NUI MAYNOOTH

Ollscoil na hÉireann Má Nuad

The spatial structure of mobile communication networks

Fergal Walsh, B.Sc.

Supervisor: Dr. Alexei Pozdnoukhov

Thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

National Centre for Geocomputation, Faculty of Science
National University of Ireland, Maynooth
Maynooth, Ireland

June, 2013

Abstract

There has been a recent surge of interest in the relationship between the spatial and topological structure of communication networks with the availability of large scale anonymous datasets on the communication and mobility patterns of individuals. These datasets, captured as a by-product of modern communications technology, provide a detailed view of the daily interpersonal interactions of millions of people. Mobile phone call logs in particular offer an unparalleled source of information given their personal portable nature and ubiquity in modern society. The use of mobile phones has become so common that these datasets are no longer merely communication logs but close approximations of the network of interpersonal relationships that forms society. The analysis of these proxy networks has the potential to uncover knowledge about society at a scale never previously possible.

Networks, and social networks in particular, have been the subject of investigation for more than a century with a rich corpus of theory and methods now available to researchers. Computational approaches to the study of networks are more recent but there are now a wide variety of structural analysis methods that have been developed and applied across many different disciplines and subject areas. The study of interactions across space has developed in parallel with theory, methods, models and a variety of applications.

Recent studies of these proxy networks have tended to use computational approaches for analysing community structure and modelling spatial interactions without much regard for the theory upon which they were built. The underlying assumption has been that all phenomena that can be represented as networks can be analysed with the same methods. In this thesis we demonstrate that this is not the case and identify a number of problems and misinterpretations that can arise when inappropriate methods or network representations are employed. Through a detailed theoretical and empirical analysis we identify appropriate combinations of network representation, spatial scale, and analysis methods for studying the spatial structure of communication networks. Using these findings we demonstrate the potential of such analysis when the appropriate methodology is employed.

Dedication

This thesis is dedicated to my parents.

Acknowledgements

I would like to express my sincere gratitude to my advisor Dr. Alexei Pozdnoukhov for his guidance and support over the past three and a half years. I have learnt a great deal from him in the course of this research and other collaborative efforts, from writing papers to conquering mountain summits.

I wish to thank my colleagues, both past and present, at the NCG where I have always felt welcome and a great sense of community over the past 7 years. I would particularly like to thank Ann-Marie and Melina for their dedicated logistical and emotional support over the past years. I wish to also specifically thank Christian and Aonghus for the numerous discussions and arguments I have had with each which have helped shape the direction of this thesis. Additionally I would like to particularly thank Paul for his guidance and advice and friendship since my early years at NCG.

To my fellow PhD students with whom I have shared many fond memories and frustrating times, I express my sincere thanks for their friendship. I would particularly like to thank Carson, Cathal, Felix and Rory for their helpful advice, en/discouragement and general geekiness. Thanks are due to Carson also for providing the valuable perspective of someone one year ahead and importantly proving that it is possible to finish. Thanks to Aine, Conor and Ishwari for their less geeky friendship and more emotional understanding.

I wish to thank Enrique Frias-Martinez for inviting me to spend time in Telefonica, Madrid, where some of the ideas in this thesis were developed. I am forever indebted to Niall and Satoko for their generosity, kindness and friendship during my time in Madrid.

I gratefully acknowledge Dr. R. Farrell and the support of Meteor for providing the data used in this research, in particular Helene Graham, John Bathe and Adrian Whitwham.

I would also like to thank my examiners Martin Charlton and Dr. Francesco Calabrese for their insightful comments which have helped to improve the final version of this thesis.

I wish to thank my family for their support over last number of years. To my parents, Jim and Jeanne, my brothers, Patrick and Cormac, and my sister, Fionnuala, for their encouragement, understanding, academic advice and discussions. I would especially like to thank my parents for making life so much easier over the past few months with numerous meals, cups of tea and dinner time conversations about maps, communities and spatial interaction.

Finally my heartfelt thanks are due to Burçin for her unwavering support and understanding over the past three years as a colleague, friend and partner in life. She has helped prove that the effect of distance and cultural borders is not always as strong as it may seem and not all interactions need decay with distance.

Funding

Research presented in this thesis was funded in part through the Stokes Lectureship programme and Strategic Research Cluster (SRC) grant (07/SRC/I1168) by Science Foundation Ireland (SFI) under the National Development Plan (NDP), faculty research awards from Google and IBM, the National Centre for Geocomputation and the NUI Maynooth Postgraduate Travel Fund. This support is gratefully acknowledged.

List of previous publications

Portions of the work discussed in this thesis have previously been published in the following peer reviewed papers;

- Walsh, F. and Pozdnoukhov A. (2011). Spatial structure and dynamics of urban communities. Pervasive Urban Applications workshop at PERVASIVE'2011. San Francisco, CA, USA.

- Walsh, F. and Pozdnoukhov, A. (2011). Understanding the evolution of spatial structure in urban communities. In NetMob2011: Second conference on the Analysis of Mobile Phone Datasets and Networks, Cambridge, MA, USA.

- Walsh, F. and Pozdnoukhov, A. (2011) Spatial alignment of physical and social ties in daily dynamics of urban communities. XNet - Complexity Networks, Satellite Meeting of the European Conference on Complex Systems, ECCS'11, Vienna, Austria.

Contents

List of Figures

Note to the reader

High resolution versions of all figures included in this thesis are viewable at `http://www.fergalwalsh.com/thesis/`.

Reference maps of Ireland are included in Appendix A to aid the reader's interpretation of the analysis discussed in Chapter 6.

List of Tables

Chapter 1

Introduction

1.1 Introduction

Human society is a complex network of more than six billion interconnected individuals. People are connected through family ties, friendships, business relationships, and many other types of social links. While some of these links cross continents the majority of social links are local, between people in the same country, city or neighbourhood. The probability that two individuals randomly selected from the entire population of the Earth are connected is extremely small. It is slightly higher if we select from within a single country and much higher again if we focus on one small town. We know that the probability of a connection between people decreases with the distance separating them, but we also see other factors affecting this probability.

The likelihood of connections between people living only a short distance apart can be extremely low if there is a border between them. People from different countries are generally less likely to communicate than people in the same country even if the separation distance is less. We see the same effect at every scale from country to town level. Sometimes these borders are literal physical borders impeding communication, other times they may be linguistic or cultural borders. There are also social borders which have no physical manifestation or officially defined boundary but can be just as real nonetheless. All of these borders affect the formation of social ties and thus the structure of the network of society.

Until very recently the effects of distance and borders on the network structure of society were little more than hypothesis with only anecdotal evidence or small scale empirical studies of social networks. The lack of data was the major hindrance as, unlike demographic statistics, there is no national friendship census in any country. Recently, however, this situation has changed somewhat. The logs collected and stored as a by-product of modern communications technology have over time become an extremely detailed quantitative *approximation* of a large sample of the network of society. Mobile phone call logs have been of particular interest because not only do they capture interpersonal relationships but also person-space relationships as every single communication is annotated with high resolution spatial information, as a by-product of the operation of phone network.

The recent availability of these large scale geo-referenced datasets on human mobility and communication has led to a renewed interest by computational scientists in the relationship between society and space. There has been a surge in the last decade in the number of studies taking a data driven computational analysis approach to studying social systems and large scale human mobility and interaction datasets. The new field that has emerged from this is now referred to by some as computational social science (Lazer et al., 2009).

While a multitude of techniques have been used to study these datasets from various perspectives, two computational approaches have proven particularly popular for analysing the interactions in geographically aggregated datasets. Community detection and spatial interaction modelling are two distinct approaches to studying interactions on networks. Their respective fields have developed independently in separate disciplines of science yet recently they have both been adopted by other disciplines for studying the same phenomenon, the relationship between society and space. Community detection, with its theoretical roots in sociology and computational roots in computer science, has traditionally been used to identify social groups in networks of individuals but more recently has been applied to networks of spatially aggregated communications and movements with the purpose of finding the real geography of cities and countries. Spatial interaction modelling, from the disciplines of human geography, transportation planning, and regional science, is a framework for modelling the interactions between spatial units, most often interregional migration or commuting, but which has also been applied recently to communication networks.

The underlying assumption of community detection is that the observed interactions are a direct result of an unobserved partition of the network into groups which affects the interaction probabilities between nodes. The most commonly used spatial interaction models, on the other hand, are based on the premise that the interaction between locations is dependant only on the attractiveness of those locations and the distance between them.

Recent studies (Blondel et al., 2010; Ratti et al., 2010; Calabrese et al., 2011a; Walsh & Pozdnoukhov, 2011) have demonstrated the existence of strong spatial community structure in society, through the analysis of spatially embedded communication networks, and thus uncovered new geographies and redrawn the maps of countries. Other studies (Krings et al., 2009; Lambiotte et al., 2008; Simini et al., 2012; Kang et al., 2012) have shown that the interactions in these networks can be modelled with simple global spatial interaction models, leading to claims of the discovery of universal laws of human interaction. Independently these findings offer interesting insights into the relationship between society and space but taken together they raise more questions than they answer.

Is community structure really merely a consequence of an underlying global interaction process? Are there true communities in these networks where people interact more because of some shared attribute such as language, religion, social class or living on the right side of the fence? How can we tell the difference between true communities and those resulting from the spatial distribution of population? If there are true communities in these networks how do we explain the supposedly successful modelling of communication with global spatial interaction models?

In order to answer these questions it is necessary to thoroughly investigate the theory that underpins these methods and models and carry out rigorous tests to ensure correct application and interpretation of the results. Furthermore, as with any geographical analysis, it is important to consider the spatial scale of analysis and the effects of spatial aggregation. High resolution data allows us to test these effects and choose scales which don't mask local differences but ensure statistical significance.

1.2 Research objectives

The aim of this thesis is to understand the effect of both distance and community structure on the probability of interaction between individuals in society. The research goals of the thesis can be defined as follows:

1. to assess the viability of using mobile phone communication logs to construct proxy networks of society

2. to identify the most appropriate methodologies for investigating the properties of spatially aggregated social networks

3. to understand the effect of distance on tie probability in spatially embedded social networks

4. to understand the concept and implications of community structure in spatially aggregated social networks

Just like the complex networks we wish to study, these four goals are intertwined and have overlapping portions. They cannot be understood in isolation but must be considered as parts of a whole. The understanding of the effect of distance on tie probability requires an understanding of the effect of community structure on social interaction. The investigation of these issues requires the identification of appropriate methods for measuring these phenomena and this can only be achieved with consideration of the context in which these interactions take place.

1.3 Thesis structure

The thesis is structured around the four research objectives defined above. The initial two chapters explore the use of mobile phone data as proxy networks for society, with an examination of existing methodologies and a discussion of the network structure of society. The following two chapters explore the theoretical underpinnings and previous applications of community detection methods and spatial interaction models. In the following chapter we discuss our analysis of a real world network and highlight the pros and cons of existing methodologies. We discuss alternative approaches and demonstrate their superiority. In the final chapter we discuss our findings and their implications for future research.

Chapter 1: Introduction Here we introduce the context of this thesis and outline the research objectives.

Chapter 2: Mobile phone data In this chapter we discuss the use of mobile phone data for geographic analysis of society. We describe the origins of this data and discuss the previous works that have used mobile phone data for studying issues related to society and space. We highlight methodologies and results that require further attention.

Chapter 3: Communication networks as proxies for society Here we discuss society from a network perspective exploring the properties that make it both similar and different to other types of network. We discuss the relationships between communication and social relationships and identify the challenges and potential solutions involved in constructing proxy networks of society from mobile phone call logs. We particularly focus on the spatial and temporal aspects of these networks. We highlight several issues which may affect the application of both community detection and spatial interaction methods to these networks.

Chapter 4: Community detection In this chapter we discuss the concept of community and the techniques used for finding communities in networks. We review in detail previous studies on communities in mobile phone networks, paying particular attention to those uncovering the spatial structure of communities.

Chapter 5: Spatial interaction modelling Here we discuss techniques for modelling interactions in space. We describe the development of this field and the theories that underpin it. We discuss previous studies that apply spatial interaction models to mobile communication data. We argue that existing spatial interaction theory is incompatible with the processes that underpin modern human communication, especially mobile phone communication.

Chapter 6: Analysis In this chapter we conduct analysis of a nationwide mobile phone dataset. We apply some of the methodologies used in previous works and demonstrate some of the shortcomings associated with the models used. In testing these models and explaining their issues we uncover and highlight some novel findings on the relationship between distance and communication ties. We apply the methodologies from the previous literature on spatial community detection of societal networks and rigorously test the results to assess their significance. We highlight a number of potential problems with existing methodologies and propose alternatives, linking back to the spatial interaction analysis.

Chapter 7: Conclusions In this final chapter we review the findings of the previous chapters and discuss their implications for future research. We discuss the findings of our experiments and provide recommendations for future data driven analyses of the spatial structure of society.

Chapter 2

Mobile phone data

2.1 GSM networks

The GSM (Global System for Mobile Communications) standard was launched in 1991. It is now used by operators in over 200 countries, forming a global network connecting more than 5 billion people (Wireless Intelligence, 2010). GSM uses a cellular network structure with base station antennae providing coverage over cells which form a lattice. The base transceiver station (BTS) is the primary structural unit of the GSM network. It is a physical entity consisting of network switching hardware and multiple transceiver antennae. Each transceiver usually covers a sector of 120 degrees so there are at least three transceivers per BTS. When a call is made from a mobile phone it connects to the nearest available cell antenna. A central server also keeps track of the current cell of every mobile phone in a network so it knows how to deliver incoming calls and messages. It is important to note that while the location of the antenna is known, the exact area of any particular cell is unknown. Radio waves are disrupted by buildings and physical features of the landscape so this impacts the coverage area of the antenna. Each antenna also has a maximum capacity so more antennae are needed in densely populated areas so the cell size will be smaller than in sparsely populated areas. This means that a phone can be geolocated to a smaller area with a higher level of confidence in more densely populated areas. Cell areas are also not static because radio waves can be disrupted by precipitation and because the demand on each antenna varies over time. This means that a phone that stays in a static location may not connect to the nearest cell all the time and may fluctuate between a few different cells. Despite this uncertainty it is still possible to estimate the coverage area of each cell with reasonable accuracy.

2.2 Call detail records

Network operators maintain a database of every communication event of every subscriber which is used for billing purposes. This is known as a Call Detail Record (CDR). Each customer has an identification number which is unique throughout the entire GSM network. Each record in the CDR is associated with two of these identifiers, one for the network operator customer and one for the other party. Each record also details the time, type of communication, phone model identifier, duration and the starting and ending cell identifiers for that customer. If a communication event occurs between two customers of the same operator then they will each have a record in the same CDR, corresponding to the two ends of the communication. For example, one record may detail a call made by customer A to other another user B while a second record will detail a call received by customer B from a user A. If they are customers of separate operators then there will only be one record for each event in each CDR. If a customer makes a call while roaming, the communication will still be logged in their own operator's CDR but with the country and operator code they are roaming with instead of a cell ID. When a CDR dataset is studied for research purposes the unique identifiers are encrypted so

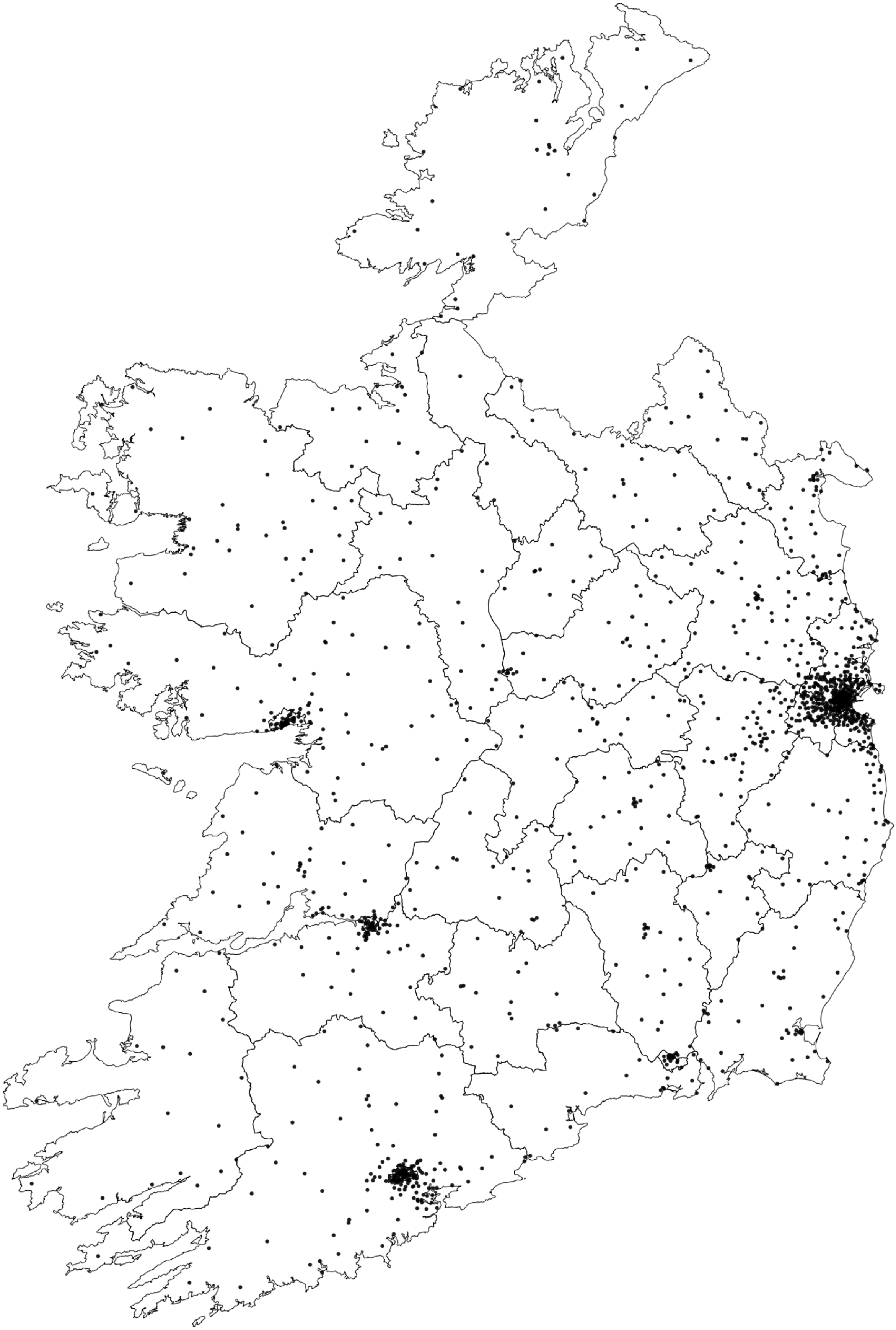


Figure 2.1: Map of the location of cell towers across Ireland for one operator.

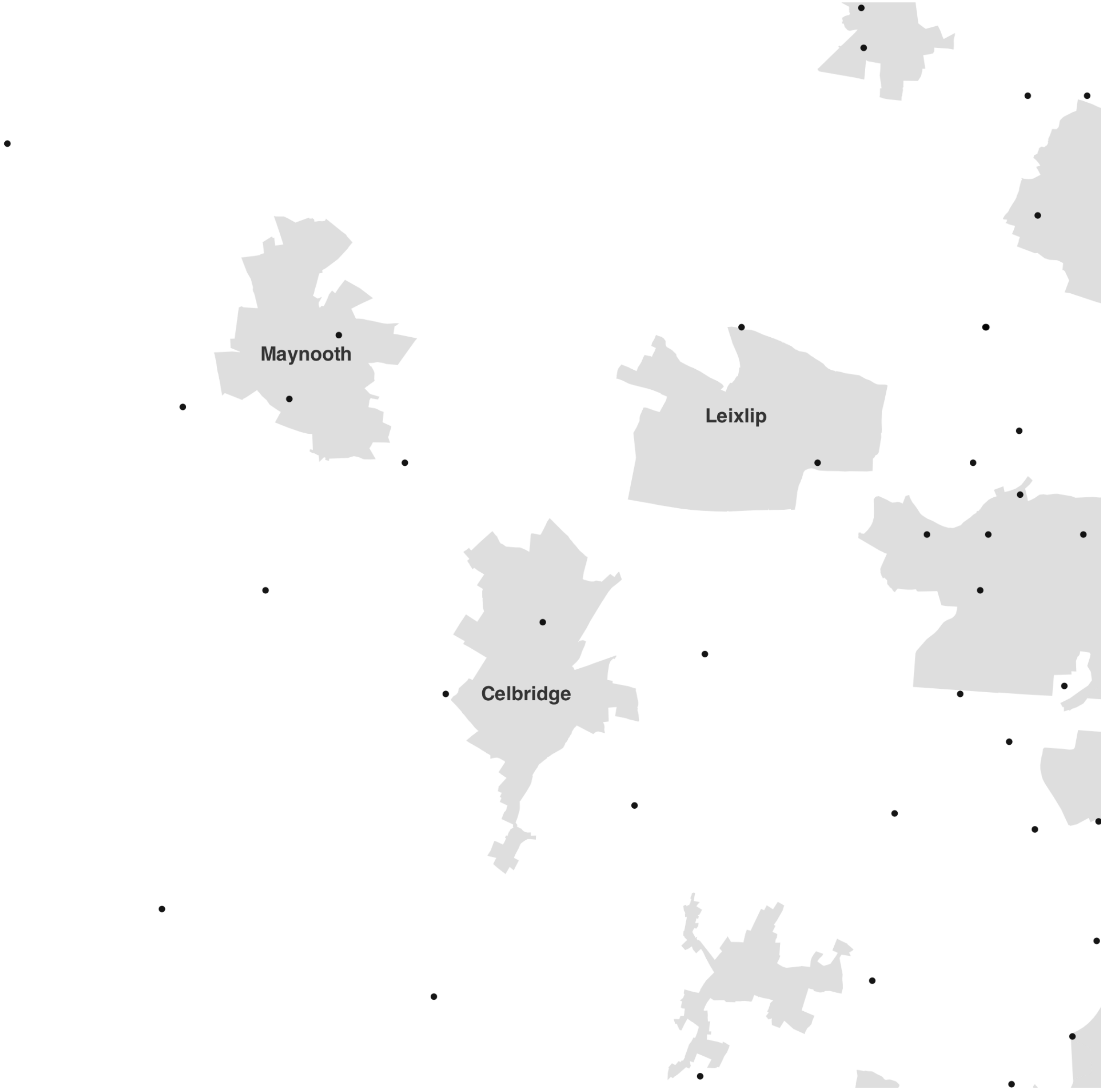


Figure 2.2: Close up map of the location of cell towers in north Kildare for one operator.

that the records can not be linked back to any individual for privacy reasons. However, each identifier is only encrypted once so each customer has the same identifier throughout the entire dataset.

A sample of a CDR table of the type used for analysis in this thesis can be seen in Table 2.1. In this case originating and terminating calls are recorded in separate tables. For calls where both users are customers of this operator there is a CDR entry in each table (entries denoted a, b and c). In this case the Caller, Callee, Date, Time and duration columns are identical in the two tables but the the TAC (phone model identifier), Cell 1 and Cell 2 columns are different as in each case they refer to a different customer.

The network operators also monitor and record usage statistics at a cell level for network performance monitoring. A standard unit of measure of load on a cell antenna is the Erlang. One Erlang is defined as one hour of usage by one phone. If there are 90 phones within one cell and each of them makes a call for 5 minutes then that will be 7.5 Erlangs. While Erlang data can give an indication of bandwidth usage in each cell, it cannot be disaggregated in any way to work out the numbers of active phones in the cell.

CDR derived networks have a number of properties that make them unique among social network datasets. The sheer ubiquity of mobile phones and the fact that they have become the primary mode of interpersonal communication means that the logs left behind by their use provide a very complete picture of nearly every single person's communication network. The existence of a link in this network is very different from a 'friend' on online social networks like Facebook or Google+ or 'followers' on Twitter. Every link in a CDR network is weighted by the frequency of interactions. Interactions have a cost associated with them, both in terms of monetary value and time, and are therefore the result of a concious act of at least one of the people involved. Each link is also directed which makes it easy to filter out the spammers, mass marketers and

Calls Originating

TAC	CALLER (customer)	CALLEE	DATE	TIME	CELL1	CELL2	DURATION	
03570036	94C7851EE4AD2	4EB3F660EB9A1	20101201	000137	25292	25292	00038	
73559025	807A57D237F8C	19FD240B7A823	20101201	000212	22375	22375	00004	
23539035	9A983E045BFB3	01AD8D7C5FA40	20101201	000207	21653	21655	00008	a
13578035	7D583C882AD19	033E879997FF9	20101201	000214	21744	21741	00002	
03570036	06451F38D5F43	31942520E4385	20101201	000207	95321	95321	00009	
13551024	1B03CEE3CD1B1	34D86C641FBC4	20101201	000210	02242	02242	00006	
23504033	C96BA6E7F3A16	5534EC677A68C	20101201	000133	21665	21665	00043	b
33502044	3045CABB438F0	853ABC8471764	20101201	000145	72155	72155	00030	
63583037	1C28E8716AB57	C94376C62D350	20101201	000214	21312	21312	00002	c
13571045	FA4CB3C6C43CF	B26F2BAE8CE84	20101201	000212	31425	31964	00005	

Calls Terminating

TAC	CALLER	CALLEE (customer)	DATE	TIME	CELL1	CELL2	DURATION	
03507063	2F124776DB5CF	89AEC04F96C6B	20101201	000114	12086	13116	00036	
72595053	1158C7C886F48	4C4FF18E1EBC4	20101201	000127	14166	19033	00024	
30830735	9A983E045BFB3	01AD8D7C5FA40	20101201	000207	13005	13002	00008	a
13587053	D47445080902B	2D9935E82FA0A	20101201	000244	10051	10051	00008	
03507063	0A7911E508AE5	C84C6879881A7	20101201	000249	14015	14015	00003	
12515043	741652FC04720	0586AFC807B54	20101201	000139	15674	15674	00013	
30649335	C96BA6E7F3A16	5534EC677A68C	20101201	000133	10166	10166	00043	b
34520043	043B8D86A8019	CAA77169FB576	20101201	000141	14072	14072	00011	
30357245	1C28E8716AB57	C94376C62D350	20101201	000214	12164	12166	00002	c
14517053	F9B29E97055FF	BEBF2B94D529F	20101201	000139	15122	15122	00014	

Table 2.1: Sample of Call Detail Records (CDRs) (artificially generated data)

over the phone services. The spatial and temporal attributes of each interaction provide a further dimension to this complex network. Given the temporal dimension, it is possible to study the evolution of the network over a period of months or years but also to look at the fine scale temporal dependencies affecting interactions. The spatial dimension provides information that allows researchers to investigate global issues such as the effect of distance on communication, but also the differences in communication patterns in different areas of space. While online social networks provide coverage of a larger geographical region than CDR networks, the effect of this is unimportant once studies are done using datasets from multiple operators in different geographical regions. The vast majority of communication is highly local, with the majority of most peoples' regular contacts living within a few kilometres, let alone within the same country. The lack of a single operator in each geographical region is also generally not too much of an issue as the market is usually dominated by a few big players in each country so it is not very segmented. Furthermore, the GSM network allows inter-operator interactions and thus the CDR logs for one operator can still provide a lot of information about customers of a different operator. It is, however, important to be aware of the spatial and demographical biases associated with the market share of datasets used for research purposes.

2.3 Previous work

In this sections we will review the literature of studies using call detail records (CDRs) from mobile phones. These studies cover a wide range of applications from identifying travel patterns of tourists to modelling the spread of epidemics. Other studies are less application focused and instead concentrate on identifying the laws and universal truths that underlie human behaviour and social systems.

2.3.1 Communication patterns and social structure

Lambiotte et al. (2008) study the geographical dispersal of the communication networks of 2.5 million customers from a 6 month CDR dataset. They create a network where the nodes are customers and pairs of customers are connected by an edge if they have at least a minimum number of reciprocated calls. They show that the probability that two individuals are connected decreases with the square of the distance between them. The distances are measured from the home addresses on file for each user rather than using locations determined from the data. It could be interesting to determine the regular locations of each individual and use these somehow in the distance measure. It maybe the case that the probability of calling another individual depends on the individuals' location at the time communication rather than their home location. The authors go on to consider the geographic extent of communication triangles which are connections between three people rather than just two. They find that shorter links have a higher probability of belonging to triangles than longer links up to a distance of 40km. After this distance there ceases to be any difference in the probabilities. The authors point out that this is the same distance that there ceases to be a relationship between average call length and distance, suggesting that there may be two distinct "regimes of communication" that are determined by the distance between the individuals. The short distance regime is "characterised by short communications and a high clustering coefficient" while the long distance regime is "characterised by longer communications and a smaller clustering coefficient". These probabilities are calculated across the entire population in the dataset but it may be the case that they vary for different types of individuals.

Krings et al. (2009) analyse the same dataset in in a slightly different manner. In this study a gravity model is used to model the communication flows between 571 cities in Belgium. The customers are assigned to locations based on their billing zip code, but aggregated to city level, and only calls between customers who have had at least six reciprocated calls are included. Unlike the previous study however, the total communication time is modelled, rather than the probability of a connection. Applying the standard gravity model they model the predicted communication time between cities A and B as $L_{AB} = K\frac{P_A P_B}{d_{AB}^2}$ where P_A is the calling population of city A, d_{AB} is the distance between the centroids of cities A and B.

A similar approach is taken by Kang et al. (2012) to study inter-city mobile communications in China using a gravity model. They discover a surprisingly low distance decay exponent suggesting a very marginal effect of distance on telecommunication in China. They note that it is often impossible to conduct such a study for an entire network in a country as big as China and its is often the case that only a subset of communications data is available. They show how a subnetwork can be used to determine model parameters for the entire country and assess the performance of this model.

Simini et al. (2012) introduce a new spatial interaction model, the radiation model, and demonstrate its use by modelling both mobile phone calls and hourly trips between municipalities in a unnamed western European country, in an effort to show how it can accurately model a number of different human interaction phenomena.

Each of these studies on spatial interaction using mobile phone data is discussed in more detail in Chapter 5.

One of the first published works on community detection in mobile phone networks was that of Palla et al. (2007). In this paper the authors analyse the evolution of social groups in 26 separate networks constructed from 26 two-week periods of mobile phone calls between 4 million people in an unnamed country. The weights of each edge are set based on the total cost of interaction during each two-week period and multiplied by an exponential time decay factor. They assessed the quality of the communities found by using additional information on the postal area and age of each user and found that the communities "tend to contain individuals living in the same neighbourhood, and having a comparable age...". They subsequently analysed the evolution of communities by matching the communities found in each time period. Their results suggest that in order for large communities to stay stable the membership must be in constant flux while small communities require a few strong persisting relationships in order to stay stable.

Blondel et al. (2008) use an individual level network derived from mobile phone data to test the performance of their fast community detection method. The network captures the calls between 2.04 million customers of a Belgian operator over a period of 6 months. Their method finds a six level hierarchy with 261 communities with more than 100 people at the top level, containing 75% of the customers. They find that most communities are almost monolingual with all but one of the communities with over 10000 customers having more 85% members speaking the same language.

The same method was employed by Onnela et al. (2011) to analyse the geographic constraints of communities in a mobile phone network. The network was constructed from logs on call and text message communication between 3.4 million people over one month, with a total of 5.2 million weighted edges. Their findings show that the geographic span of communities increases gradually with community size (number of members) but it increases dramatically after the size reaches 30 members. They also find that the number of spatial clusters within each community increases linearly with the community size up to 20 members.

Each of these studies on detecting communities in networks derived from mobile phone data is discussed in more detail in Chapter 4.

Zhao & Oliver (2010) also study the structure of the communication network but they take into account the temporal context of communication activity. They propose the concept of communication motifs which are topological structures with temporal constraints that occur within a network much more often than expected in a random network. The bidirectional links, triangles and stars mentioned previously are examples of such motifs. When these are considered with temporal constraints they could be used to determine different types of communication activity over short periods of time, giving insight into how an individual's communication activity changes over time.

Stoicaa et al. (2010) consider how age and gender affects the structure of each individuals ego network (the network from the perspective of that individual). They describe each ego network by the numbers of edges, vertices, isolated edges, triangles and stars and then define six distinct profiles based on these attributes. They then compute the probability that an individual of each age from 18 to 60 belongs to each profile and from this they perform hierarchical clustering of ages to determine groups of ages that are likely to have similar network structures. The results are interesting because they show there are clear differences in the communication activity structure of individuals of different ages rather than just a single type of structure for all individuals.

Blondel et al. (2010) use CDR data to construct a network of communicating places in Belgium for the purpose of identifying spatial communities. The network was formed by aggregating the individual level social network, as discovered from six months of mobile phone call logs, to the billing address municipality of each customer. Community detection on this network resulted in surprisingly spatially contiguous regions despite the fact that the method used made no assumptions about community composition based on spatial distance.

Landline phone call logs were used in a similar manner to identify regions in Britain by using “the network's characteristics to partition the geographic space underneath the network's topology” (Ratti et al., 2010). One month of data was used in this study, aggregated to 3042 spatial units of 9.5km by 9.5km. As with the previous study, mostly spatially cohesive communities were found initially without any constraints. Constraints were later enforced as the aim of the study was to extract regions which must be spatially contiguous.

Another study in the US using one month of mobile phone call data aggregated at the county level showed similar results when a community detection method was applied (Calabrese et al., 2011a). In this study, the authors create multiple networks using different aggregations for mode of communication (call or SMS) and show that different network partitions are found in each case. They also investigated the hypothesis that calling behaviour is more dependant on the actual location at the time of a call than the home location by comparing the distance between home counties to the distance between time of call counties. A stronger distance decay was found when using actual location rather than home locations.

In Walsh & Pozdnoukhov (2011) this author attempted to investigate the dynamics of spatial communities in an urban setting using mobile phone data from and Irish operator in Dublin. In this case the cell tower was used as the unit of spatial aggregation with each call creating a link between the towers used by the caller and callee at the time of the call. By detecting communities in different network snapshots the change in community composition was tracked over the course of a day.

2.3.2 Urban sensing

One of the first examples of using data garnered from mobile phone networks for urban analysis was the study of Reades et al. (2007). In this study the authors use Erlang data from Rome, Italy, covering a period of four months in 2006. Rather than using a Voronoi tessalation of the cell towers they rasterise the point data creating a grid of 262,144 "pixels". The Erlang values were recorded every 15 minutes, resulting in 96 measurements per day per pixel. In order to identify routine patterns they create a typical timeline for each weekday by averaging over all Mondays, Tuesdays, etc. The authors select six pixels corresponding to unique locations which they hypothesise would have very different time signatures. They find large differences in the signatures of different pixels and also markedly different patterns on weekends to weekdays. Additionally they perform clustering on the time signatures of all the pixels using K-means and identify spatially contiguous clusters which correspond well with the authors' intuition about the human activity in the city.

Reades et al. (2009) attempt to find 'eigenplaces' by "analysing cities using the space-time structure of the mobile phone network". They use eigendecomposition, a dimensionality reduction technique, to identify patterns of mobile phone usage over time in the city of Rome. They divide the city into pixels and represent the phone activity signature of each pixel by a weighted sum of eigenvectors. Eigendecomposition itself does not ascribe any semantics to these eigenvectors but the authors attempt to provide an interpretation by examining how the eigenvalues of each eigenvector change over space and time, taking into account prior knowledge of the use of space in the city.

