Analysing Ireland's Social and Transport Networks using Sparse Cellular Network Data

Damian Kelly†, John Doyle and Ronan Farrell

Strategic Cluster for Advanced Geotechnologies,
Callan Institute,
NUI Maynooth, Ireland.

E-mail: †dkelly@eeng.nuim.ie

Abstract — **This work utilises data from a mobile phone network to observe the communication and movement patterns of the Irish population between specific locations. A novel technique is applied to mobile phone network billing data to estimate the distribution of the phone network population across the country at key times, namely home times and work times. From these population distributions, sets of weighted transportation and communications links between locations are generated. These networks are analysed using indicators commonly employed in network theory such as clustering coefficients and betweenness centrality. Furthermore, the ability to generate such weighted links when phone network data is unavailable is investigated.**

Keywords — **Population Density Estimation, Network Analysis, Call Detail Records, Spatiotemporal Data Mining.**

I Introduction

The ubiquity of cellular network infrastructure is leading to interesting avenues of research due to its prevalence in many aspects of modern life. By viewing human movements and interactions from the perspective of a cellular network it is possible to obtain a broad understanding of behaviours which have previously been impossible to observe in their entirety. When a phone network's billing data is considered, two types of behaviour are observable; movement and communication.

A phone network's billing data provides details of the cell tower to which a phone is connected when a communication event takes place.[1] This enables the estimation of locations of interest for users, such as locations where they live and locations where they work. This knowledge allows an estimation of the strength of the transport links between locations. Hence, road network planners can exploit this information to enable better planning decisions. Billing data also enables visibility of the strength of the communication connections between individuals within the network. Furthermore, knowledge of the source and destination locations of these connections enables an understanding of the volume of traffic between locations. With this information the phone network can be analysed, enabling more informed phone network planning decisions.

As a result, this paper analyses the Irish landscape in two aspects; the transport network and the communication network. Section II highlights relevant prior work in the areas of communication and transport network analysis. Section III presents a novel technique of inferring the locations of interest for individuals. These locations of interest, along with a social network structure, are then used to infer the communication and transport networks. Section IV analyses the networks on a high-level inter-county scale and attempts to derive the networks in the absence of phone billing data. Finally, V summarises the findings of this work.

II Background

This work focuses on the estimation and analysis of the communication and transport networks from a

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

phone network's Call Detail Records (CDR). Previous work has used similar CDR data to estimate traffic parameters [1], predict communication intensities [2] and estimate population densities [3]. This work combines elements from that work to characterise the Irish road and telecommunication networks.

It has been shown in the past that many systems can be represented as a network of nodes, connected by weighted or unweighted links [4]. Social networks have commonly been represented as a network in which each node is a person and links between the nodes represent social connections. Onnela *et al.* [5] use a CDR-like dataset to illustrate the importance of weak ties to the propagation of information through a communication network. Krings *et al.* [2] also consider a CDR dataset. Unlike Onnela *et al.* they associate users with locations and aggregate links between users to links between geographical locations. They confirm that the strength of the links between locations is proportional to the populations at the locations and inversely proportional to the distance between the locations. Hence, this is referred to as a "gravity model".

Transport networks have also been considered in the past from a network theory perspective. Latora and Marchiori [6] use a network representation to evaluate the efficiency of the Boston subway system. A larger transport system is considered by Colizza *et al.* [7] who use a database of the international links between airports to simulate the international spreading of a hypothetical epidemic. In this sense, the spreading of an epidemic through a physical network can be considered analogous to the spreading of news through a social network.

Other work, which analyses transport networks in more detail, utilises more sophisticated techniques of acquiring network data, such as travel surveys [8] or video camera networks [9]. Hence, the acquisition of more detailed transport network data requires significant effort or sensing infrastructure. CDR data has been used in the past to estimate communication network configuration, but has never been utilised to generate transport network representations. Due to the omnipresence of mobile phone networks, the development of techniques to estimate individual's locations of interest from CDR data would enable high-resolution transport analysis to take place with little data acquisition effort. It would also allow communication network analysis to take place in the common scenario where users' declared home locations are inaccurate or incomplete. Hence, the aim of this work is to generate estimates of locations of interest for individuals to estimate the strengths of links between network nodes, both for transport networks and communication networks.

