Analysing Ireland's Interurban Communication Network using Call Data Records

Emmett Carolan, Séamus C. McLoone, Seán F. McLoone and Ronan Farrell

The Callan Institute,

National University of Ireland Maynooth,

Maynooth, Co. Kildare, Rep. of Ireland.

email: ecarolan@eeng.nuim.ie, seamus.mcloone@eeng.nuim.ie,

sean.mcloone.nuim.ie, ronan.farrell@nuim.ie

***Abstract*—This work utilises data from an Irish mobile phone network to provide a preliminary, but novel, analysis of the interurban communication network between twenty five of the largest cities and towns in Ireland. An intuitive technique is applied to a mobile phone operator's call detail records to identify the actual subscriber population of different urban areas with various penetration rates. Weighted communication links are generated between the urban centres based on spatial and temporal metrics of distance, and are examined for different times of the day and for different days of the week. These communication links are compared to the output of a standard gravity model in order to ascertain the latter's ability to accurately represent Ireland's interurban communication network. The results obtained are presented and discussed within.**

***Keywords* – Call Detail Records, Network Analysis, Population Density Estimation, Gravity Model**

I INTRODUCTION

The pervasive nature of modern cellular network infrastructure has facilitated many interesting areas of research due to its seemingly omnipresent nature. Cellular networks produce a vast amount of records on a country wide scale. These records make it possible to observe behaviours and trends which have been previously impossible to view in their entirety.

A phone network's billing data gives information on the cell tower to which a phone is connected when a communication event takes place.[1] The geographic coordinates of each cell tower and its coverage area are known. This enables the creation of a list of cell towers that cover specific geographic regions of interest i.e. cities and towns. Thus, by examining all the billing records corresponding to a target city/town, it is possible to build a complete picture of the target's cellular communication.

In the case of calls and SMS, information on the cell tower to which the recipient is connected is also stored. Thus, knowledge of the source and destination locations, coupled with information on which towers cover which cities/towns, enables an understanding of the communication flow between cities/towns. With this information the cellular network can be analysed allowing for better informed phone network planning decisions.

This paper carries out a preliminary examination of the Irish situation, by focusing on the relationship between city/town size, interurban communication volume and distance. Different weighted communication links between the various cities/towns are obtained, based on two metrics for distance, namely spatial travel distance and travel time taken. In addition, these links are examined for working days and weekend days, as well as for working hours and non-working hours within a day. Finally, the standard gravity model is analysed in the context of modelling the communication links between the urban centres. Results suggest that it does not provide a suitable representation of the Irish scenario.

The rest of the paper is structured as follows. Section II highlights some relevant prior work in the area of communication network analysis. Section III outlines the techniques used to estimate the number

[1] *Communication events include calls, Short Messaging Services (SMS) and data services*

of subscribers in a specific town/city along with the interurban communication flows. It also includes background information on the data analysed in this paper. Section IV analyses these flows and, finally, section V summarises the findings of this work.

II RELATED MATERIAL

This work focuses on the estimation and analysis of the interurban communication network from a phone network's Call Detail/Data Records (CDR). Previous work has used similar CDR data to estimate population densities, predict communication intensities and to estimate traffic parameters [1 - 3].

It has been previously demonstrated that various systems can be represented as a network of nodes, connected by weighted or unweighted links [4]. It is a common technique to represent social networks as a network where each node represents a person and links between the nodes indicated social interactions. Onnela *et al*. [5] utilise a dataset similar to the CDR to highlight the importance of weak ties to the propagation of information through a communication network. Several other authors have made use of large recently available phone and email datasets to study human connections and behaviours [6 - 9]. Geographical information allows for a more detailed and interesting exploration of group and individual interactions. For example, Lambiotte *et al* use a mobile phone dataset to show that the probability of a call between two people decreases by the square of their distance [10].

Interurban connections, such as passenger flows and phone messages and their dependence on separation distance, have been studied for a considerable amount of time [11, 12]. In various economic and social networks, interactions between actors such as regions and countries has led to models similar to Newton's Gravity law, where the size of the actor plays the role of mass [13]. These *gravity models* take the following form:

$$W_{ij} = K\frac{M_i M_j}{d_{ij}^n} \quad (1)$$

Here, W_{ij} is the weight of the link between nodes i and j, d_{ij} is the distance between the nodes, M_i and M_j are the masses of the respective nodes, K is simply a constant term and n represents the order employed. Note, for traditional gravity models, $n = 2$.

