections made per base station over one day ranked by their causal flow. The top 10% represent strong sources while the bottom 10% represent strong sinks.

Figure 8 shows the CDF of the total number of connections made per base station over one day as ranked by their causal flow. The top 10% represents strong sources while the bottom 10% represents strong sinks. Figure 8 illustrates that strong sources have a much larger amount of connections per day than the other groups. The median strong source base station has approximately 2.5 times the number of connections per day as the median of all base stations. Thus, strong source base stations generally use the most data and have the largest number of connections in a day.

## VI CONCLUSION & FUTURE WORK

This work explored the spatial characteristics of network usage and methods for identifying influential base stations in sub-networks. A significant amount of spatial correlation was found for base stations in close proximity, decreasing as the seperation distance increases. Signifiant spatial correlation indicates that for monitoring purposes it may only be nessecary to monitor a subset of base stations. Also, a statistically significant causal structure was found in the network between 38% of neighbouring base stations. The presence of significant causal relationships in the network indicates that load fore-casting techniques should utilise information based on the past load of neighbours for increased accuracy. A metric for qualifing interesting base stations that act as either sources (influencers) or sinks on the network was also examined and characterisitics of these sources identified. For example, it was found that influental base stations are generally more heavily utilised both in terms of traffic volume and subscriber connections than other base stations. This could possibly be due to the presence of transport routes, busy streets etc.

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 interesting trends (if 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 behaivour. Furthermore, future research focussing on using the Granger causality relationship between base stations to inform models of spectrum usage in sub-networks could have important applications for local spectrum allocation.

## 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/ns341/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] E. Carolan, S. C. McLoone, and R. Farrell, "Comparing and Contrasting Smartphone and Non-Smartphone Usage," presented at the ISSC, LYIT, 2013.

[10] B. Bessette, R. Salami, R. Lefebvre, M. Jelinek, J. Rotola-Pukkila, J. Vainio, H. Mikkola, and K. Jarvinen, "The adaptive multirate wideband speech codec (AMR-WB)," *Speech and Audio Processing, IEEE Transactions on*, vol. 10, pp. 620-636, 2002.

[11] H. Taddei, I. Varga, L. Gros, C. Quinquis, J. Y. Monfort, F. Mertz, and T. Clevorn, "Evaluation of AMR-NB and AMR-WB in packet switched conversational communications," in *Multimedia and Expo, 2004. ICME'04. 2004 IEEE International Conference on*, 2004, pp. 2003-2006.

[12] 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.

[13] J. Doyle, "Estimating Movement from Mobile Telephony Data," *NUI Maynooth*, 2013.

[14] *Third Generation Partnership Project 2 (3gpp2)*
Available: http://www.3gpp2.org/

[15] M. Chuah, W. Luo, and X. Zhang, "Impacts of inactivity timer values on UMTS system capacity," in *Wireless Communications and Networking Conference, 2002. WCNC2002. 2002 IEEE*, 2002, pp. 897-903.

[16] 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.

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

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

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