Soto & Frias-Martinez (2011) use CDRs to automatically identify land use in the city of Madrid, Spain. The data was collected over 1 month in 2009 for 1100 cell towers within the city. In a similar manner to the previous two studies, a temporal profile is created for each tower. In this case however

CDRs are used rather than Erlang data so the values are the number of calls handled by the tower during the given time interval. A resolution of 5 minutes is used, yielding 288 samples per day. Noting the difference between weekday and weekend communication they separately compute weekday and weekdend signatures and concatenate these to produce a feature vector for each tower. Applying fuzzy c-means clustering they identify clusters which have similar signatures. Validation is performed using the author's expert knowledge as no dataset of actual land use exists for the city.

Soto et al. (2011) predict socioeconomic levels in a unnamed city in Latin-America using CDR data collected over a 6 month period for approximately 500,000 users. The national statistical institute three class socioeconomic level is predicted for each cell tower coverage area by classifying a feature vector for each coverage area using support vector machines (SVM). The feature vector for each area is computed as an average over a set of behavioural features defined for each of the residents of that coverage area. Initially a total of 279 features were computed for each user, of which 38 were selected for the best model which had a classification accuracy of approximately 80%.

Cariou et al. (2010) study collective dynamics in the city of Paris using mobile phone data. They aggregate trajectories of about 1.1 million users for each of the three weekends to generate paths of movement of people from one cell to another. By comparing the paths during the two festival weekends to a normal weekend they attempt to describe and assess the impact of the festivals on the city.

Calabrese et al. (2010) also study mobile phone data in order to understand the affect of social events on a city. They attempt to forecast travel demand for planned special events (e.g. concerts, sporting events, exhibitions, etc) by learning from mobility patterns during previous events. By determining the home location of attendees at events from their mobile phone usage the authors correlate event types with origin locations. For a future event of a known type and location they predict the spatial distribution of origins.

Traag et al. (2011) also use mobile phone data to study social events, focusing on automatic detection of social events from mobile phone call datasets. They define social events as “exceptionally large gatherings of people who are ordinarily not present at a specific location” and propose a methodology to detect such events using a probabilistic modelling approach to separate those present at a given location into those who are normally there and those who usually are not. They test their method on a dataset of 14 months with 5.75 million users and successfully identify known past events. Importantly their method is not affected by seasonal variations so it correctly detects events in touristic areas without identifying false positives.

Ahas et al. (2008) use CDR data from Estonia to identify foreign tourists for the purposes of understanding origins, destinations and travel behaviours of visitors to the country. The dataset covers a period of 17 months in 2004/05 with 1.2 million unique roaming customers from 96 different countries. They examine the temporal patterns of activity from tourists and show the calling activity correlates well with accommodation statistics in many areas. Furthermore they demonstrate how the data can be used to assess the impacts of advertising in particular countries for specific destinations or events.

2.3.3 Individual behaviour

Eagle & Pentland (2006) introduce the concept of “Reality Mining” as a method of sensing complex social systems using mobile phones. Using on-device software they logged location, proximity and communication information on 100 mobile phones used by staff and students at MIT for 9 months resulting in the Reality Mining Dataset. This dataset is much richer than the typical CDR dataset because it logs much more than just call event details. In addition to determining the location of the phone from the cell towers, they setup a number of Bluetooth beacons around the labs and social spaces so that the phone could log location information at a much finer spatial scale. They also used Bluetooth to log the proximity of other phones so that they could log face to face communication as well as electronic communication. While the resulting data is much more detailed than CDRs the data collection method requires installation of specific software on the phone so this type of study is only feasible on a small scale.

Eagle & Pentland (2009) analyse this dataset to extract eigenbehaviours to “identify structure in routine”. Eigenbehaviours were the inspiration for eigenplaces and have the same theoretical basis. According to them, each individual’s behaviour can be approximated by a weighted sum of their eigenbehaviours. Again the semantics of these eigenbehaviours is open to interpretation but they can be used to cluster individuals who are behaving similarly, even without assigning semantic labels. In this case they used an individual’s location and proximity to others as features to describe behaviour. Clearly it is a misnomer to refer to these as eigenbehaviours as there is certainly more to the behaviour of a person than their location in space and proximity to others.

González et al. (2008) use a CDR dataset to study the mobility patterns of 100,000 randomly selected individuals over a 6 month period. They report that there is a high degree of temporal and spatial regularity in each individual’s trajectory with a significant probability that the individual will return to a few highly frequented locations. The location of each individual is only known at the times when they make a call so the trajectories formed from these locations may be affected by irregular “bursty” calling patterns. However the authors addressed this issue by obtaining a dataset of regularly updated locations for 206 of the individuals in the CDR dataset and found similar results. This is important because it suggests that trajectories obtained from CDR data are a viable proxy for the real trajectories of individuals (at the spatial scale of cells).

Song et al. (2010) study a similar dataset of 50,000 individuals over a 3 month period to assess the limits of predictability in human dynamics. They measure the entropy of each individual's trajectory and find 93% predictability across all users. They find that the level of predictability does not vary with the distance travelled on a regular basis but is dependant on the fraction of hourly periods for which there are no communication events and hence the location is unknown. They find that this value is 0.7 on average in their dataset and the maximum limit for predictability is 0.8, meaning that as long as the location of the individual is known at least 20% of the time, it should be possible to predict their future location. This is a useful result because it suggests that most individuals have very regular mobility patterns. Furthermore, they provide a method to assess the predictability of an individual and hence filter out those whose mobility is unpredictable due to lack of data.

Frias-Martinez et al. (2011) use CDR data to build an agent based model of the population of a Mexican city to test the impact of government interventions during the H1N1 flu epidemic. Using CDR data they build a model of the mobility and interactions of the population during the period before the flu outbreak and compare this to the observed mobility during the outbreak. While the original dataset contains 2.4 million unique individuals the final number of agents was only 25,000 after filtering to only include those individuals with frequent location updates and activity during each of the stages of the epidemic. Their simulations show that the government interventions were effective in reducing the mobility of individuals and hence the spread of the virus. They claim there method is better at predicting the movement of people and the spread of a virus because it incorporates finer grain mobility and social interaction data at the individual that is not available from the census, the traditional source of data for such simulations.

Calabrese et al. (2011b) analyse a year of CDR data in order to investigate the relationship between telecommunications and physical co-locations. They use data for 1 million individuals in an unnamed European country. They find that 90% of the pairs of individuals who call each other have visited the same cell tower at some point during the year. They also find that 70% of pairs who communicate at least once a month have visited the same cell tower at the same time, indicative of face-to-face communication. The authors also note that the probability of a shared location decreases only slightly with the distance between individuals home locations. In many ways these results tell us precisely what we should expect; that people generally only call those who they meet face-to-face from time to time. Perhaps more interestingly they also find that the distance travelled for face-to-face meetings is consistently unequal with the same individual travelling further each time.

Isaacman et al. (2011) propose an algorithm that makes use of CDR data to find important places in people's lives. They claim the method works well for individuals who make a few calls a week and those who make many calls per day. Importantly, they make a distinction between the number of calls made from a given location and the number of days the location was visited, using the latter as the basis for inferring the level of importance. Interestingly the authors were able to obtain ground truth data along with CDRs for 37 volunteers. Half of the dataset is used to train the model while the other half is used to test its performance. This adds an extra degree of integrity not seen in most studies using CDR data but we must still note that the algorithms performance may be biased towards a particular demographic, given that the "majority of the volunteers work at high-tech jobs". The authors apply their method to two large scale CDR datasets of 97,000 individuals from Los Angeles and 71,000 individuals from New York, covering a period of 78 days in 2009-10. They find that more than 75% of individuals have between 3 and 6 unique important places, where the places are separated by at least one mile. The authors then proceed to whittle these places down to a maximum of two locations, corresponding to home and work. Rather than simply choosing the location with the most visits or calls during ‘work‘ hours, they use their volunteer data to train a model to identify these locations. In fact the model does simply identify the ‘home’ location as the place with the most calls during ‘home hours’ (7pm to 7am) but the ‘work‘ location is identified by the place with most ‘work hours’ calls and the lowest percentage of ‘home hour’ calls. Using these locations they calculate the average commuting distances for each zip code and show that these match closely with the census records, providing further validation of their approach.

2.4 Conclusion

In this chapter we have discussed the structure of GSM networks and the nature of Call Detail Records (CDRs). We have seen how in a relatively short period of time (5-6 years) CDR datasets from numerous countries and operators have been used for a wide variety of research purposes. Different research groups have analysed the data from varying perspectives with some focusing on the behaviour of individuals, others focusing on aggregate patterns of the masses while others are looking at how the interactions between individuals produce the patterns we observe at higher aggregation levels.

It is interesting to note that a number of the individual level studies have highlighted the complexities of human mobility, showing that people tend to travel to a number of locations repeatedly, and even predictably. Yet we see that most of the studies at higher levels of spatial and network aggregation tend to ignore this and assume that individuals are tied to single geographical locations.

A further important observations is that we see very few instances of the same methodologies being applied in multiple countries or with multiple different datasets. This is somewhat understandable given the difficulty in obtaining access to such datasets but it leaves many questions about the general applicability of the proposed methods. We should therefore retain a level of scepticism about studies claiming to uncover universal laws and truths about human behaviour as the analysis has always been one off with a single dataset and never reproduced by others.

In the areas of community detection and spatial interaction in particular we see a number of potentially interesting results but with very little discussion of theory or testing of the significance and accuracy of results. We will explore these issues further in subsequent chapters.

Chapter 3

Communication networks as proxies for society

3.1 Introduction

With the advent of online social networking services the idea of considering society as a network has become as commonplace as the term 'social network' itself. We naturally understand the concept of network distance and can easily fathom the idea that everybody is connected to everyone else through a relatively small number of steps. Although these services have made the underlying structure of society more explicit and observable, it has long been known that society is structured like a network. Durkheim argued in 1897 that societies were composed of interrelated components like biological systems (Durkheim et al., 1997). Nadel (1957) saw societies as "a pattern or network (or 'system') of relationships" (Borgatti et al., 2009).

Traditionally sociologists and anthropologists have analysed the structure of societies through the use of survey methods on small samples of a population of interest. In recent years vast quantities of information on human interactions have been created by modern communications technology providing opportunities for a computational social science approach by physicists and other computational scientists (Lazer et al., 2009). In the former case the survey data describes a sample of the population, usually with a relatively high degree of accuracy and detail. In the latter case massive data derived networks are used as proxies for large scale social networks under the assumption that the proxy is a good approximate of the real network. The practitioners of these two approaches work with datasets of massively different scales and often have different goals in mind but ultimately deal with data describing the same phenomena - the interactions and relationships of individual people.

Borgatti et al. (2009) note that "in the physical sciences it is not unusual to regard any dyadic phenomena as a network" and "a common set of techniques is used to analyse all instances" whereas in the social sciences a distinction is made between different types of dyadic links, both in analysis and theory. They argue that by treating all networks in the same manner physical scientists ignore the differences between different types of links and information that a network can encode. In a similar vein, Butts (2009) cautions that "[to] represent an empirical phenomenon as a network is a theoretical act. It commits one to assumptions about what is interacting, the nature of that interaction and the time scale on which that interaction takes place... the appropriate choice of representation is key to getting the correct result."

In this chapter we will heed these words carefully by examining the assumptions that we are making by using mobile phone data as a proxy network and attempt to determine the appropriate choice of representation for this network. As well as discussing what is interacting, the nature of the interactions and the time scales of interaction, we will also discuss where these interactions take place, a uniquely important element of mobile phone call networks.

3.2 Spatial context

Society and the proxy network through which we observe it, the mobile phone call network, has a spatial context. We are physically embedded in geographic space and our movements are constrained by this space. The primary means by which we meet people and form relationships is co-location and face to face communication. The probability of knowing someone is dependant on the probability of co-locating with that person previously, which in turn is dependant on the distance between their usual locations. Clearly space and distance are influential on the structure of society and the patterns we observe in mobile phone call networks but is society really a spatial network? Does it have the characteristics we usually associate with spatial networks?

3.2.1 The cost of distance

Barthélemy (2011) defines spatial networks as "networks for which the nodes are located in a space equipped with a metric" and considers communication networks, including mobile phone call networks, among the examples of spatial networks. He notes that connection probability generally decreases with distance in spatial networks, even when the links themselves are not embedded in space, as is the case with social networks. In typical spatial networks both the nodes and links are embedded in space, meaning that each link has an explicit cost determined by its spatial length. In transportation and infrastructure networks the cost is determined by the financial cost of physical material while the links in mobility or commuting networks have a cost associated with the fuel or ticket price and time required to traverse a link. The links in social networks, including mobile phone call networks, are not embedded in space as the cost of a link is constant regardless of the distance (within a single country). So why then does the connection probability still decrease with distance in these networks? Barthélemy explains that "in order to minimize their effort and to maintain social ties most individuals will connect with their spatial neighbors" (Barthélemy, 2011, p. 29). This explanation fits nicely with the "principle of least effort" (Zipf, 1949) but it seems to ignore the realities of human social interaction. A much more simple and intuitive explanation is that we simply have much less chance of meeting people who live far away than our spatial neighbours. We do not purposely choose not to befriend and communicate with them because it would require more effort, rather the opportunity to befriend someone arises with decreasing frequency the further away they are. The probability of a connection in the mobile phone call network is likely to decrease with distance but it is not for reasons of minimising cost, as with other types of spatial networks. A side effect of this is that it is possible for any two people to be connected, no matter what the distance between them, as long as they spent some time in close enough proximity to form a relationship in the past. Their current distance may affect the frequency of communication however because the reasons for communication change with distance. One possible explanation is that people tend to make shorter calls more frequently to their contacts who live relatively close, possibly for the purpose of arranging face to face meetings or day to day activities but after a certain distance it becomes impractical to meet on a regular basis so communication is less frequent but for longer durations. On the other hand the frequency of phone communication between individuals may actually increase if the distance between them passes the point that allows for regular meeting, as the phone replaces face to face contact as the primary form of communication. The lesson here is that there are multiple styles of communication so we should not assume that there is one explanation that fits all cases. We must take this into account when assigning weights to the links of the network.

It is clear that mobile phone call networks and society itself are strongly affected by the space in which they are embedded and therefore should not be assumed to have a random configuration but it is equally clear that they are affected by space in different ways to other types of spatial networks. It is important that we consider the implications of this when we attempt to uncover structure or model processes on these networks.

3.2.2 Spatial clustering and hierarchical structure

The nodes of spatial networks are often clustered in space. In networks where distance plays an important role the spatial clustering of nodes also affects the topological structure. We see this in networks of all spatial scales but the effect is not necessarily the same in all cases.

The nodes and edges of transport networks are clustered hierarchically with local, commuter, intercity, and international services each forming separate spatial networks that interlink at hub nodes. In this case the spatial clustering is directly related to the network's topological structure. At each scale every node of the network does not just represent a transit station but also a larger geographic entity and a sub-network.

The neurons and the connections between them in mammalian brains also have a hierarchical spatial structure (Sporns et al., 2004). Just as transport connections are expensive to construct, neural connections have a cost so brains consist of many short low bandwidth links and few high bandwidth long distance links. These long distance links connect different regions of the brain together through hub nodes, similar to transport hubs.

There is a very strong spatial clustering effect in human society based on the distribution of population across space. As discussed above, this clustering has a strong influence on the connection probabilities between individuals, causing spatial clustering among the links of the network also. The clustering of nodes is hierarchical just like the transportation network. Indeed, our addresses have a hierarchical structure starting from the street or neighbourhood in which we live, to the city, county, province/ state, and finally the country. There is a significant difference however with the other hierarchical spatial networks discussed above. The hierarchical structure is purely spatial and separate from the topological structure. In modern society we have the freedom and ability to directly communicate with anyone we wish, in any geographical region rather than having to communicate via a hierarchical system of representatives. The mobile phone network infrastructure is also designed in the same way so it does not impose any extra structure, unlike the communication networks of the past. The consequence of this is a higher proportion of independent interregional links than in other spatial networks.

3.2.3 Spatial aggregation

The advantage of networks with clear topological hierarchical structure is the ease with which they can be simplified to more abstract forms for easier understanding of larger scale interactions. Even when networks lack a natural hierarchical topology it is common to artificially create a hierarchical structure by aggregating nodes and links together based on some other property. In spatial networks it is natural to aggregate nodes to geographical regions, thus creating simplified spatial networks with weighted links between regions. The specific regions chosen depend on the scale of analysis, varying from census areas to cities to provinces or states. The rational for such aggregation comes from Tobler's first law of geography, that "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970). In the context of communication we assume that the calling habits of individuals in the same area are related, and more related than those in distant areas. We often form generalisations about people from a particular town, county or part of a city so it is useful to be able to aggregate interactions to these levels to test the validity of these generalisations or to present findings at a scale that people are familiar with. For example it is a well known generalisation that ‘south siders’ and ‘north siders’ in Dublin city don’t talk to each other. We can put this theory to the test by aggregating all interactions in Dublin into two groups and examining the interactions between the two.

The challenges with spatial aggregation are two-fold. The particular assignment of sub units to larger units can affect the results dramatically. This is known as the modifiable areal unit problem (Openshaw, 1983). The second issue relates to the homogeneity of the aggregated units. The properties of the aggregated unit will be representative of its components only if their properties are similar to each other. Despite Tobler’s law, this is not always the case. This issue can arise in spatially aggregated networks when nodes with differing connectivity patterns are aggregated. In mobile phone call networks this can happen due to the spatial resolution limit of coverage areas which causes dissimilar individuals to be assigned to the same location, or it may be due to larger scale aggregations of dissimilar coverage areas. To assume aggregated data is representative of the individual components is to commit ecological fallacy while the converse, assuming that the properties of individual components apply to the aggregated data, is known as the fallacy of composition (Fotheringham et al., 2000). We must be aware of these issues and perform tests to ensure that we do not misinterpret results when dealing with spatially aggregated networks. It should be noted that the issues we raise here are not limited to aggregations defined manually or by geographical borders but also apply to aggregations automatically determined through community detection (see Chapter 4).

3.3 Temporal context

Mobile phone call logs record a communication process that takes place over time. Each call is a discrete interaction covering a short time period but it takes place within a multi-scale temporal context that has influenced the probability of it happening. While there are numerous factors that contribute to the probability of a call happening at a particular moment in time, the fact that it happens at all is usually due to a longer term relationship that exists between two individuals. It is generally the network of longer term relationships, society itself, that we are interested in studying and modelling but we must infer the long term structure of this network from information about these momentary interactions. In the following sections we will discuss the challenges associated with creating a static network from dynamic data, using the temporal context to identify significant relationships, infer relationship types and assign appropriate link weights.

3.3.1 Historical context

While we use mobile phone call networks as proxies for society we must keep in mind that they do not simply tell us the relationships between individuals. Instead they tell us the interactions between individuals during the time period of the dataset. It is up to us to infer the relationships from this disaggregated data. Ideally the dataset would capture the formation, evolution and every interaction of every relationship but this is never the case. We must work with incomplete snapshots covering a short time window of the real network. In the time prior to the snapshot people have formed relationships, travelled, migrated, changed schools or jobs and done various other things which affect the social structure of the network. If we are to have any hope of understanding this structure we must be aware of the presence of this unobservable historical context rather than assuming that all information explaining the structure is observable in the present network snapshot.

3.3.2 Relationship inference

The temporally disaggregated nature of mobile phone call logs provides extra information that can be used to infer the nature of relationships. While a raw count of calls is a simple measure of relationship strength it would not be very accurate as it is easily affected by abnormal activity such as many calls during a short time period. A more appropriate measure would take into account the temporal context by measuring the number of distinct days of communication, for example. This information can be used to filter out insignificant relationships. The short term temporal context of calls may also be used to infer the type of relationship, should we wish to know it. The time of day that people are most likely to communicate, the interval between communications or the duration of calls are all potentially discriminating variables. However, we must keep in mind that people have different styles of communication and phone calls are only one of numerous methods of communication so it should not be assumed that the absence of regular phone calls implies a weak relationship - it may in fact be the direct opposite, that a couple or colleagues who are together so often they do not have a need for regular calls. An aggregation and weighting scheme must be carefully chosen to try to accurately reflect the importance of links while accounting for different styles of communication. This is an example of a property of the proxy network that is not necessarily representative of the real social network. It is also an issue that does not usually arise with other forms of social network data such as surveys or online social network profiles where relationship types may be explicitly stated and are generally not inferred from interactions.

3.3.3 Temporal aggregation and homogeneity

We discussed above how it is important to consider the homogeneity of spatially aggregated units. A similar but more complicated issue arises with temporal aggregation. When we aggregate interactions across time periods into weighted links we are making an assumption that the link strength or connection probability is constant through time. If this assumption is incorrect then the topology of the network actually varies with time so information will flow through the network differently depending on the time period under study. The complication with temporal aggregation comes from the periodic nature of human activity which makes time non-linear. The effect of this is that communication activity during weekday mornings should generally be more similar to other weekday mornings than weekend mornings or weekday evenings. This issue becomes important if we wish to understand a temporally dependant process on a network where the connectivity varies significantly with time. For example, if individuals or places are strongly connected during the weekdays but very weakly at the weekends we still assume that they have a high level of connectivity on average.

3.3.4 Spatio-temporal dependency

When discussing the spatial aspects of this network above we worked off the assumption that each person is statically located in space and their interactions could be aggregated based on the places where they 'live'. In reality people move around space on a daily basis and have multiple places they frequent, for living, work, education, social and entertainment purposes. The effect of this varies depending on the scale of analysis and the size of the 'habitat' of the people under study. If we work at a scale where we consider these as separate locations and nodes of the network, we must decide how we should aggregate individual interactions over time and space. The most simple option is to ignore the temporal aspect of location and only consider each individuals home location, either using the billing address if known or determining it from the data. This approach potentially ignores important spatial context information however. An alternative is to aggregate based on the time-of-call locations, i.e. the spatial unit containing the cell tower used to make each call. This approach has the advantage of maintaining spatial information about the call but is not necessarily useful for discerning the structure of society, as the location of a person at the time of making or receiving a call is not always significant. We propose that a further alternative, moving the emphasis back to relationships rather than calls, is to create links between all significant locations of all connected pairs of individuals. In this case the actual location at the time of call is unimportant but the multiple significant locations of individuals are still considered. A link in this network between nodes i and j would signify that someone who spends a significant amount of time in location i knows someone who spends a significant amount of time in location j. As with other approaches, a weighting scheme is still required to define the importance of each link in this network.

3.4 Weighting schemes

Networks can have weights associated with their links. It is appropriate to assign weights to links when all links can not be considered to be equal. In a social network a person may have connections with many other people but they might be much closer to some than others so it is necessary to signify this with a weight. Similarly in spatially aggregated networks, the links must be weighted to represent the sum total of interaction between nodes. The choice of weights is not straightforward however. When using mobile phone call logs as our data source it may seem obvious that we should use the number of calls between individuals as the weight but this is not necessarily a good choice. As we discussed above, people have differing styles of communication so a pair of individuals with many calls between them do not necessarily have a stronger relationship than a pair with far fewer calls. Similarly, the population of nodes in spatially aggregated networks affects the link weights so a higher weight does not necessarily signify a stronger connection. The total call time in minutes has also been used as a weight in telecommunication networks but the same argument holds. The weight can be normalised based on each nodes total number of calls or average per link, or similar. This effectively creates a weighted directed network as the intensity of interaction is relative to each individual. This brings up another issue; directionality. Networks can be directed, meaning the link weight from i to j can be different than the link weight from j to i. In fact the call logs themselves are directed, as each call is initiated by a caller and received by a callee. However, there is no reason to believe that the directionality of calls is relevant to the relationship between two individuals (except for commercial services which we filter out by exactly this means). However, as we have seen now, a network can be made directed by normalising the weights, even if the underlying data is undirected (or made undirected). Alternatively we may decide that it is only the property of being acquainted that is important rather than the strength of the connection so all relationships carry equal weight, once a discrimination between relationships and one-off calls has been made. While this may make sense at the individual level, weighted aggregation and normalisation are still required when dealing with spatial aggregations of population.

3.5 Conclusion

In this chapter we have discussed the issues involved in constructing and analysing a network representation of society from mobile phone call logs. By comparing the known properties of mobile phone communication and society itself with other networks we have highlighted the similarities and dissimilarities with spatial networks. We have discussed issues related to both topological and spatial hierarchical structure and identified a number of potential problems which need to be tested for when doing analysis on spatially aggregated networks of this kind. Furthermore, we have investigated the temporal context, both short and long term, and pointed out the complex effects it has on the network structure but also the opportunities it offers in terms of extra contextual information which can be used to infer relationship types.

From the discussion in this chapter it should be clear that there are a number of steps involved in creating a network from mobile phone call logs and that the resulting network has a number of unique properties which require special attention and consideration in subsequent analysis. In the following two chapters we will revisit the theory and application of two forms of network analysis; community detection and spatial interaction modelling. Based on the discussion of this chapter, we will perform both theoretical and experimental tests on networks derived from mobile phone call logs to assess the limits of applicability of these forms of analysis to societal networks.

Chapter 4

Community detection

4.1 Introduction

Community detection refers to the art and science of finding strongly connected groups of nodes in a network. It is a popular area of research in various fields including geography, sociology, biology, physics, complex systems and recently, computational social science. Many different systems can be represented as networks or graphs and within these graphs there is often a natural grouping of nodes. The aim of community detection is to identify these groups and thus reveal the underlying structure of the network at a higher level of abstraction. Community detection methods have been used to identify groups in social networks in a variety of scales and contexts from friendship networks in clubs (Zachary, 1977) to political ties (Porter et al., 2007) to massive online social networks (Traud et al., 2011). In biology they have been used to understand functional regions of mammalian brains (Meunier et al., 2010) and to find groups of related genes (Wilkinson & Huberman, 2004). The network of science itself has been an area of interest with a number of physicists looking at the interactions between different disciplines and tracking the emergence of new ones through collaboration and citation networks (Newman, 2001; Rosvall & Bergstrom, 2010). Perhaps of most interest to the current study is the application of such methods to networks derived from mobile phone call logs, particularly those where interactions are spatially aggregated. These will be discussed in detail in the following sections.

While community detection is one of the most popular techniques for studying complex networks, there is no consensus definition of what a community in a network is (Fortunato, 2010). Methods generally try to find groups of nodes which have high densities of within group edges and low numbers of edges between groups. Newman (2012) defines a community as “a dense subnetwork within a larger network, such as a close-knit group of friends in a social network”. Others take a more restrictive view and require communities to be cliques, where every node is connected to every other node (Palla et al., 2005). We know for sure that one does not expect to find communities in random networks so the presence of communities means there is structure in the network. In fact we shall see that a number of methods detect communities by comparing the network under study with a random network. In our discussion of methods we will see that most methods do not have an *a priori* definition of community but instead attempt to find groups or a partition that best satisfies some goal. These communities are said to be algorithmically defined (Fortunato, 2010).

In this chapter we will review the most important developments in the field of community detection, focusing mostly on methodological issues initially, followed by a review of the applications of community detection methods to communications networks.

4.2 Methods

In this section we will review the literature of methods for detecting communities in graphs or networks. Rather than reviewing each of the hundreds of methods, we will focus on the major developments that shaped the field or that have been used in the analysis of mobile phone networks. We will begin with traditional methods for graph partitioning, move onto quality function based methods and then discuss more recent developments with methods based on information theory and local statistical significance. The reader is referred to Fortunato (2010) and Porter et al. (2009) for in depth reviews of the entire field.

4.2.1 Graph partitioning

Kernighan & Lin (1970) introduced one of the earliest graph partitioning methods for the purpose of dividing electric circuits into boards. Clearly the aim here is to maximise the number of links within the same board while minimising the links between boards, thus minimising the cost and complexity of interboard wiring. This is expressed as a quality function which measures the difference between the number of internal and external connections (where internal connections refers to connections between vertices in same community or board). The algorithm starts with the vertices of the graph split equally into two groups and then repeatedly swaps equally sized groups of vertices between the groups while trying to maximise the quality. The results of the algorithm are strongly dependant on the initial partition of the graph and thus the method is more often used to improve the results found by other algorithms. This method also requires the number and size of groups or communities to be specified *a priori* which is typical of graph partitioning methods. This is reasonable for the purpose of dividing a circuit among boards or a database among servers but is clearly not useful or reasonable when trying to discover the inherent structure of a network.

4.2.2 Divisive methods

Girvan & Newman (2002) take a divisive approach to community detection with a method that finds communities by iteratively removing edges that are between rather than within communities. To find these edges they calculate the betweenness centrality of each edge. This is defined as the number of shortest paths between vertex pairs that include that edge. The edge with the highest betweenness centrality is then removed and the process repeats until no edges remain. The result of the method is a dendrogram which shows how the network splits into disjoint components as edges are removed, encapsulating the multilevel community structure of the graph. One significant drawback of this and other similar methods is that there is no way to say which level of the dendrogram represents the best partition of the graph. This lack of a quality measure led to the development of the modularity metric and a refinement of the of the Girvan-Newman algorithm (Newman & Girvan, 2004). The modularity metric can be used to measure the quality of each partition of the dendrogram and thus identify the optimal one. This was a significant development in the field and heralded a new era of community detection methods based on the modularity quality function. In the next section we will discuss modularity in more detail and review some more recent modularity based methods.

4.2.3 Modularity maximisation

The modularity metric (Newman, 2006) is a partition quality measure that gives a measure of how modular or compartmentalised a network is. It compares the fraction of edges found within communities to the expected fraction, using a null model. It is defined as

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - P_{ij} \right] \delta(c_i, c_j) \tag{4.1}$$

, where m is the sum of the weights of all (undirected) edges in the network, A_{ij} is the weight of the edge between nodes i and j, P_{ij} is the expected weight of the edge between nodes i and j in the null model and $\delta(a, b)$ is 1 if $a = b$ and 0 otherwise. Q ranges between 0 and 1, increasing with the deviation from the null model. In other words we can express modularity as the normalised sum of the differences between the observed and expected edge weights within communities. The standard null model used in modularity is a random rewiring of the graph, preserving the original degree distribution. This can be expressed as

$$P_{ij} = \frac{k_i k_j}{2m} \tag{4.2}$$

where k_i is the sum of the weights of edges attached to node i, thus yielding

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \tag{4.3}$$

as the final formulation of modularity with the standard null model.

While modularity was originally proposed as a quality metric for finding the best partition among several possibilities, it has subsequently been used as a quality function, shifting the focus in the field to treating community detection as an optimisation problem. Modularity maximisation is actually impossible on real world networks as it involves trying every possible combination of nodes in communities of every possible size. Instead there exist many methods which try to optimise modularity in an efficient manner - finding a reasonable trade off between quality and computation time.

The Louvain method (Blondel et al., 2008) is currently one of the most popular approximate methods as it is capable of handling extremely large networks and has been shown to find high modularity partitions very quickly. It is a greedy agglomerative hierarchical method which means it builds partitions in a hierarchical manner from the bottom up. It starts with every node as a community on its own and iteratively tries to move a node from its current community into another community that will increase the overall modularity score. It stops this step when no single node can be moved to a different community that will increase modularity. It then aggregates all nodes within a community into a single node and repeats the previous step. This continues until there is no change in the network between two successive iterations. The efficiency of the method is due to two aspects. The first is that the change in modularity is extremely simple to calculate when moving a single node from one community to another as it only involves the edges within those communities. The second is the hierarchical nature of the method. Once the first step of the algorithm has completed, the size of the network decreases hugely if there is any community structure. With each successive step the number of nodes decreases further. The hierarchical nature of the method means that the result is not just a single partition but instead multiple partitions at different scales. This feature makes it an attractive method for analysing large complex networks with a hierarchical structure.

Limitations of modularity maximisation

While modularity offers an objective quantitative measure of the quality of a network partition, it unfortunately suffers from some limitations that bring its use into question. The two most significant of these are the resolution limit and the large number of near optimal but dissimilar solutions. The resolution limit, first pointed out by Fortunato & Barthelemy (2007), concerns the inability of modularity based methods to identify communities below a certain size. The precise limit of resolution is dependent on the total size of the network. Network partitions with small communities have lower measures of modularity than partitions with agglomerations of smaller communities. To understand why this is the case we must understand the derivation of the modularity formula.

Modularity is the normalised sum of the unexpectedness of edges within communities, where unexpectedness means the difference between the observed and the expected edge weight. The second term of the summand in Equation 4.1 expresses the expected weight of the edge i, j in the given null model. The null model of modularity is the configuration model which is a random rewiring of the network maintaining the same degree distribution. The probability of finding an edge involving node i is $\frac{k_i}{2m}$ so therefore the probability of finding an edge involving both nodes i and j is $\frac{k_i}{2m}\frac{k_j}{2m}$. This is the probability of finding a single edge between nodes i and j so the expected number of edges (or the expected weight of the edge) is $2m\frac{k_i}{2m}\frac{k_j}{2m}$ which simplifies to $\frac{k_i k_j}{2m}$. It should be clear that the expected edge weight is dependant on $2m$, the number of edges in the whole network while the observed weight is obviously a constant, independent of the weights in the rest of the network. Unless the degree of each node is directly proportional to the total degree of the network, the expected weight of each edge will be inversely proportional to the total network degree. This means that the higher the total network degree, the higher the number of edges that are considered unexpected. Given that only edges within communities affect the modularity and the gain from an expectantly present edge is much higher than the loss from an unexpectedly absent edge, the modularity of partitions with combined communities tends to be higher than that of partitions of individual small communities.

This problem is symptomatic of a more general problem involving null models and community detection. The null model is meant to represent a version of the graph under study without any community structure. In the configuration model used in modularity all nodes have equal probability of being connected yet this is often not the case in real networks. In some networks the nodes have other unobserved attributes which affect their probability of interacting with others. In social or communication networks these factors may be geographic location, language, social class or nationality. These factors create *a-priori* probabilities of edge weights which the configuration model does not account for. It is also possible for networks to have properties that limit the number of edges that a node may have no matter the network size. In Chapter 3 we discussed how this is certainly the case for social networks. The number of social ties per individual is bounded not by the available opportunities but by our own cognitive limits. This does not make it any less likely that an edge will occur between two particular nodes however. Similarly, in weighted networks the edge distribution and weight distribution may be very different, something which the configuration model can not handle. The specification of more realistic null models is very difficult however and misspecification can lead to interpretation problems. This is likely why the simple configuration model remains the most popular despite its limitations.

Expert et al. (2011) point out that distance affects the clustering coefficient and degree in spatially embedded networks, thus affecting the probability of two nodes being connected. They argue that the standard configuration null model is inappropriate in these circumstances and an alternative is required. They suggest that a “natural choice” for the null model is one based on the gravity model where

$$P_{ij} = N_i N_j f(d_{ij}) \tag{4.4}$$

where N_i is the size or importance of node i and $f(d)$ is a function which captures the negative impact of distance. It is debatable whether a model based on the gravity model is a natural or appropriate choice when only the knowledge that distance plays an important role is available. There is no theoretical basis for suggesting that the gravity model is a universally appropriate model for any spatially embedded network regardless of the process underlying the network. This issue is discussed in more detail in Chapter 5. Assuming that such a model is appropriate for the data at hand, the function $f(d)$ must be chosen so that the model correctly models a spatial network with no community structure. The function $f(d)$ must either be directly measured from the data or determined by fitting the parameters of a model to the data. In both cases the properties of a network with community structure are used to determine the parameters for a model of a network with no community structure. Clearly this will lead to an imperfect specification of the null model. A poorly specified null model affects the modularity score and resulting partitions. This leads to problems interpreting the results as it can be difficult to tell which effects are truly from community structure and which are due to model misspecification.