Studies have also been carried out on road and airline networks between cities [14, 15]. In the case of road networks it appears that the gravity model holds for the strength of interactions.

Krings *et al*. [2] analyse a CDR dataset but, unlike Onnela *et al*. [5], they associate users with locations and aggregate links between users to links between locations. Furthermore, they explore how the strength of the links between locations varies relative to separation distance and population. They find that the strength of the link between locations is proportional to the populations at the locations and inversely proportional to the distance between the locations. Hence, they conclude that the inter-city communication intensity is characterised by a gravity model.

One limitation of this work, however, is that it relies on the billing address zip code provided by the subscribers to the network operator. All users in a specific zip code are aggregated and zip codes are aggregated to form cities. However, this introduces a potential source of error as users often provide unreliable information to service providers. Customers with a bill phone are obliged to submit correct home address details but there is no such guarantee with pre-pay users. This is particularly challenging due to the growth in popularity of pre-pay plans. Bill-pay customers currently account for just approximately 10% of the users in our dataset.

Finally, Kelly *et al.* [16] carry out their work on the Irish communications network. However, they only consider this network at a coarse level.

In this paper, we employ an intuitive algorithm to estimate target cities' populations and, subsequently, provide a more detailed investigation of the Irish communications network.

III ESTIMATION OF NETWORK PENETRATION AND COMMUNICATION LINKS

This work analysis CDR data from one of the Republic of Ireland's cellular phone networks, Meteor. The network serves over 1 million customers accounting for just less than one in four of the state's 4.6 million residents. Thus, this dataset is not a complete representation of the population. Unlike previous work [16], we do not assume that the Meteor phone network penetration is constant across all regions. Instead we calculate the actual population of users resident in each city at the times of interest and scale the data accordingly.

The CDR dataset utilized in this work was obtained over one week in February 2011. The data contains anonymous user ID information with exact time and cell location information for every event which occurs on the network. The events for which information is stored are all incoming and outgoing SMS, calls, Multimedia Messaging Services (MMSs) and data connections. Data used in previous work only utilised records of outgoing calls [1]. Thus, this work has a higher visibility of users as more events are detectable. This information is used to estimate the resident population in each town and the link strength between them, as described in the following sections.

As can be seen from equation (1), it is important that an accurate estimate of the population of the two nodes (cities) be made. We cannot assume that the network for which we have data has equal penetration across all cities studied. Thus, we must find an alternative to using census data. Here, we employ the use of an intuitive algorithm, as summarised by Algorithm 1 below. A similar algorithm could be used to estimate other places of interest, for example work locations, by replacing "home times" with "work times".

Algorithm 1: The home location estimation algorithm.
1: homeEvents = extract all events which occur at "home times" and group them by user id.
2: Load list of cell towers in each target city. If a user has no events associated with a cell tower in a target city then discard the user and events. The remaining users are those of interest.
3: For each user of interest count how many events occur within each cell tower.
4: Iterate through all users of interest and determine the most frequent cell for each user of interest for each day. Designate this cell as the user's home cell.
5: Count the number of users who have a home cell in each one of the respective city lists.

Step 1 extracts all events from the MySQL database (where all the CDR data is stored) that occur during the home hours via a MySQL query. The home hours used here refer to the time between 8pm and 7am on the nights Monday through to Thursday, inclusive. These times were chosen as it is more likely that people will be at home during these periods than on weekends. All the events occurring during home time were then grouped by user id.

In step 2, a list of cell towers of interest are loaded based on the cities that are under inspection. Each city has a separate list of cell towers that service the population of said city. All users who have not made or received a call/text or used a data connection in a cell of interest are discarded with a MySQL query. Thus step 1 significantly reduces the amount of data to be processed. Step 2 further reduces the computational overhead of later steps by reducing the user set from all users to only those of interest. This is an improvement on similar algorithms proposed in [1, 16] provided that only specific cities, towns or regions in the country are of interest.

In step 3 a MySQL query generates a dataset summarising how many events a user of interest makes or receives for each cell tower for each day. Step 4 iterates through all users of interest and determines the most frequent cell per user of interest for each day. This cell is then designated as the user's home cell. This is predicated on the assumption that during the specified home hours a user initiates most of their events in their home cell. If this cell is not in the list of cell towers of interest, generated in step 2, then the user is discarded. Thus step 4 further reduces the number of users to be considered. At this point only users who have a home cell location in one of the city lists generated in step 2 remain. Finally in step 5 the number of users with a home cell tower listed in each of the respective city cell tower lists is calculated and stored.