III Estimation of Locations of Interest

Previous work which used CDR data for network analysis [2] had the advantage that all customer's home locations were known, on the level of home zip code. One disadvantage of such data is that it only indicates home locations, it does not suggest other locations, such as work or recreational locations. Another potential issue with such data is the reliability of the user's stated home address. Customers with a bill phone must submit correct home address details. However, as a result of the recent growth in popularity of pre-pay mobile phone plans, bill-pay customers account for just over 10% of the users in our dataset. Hence, there are deficiencies in the resolution and reliability of pre-pay customers' stated place of residence, as exmplified by the encountered home addresses; "Unknown Address", "Fake Street" and "Ballyfake".

Hence, before communication and transport network analysis can take place it is necessary to estimate individual user's home and work locations from CDR data, by translating the daily sets of events from each user into an estimate of home cell tower and work cell tower. Two techniques of home location estimation are considered, one which has recently been suggested in literature and another which is optimised for our comparatively large dataset. The latter technique is then extended to the estimation of user work locations.

a) Call Detail Records

This work utilises CDR data from one of the Republic of Ireland's cellular phone networks, Meteor. This network has just over 1 million customers which represents just under a quarter of the country's 4.5 million inhabitants. Hence, this dataset is not a complete representation of the population, rather suggesting the efficacy of our techniques when further mobile phone companies' data becomes available. For this early stage of the research we make the somewhat naive assumption that the Meteor phone network penetration is constant across all regions.

The CDR dataset which we employ for this work was obtained for a week in February 2011. This CDR data consists of a consistently anonymised user ID, exact time and cell location information for every event which occurs on the network. The events include all incoming and outgoing calls, SMSs, Multimedia Messaging Services (MMSs) and data connections. The data used in previous work only utilises records of outgoing calls [3], so this work has higher visibility of users since more events are detectable. This data is used to estimate home location in two ways, as described in the proceeding sections.

b) Prior Technique

The work conducted by Ahas *et al.* [3] takes an approach similar to ours in that they are using CDR data to estimate people's homes throughout the study country, Estonia. The application of the work is to determine multiple places of interest such as work and home locations with the home location densities compared with census data.

The central steps of the algorithm detailed in [3] are as follows. For each user, a list of regularly visited cells is compiled. Regular cells are defined to be cells in which calls are made on two separate days each month. Next, regular cells which are used on less than 7 days a month are removed from the list. Users with too many calls are then removed because they are likely to be organised call procedures and users with too few calls are removed since they would not have enough information to make an informed decision on their places of interest. The regular cells are organised in descending order of number days on which they are detected. For each user, the two regular cells which are used on the most days are analysed. If the standard deviation of the call times at a given cell tower on a given day is above 0.175 the cell tower is said to be the home cell. If the standard deviation is below 0.175 but with the average call time is after 5pm the cell is also said to be a home cell. Otherwise the cell is a work location.

If neither of the two most frequent regular cells are classified as home cells, the the remaining regular cells are evaluated for the relevant event time mean and standard deviation. If none of these cells fit the call time profile, this user is discarded. This recursive step of including further cell towers is the most costly part of the algorithm since a large number of cell towers may need to be evaluated for call time statistics. It is important to note that this is only a summary of the algorithm, for a more detailed outline the reader should refer to [3].

c) Our Technique

A significant implementation overhead of the original algorithm is the fact that all data for an individual user is retrieved and then the statistics of the most frequent cell towers are recursively estimated on each cell tower until the necessary profile is met. Our algorithm takes a different approach. Our algorithm is outlined in Algorithm 1. Step 1 takes the initial dataset and extracts only events which occur during the specified "home times" with a single MySQL query. The "home times" are defined to be between the hours of 8pm on a particular night and 7am the following morning on the nights of Monday through to Thursday. Only these nights are considered since it is more likely that users spend significant quantities of time in non-home locations during these times at weekends than during the middle of the week.