The second important issue with modularity is the large number of near optimal but dissimilar solutions. In optimisation problems one generally does not expect to find the optimal solution but rather one close to the optimal. It is generally assumed those solutions will be similar, converging towards a global optimal solution. Good et al. (2010) have discovered, however, that there are a large number of structurally dissimilar solutions with near optimal modularity. Furthermore, the modularity scores of these solutions are so close to each other and to the optimal solution that their differences are insignificant. This means that even if an exhaustive search procedure is used to find the highest modularity partition the result is still just one of many equally good but dissimilar solutions.

Despite these numerous criticisms and fundamental problems modularity maximisation remains the most popular form of community detection. Many authors ignore these issues because it appears to find intuitive communities in many networks and, importantly, fast optimisation algorithms are widely implemented in a number of software packages.

4.2.4 Information theory

Rosvall & Bergstrom (2008) introduce a method based on information coding theory which produces a compressed description or map of a network. The map is a two level description of the network, describing information flow within and between modules, where the same names or identifiers are reused for nodes in different modules but the module names are unique. The map that will produce the shortest description length of a random walk through the network is thus the best assignment of nodes to modules. A function that provides a measure of the expected minimal description length of a random walk through the network with a given map can then be used as a partition quality function which can be optimised in a similar manner to modularity. The authors define such a function which they call the map equation. In their original work they optimise this function using an approach based on simulated annealing. In a later paper (Rosvall & Bergstrom, 2010) they improve upon this by using a stochastic and recursive algorithm, based on the Louvain algorithm (Blondel et al., 2008), which is faster and provides more accurate results. The map equation and associated algorithms are a significant development in this field as they take a different approach to community detection by modelling the information flow process on the network rather than just clustering the network based on its topological structure.

4.2.5 Overlapping communities

So far all the community detection methods we have discussed are designed to find partitions of networks where each vertex is assigned to one and only one community. It can be argued however that such partitions are not realistic for real networks where it is possible for each node to be a member of multiple communities. A many-to-many mapping of nodes to communities is known as a coverage rather than a partition. There are a number of methods capable of detecting coverages, most notably the Clique Percolation Method (CPM) (Palla et al., 2005). The CPM is interesting not only for the fact that it finds overlapping communities but also because it is one the of the few methods that has an *a-priori* definition of community. The authors base their definition of communities on the idea that communities consist of several complete subgraphs (cliques) that have many nodes in common. They define a community as the union of all *k-cliques* that share $k - 1$ nodes, where a *k-clique* is a fully connected subgraph with k nodes. This method can also be classed as a local method because communities can be defined and found without reference to the entire network. This means that changes to one part of a network will not affect the community structure in other parts. This is in contrast to modularity based methods where a subgraph can only be considered a community if its structure is unexpected given the structure of the rest of the network. An unfortunate drawback of methods using a clique based definition of communities is that the notion of a clique does not easily extend to weighted networks. (Palla et al., 2005) suggest thresholding the edges is an adequate solution but there is no way to determine what the correct threshold should be. Coupled with the need to specify k, this means the CPM effectively requires two parameters when dealing with weighted networks.

4.2.6 Local significance

From the discussion so far we have seen that a number of useful concepts and features have been developed and incorporated into community detection methods since the early graph partitioning methods. These include quality functions, hierarchical unfolding, weighted and directed edges, overlapping communities and local community definitions. Of the methods presented so far, however, none have all of these features. Lancichinetti et al. (2011) identified this problem and devised a method to address it. They claim that their method, the Order Statistics Local Optimization Method (OSLOM), is "the first method capable to detect clusters in networks accounting for edge directions, edge weights, overlapping communities, hierarchies and community dynamics". The method works by locally optimising a quality function which measures the significance of communities. The method uses a quality function which measures the statistical significance of each community separately with respect to a random network null model. This is in contrast to the modularity quality function which measures the significance of the entire partition with respect to a null model. The overlapping communities result from the fact that the method repeatedly builds communities from random starting points until all nodes are members of at least one community and similar communities are found over and over. Once these communities have bean "cleaned up" to ensure they are minimal communities with no internal sub-structure, all the nodes and edges of each community are aggregated into super-nodes and super-edges. The algorithm is then applied to this new super-network and the process repeats until no communities are found, thus revealing the hierarchical structure of the network, in a manner similar to the Louvain algorithm(Blondel et al., 2008).

This method appears to incorporate every desirable property of community detection but unfortunately there is one that it lacks. It does not handle links from a node to itself (self links or loops). Fortunato (2010) notes that "loops are usually absent in real graphs" while Massen & Doye (2005) have proposed a modification of modularity's null model to forbid loops to make it more realistic so it would seem that the inability to deal with loops is not a limitation but an advantage in most networks. However we will demonstrate in our analysis that loops are of vital importance in aggregated networks, particularly spatially aggregated ones. It should be noted that loops are not considered at any level of the hierarchical process in this method, unlike in modularity optimising agglomerative techniques which add weighted loops to the super-nodes even if the original network does not contain any loops.

4.2.7 Finding dynamic communities

The detection and analysis of dynamic communities is a relatively recent area of study in the field of community detection, which has become more popular as the availability of timestamped networks has increased. The general aim is to find communities that persist through time and to study how they evolve. We can easily imagine the evolutionary changes that might take place in social networks when people develop new relationships, join new social groups or drift away from old ones. Communities are constantly in flux, with frequent small changes and occasional large disruptions. Similarly, the "scientific community" is itself made up of many communities of scientists whose interactions and influences can be quantified through co-authorship and citation networks. It has been observed (Rosvall & Bergstrom, 2010) that these networks are dynamic also with new fields emerging from within established ones, and separate fields merging together into larger ones. Some of these changes take decades while some happen in a matter of years.

Mobile phone networks are basically social networks with high temporal resolution. If we look at the network over long time scales we see the patterns described above but there are also other short term patterns which we should be able to identify through dynamic community detection. Some communities may only exist during certain hours of the day (e.g. a work community) while others may be more time independent (e.g. family). No matter what the

Method	Hierarchical	Overlapping	Weighted	Directed	Self Loops	Deterministic	Local	Parameters
Graph partitioning								
Kernighan & Lin (1970)	No	No	Yes	No	Yes	Yes	No	Yes
Girvan & Newman (2002)	Yes	No	Yes	No	Yes	Yes	No	Yes
Modularity optimisation								
Newman & Girvan (2004)	No	No	Yes	No	Yes	Yes	No	No
Blondel et al. (2008)	Yes	No	Yes	No	Yes	No	No	No
Random walk map length optimisation								
Rosvall & Bergstrom (2008)	No	No	Yes	Yes	No	No	No	No
Rosvall & Bergstrom (2010)	No	No	Yes	Yes	No	No	No	Yes
Rosvall & Bergstrom (2011)	Yes	No	Yes	Yes	No	No	No	Yes
Local optimisation								
Palla et al. (2005)	No	Yes	No	No	No	Yes	Yes	Yes
Lancichinetti et al. (2011)	Yes	Yes	Yes	Yes	No	No	Yes	Yes

Table 4.1: Features of popular community detection methods.

time scales are, the detection of dynamic communities always involves finding communities on independent temporal slices or snapshots of the network and identifying relationships between communities of different snapshots. Some methods treat these as completely separate steps and find a good partition of each network snapshot and then process these to find correspondences between communities in each snapshot. Other methods use the knowledge of the previously found communities while detecting the communities in subsequent time steps, incorporating the matching step into the main algorithm. Methods taking both of these approaches will be discussed further.

Greene et al. (2010) introduce a method which discovers dynamic communities by matching the static communities found at each time step. This matching is done by measuring the similarity (using the Jaccard Coefficient) of each static community at each time step to the dynamic communities found previously. If the membership is similar within a certain threshold then the static community is assumed to be a continuation of that dynamic community. If it is not similar a new dynamic community is created. When dynamic communities merge together they continue to exist as separate communities with the same membership in subsequent time steps. When a dynamic community splits into two or more parts, it may result in the creation of new dynamic communities or it may cause the re-emergence of previously seen communities. This can happen when two communities merge together and then split again after a few time steps. The method however includes a parameter which determines how long an unobserved dynamic community should be considered before it is marked ‘dead’. If the splitting occurs after this time a new dynamic community is created instead of the resurrection of the original one. An advantage of this approach is that it is completely independent of the community detection method chosen and so can be applied to the results of any other method.

This approach has its problems however. It is based on the assumption that the result of the community detection method represents *the* community structure of the network and that a difference in the results means the community structure is different. As discussed in Section 4.2.3, this is a false assumption when using modularity based methods (and others that use similar partition quality measures). Most methods do not exhaust the possible partition space and hence find suboptimal solutions that merely give *good* partitions. There can be many solutions with equally good modularity that are structurally different from each other. Even when exhaustive methods are used, it has been shown that the difference in modularity between the optimal and many suboptimal solutions is not significant yet the community membership is significantly different (Good et al., 2010). This means that the partition detected by the method on one network snapshot is just one of many good partitions. It is possible that $modularity(G_i, P_j) > modularity(G_i, P_i)$ where G_i, G_j are network snapshots and P_i, P_j are their respective partitions as found by a modularity optimisation method. Therefore if the method finds a different partition on a different network snapshot, it does not necessarily mean the community structure is different. This approach therefore has a high chance of detecting false change which forces one "to introduce *ad hoc* hypotheses on the graph evolution to justify the variability of the modular structure, whereas such variability is mostly an artefact of the approach"(Fortunato, 2010).

Rosvall & Bergstrom (2010) attempt to "distinguish between meaningful trends and statistical noise" with their method for mapping change in networks. They do this by performing significance clustering using bootstrap networks to ensure that the communities found in each snapshot are robust to noise. The significance clustering finds the significant subset of each community and provides a measure of the probability of two significant subsets being clustered together. This allows the significance of any community merging or splitting events between snapshots to be tested, thus providing a method for differentiating between noise and real change.

The authors of OSLOM (Lancichinetti et al., 2011) state that it can also be used to analyse dynamic communities because algorithm can be started from any partition, not just a random assignment of nodes to communities. They suggest using the partition for snapshot $t-1$ as the starting point for snapshot t. If the communities are similar they will be left as is but if there is significant change the cleaning up procedure will identify this and find new communities. This prevents some of the problems associated with the two-step detect and track methods as it ensures that stable communities between consecutive snapshots will be identified as such. However it still suffers from two significant limitations. Firstly, it only has a single time step history which means that it won't keep track of communities that disappear and re-emerge, as can be the case in networks exhibiting periodic temporal patterns. The second limitation is the potentially more problematic issue described above for modularity based methods. This problem potentially affects all non-exhaustive methods based on the optimisation of a quality function. Further work is required to understand if local quality measures are as susceptible to this problem however.

4.3 Community detection on telecommunications networks

Telecommunications networks are excellent examples of real world networks with rich community structure and are of interest to researchers in a number of fields, both as examples of massive complex networks to drive forward research in community detection methods and as quantitative evidence of human interactions and social structure. The recent availability of telecommunications datasets to researchers has contributed to a surge of interest in community detection methods with numerous studies making use of such datasets for various purposes. These datasets have been studied both as social networks of individuals and in spatially aggregated form. It should be obvious that communities exist in the network of individuals as people tend to socialise and live in groups rather than individually or in pairs. Within a family group it is highly likely that each family member will communicate with every other member. This type of fully connected group is referred to as a clique. Communities need not be cliques however. Generally it is too strict a requirement as a group of work colleagues or school classmates would often be considered a community even if every individual does not regularly directly communicate with each other individual. It may be less obvious that communities exist in spatially aggregated networks. We have previously seen how the likelihood of people communicating is strongly affected by the distance between them so it is clear that communication is mostly local. This fact, coupled with the uneven spatial distribution of the population, results in the presence of densely connected spatially contiguous regions. We expect communities to form around population centres and this is usually what is found but there may be further interesting patterns in the structure of these networks. Seemingly contiguous urban regions often mask a highly fractured discontiguous social structure that can be mapped through communication patterns. Community detection can also be used to reveal spatial hierarchies, which in some cases can be quite different to the official administrative boundaries of a country. In the following sections we shall review some of the recent literature on the detection of communities in both individual level and spatially aggregated networks derived from CDR data.

4.3.1 Contact networks

One of the first published works on community detection in mobile phone networks was that of Palla et al. (2007). In this paper the authors analyse the evolution of social groups using the clique percolation method (CPM) (Palla et al., 2005). The authors construct 26 separate networks from 26 two-week periods of mobile phone calls between 4 million people in an unnamed country. The weights of each edge are set based on the total cost of interaction during each two-week period and multiplied by an exponential time decay factor. As mentioned previously, the CPM requires a weight threshold, $w*$ and the clique size k to be set *a-priori*. Values of $k = 4$ and $w* = 1.0$ were used, although no indication is given as to how these values were chosen. They assessed the quality of the communities found by using additional information on the postal area and age of each user and found that the communities "tend to contain individuals living in the same neighbourhood, and having a comparable age...". They subsequently analysed the evolution of communities by matching the communities found in each time period. Their results suggest that in order for large communities to stay stable the membership must be in constant flux while small communities require a few strong persisting relationships in order to stay stable.

Blondel et al. (2008) use an individual level network derived from mobile phone data to test the performance of their fast hierarchical modularity optimising method. The network captures the calls between 2.04 million customers of a Belgian operator over a period of 6 months. Their method finds a six level hierarchy with 261 communities with more than 100 people at the top level, containing 75% of the customers. The modularity of the final partition is 0.76. A further measure of quality is possible due to the fact that there are two main languages in the country and the language of each customer is known. They find that most communities are almost monolingual with all but one of the communities with over 10000 customers having more 85% members speaking the same language. These results appear to provide strong support for the algorithm and the modularity metric. However, one must bear in mind that there are many many configurations of communities that would produce similar levels of quality, both in terms of modularity and linguistic cohesiveness.

The same method is employed by Onnela et al. (2011) to analyse the geographic constraints of communities in a mobile phone network. The network is constructed from logs on call and text message communication between 3.4 million people from an unnamed European country over one month, with a total of 5.2 million undirected unweighted edges. In order to investigate spatial relationships in the network they assign a location to each person based on the most frequently observed cell tower. This is used rather than their billing address in an effort to improve locational accuracy. Their findings show that the geographic span of communities increases gradually with community size (number of members) but it increases dramatically after the size reaches 30 members. They also find that the number of spatial clusters within each community increases linearly with the community size up to 20 members.

4.3.2 Spatially aggregated communication networks

Blondel et al. (2010) use mobile phone call data from Belgium to find regions in a network of municipalities. The network is formed by aggregating the individual level interactions, as discovered from six months of mobile phone call logs, to the billing address municipality of each customer. Two different measures of interaction are used to produce two separate directed networks. In the first case the 'relative frequency' is used where the weight of the edge from municipality A to B is given by $\frac{W_{AB}}{P_A P_B}$ where W_{AB} is the number of calls made by customers in A to those in B and P_A is the number of customers in municipality A. The authors note that this allows them "to compare frequencies between pairs of municipalities even if the pairs of municipalities have different sizes and if the market shares of the provider in the municipalities are different". In the second case the network is constructed using the average duration of calls from A to B.

Community detection on both of these networks using the Louvain algorithm results in surprisingly spatially contiguous regions despite the fact that the method makes no assumptions about community composition based on spatial distance. In the first case using the relative frequency the method finds 17 distinct communities or regions while in the second case using average duration it finds only two regions. Interestingly in the second case the regions correspond very closely to the two distinct linguistic regions of the country. In this case only 2% of all communications are between customers in different regions.

The authors acknowledge that the results may depend on the ordering of the edges in the input to the method. They check the robustness of the method by running the algorithm on 100 random permutations of the edge order for each network. For the first network they find that 91% of municipalities remain in the same communities and those that move are always the ones on the borders between communities. For the network based on average call duration they find that there is no variation of the communities found. Furthermore they note that moving any municipality from its assigned community to any other community reduces the modularity.

Ratti et al. (2010) use landline phone call logs in a similar manner to identify regions in Britain by using "the network's characteristics to partition the geographic space underneath the network's topology". One month of data is used in this study to create a network of interactions aggregated to 3042 square spatial units of 9.5km by 9.5km. The edge weights in this network are defined using the total call time between nodes A and B in an effort to take into account local population density. The authors apply spectral optimisation of modularity to this network and initially identify a partition of 23 communities with a modularity of 0.58. 13 of these communities are spatially contiguous while the others are scattered individual pixels. After applying Kernighan-Lin steps and enforcing spatial contiguity the modularity increases to 0.60 with 14 contiguous regions. The authors assert that the results show “that not only population distribution in space but also regional boundaries affect the patterns of communication”.

They acknowledge the finding of Good et al. (2010) that modularity has many local maxima and that these are likely to be structurally different to each other. In order to identify alternative local maxima they apply five variations of modularity optimisation techniques to the same network and compare the results. Each result has between 12 and 14 spatially contiguous communities and modularities between 0.606 and 0.613. The authors note that there is some variation along the boundaries but the regions are always spatially contiguous and centred around the same geographic locations. However if we look closely at the mapped outputs of each of these five methods we see that there is significant variation in the community sizes and boundaries between the results. This is even true for the best three results which have practically insignificantly different modularity scores of 0.611474 0.611636 and 0.613114. As a final output the authors take the intersection of the results of all five methods to identify 11 core regions where the nodes are always clustered together. These 11 cores contain 85% of the total population. The authors of this study also acknowledge the resolution limit that affects modularity optimisation but claim their analysis does not suffer because they are interested in detecting large regions.

Calabrese et al. (2011a) use the same modularity optimisation method as in the previous study to detect communities in network of counties in the USA using one month of CDR data. In this study separate networks are created using for calls and SMS to investigate the difference in the effect of distance on both modes of communication. As with the previous study, the total call time is used for the edge weights between counties for the first network while the number of messages between counties is used in the second network but each of these values is first normalised to account for differing market shares in different counties. The edge weight for counties A and B is given as $W_{AB}\frac{P_A}{C_A}\frac{P_B}{C_B}$ where P_A is the population of county A and C_A is the number of customers in county A. The method finds 26 communities for the call time network and 28 communities for the SMS network. While a number of the borders and aggregations of counties do change for the two modes of communication it is questionable as to whether these changes are significant or merely represent two local maxima of the modularity for similar networks.

In (Walsh & Pozdnoukhov, 2011) this author uses the Louvain method in conjunction with the tracking method of (Greene et al., 2010) in an attempt to investigate the dynamics of spatial communities Dublin. In this case the cell tower is used as the unit of spatial aggregation with each call creating a link between the towers used by the caller and callee at the time of the call. A series of network snapshots are created by aggregating over two hour periods in five weekdays of data with the Louvain method applied independently to each of these snapshots. The tracking method is then used to identify the change in each community over time. While this study appears to show that these communities do indeed change over the course of the day it unfortunately suffers from the problems described in Section 4.2.7. Later analysis has shown that the different community structure detected in each snapshot is entirely due to the Louvain method finding local optimum solutions. This is confirmed by the fact that there is actually a single partition that has the highest modularity on all network snapshots.

4.3.3 Discussion

In these three studies on community structure in spatially aggregated networks we have seen five different edge weighting schemes and three different scales of spatial aggregation units. Although one may wish to express tie strength between nodes in different ways, we must be aware of the assumptions and expectations of the algorithm being used. In these three cases a modularity optimisation approach is used so it is the calculation of modularity that must be considered.

For assigning edge weights Blondel et al. (2010) use 'relative frequency' and average duration, Ratti et al. (2010) use total call time and Calabrese et al. (2011a) use market share normalised total call time and number of SMS. The 'relative frequency' measure is supposed to make the weights comparable between pairs of nodes of different sizes but in actual fact it is biased against large nodes in the same way as Lambiotte et al. (2008)'s P_d because it divides the observed interaction by the product of the node populations. This means that the weights are not in fact comparable and distorted by the node populations in more subtle ways.

The use of average duration is also problematic because the total network degree and the node degrees are used in the calculation of the modularity measure to define the probabilities of a node having an edge. The average duration of a call divided by a sum of average durations does not actually capture the probability of anything.

Using total call time is also problematic in other ways however. The P_{ij} in modularity (Equation 4.1) is the expected weight of an edge between nodes i and j, calculated as $2m$ times the probability of a single weight between i and j. In this case it is the probability of a single second of communication. The problem is that the total call time between two spatial units is an aggregation over all the seconds of each call of each pair of customers but each second of call activity is not an independent event. Each second is dependant on a number of other seconds which make up a single call and each call is dependant on the other calls between the same pairs of customers. In this case if there is a link at all we expect the weight (number of seconds) to be quite high but the null model of modularity has no concept of dependant and independent weights so the weights may be much more widely dispersed with many edges with very low weights. The same issue applies for total numbers of text messages.

If we assign weights based on the number of social ties where each unit of weight represents a single pair of customers we no longer have the same dependency issues. This formulation is also more robust and less affected by outliers and different styles of communication which are not necessarily relevant to societal level interaction patterns (see Section 3.4). It may be argued that the number of calls or the total duration is directly proportional to the number of ties and therefore the relationship is the same. In Section 6.7 we prove this is not the necessarily case using data from an Irish national network.

4.4 Community detection and regionalisation

It must be noted at this point that the idea of partitioning space into areal units using interaction networks long pre-dates the existence of mobile phone call logs and most community detection methods. Quantitative geographers and regional planners have been developing and applying regionalisation methods for more than forty years for defining statistical output areas, travel to work areas and local labour markets (Masser & Brown, 1975; Coombes et al., 1986; Coombes, 2000; Flórez-Revuelta et al., 2008; Farmer & Fotheringham, 2011; Papps & Newell, 2002).

In both regionalisation and spatial community detection interaction data between places is used to determine a set of boundaries by aggregating areal units. While the inputs and outputs are similar in these two types of analysis, the underlying assumptions and requirements are quite different. The aim of community detection in spatial networks is to partition the space based solely on the topological structure of the underlying network. The fact that the resulting partitions are often contiguous regions is incidental.

Regionalisation methods on the other hand are designed to partition space in a way that satisfies or optimises certain criteria. While properties of the network structure are often included in these criteria, they are rarely used on their own. Regions defined for statistical analysis and reporting purposes are usually required to be spatially contiguous, non-overlapping and of a certain minimum and maximum size, dictated by the specific use case. The use of regionalisation methods for the analysis of the spatial structure of social networks would be inappropriate given these constraints and the requirement of an *a-priori* specification of region sizes. On the other hand community detection methods have recently been adapted for defining functional regions in cases where it is not desirable to specify such parameters *a-priori*. Farmer & Fotheringham (2011) used a modularity optimisation approach was used with a geographical weighting that ensured the required spatial contiguity.

While it may appear that the authors of recent studies using mobile phone call data are unaware of the existing geographical literature on regionalisation it would not be fair to say that they are merely reinventing old techniques. The approach taken and the methods used are quite different in each case because the aims differ. Community detection studies are generally more exploratory in nature than the outputs focused research of regional geographers and planners and thus have less rigid requirements and constraints.

4.5 Conclusion

In this chapter we have discussed the development of the field of community detection from the computational graph partitioning methods through to modularity modularity optimisation by comparison to null models and up to the most recent local optimisation methods. We noted that most methods do not work towards finding precisely defined communities but instead algorithmically find a division of the network that satisfies or optimises some goal. Some of these methods attempt to find discrete partitions of the network while others accept that communities are likely to overlap in many situations and instead optimise coverages. Furthermore, we noted that a number of different hierarchical methods exist, working either divisively or agglomeratively. Interestingly even within the subclass of hierarchical agglomerative methods we saw that they differ in terms of the significance of their outputs. The modularity based Louvain method produces its most significant communities at the highest level of aggregation while the OSLOM method finds the most significant communities at the lowest level. In these examples we see that each method works to optimise a slightly different goal and thus finds different types of communities.

In our discussion of modularity based methods we highlighted some of the potential issues that arise from optimisation of a global rather than local quality measure. Specifically we saw how such methods are susceptible to the resolution limit which restricts the minimum size of communities found and is especially problematic in networks with uneven degree distributions. We also noted the findings of (Good et al., 2010) that showed that the modularity measure does not have a clear optimum value for most networks, implying that there are many possible partitions with similarly high modularity scores. These two limitations make the continued use of modularity based methods highly questionable yet it still remains a popular choice. This is likely directly due to its established popularity and the fact that it is included in many software packages with efficient implementations.

We discussed two recent innovations in the field in the form of the Infomap and OSLOM methods. Infomap is innovative in the fact that it assigns communities by optimising the length of 'map' of a random walk through the network, thus modelling the flow on the network rather than simply its structure. OSLOM on the other hand is a local optimisation method with the *a priori* definition of a community as a subgraph with no internal substructure. This definition together with its strong statistical basis makes it an attractive method that should be free from the limitations of modularity based methods.

In our discussion of previous studies applying community detection methods to communication networks we noted a number of potential problems with the methodologies used. We found that most of these studies use a modularity based methods which, as we discussed, have severe limitations. We also saw that different choices of network representation were used in each study, some of which are questionable given the assumptions of the method or the assumptions of the authors regarding the phenomenon under study.

There appears to be a tendency in this field to assume that the community partition found by one of these methods represents the true underlying structure of the network. Often it is not the case that there is a single possible representation of this structure and the methods merely find a representation that optimises their quality measure. The root of this problem is in the adoption and repurposing of an engineering solution to graph partitioning for finding the underlying structure of real world social networks. In the graph partitioning case the aim is to simply minimise a cost and it doesn't matter if there are numerous ways to achieve this or if the exact minimum is not found. On the other hand in community detection we are generally concerned with finding a single solution that explains the structure of the network. In this case we do not expect there to be many different possible solutions and assume that the solution found is the only one. The move away from simple global optimisation methods towards ones with locally defined communities based on significance tests is indeed a welcome one. The use of local optimisation with a precise community definition and allowance of overlaps between communities reduces the number of possible solutions and makes it more likely that the result is significant.

In Chapter 6 we investigate these issues in detail through empirical analysis of a nationwide mobile phone call dataset from Ireland. We analyse this network at various scales with multiple techniques to assess the effect of different methodological choices in the analysis of such datasets.

Chapter 5

Spatial interaction modelling

5.1 Introduction

Spatial interaction can be defined as any movement over space that results from a human process, including commuting, migration, information and commodity flows (Haynes & Fotheringham, 1984). Spatial interaction modelling refers to the use of mathematical models to explain and or predict these flows. The study and modelling of spatial interaction has its roots in social physics, beginning with Carey's observation in 1858 that the movement of people between cities was related to both their size and separation distance, analogous to the gravitational attraction between masses, giving rise to the term gravity models Carey (1858). The models and theory have matured greatly since the nineteenth century and spatial interaction is now an important field of quantitative geography with a firm theoretical foundation, even if this is often overlooked by geographers and physicists alike (Fotheringham et al., 2000). In this chapter we will briefly review some of the main theoretical developments in the field starting from basic gravity models, through to models based on spatial choice and finally some recent developments. We will then review applications of spatial interaction models to telecommunications data and discuss the some issues identified with these studies. The interested reader is referred to the reviews of Farmer (2011); Fotheringham et al. (2000); Roy & Thill (2003) for more detailed treatments of the developments in this field.

5.2 Models

5.2.1 Gravity models

The observations of Carey (1858) and Ravenstein (1885) that the migrations of people between cities was related to both their size and separation distance led to the development of the first spatial interaction model, the gravity model, inspired by Newton's law for computing the attraction between masses. The number of trips between origin i and destination j is estimated as

$$T_{ij} = k\frac{P_i P_j}{d_{ij}} \tag{5.1}$$

where P_i represents the sizes (usually population) of location i, d_{ij} represents the distance between the origin i and destination j and k is a scaling factor. It was realised that the deterrence effect of distance may vary depending on the type of interaction being modelled so an exponent was added to the distance variable to account for this. Similarly it was recognised that the effect of population size may vary with other characteristics of the origin and destination so two further exponents were added to those variables:

$$T_{ij} = k\frac{P_i^{\alpha} P_j^{\lambda}}{d_{ij}^{\beta}} \tag{5.2}$$

According to (Haynes & Fotheringham, 1984) such adjustments may be used to account for varying average income levels when examining the flow of shopping expenditure for example. Further refinements were made to the model, including the addition of multiple origin and destination attributes.

Despite multiple refinements, the model has been criticised for a lack of a theoretical basis in individual behaviour and an unwarranted focus on analogy to physical models. Empirical research however suggested that the model was rather good, leading to much effort on developing an acceptable theoretical framework to justify its use. The principle of least effort proposed by (Zipf, 1949) provided an alternative interpretation of the gravity model but was equally devoid of a behavioural foundation. Dodd's interactance hypothesis (Dodd, 1950) was a more sociological approach and introduced a probabilistic formulation of the model.

It was the work of Huff (1959) on consumer behaviour and Stouffer (1940, 1960) on intervening opportunities that introduced a behavioural interpretation of the gravity model (Haynes & Fotheringham, 1984). These models importantly allowed for the fact that humans make choices and it is these choices that lead to the aggregate interaction patterns that we observe. While this is a much more logical basis for a model of spatial interaction, the assumptions on how people make choices were far too simple.

5.2.2 Statistical mechanics

According to Fotheringham et al. (2000) the next major advance in providing a theoretical foundation for spatial interaction models came with the work of Wilson (1967, 1970). The important contributions of this work was a reframing of spatial interaction in the principles of statistical mechanics rather than analogy to physical models and the identification of a 'family of spatial interaction models'. His use of statistical mechanics enabled the estimation of individual level interactions from aggregated information. He considered a system with macrostates which result from the combination of many microstates where the macrostate is a particular origin-destination flow matrix and a microstate is the movement of an individual from a particular origin to a destination. Each particular macrostate is constrained differently and therefore can result from a different number of combinations of microstates. This number can be calculated as

$$R = \frac{T!}{\prod_{ij} T_{ij}!} \tag{5.3}$$

where T is the sum of all flows in the system and T_{ij} is the flow from origin i to destination j. Without any further constraints, the most likely macrostate to occur is the one which results from the largest number of different combinations of microstates so the goal is to find the set of Tijs that maximises R. This can be written as

$$H = -\sum_{ij} (T_{ij}/T) \ln(T_{ij}/T) \tag{5.4}$$

which in turn can be formulated in terms of probabilities as

$$H = -\sum_{ij} p_{ij} \ln p_{ij} \tag{5.5}$$

where p_{ij} is the proportion of all trips that originate at i and terminate at j (Fotheringham et al., 2000). This is the formula for the entropy of a distribution so the problem of maximising R is now a problem of entropy maximisation. H can be maximised trivially if all p_{ij} values are equal but this is not a useful solution so additional constraints are required. The possible constraints include an origin constraint which requires the sum of the modelled flows from a given origin to equal the observed total for that origin. Similarly the destination constraint is the inverse of this. A distance constraint requires that the sum of the distances of all flows is equal to the total observed flow distances. It was the combination of these constraints that led to the development of a ‘family of spatial interaction models’ including the production constrained model, the attraction constrained model and the doubly constrained model.

While Wilson's work reinvigorated the field of spatial interaction modelling and introduced an important framework of models, the specifics of his approach have been criticised and disregarded in favour of behavioural approaches. The theoretical justification for the model was still lacking a basis in human decision making and thus was the theory was quite removed from the processes being modelled. Other issues were raised regarding the specific derivation of the entropy formulation but later work addressed these by providing alternative derivations (Fotheringham et al., 2000).

5.2.3 Aspatial information processing

The first behavioural approaches to spatial interaction modelling were based on information processing models borrowed from economics. The discrete choice model (Mc Fadden, 1973) provides a mechanism for estimating the choices made by individuals when choosing from a set of alternatives. The intuition behind these models is that a person makes a choice by choosing the option that provides maximum benefit or utility. The utility U_{ij} is a composition of some known observable components V_{ij} and an unobservable component μ_{ij} so $U_{ij} = V_{ij} + \mu_{ij}$. If we had full knowledge of the utility to be had by choosing each of the possible alternatives we could easily say with certainty which alternative would be chosen. The probability p_{ik} of individual i choosing alternative k from N alternatives is therefore

$$p_{ik} = \begin{cases} 1 & \text{if } U_{ik} > U_{ij} (\text{for all } j \in N, j \neq k) \\ 0 & \text{otherwise} \end{cases} \tag{5.6}$$

.

Instead we only know the observable components and have to assume the unobservable components are randomly drawn from a continuous distribution. We can therefore only calculate a probability of choosing each alternative over all others so

$$p_{ik} = prob[U_{ik} > U_{ij}(\text{for all } j \in N, j \neq k)] \tag{5.7}$$

.

In the spatial interaction context, the observable component of the utility of each alternative destination can seen as a function of its attributes. These attributes might include the travel cost, size or other attractive properties. A naive assumption would be that these attributes would have a linear effect on the perceived utility of a particular destination but the true effect on people's decision making may be closer to a logarithmic relationship. Small changes in cost, for example, make a much larger difference for small initial costs than higher ones. By accounting for these effects in the formulation of the utility function we are truly basing the model on behavioural theory, with an understanding of how people make decisions rather than assuming people are attracted to places because of a physical law.

5.2.4 Spatial information processing

While this model incorporates an element of human behavioural theory, it still ignores the spatial context of the decision making process. The effect of geographic space on decision making is not simply that of distance or cost. It also implicitly creates relationships between locations which we take into account when making a decision. According to Fotheringham et al. (2000), the location of a destination with regard to its competitors is important. When choosing a place to shop, for example, our decision to choose a particular store may be influenced by the presence of other stores nearby.

Furthermore, there is evidence (Fotheringham & Curtis, 1992) to suggest that we process spatial information hierarchically, choosing first between large regions and then choosing smaller subregions with in the chosen region. This is quite different from the aspatial choice process suggested by the discrete choice model and indeed is quite a necessary difference to cope with the vast number of choices typically available in a spatial choice problem. Typically discrete choice is used to model the choices made between a small number of alternatives such as brands, transport modes or political allegiance, while the potential number of choices in a spatial choice problem are much greater. Even with some hard constraints set by external factors, the number of choices can far exceed our processing capabilities. However by turning the task into a hierarchical decision problem the number of alternatives to consider is vastly reduced.

Based on the characteristics, assumptions and constraints of the spatial choice process, Fotheringham (1983) developed a spatial interaction model known as the competing destinations model. The model assumes that individuals make their choice from a limited set of possible alternatives through a hierarchical decision process rather than evaluating all possibilities. This allows for the important human ability of making suboptimal decisions by neglecting to consider a better alternative, something which previous models were criticised for. In the terms of the discrete choice model, this model can be expressed as

$$p_{ik} = prob[U_{ik} > U_{ij} + \ln p_i(j \in M)(\text{for all } j \in N, j \neq k)]p_i(k \in M) \tag{5.8}$$

where M is the limited set of possible alternatives that are considered. Of course we cannot know what alternatives are included in the set M for each individual so $p_i(j \in M)$ must be a function of the attributes of j with respect to the other alternatives and not actually require any knowledge of M. The attributes used in this function must only be those that affect whether or not a particular alternative is considered. Fotheringham (1983) suggests that the accessibility of an alternative k to all other alternatives is a suitable proxy for the probability of a alternative being a member of the limited set. The rationale behind this is as follows; the higher the accessibility or proximity of an alternative to other alternatives, the more likely it is to be considered as part of a large cluster. However, large clusters are less likely to be selected due to the psychophysical law that people underestimate the size of large objects, so an alternative in close proximity to many other alternatives is less likely to be selected. The competing destinations model combines a number of important positive attributes of spatial interaction models. Most importantly, it is based on theory of the decision making process of individuals, accounting for the fact that people don't consider every possible option and make suboptimal decisions. Furthermore, the predicted number of flows to any given destination is dependant on the existence and location of alternative destinations so the predictions will change with the addition or removal of alternatives. Finally, the model is parametric and thus has explanatory power when calibrated for modelling systems with different types of flows or different spatial extents.