The algorithm was implemented as a mix of MySQL queries and C code to minimise execution time. In a similar manner to Kelly *et al*'s algorithm [16], our algorithm could be used to estimate work time locations for specific cities/regions. This could be achieved by altering step 1 to extract events during work hours as opposed to home hours.

Algorithm 1 does not provide a quantifiable, prefect set of results. Instead, it offers a probable, but arguably accurate, estimation of the user population of, in this case, each of the twenty five target cities in our study. Thus, we have information on M_i and M_j as given in equation (1). To verify this equation, values for the link weight W are also required. To generate the interurban communications network link weight, the total communications originating and terminating in a city are aggregated together. We can then define the weight of the link, W, between two cities α and β as:

$$W_{\alpha\beta} = \sum_{i\in\alpha, j\in\beta} w_{ij} \qquad (2)$$

where w_{ij} is a link between individual users in the respective cities. The weight of the links between the twenty five cities/towns is calculated for each of the seven days, including workdays and weekend days. The weight of the links between the cities/towns is also calculated for two times of interest during each day, namely work times (9am-4pm Monday-Thursday and 9am-3pm Friday) and home times (8pm-7am Monday-Thursday). The weight of the links is also calculated for daytime weekend (7am Saturday & Sunday to 8pm Saturday and Sunday) and night time weekend (8pm Saturday and Sunday to 7am Saturday and Sunday).

All of the above calculations are performed for three different metrics of link weight – number of calls between cities, total call time in seconds between cities and number of SMS between cities.

IV COMMUNICATION NETWORK EXPLORATION

As previously mentioned, Kelly *et al.* [16] performed a high level investigation of the gravity model on Ireland's communication network. They found that the gravity model approximates the actual data under their specific aggregations. They aggregated results over the period of a week and were dealing with the country as a whole. However, aggregation of time and population masks different degrees of compliance and non-compliance, depending on several factors. Equation (1) can be rearranged as follows:

$$d_{ij}^{n} = \frac{K\, M_i M_j}{W_{ij}} \qquad (3)$$

Using equation (3), results obtained can be tested for degrees of compliance with the gravity-style model, using linear regression.

Two different measures of distance were used when testing the gravity model. The first was the spatial metric, given by the travel distance between the centres of the two cities, and the second was the temporal metric, given by the travel time by road between the two cities. This latter metric offers an alternative and indirect measure of distance. Consider, for example, three towns X, Y and Z. Towns X and Y could be equidistant from town Z but it could take twice as long to travel to X from Z in comparison to travelling to Y, for various reasons including, road type, road quality, traffic volume, etc. This concept is readily captured by the temporal metric.

Results were obtained for all 25 towns/cities in this study. However, due to lack of space, only 2 set of the results will be shown. Furthermore, note that in all figures the correlation value R^2 represents the quality of fit to the given data points, where a value of 1 implies a perfect fit and a value of 0 implies no correlation between model and data.

Figure 1 compares one week of data plotted for distance as measured by both the spatial distance and by the travel time, for the cities of Cork and Dublin. The results illustrate that the gravity model performs better when distances are obtained using the spatial metric. This result is repeated for all the cities examined in the study with the agreement between the gravity-style model (represented here by a linear fit) and the results being on average 15% less when the temporal metric for distance is used.

The agreement between the results and the gravity-style model varies both with the day of the week and the time of the day. Figure 2 shows how the results change between the working week and the weekend for the cities of Cork and Dublin respectively. Overall, the gravity model performs worse for cities during the weekend with, on average, 10% less agreement between observation and the model, when compared with the working week. One possible explanation for this is the large amount of Irish people who work/study in the cities during the week and move back to the small towns/rural areas where they grew up on the weekends. There is also a small change in the agreement with the model based on the time of day. During the daytime/evening there is a slightly larger agreement between the gravity model and the results than at night. The effect is smaller than the weekday/weekend shift and is probably a result of non-residents being present in the city during daytime hours on weekdays and returning home outside the city at night.