Algorithm 1 Our home location estimation algorithm. A similar algorithm is used to estimate work-time locations which uses "work times" instead of "home times".

```
1: homeEvents ← all events which occur at "home times"
2: homeEventsSummary ← summarise homeEvents in terms of distinctWeekday, distinctCell, corresponding eventCount
3: users ← distinct users from homeEventsSummary
4: for user ← each users do
5:    for day ← each distinctWeekday do
6:       distinctCells,eventCounts
         ← homeEventsSummary(user,day)
7:       dailyCells[day]
         ← distinctCells(argmax(eventCounts))
8:    homeCell ← mode(dailyCells)
```

Step 2 uses a MySQL query to efficiently generate a dataset which summarises the number of events a user makes or receives for each cell tower on each day. Step 3 compiles a list of users to be evaluated in step 4. Step 4 iterates through all users and determines the most frequent cell for each day. Then step 8 uses a majority vote of each day's most frequent cell to determine the home of each user. This is based on the assumption that during the specified home hours a user makes the most of their events at a home location. Hence, steps 1,2 and 3 are executed using MySQL queries to reduce the data remaining to process in the proceeding steps. Steps 4 to 8 are implemented in Python for simplicity but could be implemented in C or C++ to minimise execution time.

To determine the work-time locations an algorithm similar to Algorithm 1 is implemented which utilises work times rather than home times. We define work times to be 9am-4pm on Monday to Thursday and 9am-3pm on Friday. It should be noted that we consider these estimates to be work-time locations rather than actual work places since it cannot be assumed that all individuals work during these times. The dataset will include people who work at home, people who work irregular shifts and unemployed people, for example. Instead it is intended to provide an estimate of large-scale movements between locations which are typically a result of work-time schedules.

d) Technique Comparison

Both techniques are applied to the available Irish CDR data. County-level home and work locations are estimated by approximating an individual's home or work county to be the county which

their home or work cell tower falls within.

To appraise the performance of the algorithms the home location estimates are compared with the CSO data. Unfortunately, the most recent CSO data was obtained in 2006 and the CDR data for these experiments was obtained in 2011, so there will be unavoidable discrepancies in the results. It should be noted that population projections exist after 2006 but they are not on the geographical resolution necessary for this study, hence the population density estimates are directly compared with the 2006 census data.

The proportion of the total population residing in each area is considered for all data sources rather than people count, since people counts are not directly comparable across data sources, due to the fact that the phone network only represents a subset of the entire CSO population. Hence, for this early iteration of this work we must assume that the phone network users are a uniformly sampled subset of the entire population. Accordingly, when comparing the accuracy of home county densities we compare the proportion of the population within the counties rather than people counts. The error for each technique is the difference between CSO population proportion and the estimated population proportion, normalised by the CSO population proportion.

Table 1: A technique accuracy comparison for county-level home location estimation.

	Mean Error	Correlation [3]
Ahas Tech	0.3555	0.9852
Our Tech	0.3614	0.9843

Table 1 summarises the mean errors for the Ahas technique and our technique for county-level population estimates. It can be seen that the difference in performance between the two techniques is negligible. The correlation measure used by Ahas *et al.* is also included in this table. This measure simply indicates the correlation between CSO data and the population predictions per area. Even though we are using census data which is five years older than the phone data, we obtain a correlation value similar to previous work which uses census data from the same period. It is assumed that the work home location predictions have similar fidelity. More detailed information on these techniques can be found in [10].

e) Communication Network and Transport Network Estimation

The result of the location estimation algorithm is a home location and a work location for each user. With this it is possible to construct an Origin-Destination (O-D) matrix for the home-work routes, similar to that employed by [8]. The phone network has 1346 distinct cell tower sites, hence, the O-D matrix is a 1346x1346 matrix. This O-D structure is then converted to a network structure in which each node represents a cell tower location and each edge represents a transport route. The weight of each edge is the number of people who take that route. Montis *et al.* obtained their network of 375 locations using a survey of people's travel behaviours, whereas we obtained our network of 1346 nodes using CDR data which is considerably easier to obtain.

To generate the communication network we summarise all links between all pairs of users for a selected day in the same week in February for which the home locations were estimated. All user pairs which only have a single event are discarded, since these are possibly erroneous communications. We then determine the corresponding home locations for both users in each link. Then the weight of link between locations is the number of minutes of calls between the people who live in those areas. A separate dataset is also created where the edge weights are the number of SMSs, which is compared to the call network in Section IV. We aggregate by home location rather than the location that the calls were made from since the likelihood of a link between users is more dependent on their home locations than the location they happen to be in when the call occurs.