5.2.5 Model Calibration

In all of the models discussed so far there are parameters which need to be estimated to fit the dataset under study. An exponent of -2 for the distance parameter in a gravity model rarely gives the best fit and so multiple possibilities must be examined. This parameter can be estimated by a simple search process such as a binary search or golden section search. However such a simplistic model with only a distance parameter is not very common or useful so there are usually exponents on each of the production and attraction parameters. In this case the parameters have traditionally been estimated by linearising the model equations in terms of their parameters and using ordinary least squares (OLS) regression (Fotheringham et al., 2000). Equation 5.9

can be linearised by taking the logarithms of each side, resulting in:

$$\ln T_{ij} = lnk + \alpha \ln P_i + \lambda \ln P_j + \beta \ln d_{ij} \tag{5.9}$$

While the values of $lnk, \alpha, \lambda, \beta$ can easily be estimated using OLS regression it has been pointed out by Flowerdew & Aitkin (1982) that there are a number of problems with this approach. These problems stem from the use of the OLS in a log space and the fact that OLS is used under the assumption that the dependent variable (the flows) is normally distributed. Count data such as number of commuters, migrations, phone calls or seconds of call time between locations i, j are always non-negative integers so an assumption of a normal distribution is inappropriate. Flowerdew & Aitkin (1982) therefore suggest that a model based on the Poisson distribution is more appropriate when modelling count data. The parameters of Poisson regression can be estimated using a maximum likelihood procedure.

5.2.6 Local modelling

The discussion so far has focused on global approaches to spatial interaction modelling, where the aim is to find the set of parameters that give the best fit for the model for all flows in the study area. The idea here is that these parameters capture the differences between different types of flows and different study areas. There is no allowance however for variation in parameters within a single study. There is an assumption underlying this approach to modelling that the process generating the flows is stationary in space. This implies that people make spatial choice decisions in same way in all parts of a country. If the process exhibits spatial non-stationarity the model will poorly predict and provide misleading explanations in some or all areas as it only captures an average. This issue is not unique to spatial interaction modelling and has been raised numerous times over the past three decades (Openshaw et al., 1987; Fotheringham, 1997; Lloyd, 2010), with a call for a move towards local forms of analysis that focus on the differences across space rather than global forms focusing only on the similarities. In the context of spatial interaction models a simple step in this direction is to map the model residuals. This at least allows one to identify spatial patterns and test for spatial non-stationarity.

A more useful alternative is to actually model different locations separately, generating a set of parameter estimates for each location. Haynes & Fotheringham (1984) discuss the use of origin- and destination-specific models to do precisely this. These model the flow from a single origin to multiple destinations or multiple origins to a single destination. By fitting such a model to each origin or destination we get a set of parameter estimates for each location that can be compared with each other or with an average, thus allowing the identification of differences across space in interaction behaviour.

5.2.7 Recent work

Somewhat surprisingly, a recently proposed model brings us back to physical analogy, explaining interactions in terms of particles and once again neglecting to consider how individuals make decisions in space. Simini et al. (2012) proposed a new parameter free model of spatial interaction, derived from first principles, that they call the radiation model as it can be formulated in terms of radiation and absorption processes. The model is similar to Stouffer's intervening opportunities model in that the number of flows T_{ij} from i to j is dependant on the total number of opportunities s_{ij} available at locations within the same distance as between i and j. The predicted flow T_{ij} from i to j is

$$T_{ij} = T_i \frac{m_i n_j}{(m_i + s_{ij})(m_i + n_j + s_{ij})} \tag{5.10}$$

where m_i and n_j are the populations of the origin and destination respectively and T_i is the total number of flows from origin i. Unlike the intervening opportunities model, this model is parameter free, leading to claims of its universality. This is not necessarily a positive attribute though as it restricts the model's usage to prediction rather than explanation.

5.3 Modelling telecommunications data

Intercity telecommunication flows can be seen as a proxy for personal and business relationships, information exchange and even financial transactions and have therefore been of interest to geographers and economists for a long time (Guldmann, 2004; Taylor, 1994).

The recent availability and high spatial accuracy of mobile phone datasets for research has led to a renewed interest in interactions and information flow across space. Given the wealth of research on modelling spatial interactions and developing a theoretical basis of spatial choice, it is surprising to note that the recent work in this area is based on the most primitive of gravity models.

5.3.1 Distance decay and gravity models

Lambiotte et al. (2008) purportedly show that a gravity model can accurately model the probability that two people are connected in a mobile phone network in Belgium. They define 'connected' as having at least six reciprocated calls in the six month period under study. They compute the probability that two individuals separated by distance d are connected as $P_d = \frac{L_d}{N_d}$ where L_d is the number of pairs of connected individuals separated by distance d and N_d is the total number of pairs of people separated by distance d and where the separation distance is defined as the distance between the centroids of the zip codes of their billing addresses. They report that this probability "P_d is very well approximated by a gravity model $\sim d^{-2}$, over a large range of distances."

While this may not initially look like a gravity model we can easily get the formulation in equation 5.9 from this result. The expected number of pairs of connected individuals T_{ij} between locations i and j is $N_i N_j P_{d_{ij}}$ or the possible number of such pairs multiplied by the probability of a pairing at the given separation distance. Substituting $d_{ij}^{-2}k$ for $P_{d_{ij}}$ gives us $T_{ij} = \frac{N_i N_j}{d_{ij}^2}k$, which is precisely the same as the gravity model in equation 5.9.

We may question if it is actually accurate to refer to this as a gravity model however as it simply says that the probability of two people being connected decreases with the square of the distance between them without being influenced by the population or size of their respective locations. The product of the terms N_i and N_j only serves to give the maximum number of possible connections so that we can calculate an expected number of connections with a probability. This is in contrast to typical use of the gravity model, for example in transportation planning or migration prediction, where the mass variables represent the attractive forces of the locations and actually influence the probability of a trip or migration.

The authors also admit that this analysis assumes the system is homogeneous and acknowledge that this is certainly not the case in Belgium where the country is geographically divided into two regions where people speak different languages. They show that the link probabilities for three particular cities and observe clear north south asymmetries but they do not modify their modelling approach to account for this or show the performance at a local level. One has to wonder how such a model could possibly provide accurate prediction or explanation of connection probability given this extreme inhomogeneity.

Onnela et al. (2011) present similar statistics in their analysis of communication and social ties in an unnamed European country. They use logs on call and text message communication between 3.4 million people with a total of

5.2 million undirected edges. Using the location of most activity to determine a location for each customer, they calculate the probability of a tie between two people at a given distance in a similar manner to Lambiotte et al. (2008). In this case a tie refers to a reciprocated call or SMS between two customers. They claim the probability decays with distance following a power law of $d^{-1.58}$ for voice ties and $d^{-1.49}$ for SMS ties. They attribute the differences with the results reported by Lambiotte et al. (2008) in Belgium to the technique used to estimate the exponents, the greater statistical power due to larger distance ranges, the use of maximal phone use locations rather than billing addresses and the difference in population densities between the two countries.

Krings et al. (2009) use the same dataset as Lambiotte et al. (2008) but analyse it in a slightly different manner. In this study a gravity model is used to model the communication flows between 571 cities in Belgium. Again the customers are assigned to locations based on their billing zip code, but aggregated to city level, and only calls between customers who have had at least six reciprocated calls are included. Unlike the previous study however, the total communication time is modelled, rather than the probability of a connection. Applying the standard gravity model they model the predicted communication time between cities A and B as $L_{AB} = K\frac{P_A P_B}{d_{AB}^2}$ where P_A is the calling population of city A, d_{AB} is the distance between the centroids of cities A and B and the "gravitational constant K can be estimated with a simple best fit of the data". Unfortunately the distance decay exponent of 2 seems to have been set *a priori* so the single tunable parameter K captures both the modifications to the distance and size parameters, limiting the explanatory power and comparability of the model. Again the same caveats as the previous study apply here. The model is only shown to be accurate on a global scale and there is no mapping of residuals to understand whether there maybe spatial patterns involved, as one would expect given the unique human geography of the country. Interestingly the authors also report that the average intensity of intercity communication decreases with the square of the distance between the cities, or $L_d = cd^{-2}$ where L_d is the average intensity of links L_{AB} where $d_{AB} = d$ and c is a constant. This is a rather confusing result as it suggests that the populations P_A, P_B have no effect on the average intensity, thus contradicting the other result that the gravity model, incorporating P_A, P_B, is a good fit.

Kang et al. (2012) take a similar approach to Krings et al. (2009) to study inter-city mobile communications in China using a gravity model and discover a surprisingly low distance decay exponent suggesting a very marginal effect of distance on telecommunication in China. They note that it is often impossible to conduct such a study for an entire network in a country as big as China and its is often the case that only a subset of communications data is available. In this case data is only available for a single province in north east China, Heilongjiang, for one month. They define three different networks based on different modes of communication. The largest is the intercity mobile communication network (IMCN) of aggregated calls between all major cities, for which they lack data. The second network they define is the outgoing telecommunication subnet (OSN) which consists of all outgoing calls from each of the 13 cities within Heilongjiang to each of the 283 other major cities in China. Finally, they define the roaming telecommunication subnet (RSN) as the network of calls between each of the 283 cities outside Heilongjiang where at least one of the customers is registered in Heilongjiang and hence the data is available to them. The authors note that "information and communication technology thrives in the past decades and remarkably eliminates the geographical constraints upon interactions between individuals." They seek to test this hypothesis by fitting a standard gravity model to the RSN and indeed find a very low distance decay exponent of 0.5, which they suggest confirms that "the distance constraint is marginal in the IMCN". This claim is highly questionable however given the nature of the RSN network. It is a network of the interactions of people from Heilongjiang province while roaming in other provinces. The individuals involved are therefore displaced from their true origin location, which may well explain the observed marginal effect of distance as the spatial configuration of the network has been distorted. While the authors do acknowledge that the properties of the RSN may not be representative of the IMCN and provide a test for this, their methodology is again questionable. Given that the only data available to them is from Heilongjiang province, they must use this data to test the model which results in a testing bias as the same population subgroup is used both for fitting and testing the model while extrapolating the results to the population of the entire country.

As discussed in Chapter 4 Expert et al. (2011) use a gravity like model with mobile phone call data in a slightly different manner. They use the model as the basis for a null model in a spatial version of the Newman-Girvan modularity quality measure for community detection in spatial networks. The model includes a distance influence function rather than a distance decay exponent, negating the need to find the best fit parameters. The number of calls between i and j is given by $T_{ij} = N_i N_j f(d_{ij})$ where $f(d) = \frac{\sum i,j|d_{ij}=dA_{ij}}{\sum i,j|d_{ij}=dN_iN_j}$ which is the weighted average of the probability for a link between nodes at a distance d. $f(d)$ in this model is identical to Lambiotte et al. (2008)'s P_d. The purpose of this model is to capture the effect of space so that relationships determined by factors other than space can be identified. Given that the aim is not to accurately model the interactions it is impossible to verify if the model really does capture the full effect of space and only the effect of space. The distance influence function $f(d)$ is by definition an average of observed probabilities so it is certainly affected by factors other than distance or space. Likewise if model parameters were fit to the data in the usual manner the same issue would arise as the model can only capture an average of the observed flows which is not necessarily in any way similar to the communityless spatial network sought.

5.3.2 Radiation model

Simini et al. (2012) use their radiation model to model the mobile phone calls between municipalities in a unnamed western European country, along with hourly trips in the same country and commuting, migration and freight in the USA, in an effort to show how it can accurately model a number of different human interaction phenomena. While the authors claim the model "offers an accurate quantitative description" of these various types of flows, the baseline model used for comparison is the gravity model and the performance is only reported at a very broad level. Apart from the lack of strong evidence for the claim of an accurate quantitative description, there is no evidence provided to support the idea that people make decisions about communication in the manner suggested by the model. While it may be reasonable to suggest that intervening opportunities may be important in the choice of destination for commuting or migration, no such argument may be made for mobile communication.

5.3.3 Discussion

While the number of studies on spatial interaction in mobile phone call networks is limited, we see some worrying trends emerging. As with the earliest studies of migration and commuting, these studies take a social physics approach, preferring physical analogy over behavioural theory. The focus has been on global fitting of simple gravity models. As we have seen in the development of the spatial interaction literature, these models have long since been disregarded due to their lack of a behavioural interpretation, their focus on global similarities, and their poor local fit. The usefulness of these models is therefore rather questionable. Where a more behavioural theory based approach has been attempted (Simini et al., 2012), the differences between communication and mobility behaviour have been ignored, rendering the behavioural underpinning meaningless. Furthermore, the methodologies and interaction networks employed in each study differ significantly in incompatible ways, yet the findings are all presented as models of interregional mobile communication flows. In this section we will address each of these issues in turn and make use of the Irish mobile phone call network to assess the significance of different methodological approaches.

Geo-referencing and spatial aggregation

Lambiotte et al. (2008) and Krings et al. (2009) each use the same raw dataset, creating connections between the regions containing the billing address of each customer. In this case the location of the customer at the time of call doesn't matter. Onnela et al. (2011) use the location of maximal phone activity to assign a single location to each customer. On the other hand the studies of Simini et al. (2012) and Kang et al. (2012) both use time of call locations. Clearly these two approaches result in different interaction networks and models unless we assume that people do not make calls outside of their home region, which we know not to be true. In each case the model is based on the properties of the origin and destination regions so we get different interpretations of the results.

In the first case the model can be interpreted from two different perspectives. The straightforward one is that the number of calls between residents of regions i and j is dependant on the attractive properties and distance between i and j. The more complicated interpretation is that the number of calls between regions i and j is dependant on the attractive properties and distance between the home regions i' and j' of the individuals making calls from regions i and j. If we wish to frame these models as models of the number of calls or intensity of interactions between regions, as in (Krings et al., 2009), then we must use this second interpretation.

The second modelling approach using time-of-call locations and a gravity model tells us that the number of calls between regions i and j is dependant on the attractive properties and distance between i and j.

Finally, the approach using time-of-call locations and the radiation model tells us that the number of calls between regions i and j is dependant on the attractive properties of i and j and all other regions within the same distance as between i and j.

Behavioural interpretation

It should be clear that a rational behavioural interpretation of any of these models is impossible. They lack any variables which capture the likelihood of two individuals knowing each other and having a reason to communicate. Unlike commuting or migration we cannot say that an individual may be attracted to communicate with another because of the attractive properties of their location. It is even more unreasonable to think that the properties of the location which they are at at the time of a call could solely determine this.

If these models are so devoid of behavioural theory we must account for why they *do* accurately predict mobile communication interactions, as evidenced by the reasonable goodness-of-fits and correlation plots reported.

Accuracy measures

The effect of the misuse of these models is not immediately apparent because the accuracy is generally reported with global measures of goodness of fit and correlation plots which mask the local misfit. If we were to map, or plot all the predictions of these models we would likely find that they do not actually represent the observed interactions to any degree of accuracy. Lambiotte et al.

(2008) show that the probability of two individuals being connected at a given distance is “well approximated” by d^{-2} by plotting the probability against d^{-2} for multiple values of d. The pairwise interactions of 1145 nodes (postal areas) are binned into 5km bins and cut off at 100km, so a total of 654940 probabilities are represented by 20 data points on the chart. Given the degree of aggregation, the results and claim of good approximation are highly questionable.

Krings et al. (2009) show the accuracy of the gravity model with a plot of observed intensity against the intensity predicted by the model. The values appear to be binned, but error bars are shown to give some indication of standard deviation from the mean observed value. The errors are shown symmetrically on a log scale however, indicating that they were calculated from the logged values. This means that the error is actually very large for pairs of cities that have large estimated intensity, contradicting the reported result.

Simini et al. (2012) show the accuracy of their radiation model by plotting the observed number of calls against the predicted number of calls for all pairs of municipalities. When the data is binned the mean values of each bin do indeed fall close to the line corresponding to a perfect fit but the median values do not fit so closely and the interquartile ranges are large, although this is not immediately apparent because of the log scales used.

Calls or social ties

A further important issue is the focus on modelling the number of phone calls rather than the number of social ties or relationships. Each of the studies is ostensibly using mobile phone call flows as an indicator of the interconnectivity of the regions under study, yet only Lambiotte et al. (2008) and Onnela et al. (2011) actually differentiate between numbers of phone calls and social links. The rest of the works are based on the implicit assumption that a higher number of calls means a stronger tie between the two regions. There are two issues with this however. The first is that each call is not an independent event, but rather one of many interactions between the same pair of individuals. This is a problem from a modelling perspective. Secondly, we cannot assume that the number of calls per pair is a constant. As discussed in Chapter 3 the actual number of interactions between a pair very much depends on the communication style of the individuals involved and is not necessarily representative of the intensity of the link.

Chapter 6

Analysis

6.1 Introduction

In this chapter we will perform quantitative analysis on data from a mobile phone network to investigate some of the issues discussed in the previous chapters. We will begin with a description of the dataset and provide some descriptive statistics which will guide our further analysis. Following this we will apply some of the methodologies of the previously discussed analyses on spatial interaction and community detection. We will discuss issues we find with the existing approaches and propose and test alternatives.

6.2 Data description

The dataset used in this chapter comes from a national operator in Ireland with approximately 1.5 million customers. The period under study covers 70 non-consecutive days between December 2010 and February 2011. The CDRs include all calls and SMSes for every customer during this period. In all subsequent analysis we do not differentiate between call and SMS communications unless explicitly stated and for ease of discussion we simply refer to both types of communication as calls. The data includes records for calls involving customers of other networks but obviously no location information is available for the other party. For most of the following analysis these records are ignored as we require spatial information. In the following section we will provide statistics on the number of calls that this effects.

The network comprises approximately 9000 uniquely identified transceivers. These transceivers are physically grouped together at approximately 1400 unique locations, which we refer to as towers for ease of discussion even if not strictly true. These towers are spread throughout the country but unevenly distributed, with higher densities in urban areas. The CDRs include the transmitter identifiers but the spatial information is limited to towers so we replace all transmitter identifiers with tower identifiers.

For much of our analysis we aggregate towers to approximately 300 towns (urban areas with populations greater than 1000 people, see Figure A.2) for a number of reasons. As discussed in Section 2.1, the accuracy of spatial positioning using towers is limited as the coverage areas of the towers are not well defined. This adds a lot of noise to the data and gives a false sense of spatial accuracy. By choosing towns as our spatial unit we remove much of this noise without too much loss in spatial accuracy. The towns are defined as continuous areas of settlement so by definition they are non adjacent, removing much of the problems associated with the modifiable areal unit problem. However we risk missing interesting patterns by considering cities as single homogeneous units so we divide them into sub regions using the electoral area boundaries, which unfortunately are continuous and adjacent. A further advantage of using officially defined regions is that census statistics are readily available. Finally the use of well known boundaries improves ease of interpretation. It must be noted that the use of non-continuous regions means we ignore data generated from calls taking place in rural areas outside of these towns. We will quantify the effect of this in the next section.

It must be noted that the dataset used in this analysis comes from the third largest and third oldest operator in Ireland with approximately 20% of the market during the period of study (Commission for Communications Regulation, 2011). While statistics are unavailable on the demographics of the customers, it is commonly known that this operator aims its services and advertising towards younger individual consumers rather than business customers. The particular demographics of the individuals included in the dataset will obviously have an impact on the results so this must be considered when interpreting the results presented in this chapter. Furthermore the operator has offered free calls and SMSes between its own customers for many years so the proportion of intra-network communications may be different to other operators in Ireland and elsewhere.

6.3 General statistics

6.3.1 Calls and contacts

The dataset includes records of approximately 1.5 billion calls between nearly 12 million unique numbers. Of these 12 million unique numbers, 1.5 million are customers of the operator under study. We remove calls from all customers who have no reciprocated ties (at least one call to and from the same person) or those who have more than 1000 ties, leaving approximately 1 million customers. Between these customers there are 10.7 million pairs and 817 million calls. Of these pairs, 68% (7.3 million) are reciprocated ties. This 68% account for 99% (809 million) of the calls however. Rather than considering only the number of calls between pairs however we also look at the number of distinct days upon which there is contact between a pair of customers. We find that only 56% of pairs have contact on two or more days, 23% on 10 or more days and 7% on 30 or more days. True to the Pareto principle, the 23% of contacts who communicate on at least 10 days account for approximately 89% of all in-network calls and the 7% account for 65%. Additionally 79% of all customers have at least one contact with whom they communicated with on 10 or more days. This decreases to 55% for 30 days.

The national median number of reciprocated ties per customer is 26 with 9 of these in-network ties (contacts who are also customers of this operator).

6.3.2 Locations

Most of the studies discussed previously use the customers' home location to geolocate calls. In some cases these are provided by the operator along with the CDRs and in other cases they have to be determined from the data. This dataset does not include any personal information aside from the CDRs so we must determine home locations from the data. We will later expand this to find multiple significant locations for each user, as discussed in 3.3.4. We consider the most significant location for each customer to be the tower at which they have made or received calls on the most number of days. While this may not strictly correspond to their actual home location, it is the location with which they are most associated. For comparison purposes we also find the tower with the most number of calls and the towers with the most number of unique weekdays and weekend days. We find that 85% of customers have the same home tower using either the total number of calls or total number of days metrics. Additionally, 75% of customers have the same home tower using weekdays or weekends. 65% of customers have the same home tower for each of the four metrics. In Figure 6.1 we show a cumulative distribution of the number of days each customer is recorded at their home location. The median is 50 days.

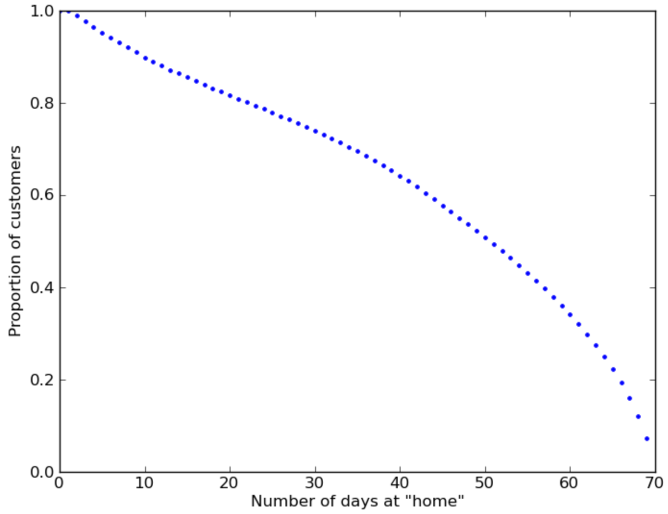


Figure 6.1: Cumulative distribution of the number of days each customer is registered at their home tower.

As we have previously mentioned, much of the further analysis will be conducted at the town level rather than at the level of individual towers. 99% of all towers have at least 1 customer whose home location is that tower. 72% of these towers are contained within one of the 300 towns with a population of 1000 people or more. These 72% are the homes of 81% of the customers. If we assign people's homes as the town which they have visited on the most number of days, with a minimum of 10 days, we can assign 91% of customers to one of the 300 towns. Subsequent mentions of 'home towns' refer to this assignment of customers to towns.

6.4 The effect of distance on the number of connections between towns

Now that we have an assignment of customers to towns we can investigate the spatial properties of communication in Ireland. First we look at the ratio of intra-town to inter-town ties. We find that 51% of ties are intra-town (the two customers share the same home town). These 51% account for 54% of the total number of calls, demonstrating that on average the highest intensity ties are between those who are spatially close. We attempt to investigate the effect of separation distance on the remaining 49% of (inter-town) ties using two different approaches, used by Lambiotte et al. (2008) and Simini et al. (2012) among others.

Using the definition from Lambiotte et al. (2008) of the probability of a connection between two individuals separated by distance d (Section 5.3.1), we attempt to quantify the precise effect of distance on the probability of a communication tie between two individuals. Once again, they define this probability as $P_d = \frac{L_d}{N_d}$ where L_d is the observed number of pairs of connected individuals separated by distance d and N_d is the total number of pairs of people separated by distance d and the separation distance d is binned into 5km bins. The distance is measured from the centroids of each town. Given the symmetric adjacency matrix of inter-town communication ties A, we calculate L_d as $\sum_{ij|d_{ij}=d} A_{ij}$ and N_d as $\sum_{ij|d_{ij}=d} N_i N_j$ where N_i is the total number of customers whose home location is town i. Using 5km bins we have 100 probabilities, with a maximum inter-town distance of approximately 500km. Plotting P_d against d on log axes we see in Figure 6.2 that it does not have the straight line signature expected of a power law. In fact with a linear distance axis we can see that the values do not always decrease with greater distance but actually increase at some distances. This suggests the probability of a connection between individuals does not decay as smoothly in Ireland as Lambiotte et al. (2008) suggest it does in Belgium.

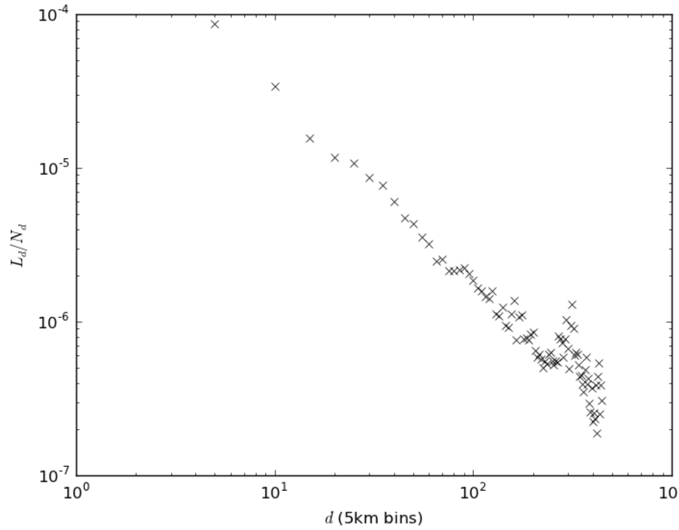


Figure 6.2: Observed probability of a connection between individuals separated by distance d. (Using the probability definition of Lambiotte et al. (2008))

Given that P_d is an average over all values in 5km bins we should test how well it approximates the unbinned values. We investigate this by plotting box plots of the values of each bin along with the mean value line in Figure 6.3. We find that there is a lot of variance at all distances and the mean value (P_d) is consistently below the median after 100km, and often below the 25th percentile. Furthermore after approximately 100km we see that there is no decay in the median value, suggesting the fluctuations we see in the mean are the result of outliers. The large variance and presence of many outliers shows that distance does not affect the probability for all pairs of towns equally.

Intuitively we do not actually expect the probability of connections to decrease perfectly with distance. Rather we expect that there is a higher probability of connections between larger towns. We could hypothesize that the outliers correspond to town pairs with higher populations and more observed number of connections. In actual fact if we plot the top 100 town pairs (green dots) in terms of observed number of connections we see that none of them are outliers and and the majority fall below the median probability for their separation distance. Furthermore if we plot the bottom 100 town pairs (red dots) we see that many are above the median and a few are even outliers in the positive direction.

This goes against our intuition so we return to the definition of P_d to understand how it is calculated. Lambiotte et al. (2008) refer to P_d as a probability because it is a measure of the observed number of occurrences of an event (a tie at a given distance), divided by the total possible number of such events. They define the possible number of ties at a given distance as $\sum_{ij} N_i N_j, d_{ij} = d$, because the possible number of connections between two towns is $N_i N_j$, or the total number of possible pairs of people. While this is indeed the theoretical maximum number of ties, it is much larger than the realistic maximum given the realities of human social interaction. The number of ties per person cannot possibly increase unbounded with the number of possible opportunities for interaction. It is well known that the cognitive limit on the number of social ties any individual may maintain, due to the size of our neocortex, is approximately 150 (Dunbar, 1993). We noted earlier that the median number of reciprocated communication ties per person in this network is a mere 26, 9 of which are likely to be in-network contacts. Furthermore we saw that 51% of all ties are intra-town. We should therefore not expect to find more than four or five inter-town calls per person between any pair of towns.

To provide a simple example, suppose we have three towns, A, B and C, equally separated in space with populations of 100, 200 and 300 respectively. If the people of town A make ties with equal probability with people from towns B and C we can imagine that each resident of town A has 2 contacts in B and 2 in C. So $L_{AB} = 2 \times 100 = 200$ and $L_{AC} = 2 \times 100 = 200$. Now if we measure the probabilities of ties we find that $P_{AB} = \frac{200}{100 \times 200} = 0.01$ and $P_{AC} = \frac{200}{100 \times 300} = 0.0066$. Clearly there is something wrong here as we assigned ties with equal probability between B and C yet the supposed observed probabilities are different. One could argue that this is because we only assigned ties from A to B and C, and not in reverse but social ties are undirected and it is obviously impossible for there to be more ties between B and A than between A and B. If we had assigned ties from the perspective of the larger town, B, then there would be $2 \times 200 = 400$ ties between A and B which would mean 2 ties per person in B but 4 ties per person in A. This is still possible but it doesn't scale. Now let us imagine town D is larger again with a population of 10,000 people. If we assumed that each person in town D had 2 ties in town A there would be $2 \times 10,000 = 20,000$ ties between the two towns, with an average of 200 ties per person in town A. If we multiply the populations the expected number of ties would be $100 \times 10,000 = 1000,000$, with an average of 10,000 ties per person in town A. Clearly the only reasonable approach is to assume the number of ties between two towns scales in proportion to the smaller population.

It should now be easy to see why the curve of P_d does not fit with our expectations. The value of the denominator, N_d, is unrealistically large, especially for separation distances involving large population centres, as it scales at a faster rate than the observed number of connections, L_d. N_d scales with the product of the populations whereas L_d would scale in proportion to the size of the smaller town if there were no effects of distance or population attraction. It therefore appears that the ratio L_d/N_d is not an appropriate measure of the observed tie probability at a given distance and some better alternative is required that is not biased against large (or small) populations.

As we have already discussed, the number of ties between any two towns must be bounded by the population of the smaller town and Dunbar's number (150). We have seen that the limit is much lower than 150 in this network, and we further know that approximately half of the ties are within the same town. However we do not know the effect of distance on the number of interactions as that is precisely what we are trying to measure so rather than trying to predict a value for the number of ties between any particular towns, we can simply say that the number of ties must be proportional to the smaller town's population. We can not formulate a probability with this information as we don't know the possible number of ties but we can derive a comparable measure of inter-town tie strength, which is simply

$$R_{ij} = \frac{L_{ij}}{min(N_i, N_j)} \tag{6.1}$$

where L_{ij} is the number of ties between towns i and j and N_i is the population of town i, as before. We do not include Dunbar's number or any similar value representing the cognitive limit per person because we must assume this value would be constant across space so it would simply scale all the values. It should be noted that the absolute value of this ratio is unimportant. Rather it is most useful to look at the value in comparison to others, as we now do.

We plot the boxplots for 5km bins as before in Figure 6.4 along with the top and bottom 100 town pairs. We see that now nearly all of the top 100 are above the median value and many are outliers while much less of the bottom 100 are above the median value. We should also note the difference in the y scale between the two plots so the variance is multiple orders of magnitude less in this case.

We will return to this measure in Section 6.7 where we will use it to help assess the quality of node groupings found through community detection.

An alternative method of investigating the effect of distance on connection probability, as employed by Simini et al. (2012), is to calculate the fraction of connections that occur at a given distance or $P(d_{ij} = d)$. This is different to the definition of Lambiotte et al. (2008) because it is the probability that the separation distance of a given tie is d. This has the advantage of being a proper probability distribution rather than a set of independent probabilities

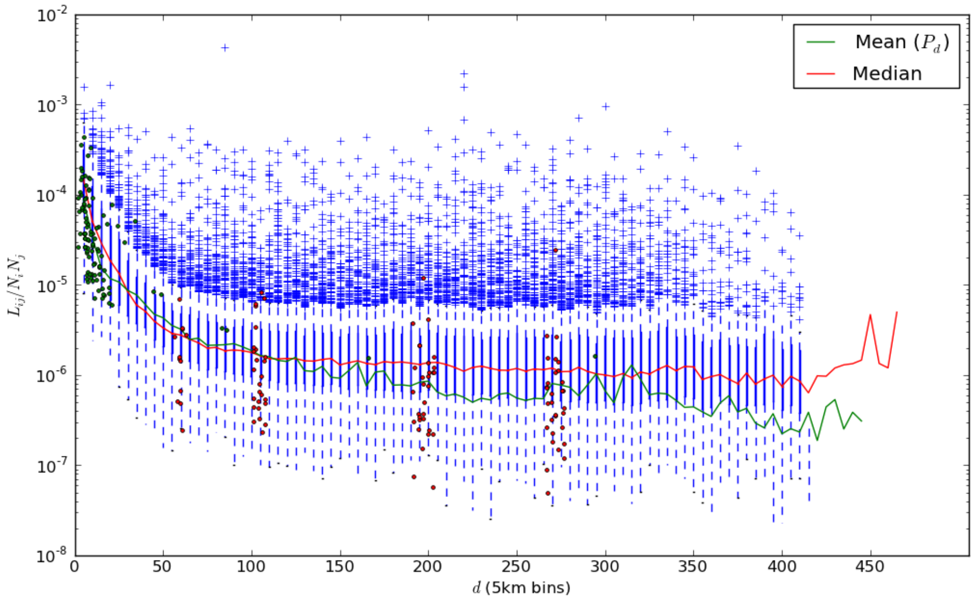


Figure 6.3: Observed probability of a connection between individuals separated by distance d with each 5km bin represented as a box plot. (Using the probability definition of Lambiotte et al. (2008))

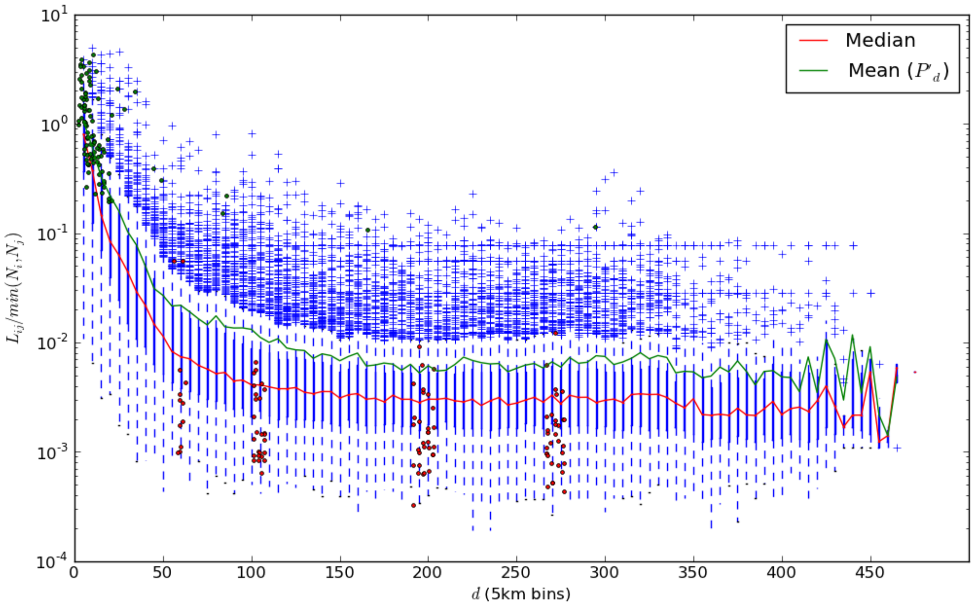


Figure 6.4: Ratio of the observed number of calls between towns i, j to the population of the smaller town, with each 5km bin represented as a box plot.

but it is still very affected by the distribution of population. If there are less people separated by a distance of 100km than 10km then there will be a smaller fraction of calls at 100km than 10km, irrespective of any distance decay effects. We must compare the observed probability distribution with a random distribution to understand the true effect of distance. The probability that a randomly created edge has a distance d is the fraction of possible edges that are of distance d, which is $\frac{\sum_{d_{ij}=d} N_i N_j}{\sum N_i N_j}$. We plot this random distribution along with the observed distribution $(\frac{\sum_{d_{ij}=d} L_{ij}}{\sum L_{ij}})$ in Figure 6.5.