The gravity model was also tested for three different types of communications links, namely total number of call connections made, total connection

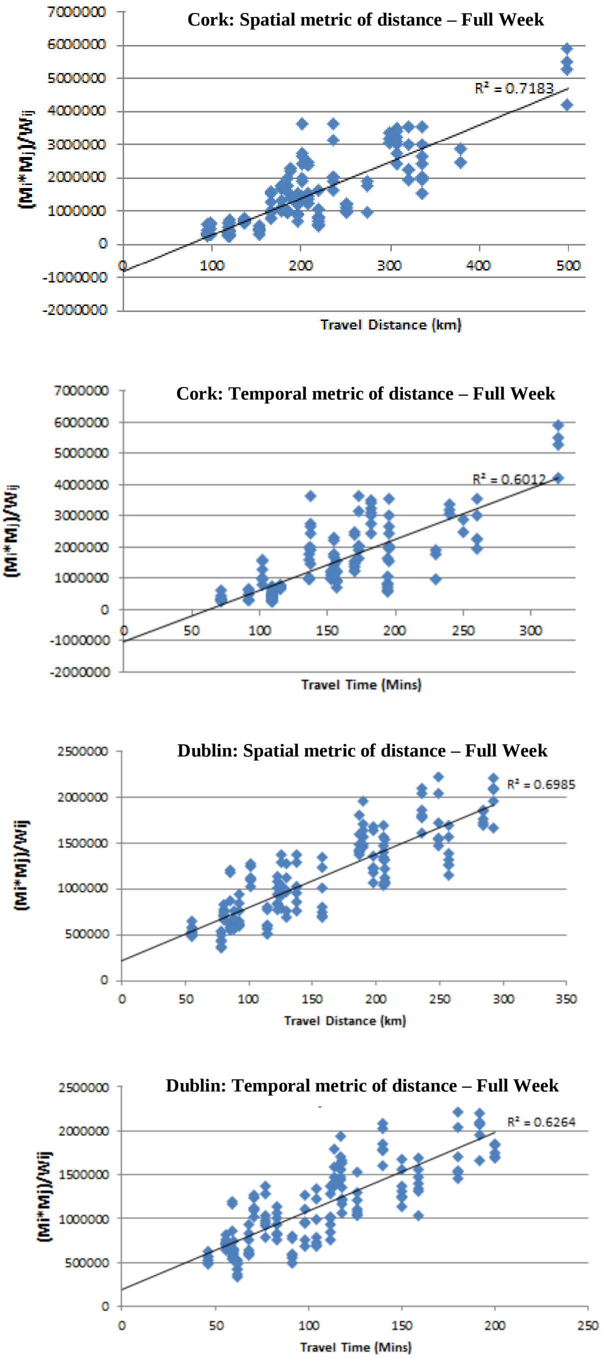

Figure 1: One full week of data plotted with two different measures of distance, for the cities of Cork and Dublin.

time of all calls and total number of SMS sent and received, the latter being used in all the figures shown. The greatest agreement with the gravity model was found when the total number of SMS was used. This result is repeated for all the cities examined in the study with the agreement between the gravity model and the results being on average 17% less when total number of connections or total call time is used. The reason for this difference is not