IV High-Level Network Exploration

For a high level exploration of the Irish communication and transport network we consider the network on the county level. The original nodal structure with 1346 nodes is aggregated to a 26 node structure where each node corresponds to each county in the Republic of Ireland.

a) Network Analysis

The two main aspects in which we compare the communication and transport networks are clustering coefficients and betweenness centrality. Betweenness centrality for a given node indicates what proportion of routes between arbitrary nodes are routed though the given node. Hence, removing a node with high betweenness centrality will reduce the efficiency of the network, or the ease at which one node can be reached from another node. For a road network this would reduce the speed at which epidemics may spread and for a social network it would affect the speed at which news spreads.

For our datasets we evaluated the betweenness centrality for the top 20% strongest links. Intuitively the most central county for both transport and communication networks is Dublin. The next most central counties for the communication net-

work are the large cities such as Cork and Galway. For transport network, on the other hand, the most central counties are the counties which are spatially central to the country, such as Meath, Westmeath and Waterford. Hence, this confirms that communication networks are independant of the spatially intermediate locations, whereas the transport networks depend heavily on the intermediate links between counties.

Clustering coefficients represent the number of links a node has to its neighbours as a ratio of the total number of neighbouring nodes. Hence, it indicates the level of cohesiveness around each node [8]. Unlike betweenness centrality, the clustering coefficients follow the same profile for both the communication and transport networks. For both types of network, the highest clustering coefficients are for Dublin. The next highest coefficients are for the counties surrounding Dublin, such as Meath and Kildare. Following that the next highest coefficients are for the major cities, Limerick, Cork and Galway. The parity between coefficients for both types of networks suggests that one network caused the emergence of the other, most likely the clustering of the geographical network caused the clustering of the social network.

b) Urban Gravity Model

This type of network analysis is only possible when the interactions between locations is detectable. Hence, it may be difficult to perform the same analysis for countries where CDR data is unavailable. Hence, this section attempts to estimate the link strengths from auxiliary information. CDR data was used to generate the link strengths between locations. Previous work represented transportation link strength [9] and communication link strength [2] between locations as a "gravity model";

$$W_{ij} = K\frac{M_i M_j}{d_{ij}^2}, \quad (1)$$

where W_{ij} is the weight of the link between node i and node j, d_{ij} is the distance between the nodes and M_i and M_j are the masses of the nodes. The masses of the nodes are the total populations at those nodes, calculated from CDR data. For the communication network M_i and M_j are both calculated from the home locations, whereas for the transportation network M_i is calculated from the home locations and M_j is calculated from the work locations.

Hence, when CDR data is unavailable, M_i and M_j can be calculated for the communication network from census data. For the transportation network, M_j must be calculated from survey data, such as the Places of Work Census of Anonymised Records (POWCAR) data. Figures 1.(a) and 1.(b) illustrate the efficacy of this model for the communications network and transportation network respectively when this model is applied. The constant, K, is estimated using linear regression and the black dashed line indicates the mean actual weights for the corresponding weights predicted using this model. For both types of networks, the line does not fully follow the theoretically perfect model, represented by the solid black line, hence the model would benefit from modifications.

c) Optimised Gravity Model

Previous work [4] utilised different exponents for different types of networks. Hence, we can rearrange Equation 1 to the form;

$$W_{ij} = \exp(a_0 + a_1 \log M_i + a_2 \log M_j - a_3 \log d), \quad (2)$$

where $a_0 = \log K$. This allows the distance exponent, a_3, to be optimised, while also introducing location mass exponents, a_1 and a_2. These mass exponents provide extra degrees of freedom which can account for the disparity between our estimated location masses and the true location masses.

These parameters are optimised using linear regression and the performance for the new weight prediction model is presented in Figure 1.(c) and 1.(d). Hence, the model in Equation 2 provides a appropriate fit than that in Equation 1. The optimised distance exponents are 1.68 for the communication network and 2.16 for the transportation network. The communication distance exponent of 1.68 is calculated using network weights derived from call intensity. As stated earlier, the weights could also be derived from SMS intensity. When SMS-derived link weights are employed a distance exponent of 1.86 is generated. This indicates that SMS communications generally occur over greater distances for this dataset. Hence, estimation of call and SMS link strength should be performed independently and with different parameters.