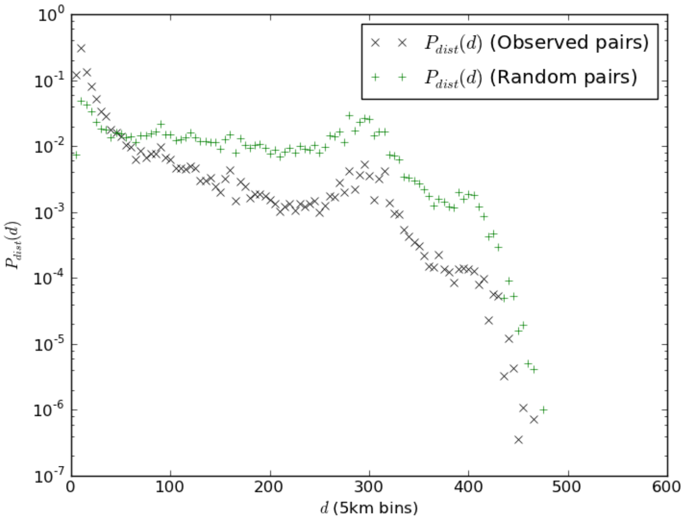


Figure 6.5: Observed probability of a connection at distance d and prior probability of a connection at distance d given the population distribution. (Using the probability definition of Simini et al. (2012))

We see that these variables are correlated and much of the variation in the observed values can therefore be explained by the spatial distribution of the population.

6.5 Gravity and radiation models of communication

As discussed in Chapter 5, Simini et al. (2012) apply a radiation model to call data from an unnamed European country and report a good fit. While we previously discussed the limitations of this approach, we apply it to this dataset for comparison purposes. We show results in Figures 6.6 and 6.7 for inter-town contact pairs and calls respectively. The results are presented in the same manner as the original paper for ease of comparison.

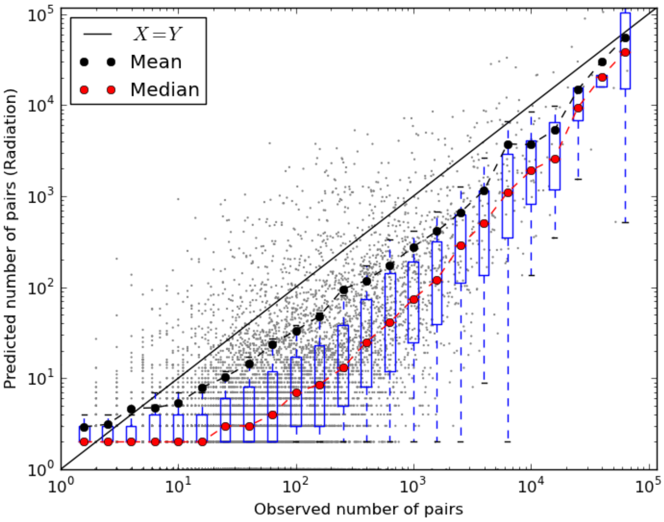


Figure 6.6: Observed number of inter-town contact pairs versus predicted number using the radiation model.

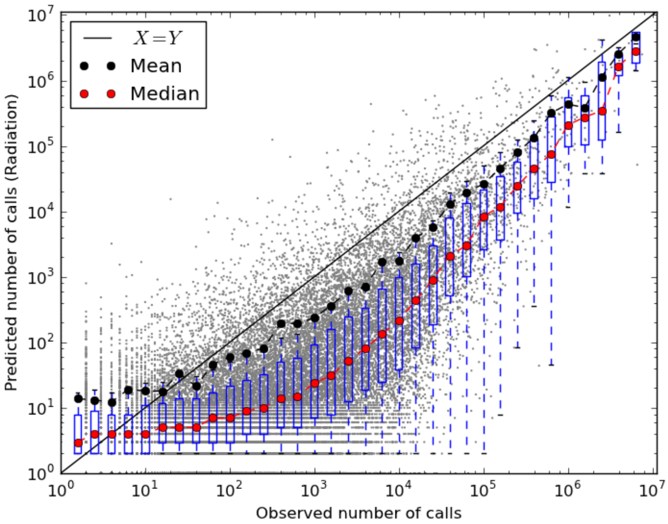


Figure 6.7: Observed number of inter-town contact calls versus predicted number using the radiation model.

For both datasets we can see that the model under predicts the vast majority of the data. We also note the there is a large amount of variance in the bins for all link sizes. Although it appears to be a better fit for the larger links to the right of the plots we must be aware that the log scale distorts the size of the errors. These results prove that the model cannot accurately predict communication flows or social ties in all circumstances and it is certainly not as universal as the authors claim.

We also show results for the basic gravity model used in Krings et al. (2009) in a similar manner. These also show a poor fit with very wide variance for links of all sizes (Figures 6.8 and 6.9).

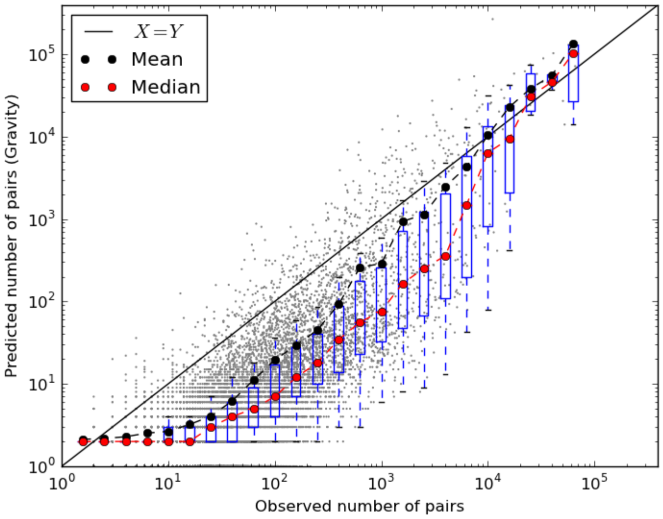


Figure 6.8: Observed number of inter-town contact pairs versus predicted number using the gravity model with d^{-2}.

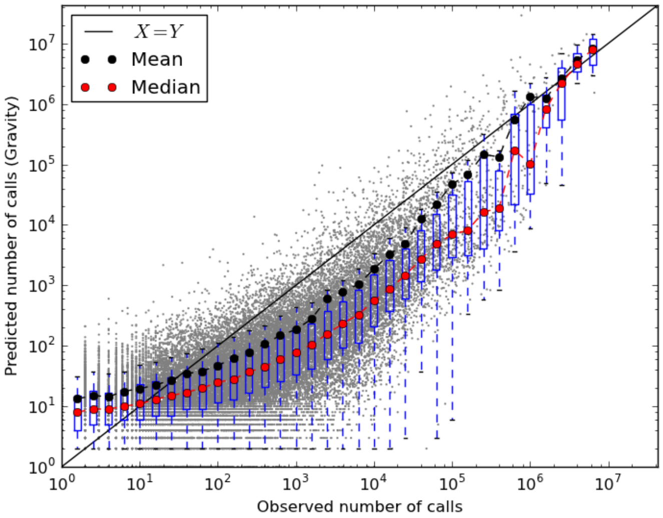


Figure 6.9: Observed number of inter-town contact calls versus predicted number using the gravity model with d^{-2}.

We do not attempt to find a better fit with either of these models because we believe they are unsuitable for the data at hand due to their simplistic nature, as discussed in Section 5.3.1. In the following section we will demonstrate that the nature of social ties is much more complex than can be modelled with such models.

6.6 Multiple significant locations

As we previously discussed in Section 3.3.4, the idea of a single 'home' location for each individual is a too simplistic view of the world so we will identify multiple significant locations for each customer. Rather than setting the number of such locations per person, we will define a significance level and allow each individual to have any number of such locations. We will then find the proportion of ties where both customers share a significant location. The purpose here is to measure the role which 'home towns' actually play in communication. When analysing communication behaviour we generally place each individual at their determined home location and model the interactions between homes. By doing this we neglect the fact that people regularly travel to other locations for work, education, leisure and entertainment purposes and meet others at those locations who may also come from different origins. In this case properties such as the separation distance or populations of the respective origin towns may have little relevance to the likelihood of a communication or social tie.

We identify the significant locations for each customer by finding all the towns from which they have made or received calls on at least a given number of days. The five number summaries of the number of significant towns per person for a range of thresholds from 1 day to 70 days(every day) are given in Table 6.1. With 70 days of data, 10 days corresponds to an average of one call a week from that location. We can see that at least 50% of customers have more than one location they have visited on at least 15 days.

With the 1 day significant locations for each customer we can repeat the analysis of Calabrese et al. (2011b) to find the percentage of pairs of communicating customers who have also visited the same town, though not necessarily at the same time. In the case of Portugal this value was found to be 93% over the course of a year. In Ireland we find it to be even higher at 97% for this

Days	Number of towns				
	Min	LQ	Median	UQ	Max
1	1	4	9	15	171
5	0	2	3	5	58
10	0	1	2	4	35
15	0	1	2	3	31
20	0	1	1	2	23
25	0	1	1	2	19
30	0	1	1	2	15
35	0	0	1	2	15
40	0	0	1	1	12
45	0	0	1	1	10
50	0	0	1	1	9
55	0	0	1	1	9
60	0	0	0	1	9
65	0	0	0	1	6
70	0	0	0	0	3

Table 6.1: Number of significant towns per person with increasing thresholds on the number of days defining significance.

period of 3 months. This suggests there is a strong link between the places that people visit and the people who they call. We should be careful about the conclusions we draw from this however as these visits to shared towns may only be one off and not at the same time. Instead we may look at the percentage of those pairs who share real significant locations in their lives.

Setting a threshold of 10 days (an average of once a week) we see that 88% of contacting pairs share at least one location that they have each visited at least 10 times. Of the pairs who have communicated on at least 10 days (23% of the total), 90% of them share a 10 day significant location.

Intuitively we would expect that pairs whose home towns are closer are more likely to have shared significant towns. We find that this is indeed the case, but the proportion does not decay continuously with distance. Figure 6.10 shows that 95% of contacts whose home towns are less than 10km apart have shared towns. This decreases to approximately 55% at 70km and then levels off, with a median value of 52%. This is an important finding because it shows that most of the frequent contact pairs whose homes are separated by large distances still have significant locations in common.

Intriguingly this suggests both that distance is no longer a limiting factor for communication, and at the same time that it is extremely influential on the probability of communication. Firstly we see that individuals who are separated by hundreds of kilometres still maintain contact and many of those also regularly visit the same locations, suggesting that the barriers of distance can be overcome with modern communication and transportation technologies. On the other hand, however, we see that 90% of the pairs of people who regularly communicate also visit the same locations regularly, meaning that their effective separation distance is actually zero thus confirming that people generally only communicate with those who they are spatially close to on a regular basis.

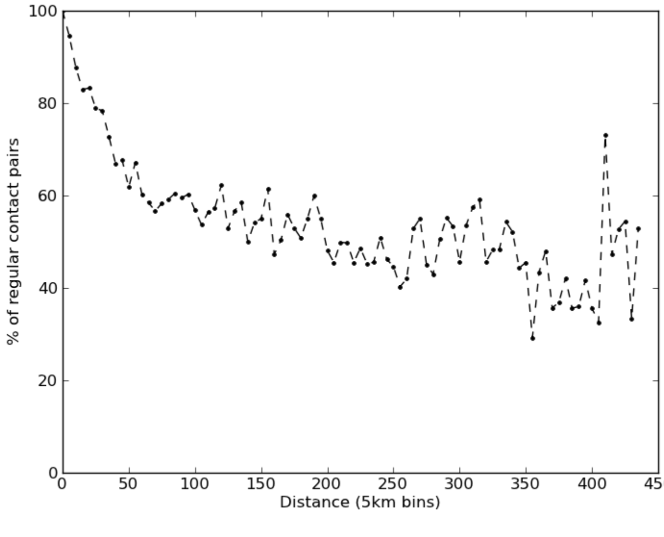


Figure 6.10: Percentage of regular (>= 10 days) contact pairs with home towns separated by distance d who have 10 day significant locations in common.

6.7 Communities

In this section we investigate the spatial community structure of this dataset using the methods discussed in Chapter 4, effectively partitioning the space using the characteristics of the social network derived from the communication network. We begin by applying the modularity optimisation approach used in the previously discussed studies of Blondel et al. (2010); Ratti et al. (2010); Calabrese et al. (2011a). We initially apply the method to the network of social ties aggregated to the tower level. We then introduce some tests to try to understand what the resulting partitions actually tell us. Following this, we apply the same technique at the town level and apply the same tests. In order to investigate the applicability of other methods we also apply two alternative non modularity based methods at both scales and again apply the same tests. Finally we discuss the advantages and disadvantages of each approach.

6.7.1 Modularity optimisation partitioning

We start with the lowest possible level of spatial aggregation, the network of towers. In this network the weight of an edge i, j represents the aggregate level of interaction or interconnectivity between the people whose homes are served by tower i and those served by tower j. Self edges or loops are included and are important as the probability of a connection between two people served by the same tower is high. As we discussed in Section 4.3.3, when optimising a network with modularity the weights on edges must be independent so aggregates such as total call time or average call duration should not be used. The most appropriate choice for edge weights is therefore a single weight per pair of connected individuals where we define a pair to be connected if they have reciprocated calls. The weight of the edge i, j is simply the number of connections between the group of individuals served by tower i and the group of individuals served by tower j.

Following the methodology of Blondel et al. (2010), we use the Louvain method to optimise modularity (see Section 4.2.3). The method finds three levels of hierarchy with 63, 19 and 14 communities and modularity scores of 0.66, 0.699 and 0.702 respectively. We map the community assignments in Figure 6.11 and note that as with the studies of Blondel et al. (2010); Ratti et al. (2010); Calabrese et al. (2011a), the communities at each level of the hierarchy are mostly spatially contiguous despite the lack of spatial information in the input to the method. Furthermore we note that in many cases the communities at the lowest level correspond quite well with the county boundaries, although some of the larger communities span multiple counties (3 and 12 in particular) while other counties contain many smaller communities. We can use the modularity measure to quantitatively assess which partition of the network is better. We find the counties partition to also be highly modular with a score of 0.622 but this is still of lower quality than the partition found at the lowest level by the algorithm. If we start the algorithm from this point it reaches a final result with 12 communities and a modularity of 0.699. We can conclude that the algorithmically found partition is better, in the sense of modularity at least.

6.7.2 Hierarchical structure

As we investigate the results in more detail we see that Dublin is divided into 15 separate communities at the lowest level, which then merge into 4 communities at the highest level (Figure 6.14). There is a clear north-south and east-west divide visible which corresponds well with the known social divides. The north-south divide is perhaps more well known because it corresponds to the literal physical divide of the River Liffey running through Dublin city centre. As Corcoran (2004) notes, "[the] relations between the city's northside

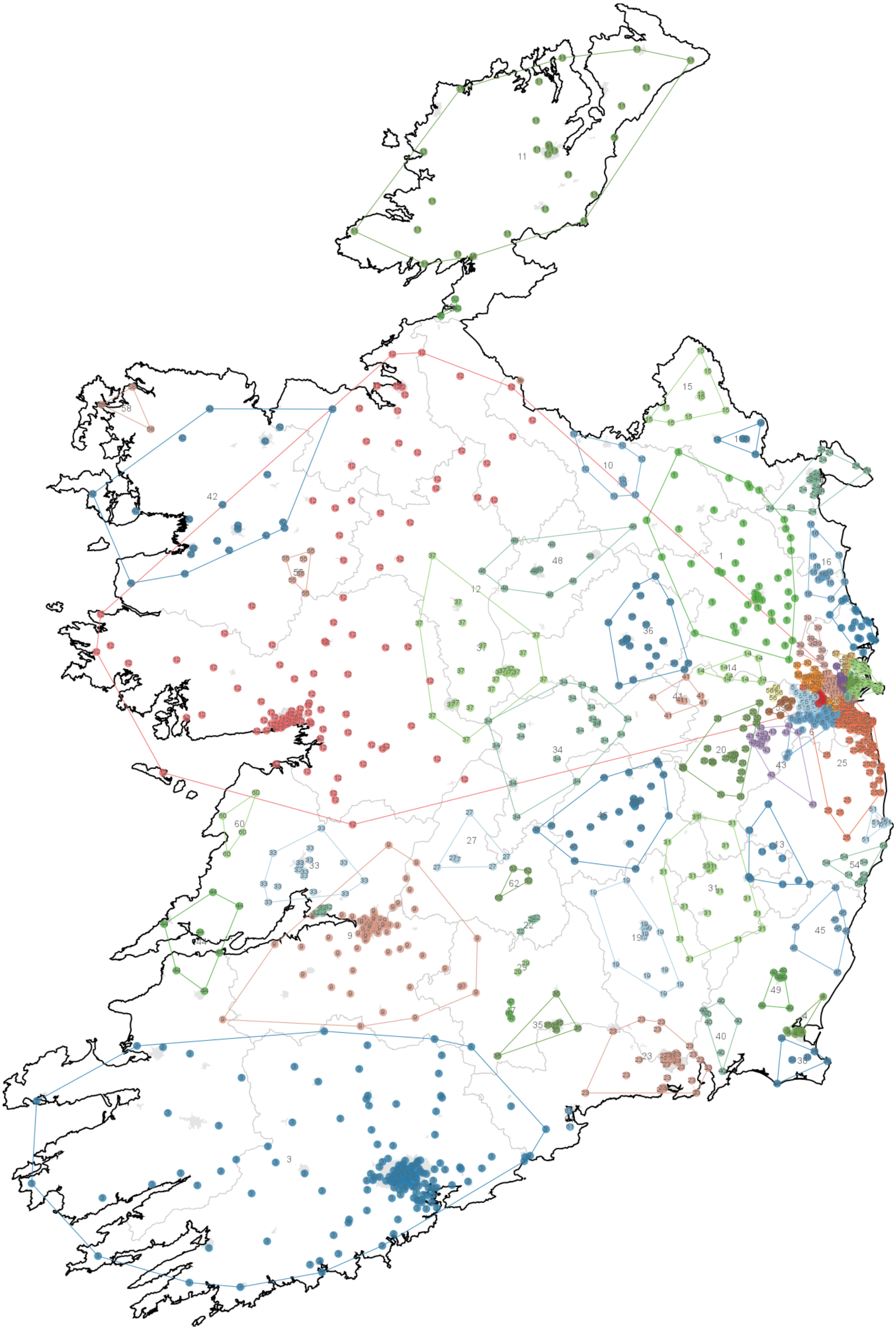


Figure 6.11: Modularity optimisation of the tower network. Level 1

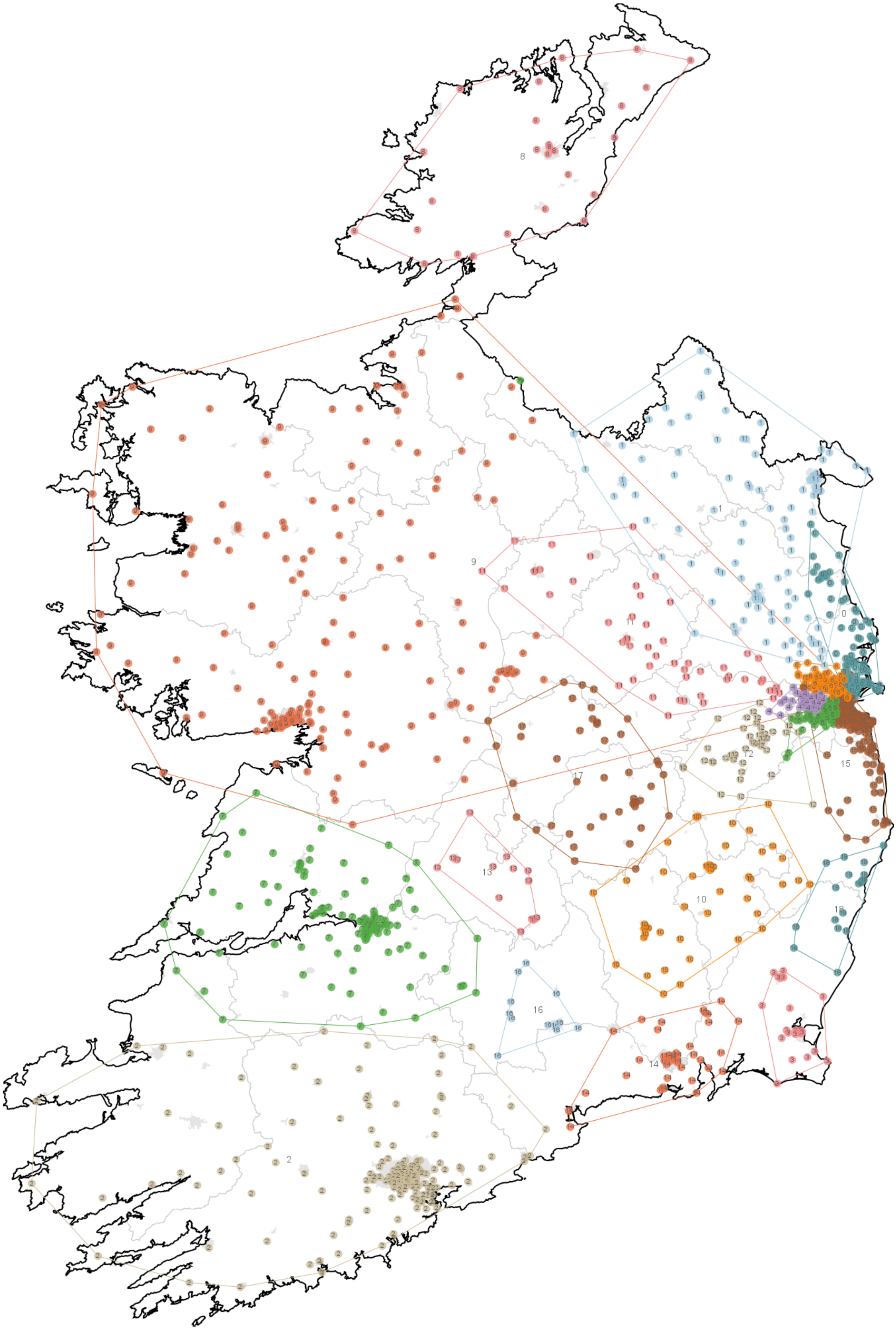


Figure 6.12: Modularity optimisation of the tower network. Level 2

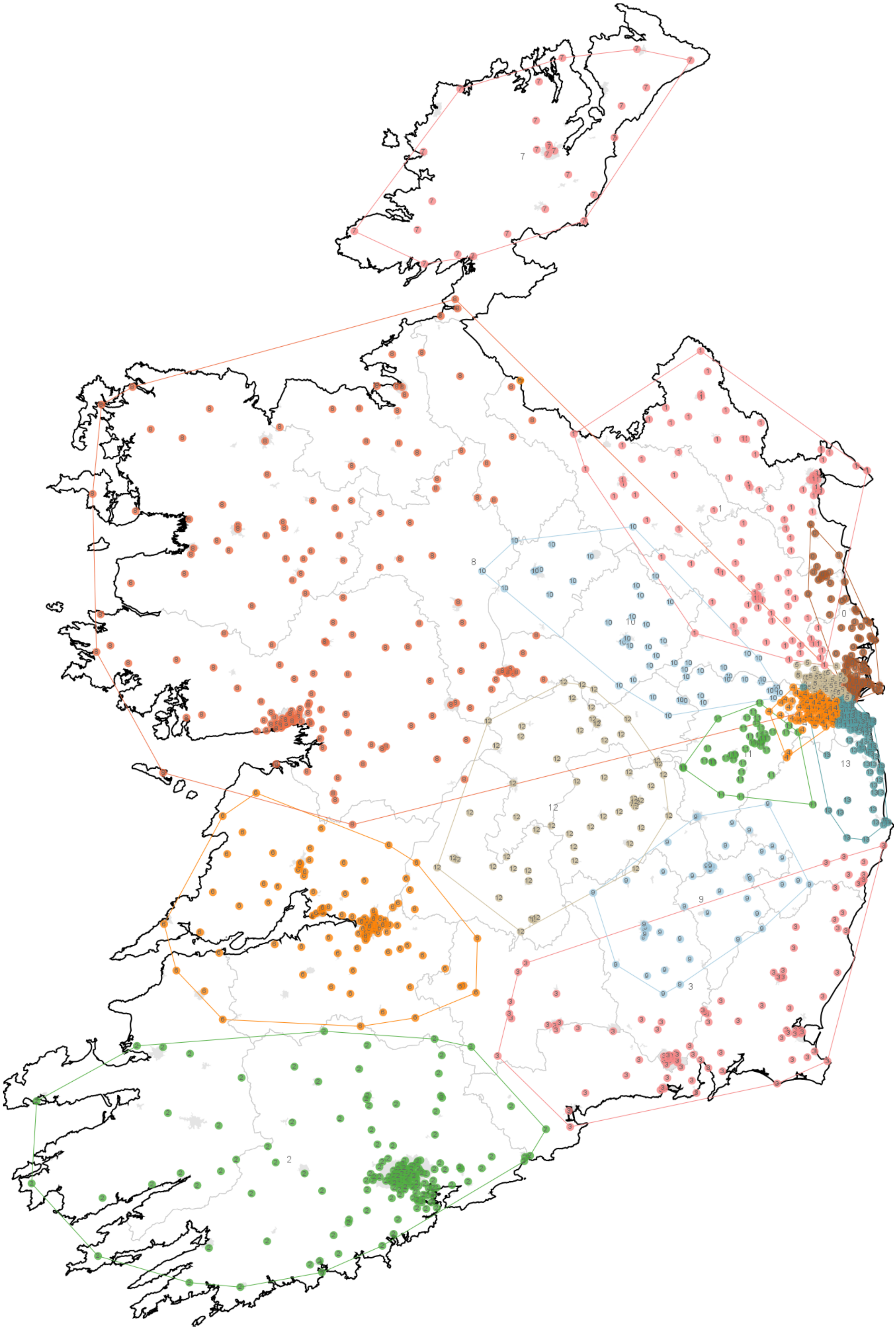


Figure 6.13: Modularity optimisation of the tower network. Level 3

and southside are somewhat uneasy.". The east-west divide is less evident in the general psyche of the people but is a real socio-economic divide with the coastal eastern parts of the city much more affluent than the western parts of the city and its suburbs. These results seem to suggest that both divides truly do exist at an interpersonal level.

We may note that the cities and counties of Cork & Kerry, Limerick and Galway all emerge as single homogeneous entities (communities 3, 8 and 11) at the lowest levels of the algorithm's output, suggesting that they are indivisible. One could perhaps use these results to argue that these cities have retained their village like social structure and are much more socially cohesive than Dublin. Before we get carried away with these arguments however, we must remind ourselves of the issues raised in Section 4.2.3, particularly regarding the so-called resolution limit.

As Fortunato & Barthelemy (2007) warn, there is an inherent limit in the size of communities that modularity optimising methods are capable of detecting. We must therefore explore the results in more detail to understand if we really have found the complete hierarchy of community structure. We will do this by applying the same community detection method to the subgraph of each of the communities found at the lowest level. We will take the community containing the city of Cork as an example. We apply the Louvain method to this subgraph and find a 2 level hierarchical structure with 24 and 12 communities respectively and modularity scores of 0.479 and 0.488. Once again we see in Figure 6.17 that the communities at each level are spatially contiguous.

This appears to contradict the previous finding that this was a single homogeneous community with no substructure. As we explain in Section 4.2.3, this happens due to the fact that modularity is a global quality measure. Modularity is calculated relative to the entirety of the current graph so edges that

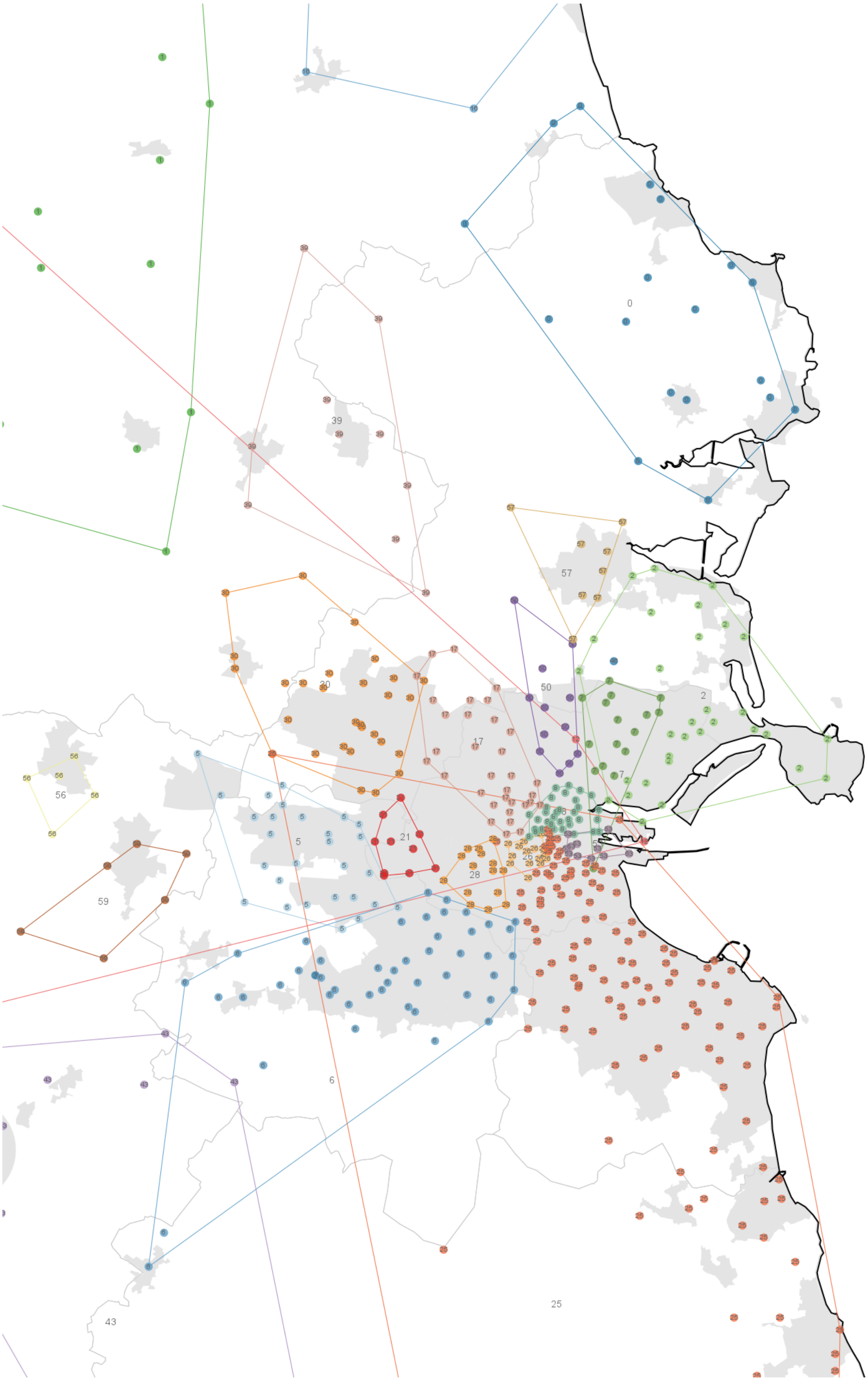


Figure 6.14: Modularity optimisation of the tower network. Community assignments in Dublin. Level 1

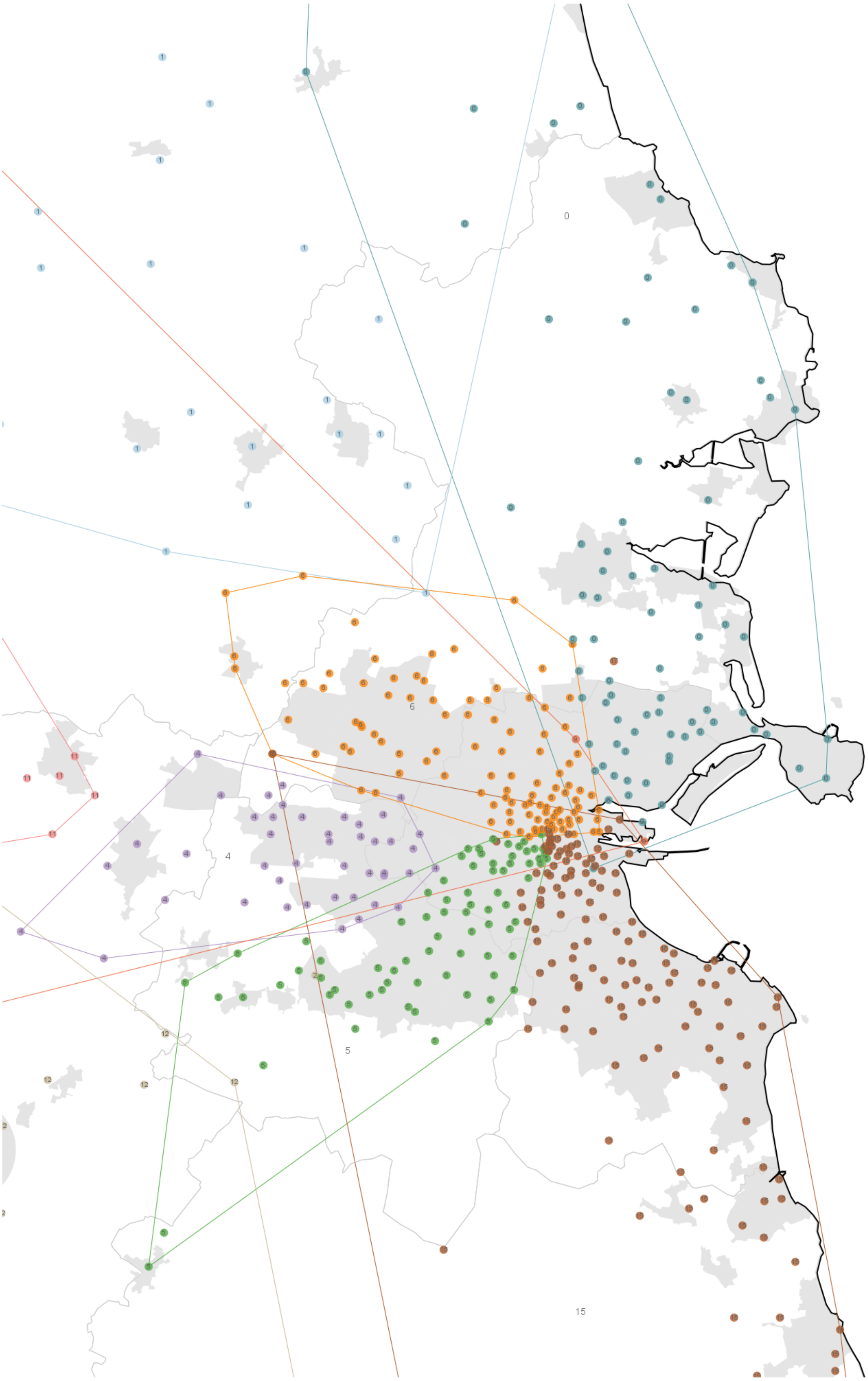


Figure 6.15: Modularity optimisation of the tower network. Community assignments in Dublin. Level 2