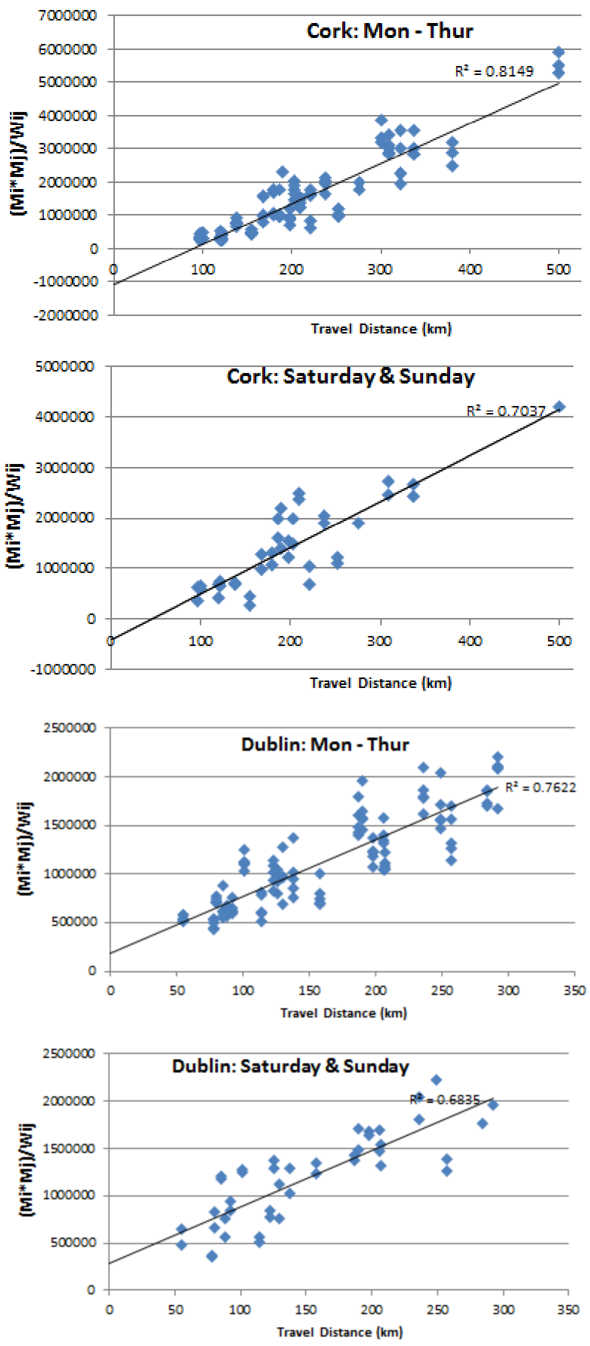

Figure 2: Change in communication patterns between working days and weekdays, for Cork and Dublin.

immediately obvious. Arguably, it could represent an underlying difference in communication behaviour between calls and SMS. However, it could also be a result of users sending on average over 4 times more text messages than making calls. Consider Figure 3 which shows the small town to small town communication. Here, it is clear that the weaker communication links (due to a smaller W_{ij}) leads to a poorer model. This small town to small town effect is even more pronounced when SMS is not considered, as the links are significantly weaker due to a large reduction in communication events.

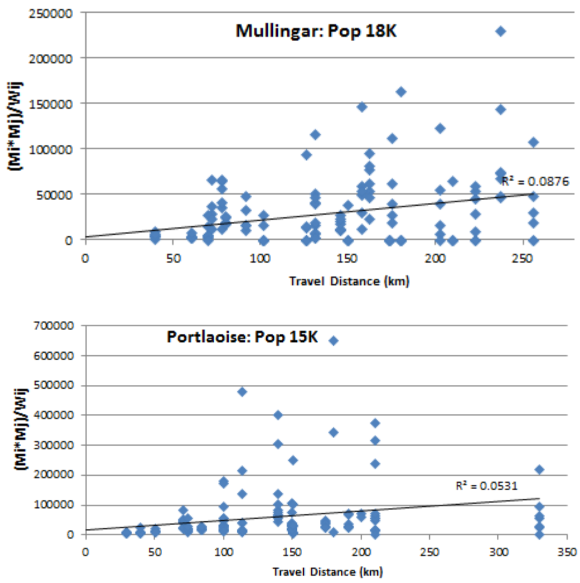

Figure 3: Small town to small town communication over one week, for Mullingar and Portlaoise.

Figures 1 and 2 strongly support a value of 1 for order n in equation (3). In other words, the results favour a linear model (as shown in the plots) over the use of the traditional gravity model of order $n = 2$. There are several possible reasons for this. Firstly, the Republic of Ireland's urban areas are separated by relatively small distances. This allows people to work/study in one part of the country while maintaining strong links with their relatively close places of birth. This large degree of mobility between urban areas would not be possible in a larger country.

Secondly, the relevance of the gravity model greatly depends on having at least one large population centre at either end of the communication link. There are two main interurban communication scenarios considered. The first is when a large population is present on either side of the link (large population communicating with small, small to large and large to large). This always provides the best fit with the gravity model (see Figure 1) even when taking into account variations due to the time of the week (see Figure 2) or time of the day. The second population scenario is where there is no large population centre on either side of the link (smaller town to smaller town). This primarily affects the smaller towns with populations of less than 50,000 inhabitants (see Figure 3). This scenario is prevalent in Ireland due to many of Ireland's urban areas being relatively small by international standards. The Republic of Ireland only has five cities with a population greater than 50,000 inhabitants. Thus for the remainder of the Republic's urban areas the gravity model is a poor choice for modelling interurban communication.

This is a key difference between our study and that of Kelly *et al.* [16], which shows an approximate national agreement with the gravity model. In their conclusion, the authors state that their "*work has focused on county-level interaction"*. Of the twenty six counties of the Republic of Ireland covered in their study only two have a population of less than 50,000, with most having significantly more [17]. Thus, the gravity model is only relevant when dealing with sufficiently large populations, either concentrated in a large urban area or more widely spread out over a larger region.

V CONCLUSIONS AND FUTURE WORK

The purpose of this paper was to provide a preliminary analysis of the Irish interurban communication network. The gravity model was selected and tested as one possible model for the Republic's interurban communication. The performance of the model was found to vary largely based on the type of link chosen/the time of the week and to a lesser extent the time of day. The results obtained support a linear relationship over that of a gravity model.