V Conclusions

This paper outlines work on the estimation of Irish transportation and communication network representations from mobile phone network CDR data. Central to this goal are techniques to infer locations of interest from mobile phone users by analysing their phone usage trends at specific times. This allows the analysis of the networks to establish the importance of specific locations to the propagation of phenomena such as epidemics and information. This work also investigates techniques of inferring such network representations when mobile phone network data is unavailable.

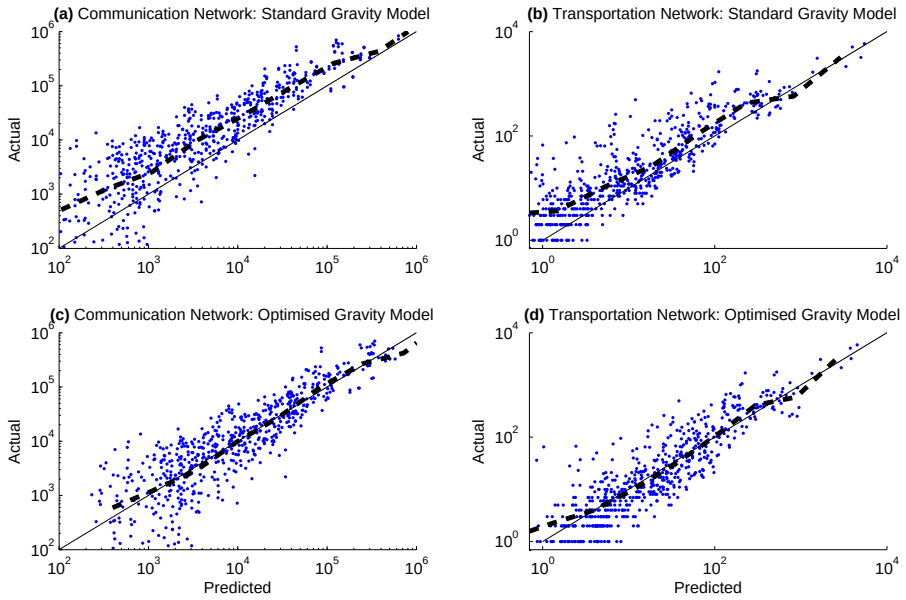

Fig. 1: Comparison of actual link intensities and corresponding model predictions. The solid diagonal line represents the perfect model and the dashed line represents the mean of the actual link weights corresponding to a given prediction.

Hence, such analysis could take place using census data in the future.

Even though this work has focused on county-level interaction for intuitiveness of results, it could take place at the sub-county level, with resolution limited only by cell tower density. However, a significant limitation of this work is that it utilises phone data from a single phone network. As a result the data represents a potentially biased subset of the population. Hence, future work seeks to improve the representativeness of the network estimates by scaling the population estimates according to geographically linked demographic information.

References

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

[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, 2009.

[3] R. Ahas, S. Silm, 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.

[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 strengths in mobile communication networks. *Proc Natl Acad Sci U S A*, 104(18):7332–7336, May 2007.

[6] V. Latora and M. Marchiori. Is the boston subway a small-world network ? *Physica A*, 314:109–113, 2002.

[7] V. Colizza, A.Barrat, M. Barthelemy, and A. Vespignani. Prediction and predictability of global epidemics: the role of the airline transportation network. In *Proceedings of the National Academy of Sciences USA*, number 103, pages 2015–2020, 2006.

[8] A. de Montis, M. Barthelemy, A. Chessa, and A. Vespignani. The structure of inter-urban traffic: A weighted network analysis. *Journal of Environmental Planning B*, 34:905–924, 2007.

[9] W. Jung, F. Wang, and H.E. Stanley. Gravity model in the korean highway. *Europhysics Letters*, 81(4):48005, 2008.

[10] D. Kelly, J. Doyle, and R. Farrell. Inferring population distributions from sparse cellular network data. In *KDD*, 2011. under review.