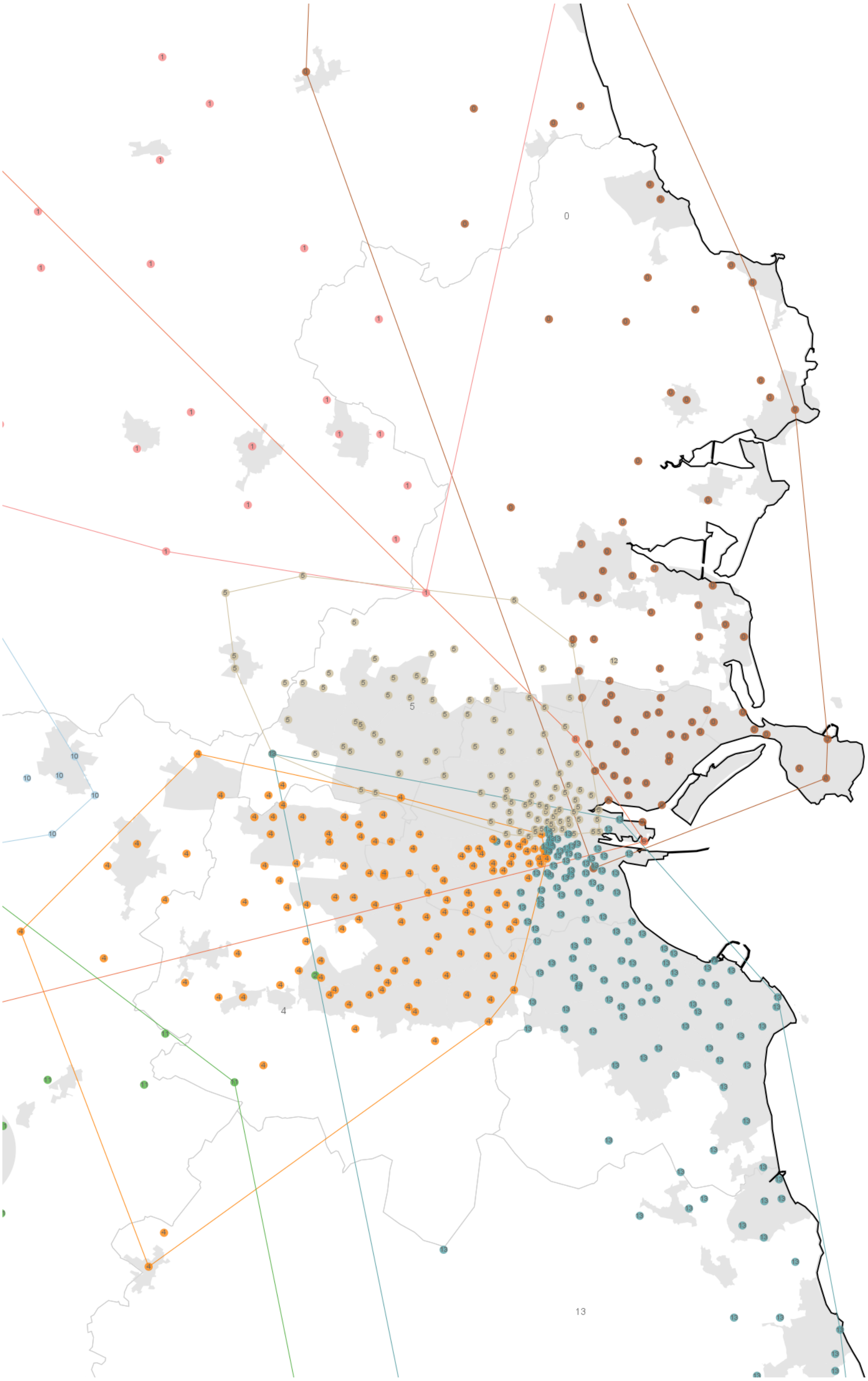


Figure 6.16: Modularity optimisation of the tower network. Community assignments in Dublin. Level 3

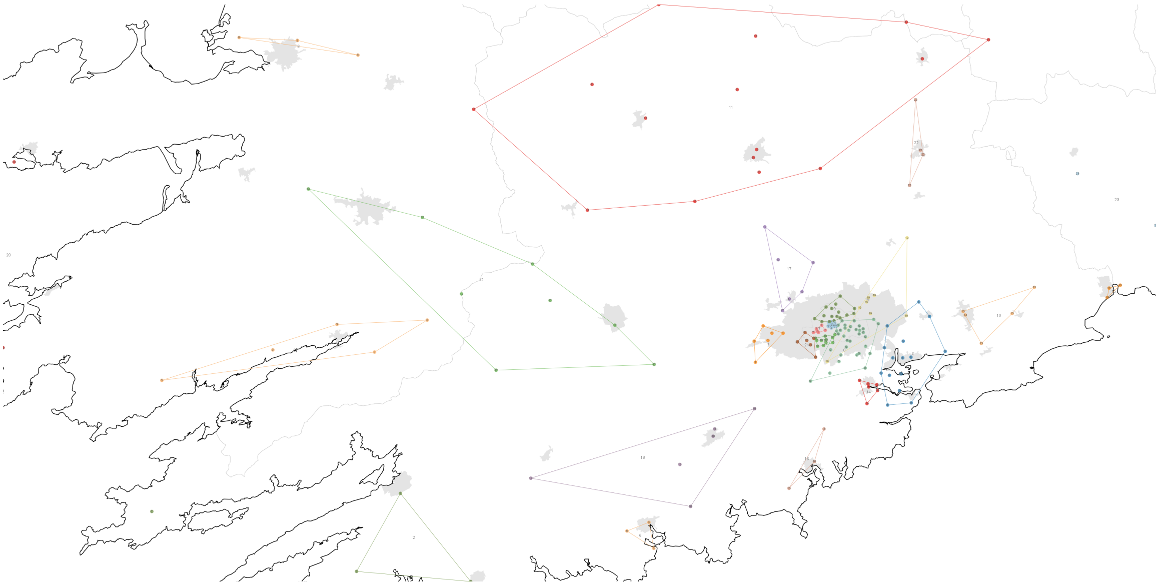


Figure 6.17: Modularity optimisation of the Cork/Kerry community subgraph. Level 1

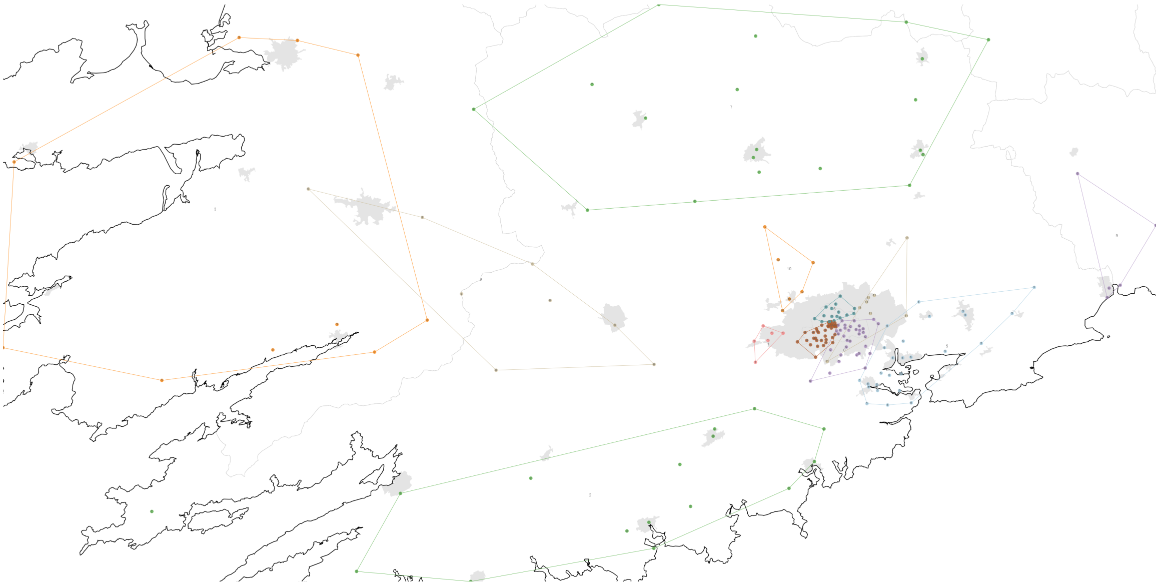


Figure 6.18: Modularity optimisation of the Cork/Kerry community subgraph. Level 2

may appear to have unexpectedly high weights in the context of the larger graph may actually have unexpectedly low weights in the context of the subgraph. This example clearly illustrates the fact that the communities found by modularity optimisation have no *a priori* definition, rather they only have a definition with respect to a given partition of a particular graph. However we may still test whether the communities we find in the subgraph could be classed as 'real communities' of the full graph. Fortunato & Barthelemy (2007) define a 'real community' within the modularity framework as a group of nodes which have a larger than expected number of internal connections. A subgraph s is a 'real community' if $\frac{l_s}{L} - \left(\frac{d_s}{2L}\right)^2 > 0$, where L is the total number of edges in the network, l_s is the number of edges within the subgraph s, and d_s is the total degree of the nodes in s. We perform this check for each of the communities found at the lowest level of the hierarchy in the Cork/Kerry subgraph and find that each of them satisfies this condition, meaning they are 'real communities' in the context of the full graph also.

This result tells us that the hierarchy found by the Louvain method is not a true hierarchy as it fails to find some low level communities that are considered real communities within the modularity framework. The authors of the Louvain method claim that the agglomerative nature of the method circumvents the resolution limit problem (Blondel et al., 2008), but this is clearly not the case. The modularity measure itself is biased towards larger communities so it is unlikely that any modularity optimising method could circumvent the issue. Despite the appeal of the hierarchical unfolding of the Louvain method and other similar methods, it appears we must consider the intermediary levels as mere artefacts of the process of the algorithm rather than as significant outputs.

6.7.3 Partition quality measures

Given that we have determined the intermediary levels of the hierarchical results to be unreliable, we now turn our attention to the final level. Indeed it is the top level of the hierarchy that is the result of the modularity optimisation process and therefore closest to the best partition of the network, at least within the modularity framework. We can assess the quality of this partition of the network in a number of different ways. Some of these measures are a-spatial and apply generally to networks while others take account of the unique nature of spatially embedded networks.

Modularity

The modularity score itself is a measure of partition quality. This value tells us how different the partition is from a random assignment of nodes to communities. In a range of 0 to 1, the score of 0.702 found at the final level of the hierarchy is considered a high value and therefore indicative of definite non random structure. This single value isn't very informative however so we look at other measures.

Spatial contiguity

We assess the spatial contiguity of the partition through simple visual analysis by plotting each tower at its spatial coordinates with a colour determined by its community assignment. We see that the majority of towers are adjacent to towers of the same colour with only very few spatial outliers. These are likely to be towers with very low 'populations' and therefore statistically noisy.

Cohesiveness

We can measure the cohesiveness of each community by calculating the fraction of in-community links for each tower of each community. The fraction of in-community links for node i is given by $\frac{\sum_j A_{ij}\delta(c_i,c_j)}{\sum_j A_{ij}}$ where $\delta(c_i, c_j)$ is 1 if i, j are in the same community. We calculate this value for every node (tower) and find the median to be 0.62 with a lower quartile of 0.51 and upper quartile of 0.71. We show boxplots for each community separately in Figure 6.19 to assess the variation between communities.

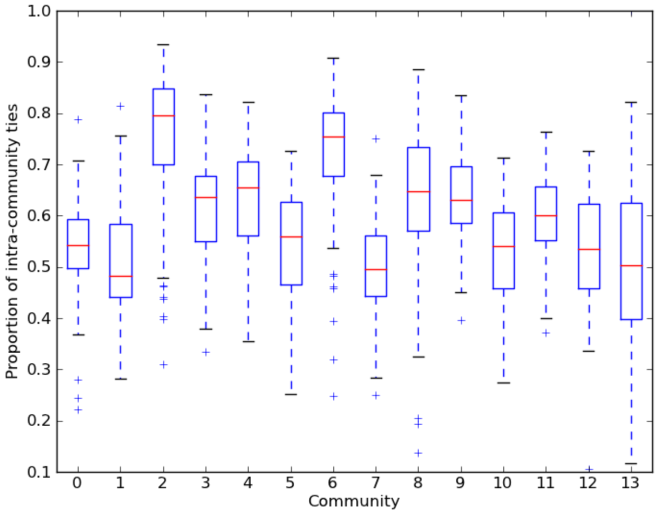


Figure 6.19: Proportion of intra-community links for each community.

We see significant differences in the medians and interquartile ranges between different communities with community 1 having a median less than 0.5 whereas community 3 has a median value greater than 0.8. We also see large ranges within some communities and with a number of very low value outliers. We should point out here that a value of less than 0.5 means that less than half of the ties from that particular tower are with other towers in the same community. We map these values in Figure 6.20 to check for any spatial patterns and find that there is indeed clear spatial autocorrelation.

We see that the nodes with the highest values are those that are densely packed at the spatial centre of the community while those nodes closer to the community boundaries generally have lower values. In some ways this is to be expected as the number of same community nodes within close spatial prox-

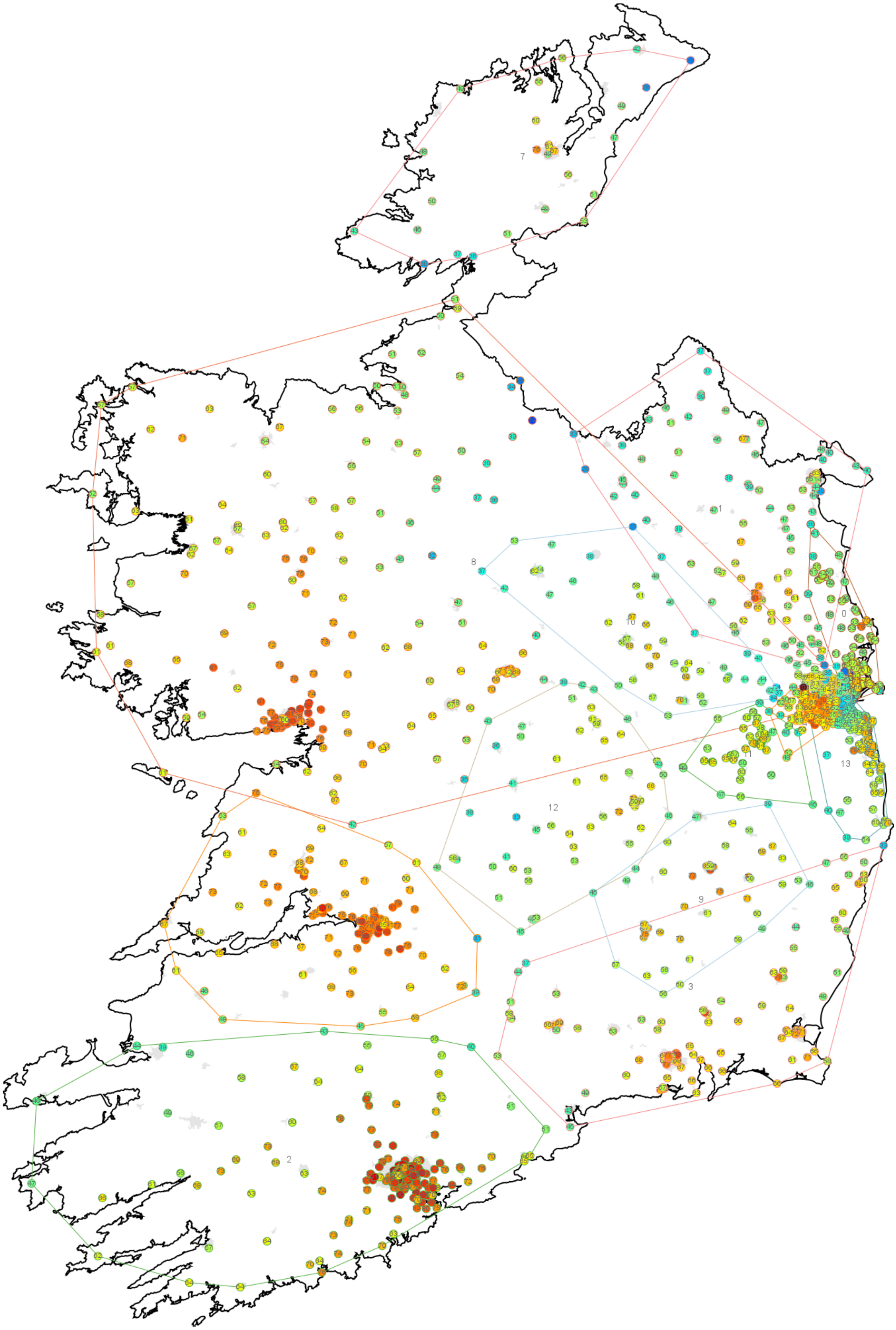


Figure 6.20: Proportion of intra-community links for each tower. Community boundaries are shown with solid lines.

imity is less for nodes close to the borders so more of the links are with nodes outside the community. On the other hand we expect there to be a difference in the interaction probabilities between nodes within the same community and those in other communities, even if the distance is the same. If this is not the case then communities do not actually tell us anything useful.

Interaction probability

In our next measure we test if there is any difference in the interaction probabilities for nodes in the same community and those in separate communities, over a range of distances. As we discussed previously in the context of spatial interaction, the concept of interaction probability is not easy to define but again here we may instead use the ratio of the number of ties to the smaller of the two tower populations. For a given distance we take the median of this ratio for all pairs and refer to it as the normalised median interaction level.

We calculate the normalised median interaction level in 5km bins separately for intra-community and inter-community ties. We plot this for all pairs of nodes in the network (including pairs with no ties) and see in Figure 6.21 that the value for intra-community ties is consistently higher than for inter-community ties. While this appears promising it is still a global measure so we also plot the same values for each community separately in Figure 6.22. Once again we see the same pattern with the intra-community ties have consistently higher values for all communities. These results suggest that the communities do capture the discontinuous effect of space and are not merely manifestations of a continuous spatial process.

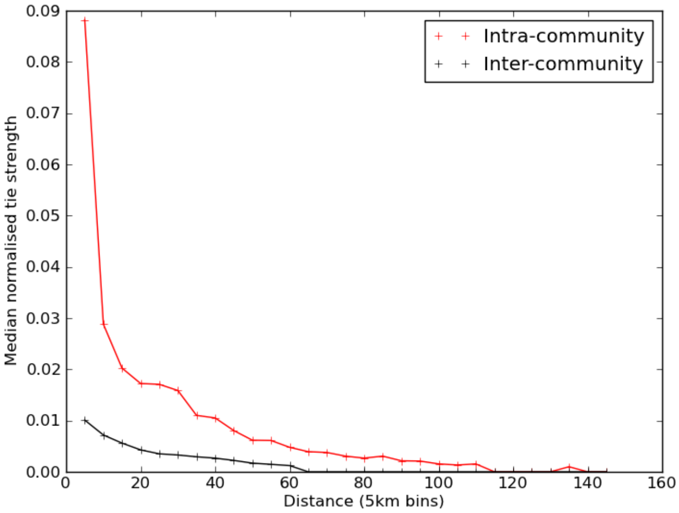


Figure 6.21: Normalised median interaction levels for inter-community (black) and intra-community links (red).

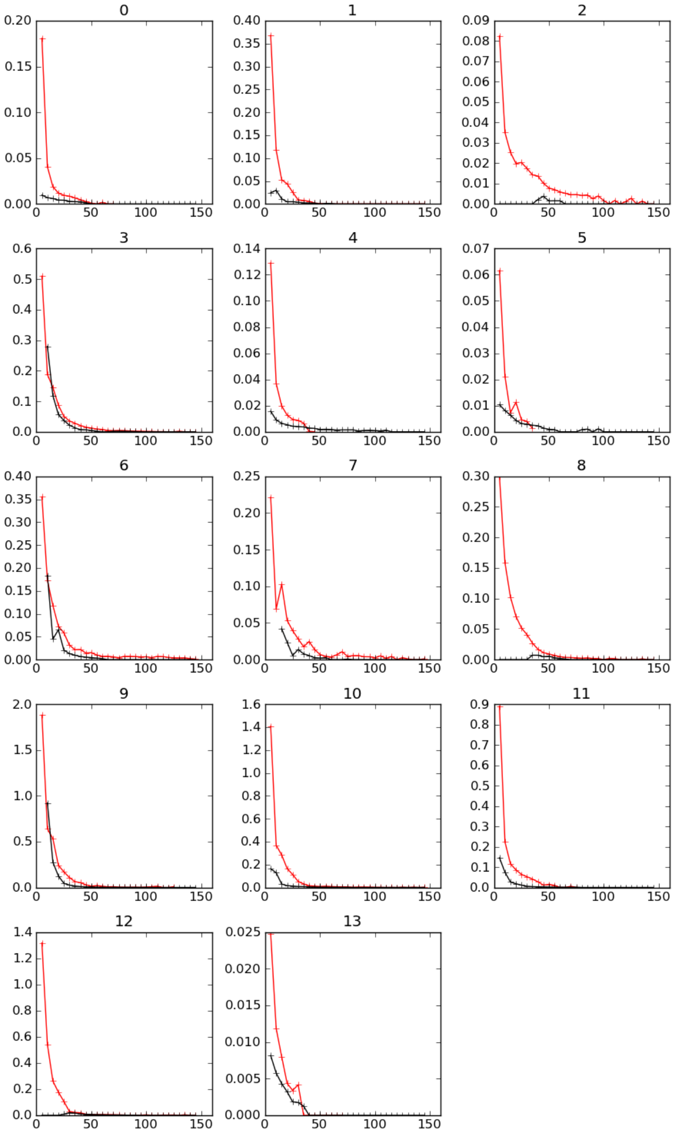


Figure 6.22: Normalised median interaction levels for inter-community (black) and intra-community links (red) disaggregated by community.

In order to show that this test is robust, we perform the same analysis on two simulated network partitions. First we use a random assignment of nodes to 14 communities (the same number found at the top level by the method). Again we calculate the normalised median interaction level for each distance bin and plot them in Figure 6.23. We can see that the levels are nearly identical for intra-community and inter-community ties in this case, as we would expect given that the community assignments are entirely random. For the second test we partition the nodes of the network using k-means on their spatial coordinates with $k = 14$. This produces spatially contiguous clusters which are more similar to the communities found than the random assignment but still lack any structure related to the social interactions. We see in Figures 6.24 and 6.25 that the intra-community values are sometimes higher and sometimes lower than the inter-community ones. This can be explained by the fact that some of the clusters found by spatial clustering are actually quite similar to the communities found through network partitioning, while others are radically different (Figure 6.26). In general though the test seems to be quite useful at identifying the difference between the soft divides of purely spatial clusters and the harder divides of social communities.

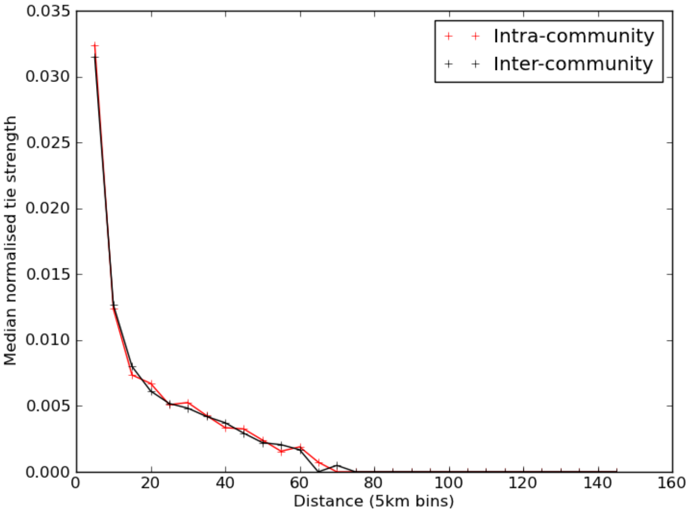


Figure 6.23: Normalised median interaction levels for inter-community (black) and intra-community links (red) for a random partition.

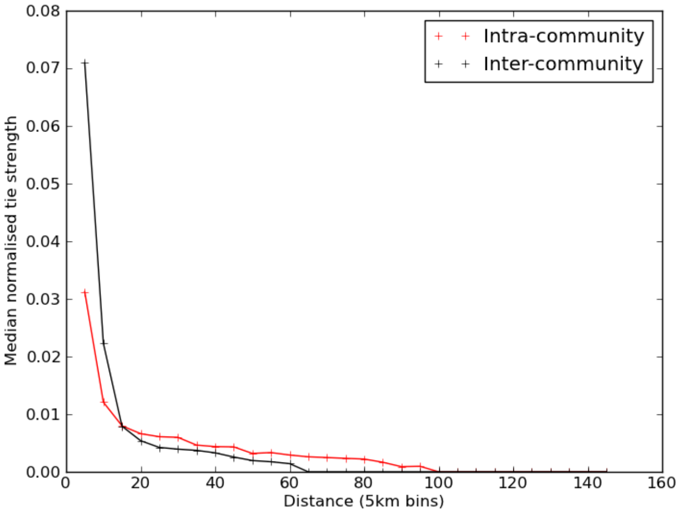


Figure 6.24: Normalised median interaction levels for inter-community (black) and intra-community links (red) for a spatial k-means partition.

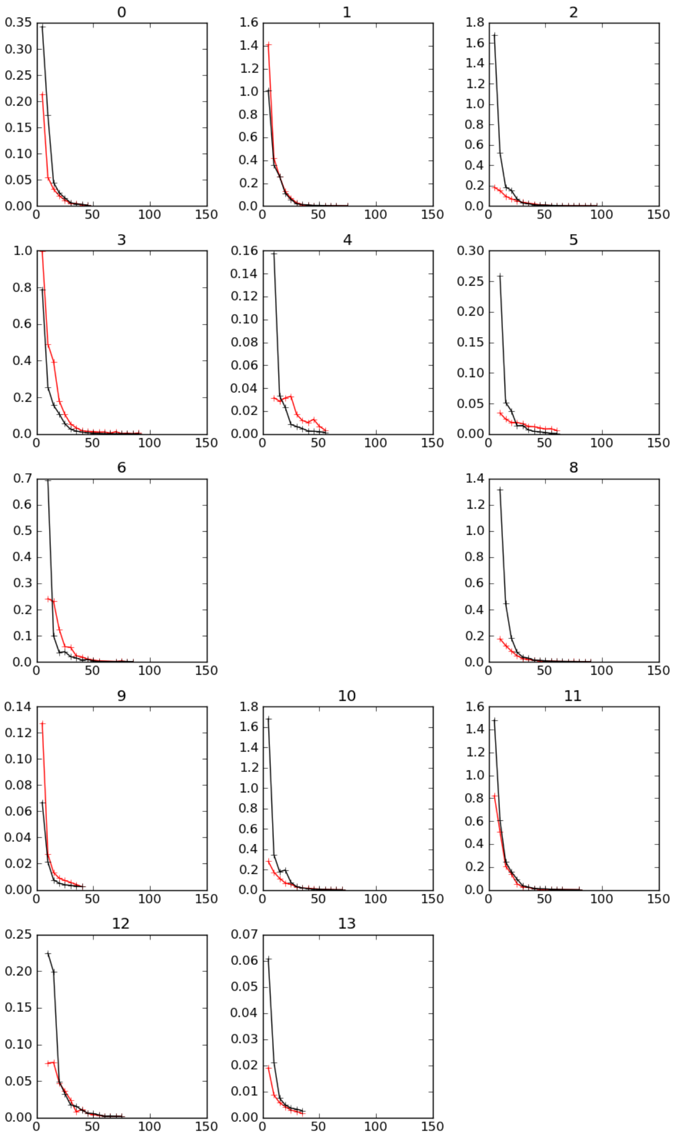


Figure 6.25: Normalised median interaction levels for inter-community (black) and intra-community links (red) for a spatial k-means partition disaggregated by community.

6.7.4 Modularity optimisation partitioning of towns network

We now apply the same methodology to the network of towns that we previously analysed. Once again a single weight on an edge i, j represents the existence of a reciprocated connection between individuals with home towns i

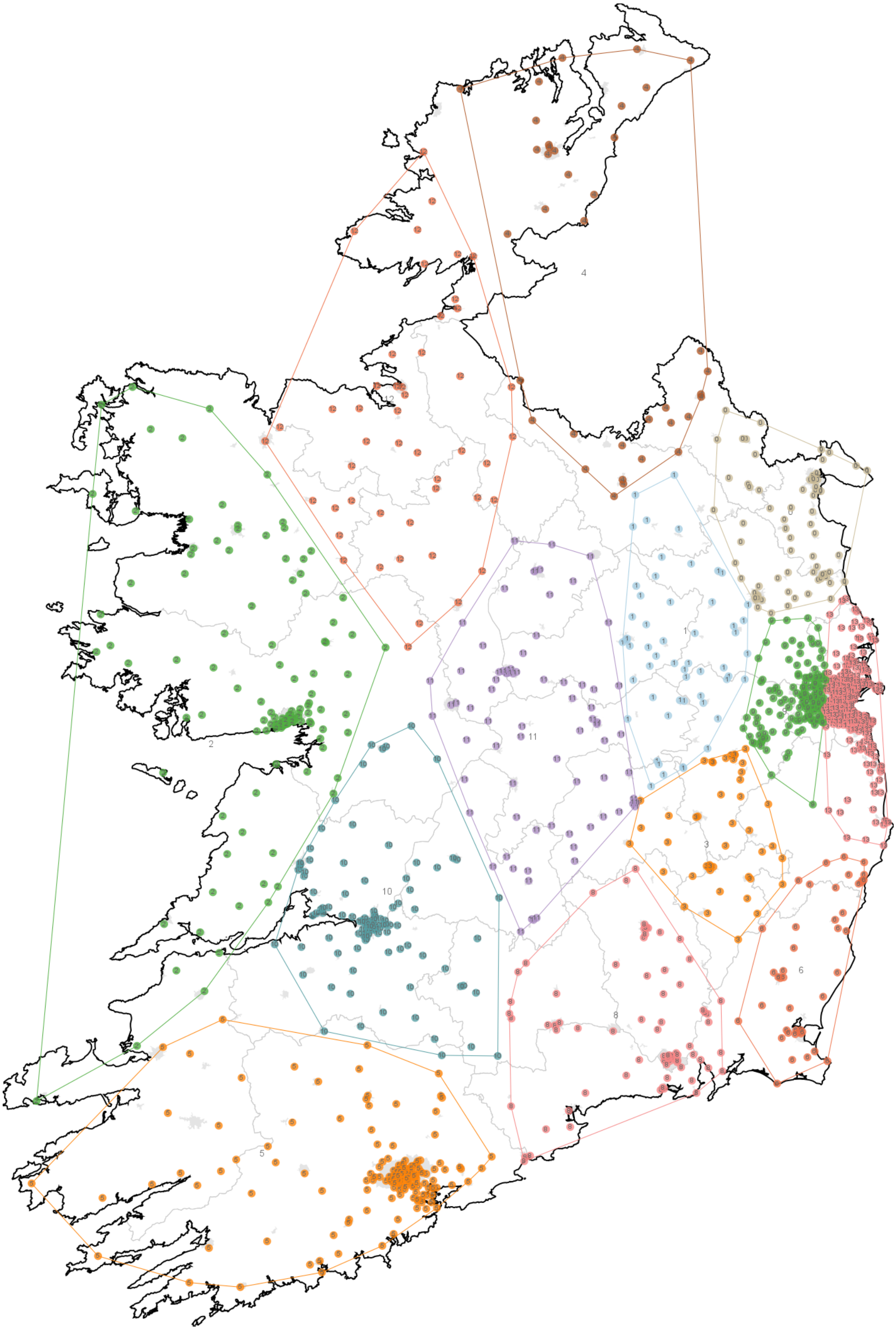
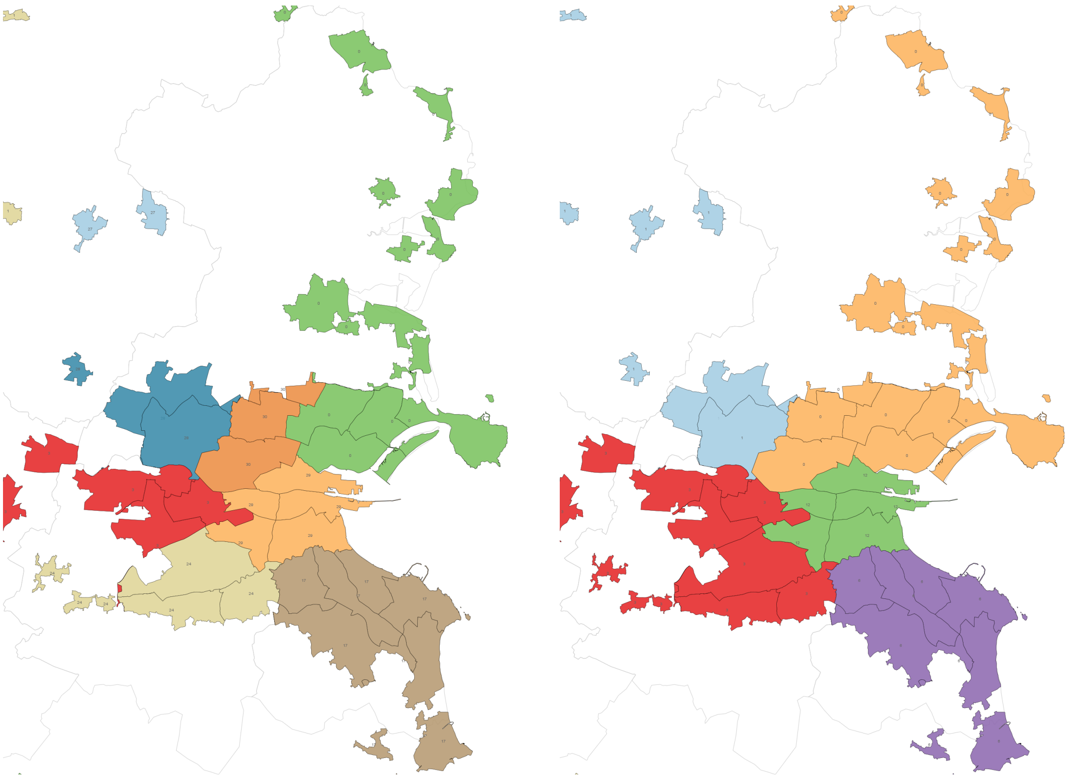


Figure 6.26: Spatial k-means clustering of towers into 14 clusters.

and j. As with the tower network we include self links. If we were to place each town in a community on its own the modularity would be 0.49 which signifies a highly non-random structure. This is due to the strength of the self links. The Louvain method identifies a two level hierarchy in this network with 31 and 13 communities on each level and modularities of 0.63 and 0.65. As shown in Figures 6.27 and 6.29 the communities are once again spatially contiguous and somewhat resemble the communities found in the tower network. A notable difference however is that in this network Dublin is divided into 5 communities with a clear central community that was not in the results for the tower network. Additionally the northern part of Donegal is part of the Galway/Sligo community whereas it is a separate community in the tower network. The significance of these changes is very difficult to assess however.