The gravity model may be more suited to static landlines than mobile phones. The simplicity of the model does not take into account the highly mobile nature of the Irish population. This is probably exacerbated by the relatively small size of the Republic. This small size facilitates people working/studying in one area during the week while maintaining strong links to their place of origin.

The gravity model was found to be only helpful when dealing with large population centres of more than 50,000 inhabitants. As the Republic of Ireland only has five cities with a population of 50,000 inhabitants or more, it is deemed a poor choice for modelling interurban communication between the country's smaller urban centres.

In future interurban work, smaller population centres should be amalgamated into larger groups or a more sophisticated model should be employed.

VI ACKNOWLEDGEMENTS

This work has been supported through the SFI Centre for Telecommunications Research (SFI-CE-I1853). We gratefully acknowledge the support of Meteor for all their assistance with this project.

VII REFERENCES

[1] R. Ahas, S. Slim, O. Jaumlrv, E. Saluveer, and M. Tiru, "Using mobile positioning data to model locations meaningful to users of mobile phones", *Journal of Urban Technology*,17(1):3 – 27, 2010.

[2] G. Krings, F. Calabrese, C. Ratti, and V. D. Blondel. "Urban gravity: a model for inter-city telecommunication flows", *Journal of Statistical Mechanics: Theory and Experiment*, 2009(07):1-8.

[3] N. Caceres, J.P. Widberg, and F.G. Benitez. "Review of traffic data estimations extracted from cellular networks", *IET Intelligent Transport Systems,* 2(3):179-192, 2008.

[4] A.L. Barabsi and R. Albert. "Emergence of scaling in random networks", *Science*, 286:509-512, 1999.

[5] J.P. Onnela, J. Saramki, J. Hyvnen, G. Szab, D. Lazer, K. Kaski, J. Kertsz, and A.L. Barabsi." Structure and tie strength in mobile communication networks", *Proc Natl Acad Sci USA*, 104(18):7332-7336, May 2007.

[6] P.S. Dodds, R. Muhamad, and D.J. Watts. "An Experimntal Study of Search in Global Social Networks.", *Science*, 301(5634):827-829, 2003.

[7] M.C. González, C.A. Hidalgo, and A.L. Barabási. "Understanding individual human mobility patterns.", *Nature*, 453(7196):779-782, 2008.

[8] J.P. Onnela, J. Saramaki, J. Hyvonen, G. Szabo, M.A. de Menezes, K. Kaski, A.L. Barabasi, and J. Kertesz. "Analysis of a large-scale weighted network of one-to-one human communication." *New Journal of Physics,* 9(6):179, 2007.

[9] Mukund Seshadri, Sridhar Machiraju, Ashwin Sridharan, Jean Bolot, Christos Faloutdod, and Jure Leskove. "Mobile call graphs: beyond power-law and lognormal distributions." In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining,* 596-604, NY, USA, 2008.

[10] R. Lambiotte, V.D. Blondel, C. de Kerchove, E. Huens, C. Prieur, Z. Smoreda, and P. Van Dooren. "Geographical dispersal of mobile communication networks", *Physica A: Statistical Mechanics and its Applications,* 387(21):5317-5325, 2008.

[11] G.K. Zipf. "Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology". Addison-Wesley Press, 1949.

[12] Wayne K. Davies. "Urban Connectivity in Montana", *The Annals of Regional Science,* 13(2):29-46, 1979.

[13] V. Carrothers. "A historical review of the gravity and potential concepts of human relations," *Journal of the American Institute of Planners,* 22:94-102, 1958.

[14] W.S. Jung, F. Wang, and H.E. Stanley. "Gravity model in the Korean Highway.", *EPL-Europhysics Letters,* 81(4):48005-48005, 2008.

[15] A. Barrat, M. Barthelemy, R. Pastor-Satorras, and A. Vespignani. "The Architecture of complex weighted networks.", *Proceedings of the National Academy of Sciences,* 101(11):3747-3752, 2004.

[16] D. Kelly, J. Doyle, and R. Farrell. "Analysisng Ireland's Social and Transport Networks using Sparse Cellular Network Data", *ISSC 2011, Trinity College Dublin, Ireland*.

[17] Central Statistics Office (2011), "Census 2011 Preliminary Results*"*. Available: http://www.cso.ie/en/media/csoie/census/documents/Prelim%20complete.pdf, last accessed 20/03/2012.