(a) Level 1

(b) Level 2

Figure 6.29: Modularity optimisation of the towns network. Community assignments in Dublin.

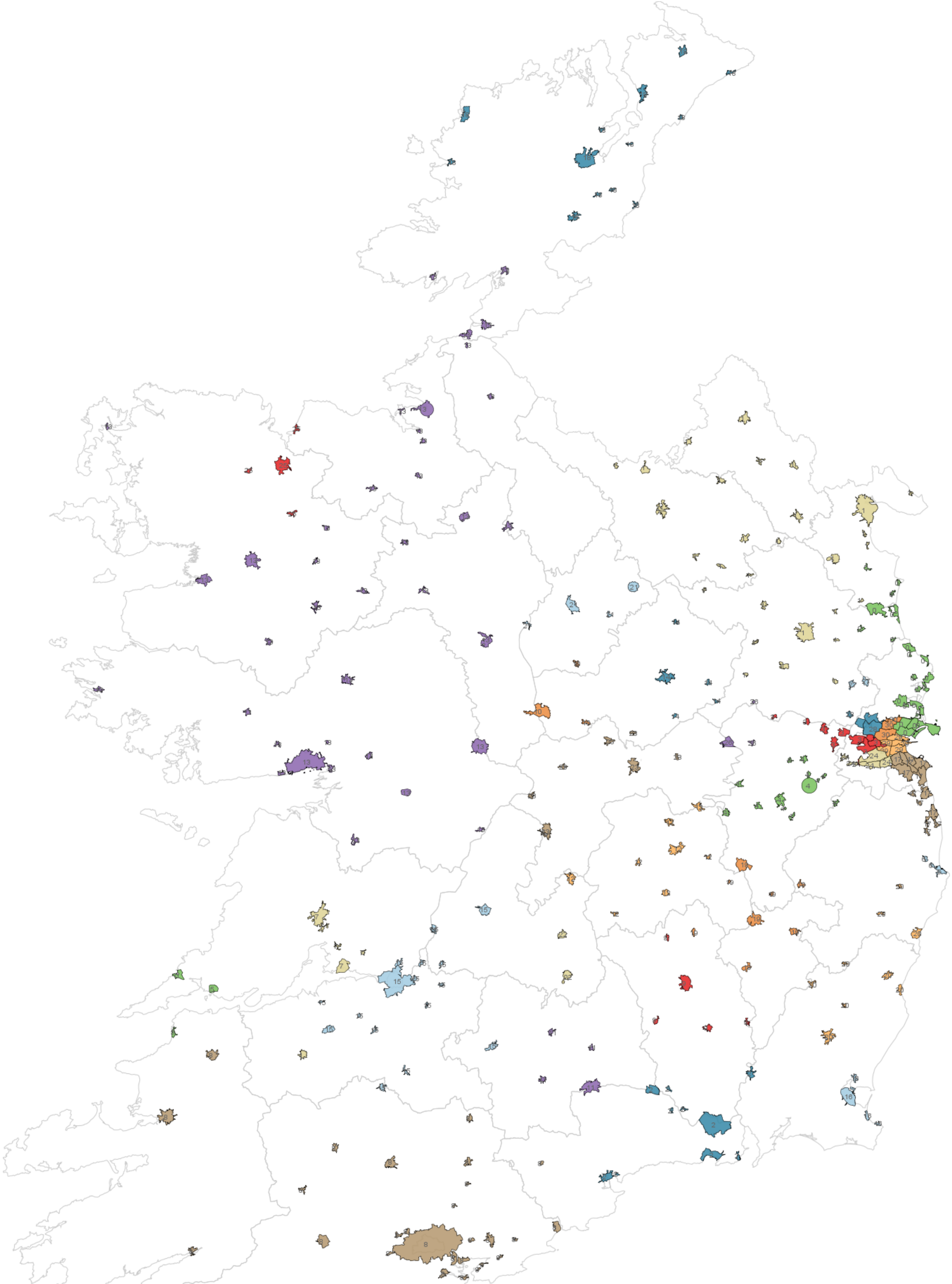


Figure 6.27: Modularity optimisation of the towns network. Level 1

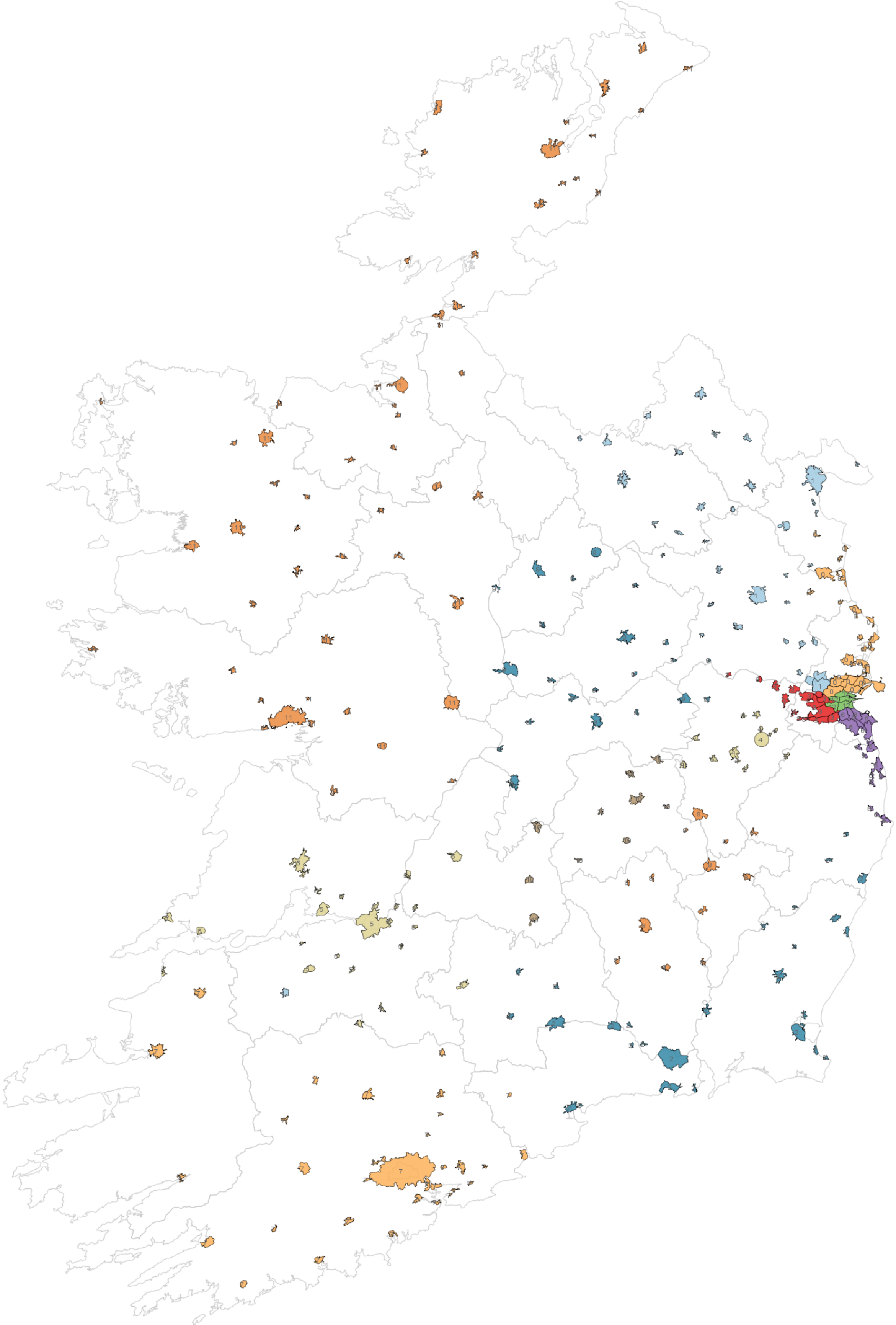


Figure 6.28: Modularity optimisation of the towns network. Level 2

We noted previously that the global proportion of intra-town ties is 51%. This increases to 70% for intra-community ties at the first level and 75% at the final level. In Figure 6.30 we show the proportion of intra-community links disaggregated for each community. The median value of 6 of the 13 communities is actually below 50%, implying that the majority of ties involving people in these communities are with people in other communities.

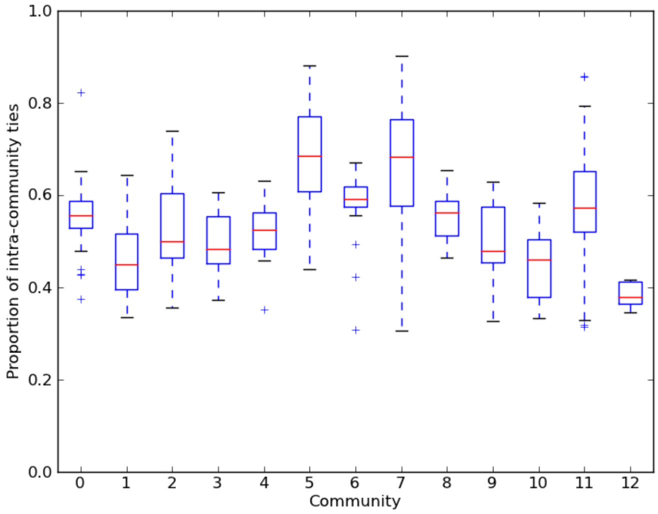


Figure 6.30: Proportion of intra-community links for each community.

In Figures 6.32 and 6.33 we plot the normalised median interaction levels for this network and the top level communities. We see that generally the intra-community links have higher values but this is not always the case. For particular communities (esp. 5, 7 and 12) we see very different patterns suggesting that at short distances the weights to towns in other communities are likely to be higher than weights on links in the same community at similar distances. The differences in these results to the tower network can partly be explained by the differences in spatial densities and the overall number of nodes. These two factors mean a lower number of samples per distance bin and thus more noise in the statistics.

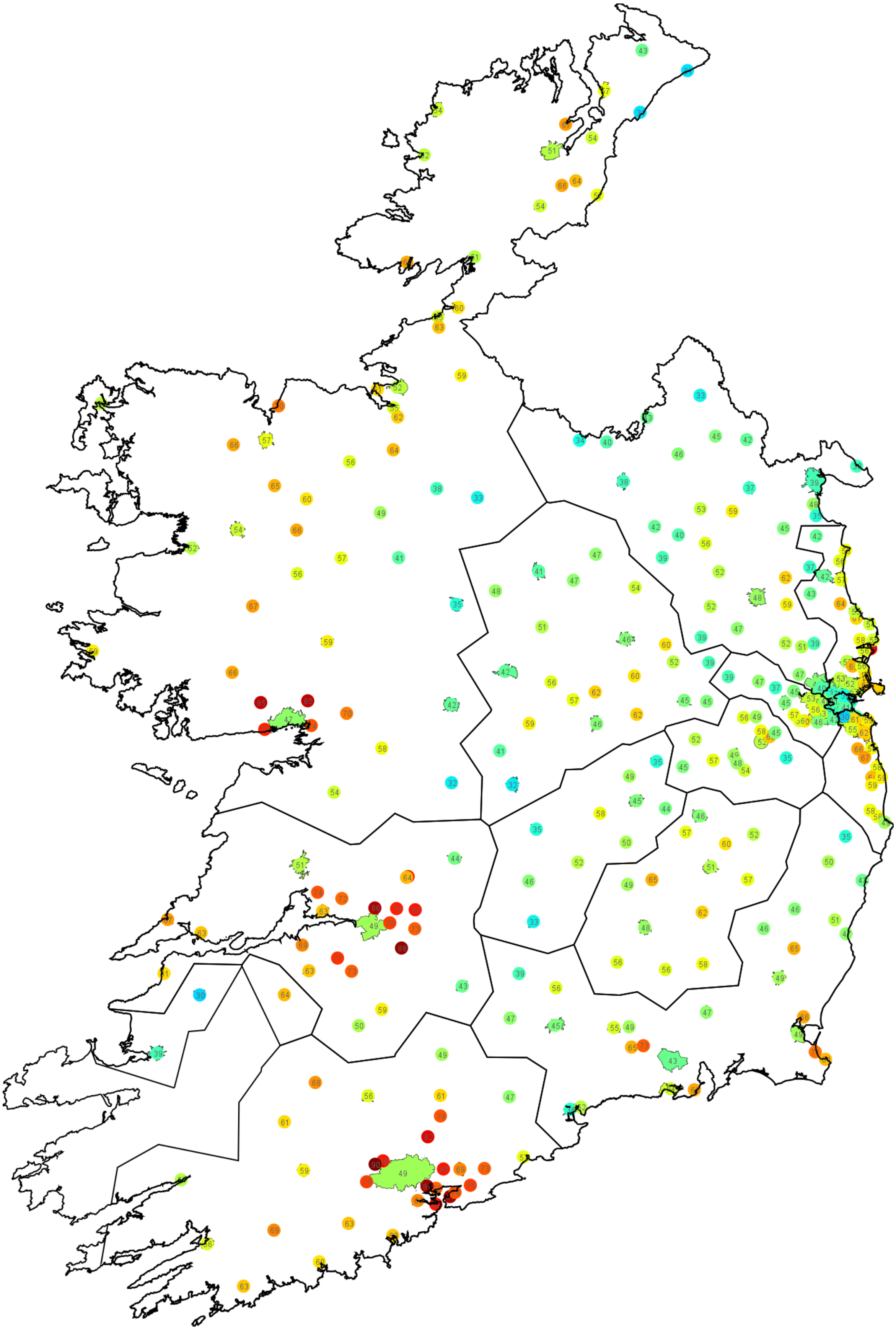


Figure 6.31: Proportion of intra-community links for each town. Community boundaries are shown with solid lines.

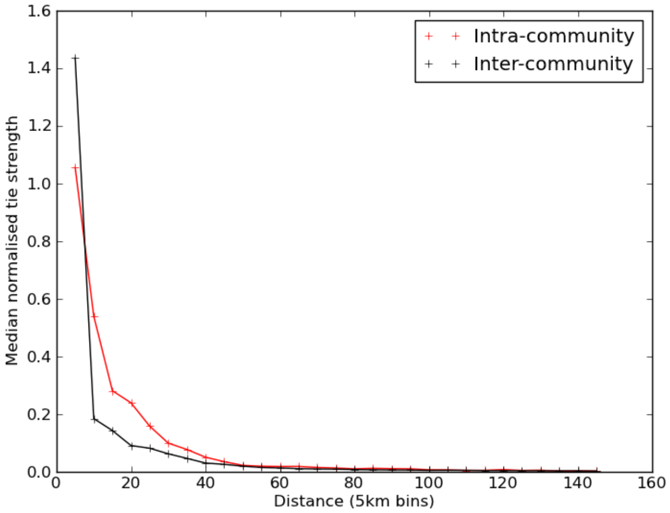


Figure 6.32: Normalised median interaction levels for inter-community (black) and intra-community links (red) in the towns network.

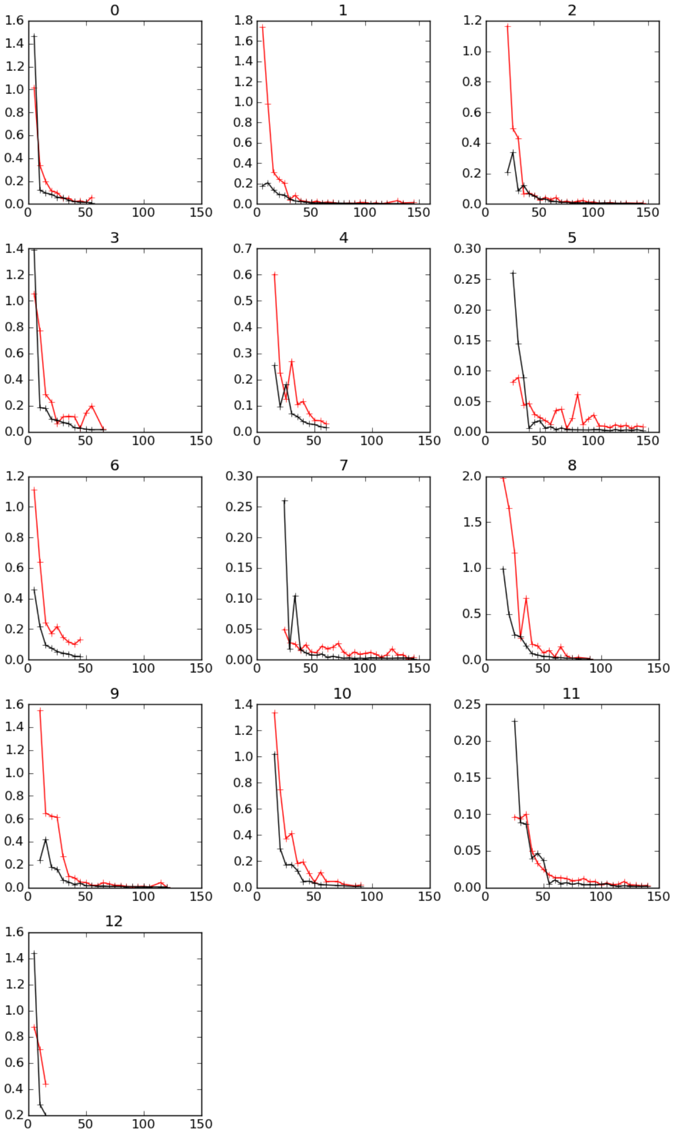


Figure 6.33: Normalised median interaction levels for inter-community (black) and intra-community links (red) disaggregated by community in the towns network.

6.7.5 Random walk map length optimisation partitioning

In this section we apply the Infomap method of Rosvall & Bergstrom (2008, 2010, 2011). As discussed in Chapter 4 Infomap assigns nodes to communities with the goal of minimising the description length of a random walk through the network. Like modularity optimisation methods, the result is a partition of the network into discrete communities and the quality function is a global measure. Here we present results for the method introduced in Rosvall & Bergstrom (2010) because it has the nice feature of providing a level of statistical significance for the assignment of each node to a community. (The later hierarchical method (Rosvall & Bergstrom, 2011) found very similar results to this method with no hierarchical structure in either network.)

Tower level network

We apply the undirected version of the method to the tower network with 100 bootstrap networks and 10 attempts per bootstrap. The results show 37 distinct communities with near complete spatial contiguity in Figure 6.34. In these maps nodes that were not assigned to the same community at least 90% of the time are shown in a lighter colour. We see occurrences of such nodes mostly at the spatial borders between communities, implying that in some cases these borders are 'soft'. This corresponds well with our similar findings for the modularity optimised partitioning. We once again see in Figure 6.35 that Dublin is divided into 7 different communities with clear North-South and East-West divisions. We see a concentration of insignificantly clustered nodes in the south inner-city, indicating an area of overlap between communities.

The biggest difference between the results of this method and the Louvain method is in the number of communities found. The Louvain method found 14 communities at the final (optimal) level whereas here we find 37. This seems more reasonable if we are to think of communities as homogeneous units without any substructure but we still see some very large (spatially) communities in the west of the country, particularly around Cork, Limerick, Galway and Mayo which each contain areas of high densities of towers. It appears therefore that there is a resolution limit effect in these results also.

We apply the distance decay tests to these results and see in Figure 6.36

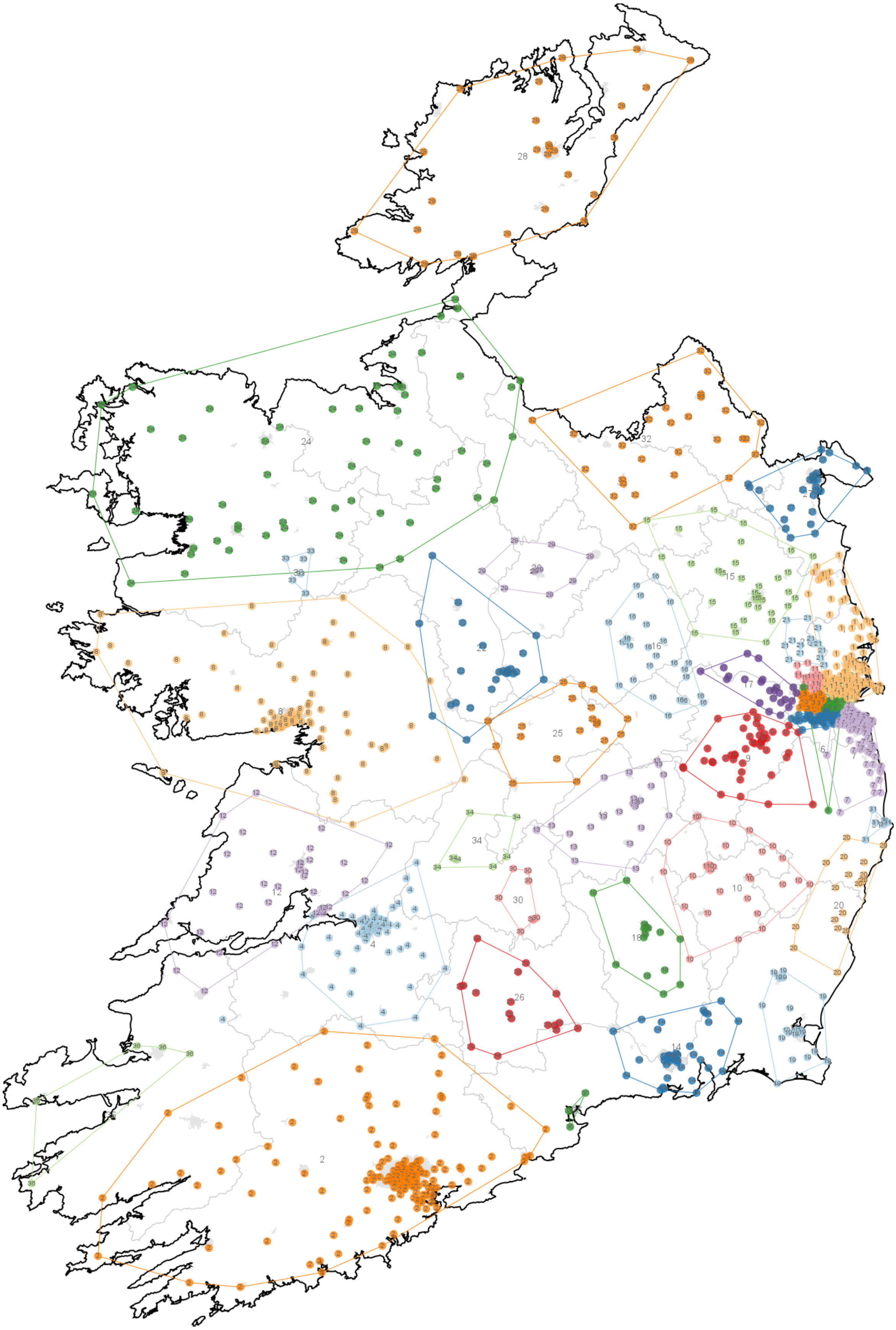


Figure 6.34: Communities identified in the tower network using Infomap.

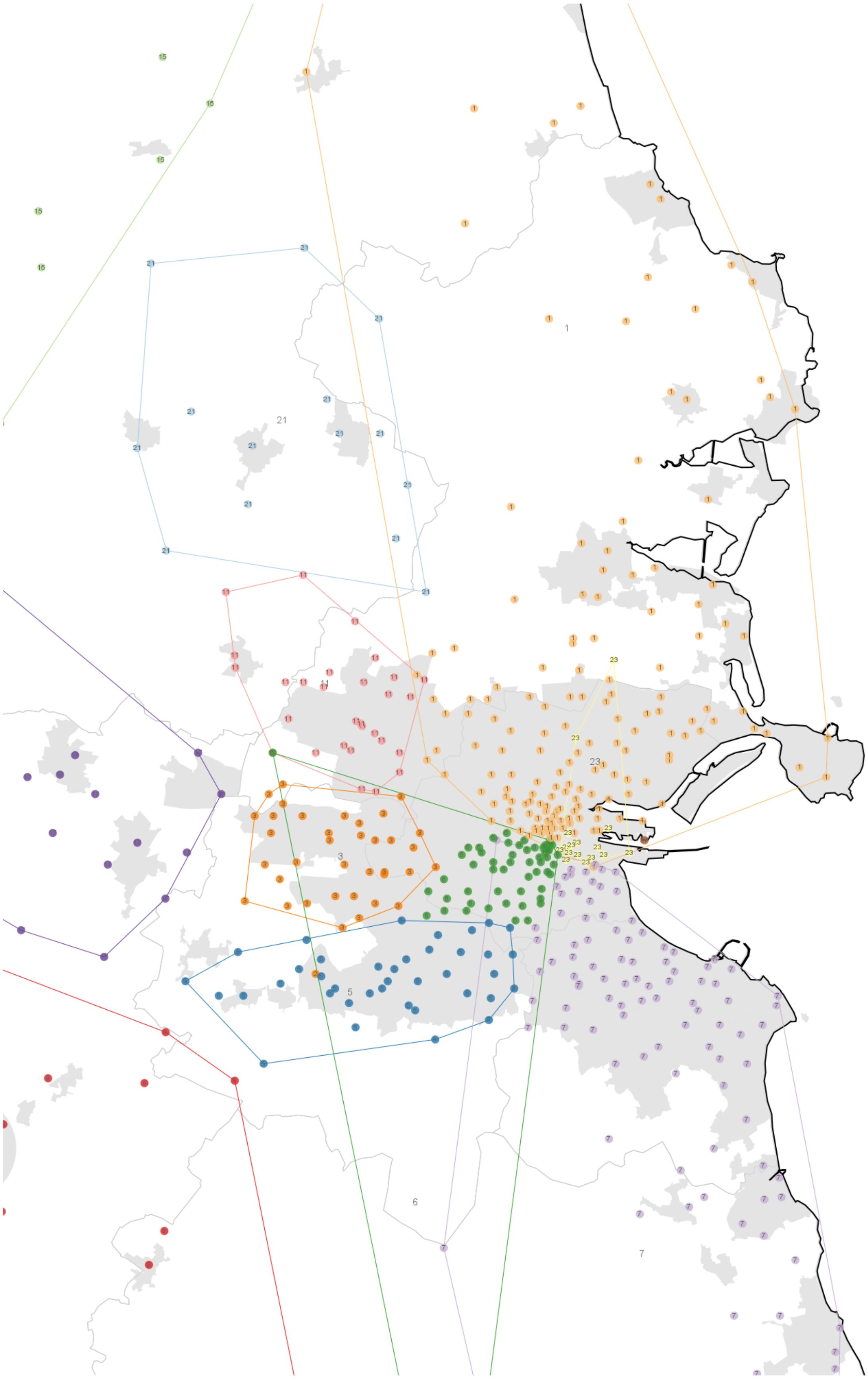


Figure 6.35: Communities identified in the tower network using Infomap. Community assignments in Dublin.

that on average the intra-community ties have a higher median normalised interaction level than inter-community ties at the same distance. When we look at the per community graphs in Figure 6.37 we see that most communities follow the same general pattern but there are a few where there is no discernible difference between the intra- and inter-community ties and some cases where the opposite effect is seen. In these cases there may be a more natural distance decay effect or it is possible that some cell towers are serving spatially and socially separate communities but we cannot detect this as the cell tower is the lower limit of spatial resolution.

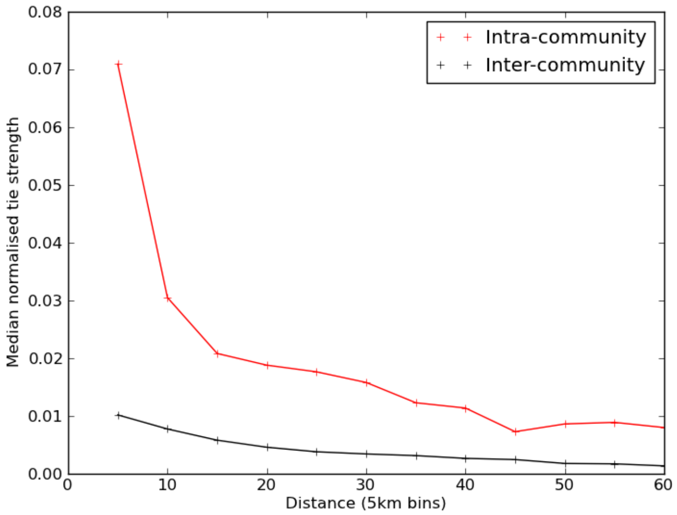


Figure 6.36: Normalised median interaction levels for inter-community (black) and intra-community links (red) in the tower network with communities found by Infomap.

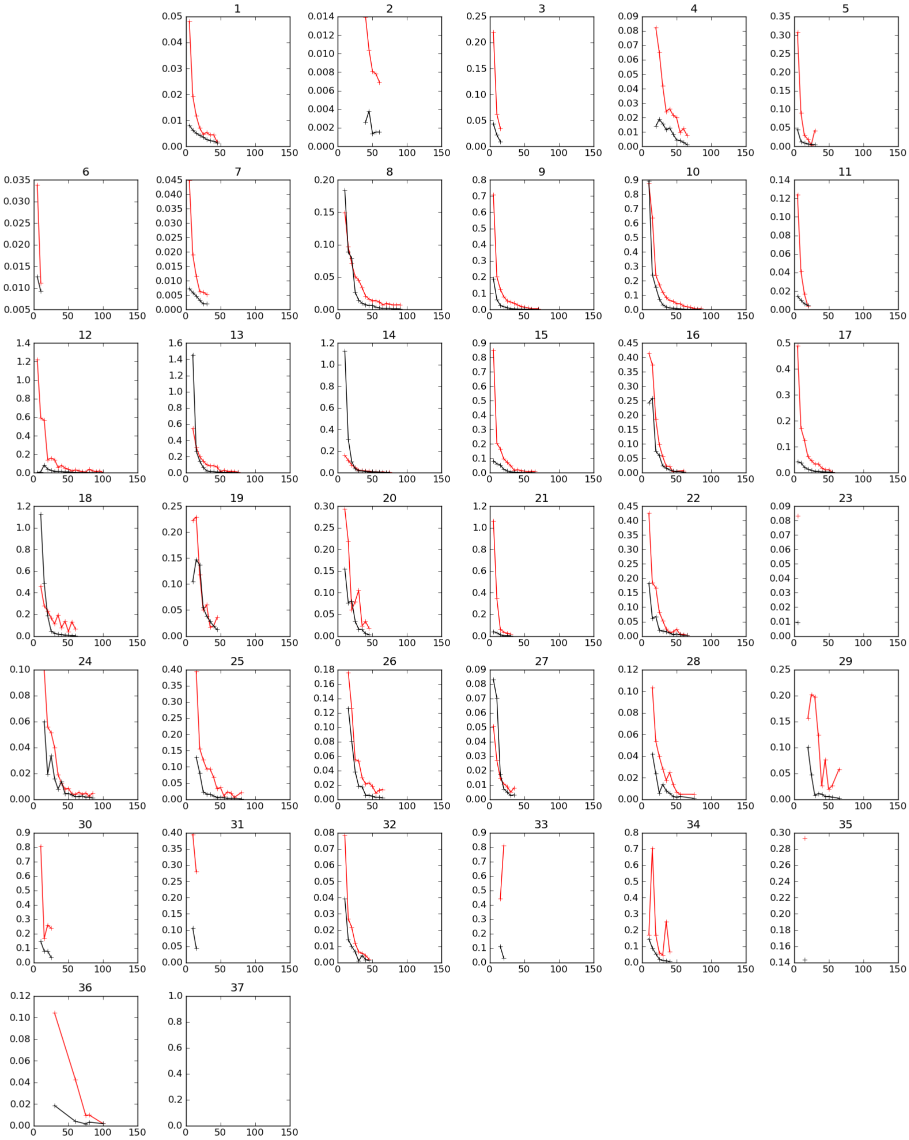


Figure 6.37: Normalised median interaction levels for inter-community (black) and intra-community links (red) disaggregated by community in the tower network with communities found by Infomap.

Town level network

When we apply the same method to the town aggregated network we obtain rather different results from both the tower level network and the modularity optimised partition of the town network. With modularity optimisation we found that the nodes of the towns network were generally grouped in the same way as the tower network. However with Infomap we find that while there was a large number of communities in the tower network, there are only a few in the towns network with previously separate entities now merged together at this level (see Figure 6.38). Particularly striking is the fact that most of Dublin, Meath, Louth, Monaghan and Cavan is a single community with only the south east part of Dublin forming a separate community with Wicklow. In the modularity optimised partition Dublin was divided into 5 distinct regions of much smaller size. Additionally, each of the three cities of Cork, Limerick and Galway are the centres of large communities covering multiple counties.

Applying the distance decay measures to this partition we see that in general the intra-community tie strength is greater up to approximately 50km but after this distance there is no discernible difference between intra- and inter-community ties (see Figure 6.39). This implies that these large communities are actually amalgamations of smaller communities but the method fails to identify these as independent communities.

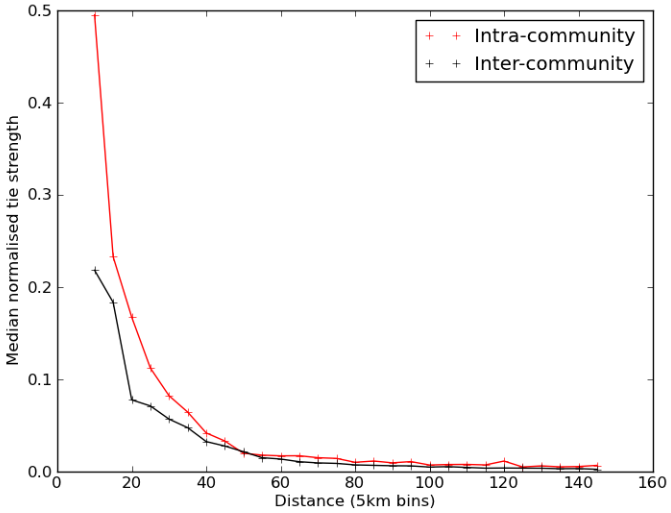


Figure 6.39: Normalised median interaction levels for inter-community (black) and intra-community links (red) in the towns network with communities found by Infomap.

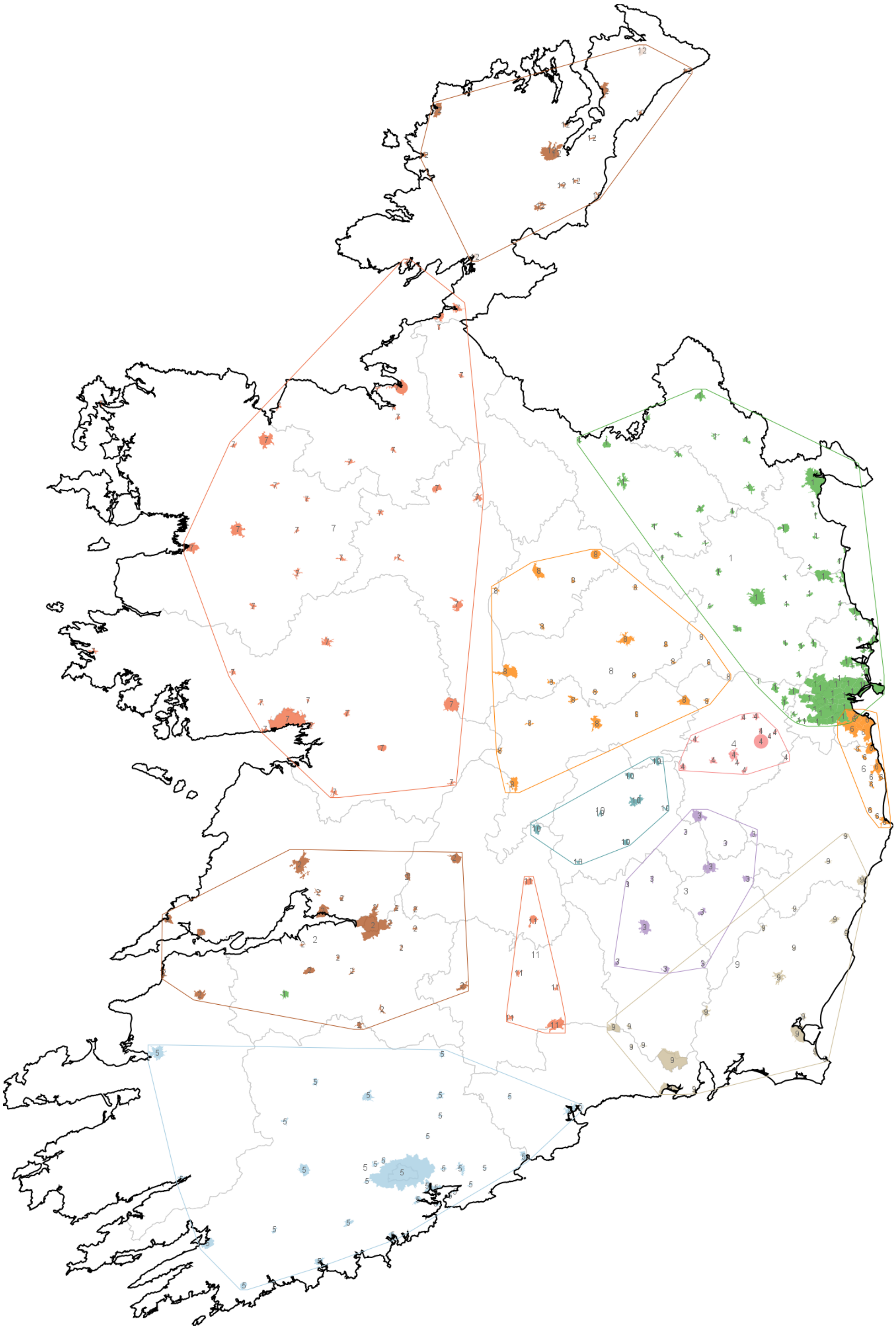


Figure 6.38: Communities identified in the towns network using Infomap.

6.7.6 Local optimisation technique

In this section we apply an alternative local optimisation method, OSLOM (Lancichinetti et al., 2011), to test the hypothesis that the large communities found by both the Louvain and Infomap methods are due to the global optimisation approach of these algorithms coupled with the uneven distribution of population in this network. In theory a local optimisation approach should be more suited to this network and should help us see past the resolution limits of the other methods. We use the undirected, unweighted version of the algorithm where each edge weight is considered as a separate independent edge. We choose this version because each edge weight represents a single tie between two individuals.

Tower level network

We apply the method to the tower level network using the partition found by Infomap as a 'hint' for the method. This is a recommended way to speed up the algorithm, making it more likely to find a good solution in a reasonable number of iterations (Lancichinetti et al., 2011). Importantly, the method does a full analysis of the hint partition to clean up and explore the substructure of each community to produce a new lowest level coverage (it allows multiple community assignments per node). While the hint partition from Infomap contained 37 communities, this method identifies 183 communities of two or more nodes. As we might expect given the high number of communities, each community is generally smaller than those found by the Louvain and Infomap methods (see Figure 6.40). Most significantly however, the counties and cities of Cork, Limerick and Galway all divide into a number of separate communities (see Figure 6.41) proving our original hypothesis that there is community substructure present that the other methods fail to identify. In Dublin we see a large number of small communities which appear to be subdivisions of the previously found north-south, east-west divisions. With each city centre we see a high number of nodes that belong to multiple communities (denoted by multicoloured nodes). It is quite likely that we are seeing the effect of the limited spatial resolution of cell towers here where people from separate areas and communities are using the same tower. We also see the same effect in some more rural areas where the tower density is low. Overall we find that 13% of the nodes belong to more than one community and 16% of these belong to more than two (shown in grey on the maps).

An important feature of OSLOM is that the communities are guaranteed to have no internal substructure so all communities satisfy the same definition of being indivisible units. The method still allows us to explore the hierarchical structure of the network however, without altering this definition. A new network is created were each of the previous communities is a single node that cannot be divided and the aggregated links between communities are the links between these nodes. Importantly, self links are removed from these new nodes. If they were not removed the method would assign each node to its own community and stop immediately.

In the second level results we find 40 communities (see Figure 6.42), which are generally more similar in size and composition to the communities found by the Infomap method. Importantly we still see multiple communities within each city and each of the most populated counties. While most of the communities are spatially contiguous, we do see some communities with spatially disjoint parts at this level. The most interesting of these is perhaps community 35, which covers the south east of Dublin city and also parts of south Wicklow, Kilkenny and Wexford. It is difficult to explain however why this is the case.

In the next level of the hierarchy we find 11 communities where each one covers multiple counties (see Figure 6.43). Interestingly we find four separate communities radiating out of Dublin to the north, west, and two to the south.

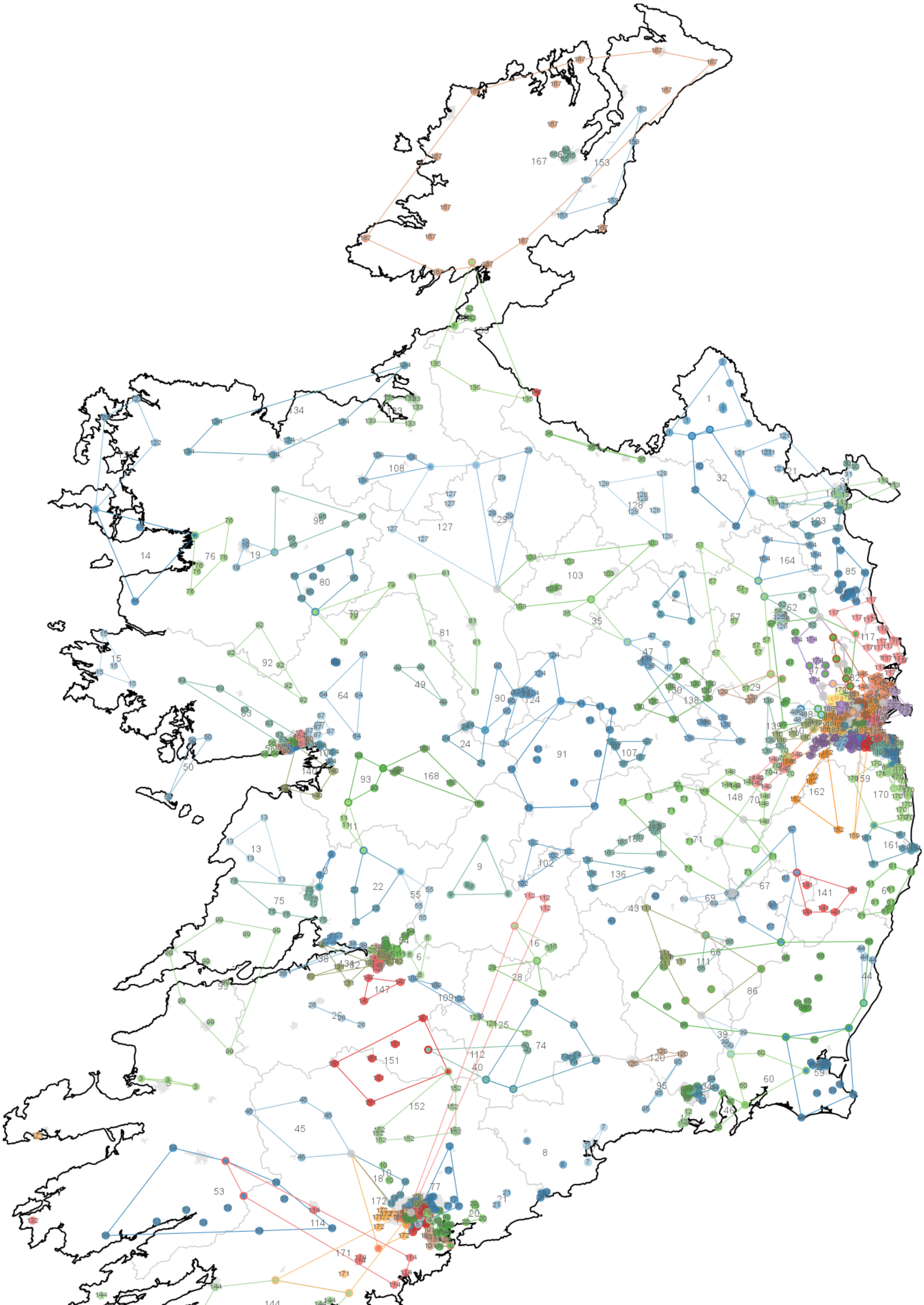
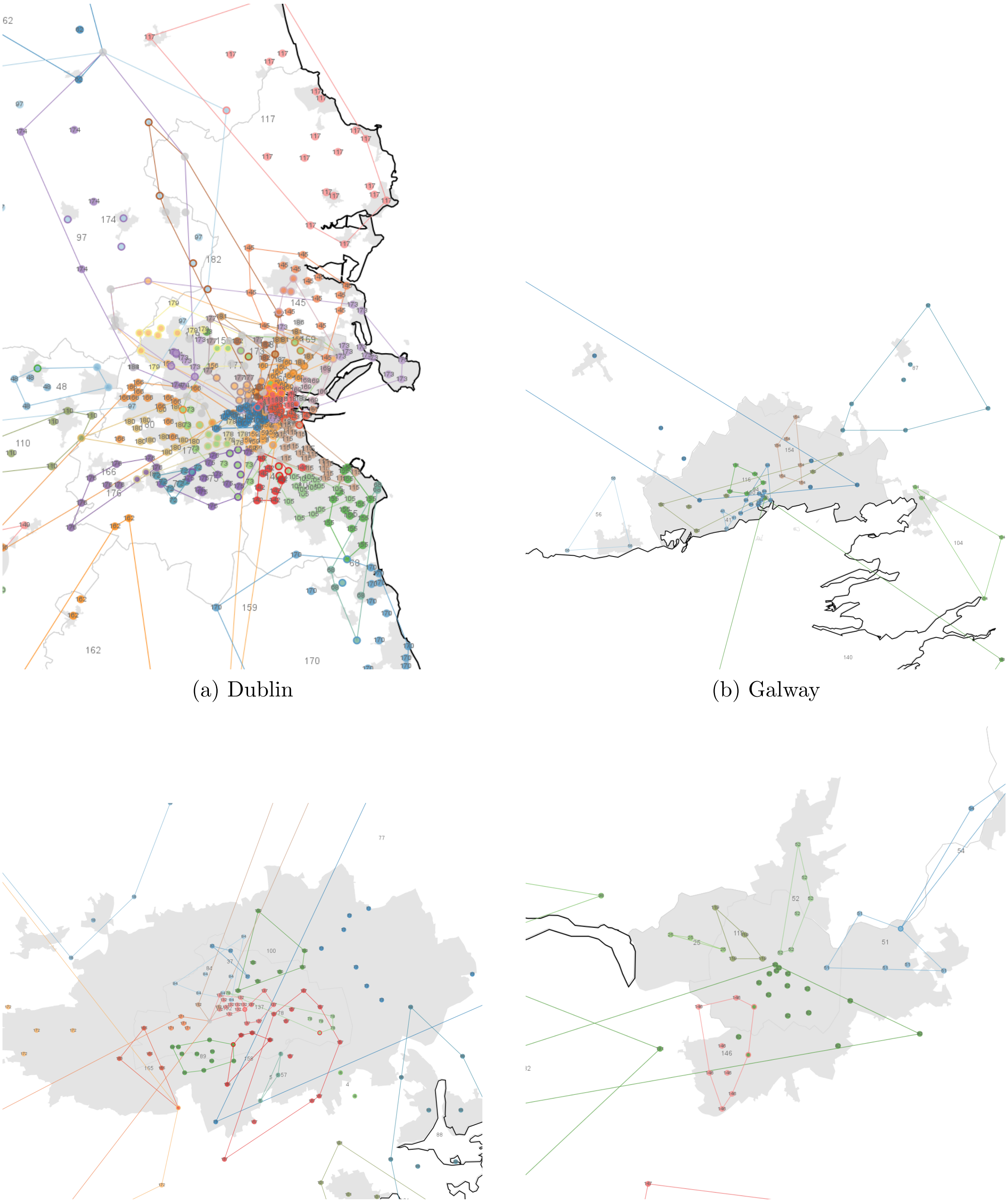


Figure 6.40: Communities identified in the tower network using OSLOM at hierarchical level 1.



(a) Dublin

(b) Galway

(c) Cork

(d) Limerick

Figure 6.41: Communities identified in the tower network using OSLOM at hierarchical level 1. Close up view of the four main cities.

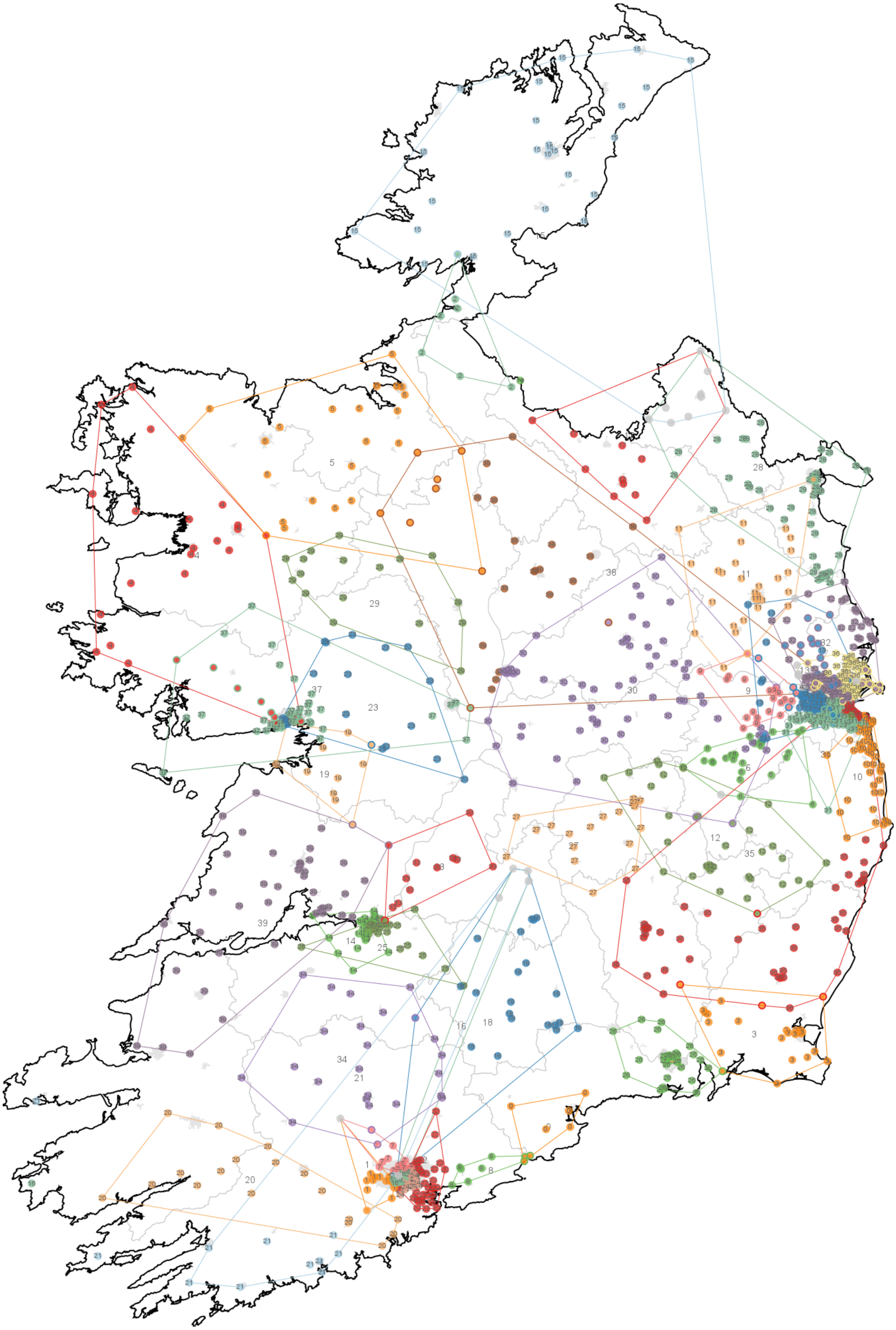


Figure 6.42: Communities identified in the tower network using OSLOM at hierarchical level 2.

These generally follow the directions of the main transport routes into the city so perhaps are due to commuting patterns. We see similarly large communities around Cork, Limerick/Clare, Galway, Sligo and Donegal.

In the final level of the hierarchy we find a rather surprising result of only two communities with most of the county of Cork as one community and the rest of the country as the other community (see Figure 6.44). While some people in Cork (and some outside) might agree with this idea (it is jokingly referred to as the People's Republic of Cork by some!), the significance of communities at this level of aggregation is questionable.

We apply the same distance decay measures to these results at each level and find that in the first three levels there is clear separation between the intra- and inter-community tie strengths (see Figure 6.45). In the final level there is no clear separation, which is not surprising given the size of the communities. This provides further confirmation that the results really capture the underlying structure of society.

Town level network

We now apply the same local optimisation method to the towns network to investigate the effects of spatial aggregation with this method. Once again the algorithm is started using Infomap to provide an initial partition. The method finds a three level hierarchical structure with 51, 12 and 3 communities respectively. It is understandable that there are fewer communities at the lowest level of the towns network because there are fewer nodes altogether as we have both aggregated and discarded some towers. However in the tower level network we saw some communities that contained the towers of just one or two towns but in this network where are very few of these small groupings and instead see more communities of a larger size. Furthermore we see that Galway city shares a community with Sligo town and Cork city shares a community

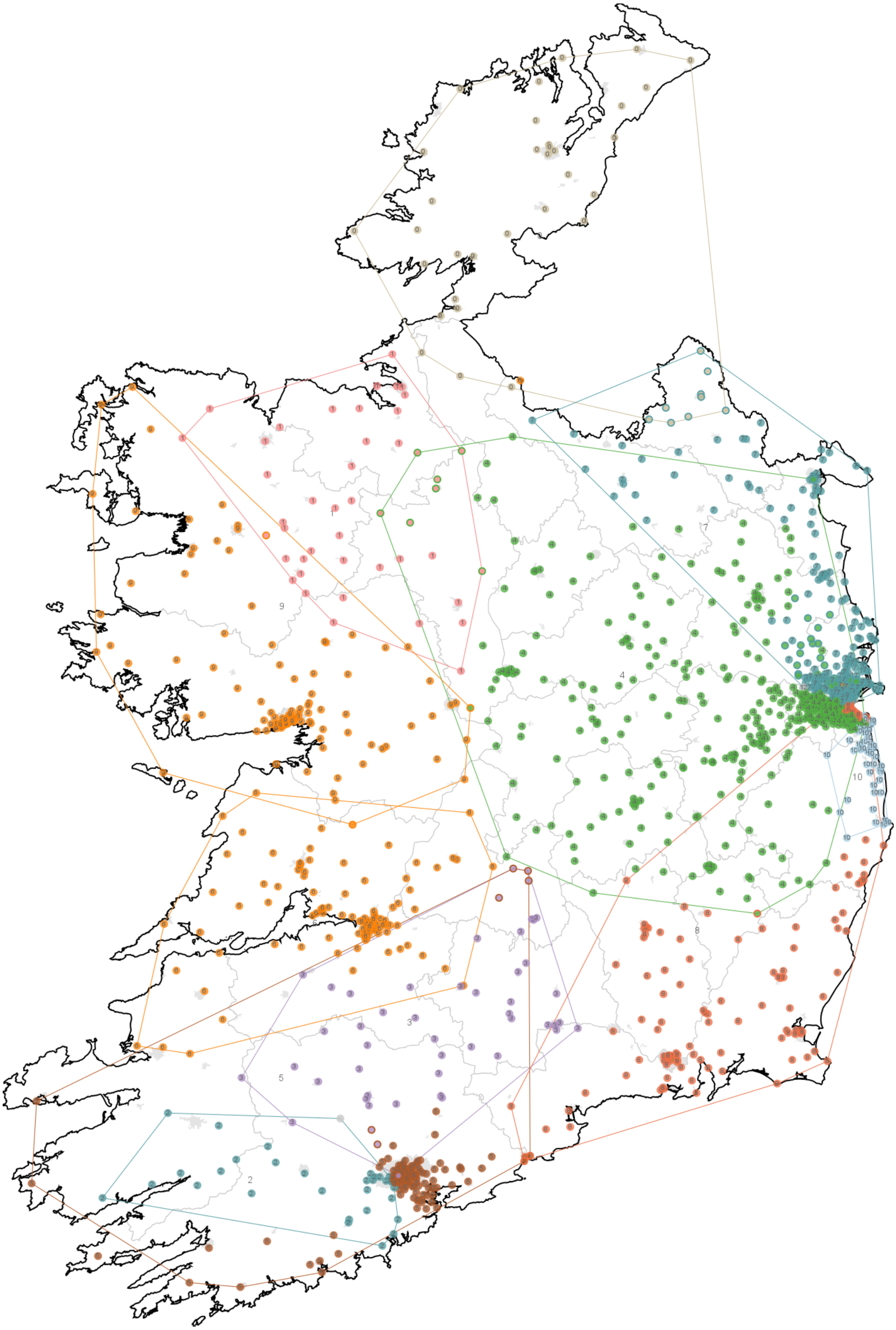


Figure 6.43: Communities identified in the tower network using OSLOM at hierarchical level 3.

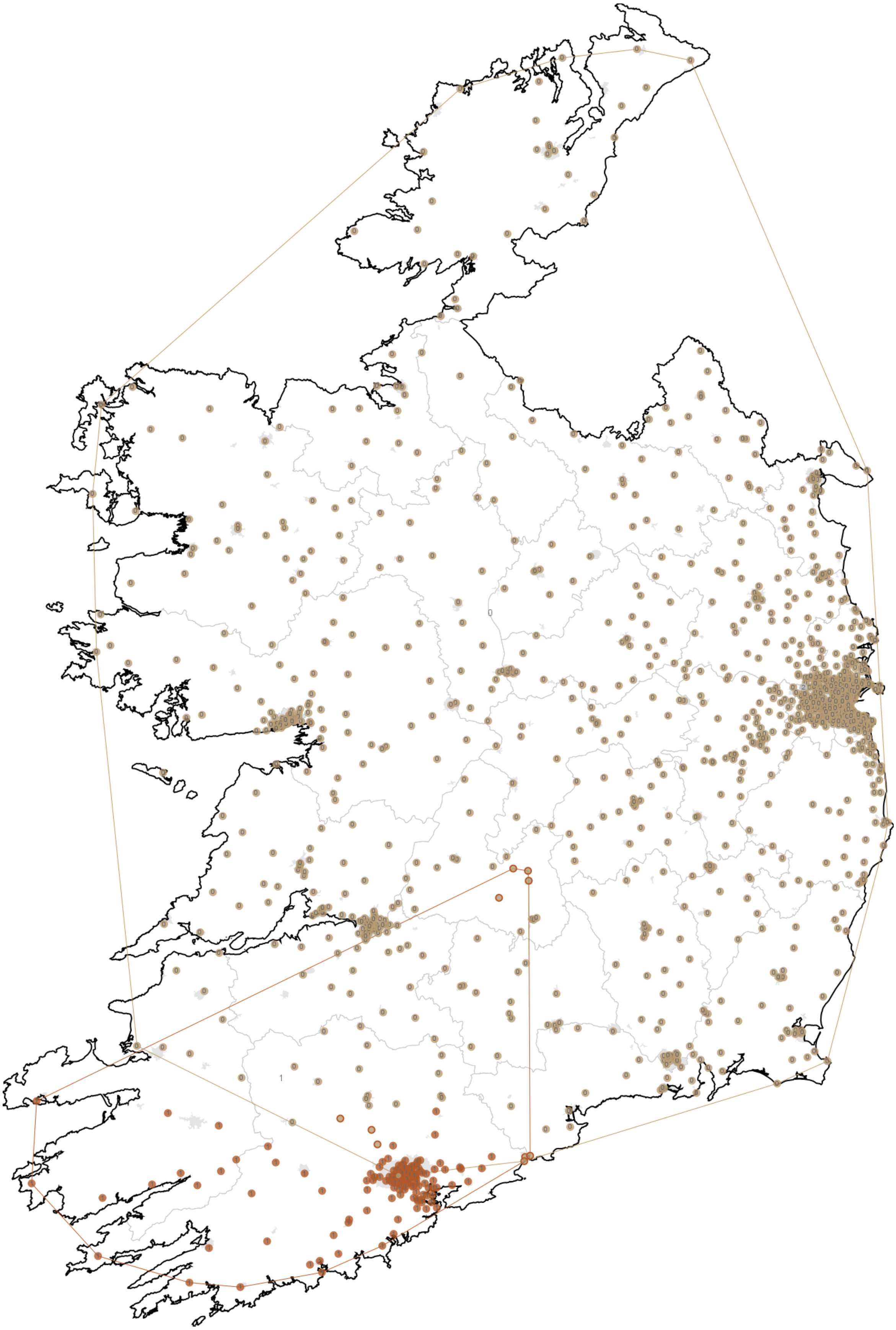
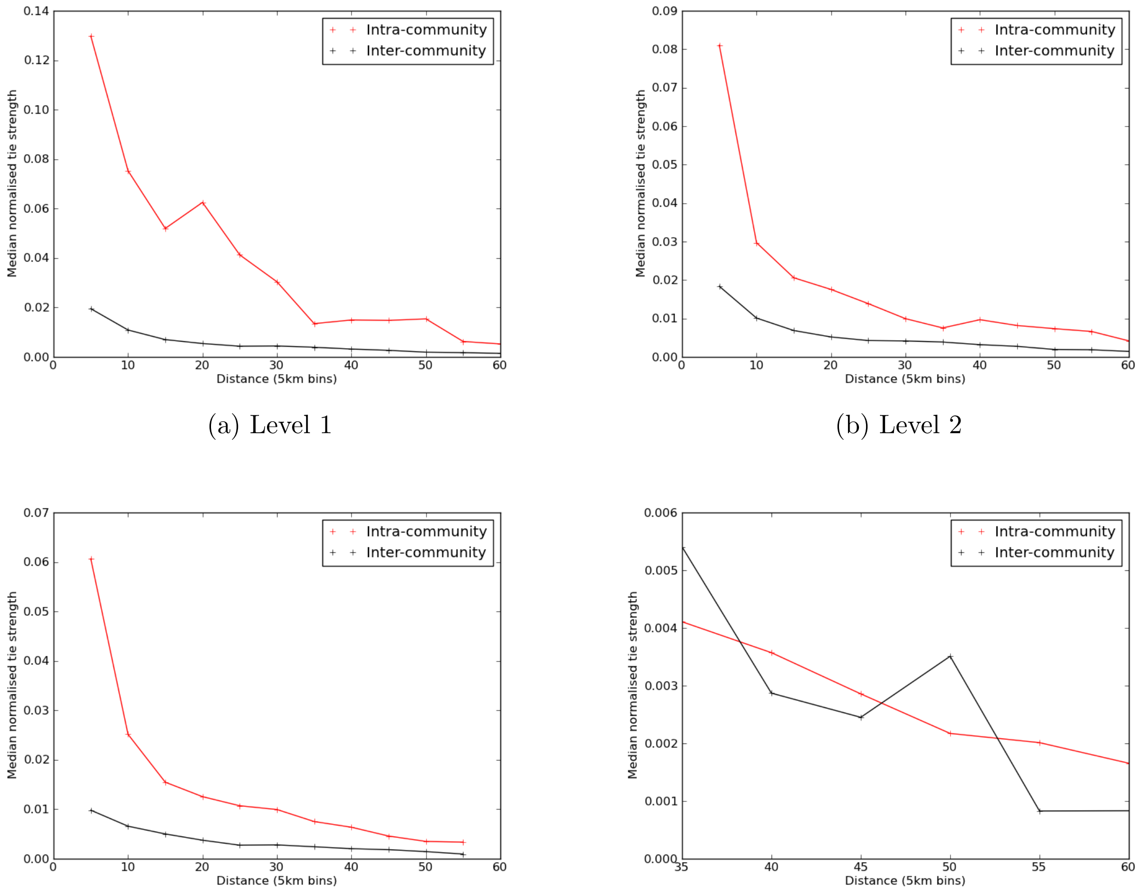


Figure 6.44: Communities identified in the tower network using OSLOM at hierarchical level 4.



(a) Level 1

(b) Level 2

(c) Level 3

(d) Level 4

Figure 6.45: Normalised median interaction levels for inter-community (black) and intra-community links (red) in the tower network with communities found by OSLOM.

with Limerick city whereas at the towers level the towers of the cities mostly joined with the towers of their county in the higher levels of the hierarchy.

These effects can be explained by the fact that OSLOM ignores self links and only considers links between nodes. It therefore cannot ‘see’ that these large nodes are in fact very significant communities in their own right. This highlights an important resolution limit with OSLOM that was not so obvious in the tower network. Intuitively we would expect that the smallest possible communities that can be found would be individual nodes. It is obvious that nodes cannot be split further but we would expect that if a node is much

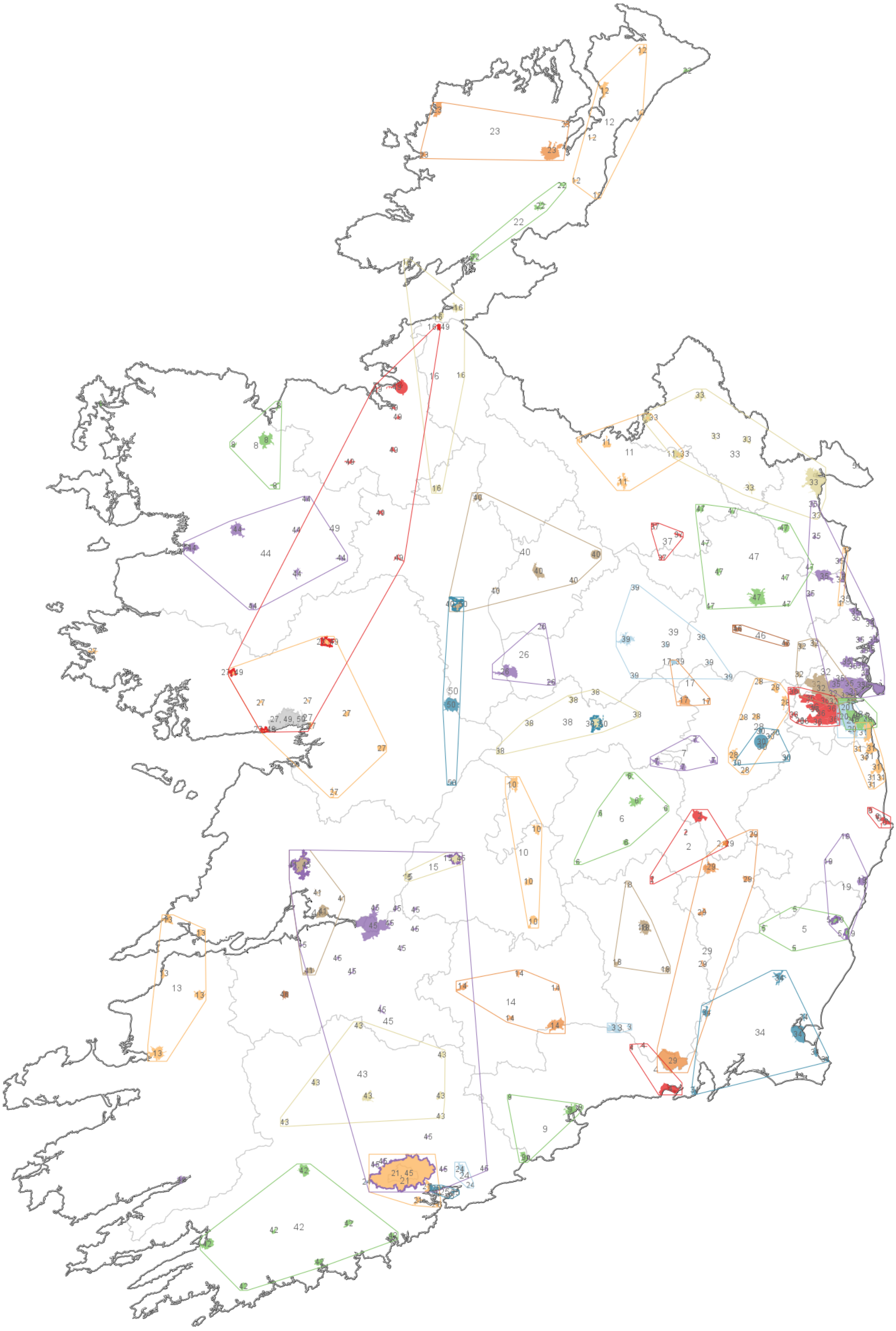


Figure 6.46: Communities identified in the towns network using OSLOM at hierarchical level 1.

more strongly connected to itself than other nodes then it should be in its own community at the lowest level. This is not the case because the method does not consider self edges. In the tower network this was not too much of an issue because most populated areas are covered by a number of towers but at larger levels of aggregation it becomes a more significant issue.

We must conclude therefore that the optimum choice of spatial aggregation unit is a level below the level at which self links are significant, where possible. Obviously if town aggregated data was all that was available in this case there is nothing more that we could do but it would be vital to point out this limit.

6.8 Conclusion

In this chapter we have performed quantitative analysis of the communication patterns in a national mobile phone network. We differentiate between reciprocated and unreciprocated contact pairs and measure the significance of each contact pair based on the number of days of activity. We show that while only 23% of pairs are active at least once a week, these pairs account for 89% of all calls. Applying the same measure of significance to the ties between customers and locations we found the significant locations in peoples lives. Initially we found a single tower and town location for each customer which we denote their home location. We later extended this to multiple significant locations. We investigated the effect of distance on the probability of a tie between customers and found this does not decay as smoothly as reported in other studies. On further examination we found that the metric used here and in other studies does not actually measure this probability and is inappropriate for use with social networks given the limits on the number of social/communication ties any one person may have. We proposed a reworking of this measure and showed that there is a distance decay in general up to a certain distance but there is

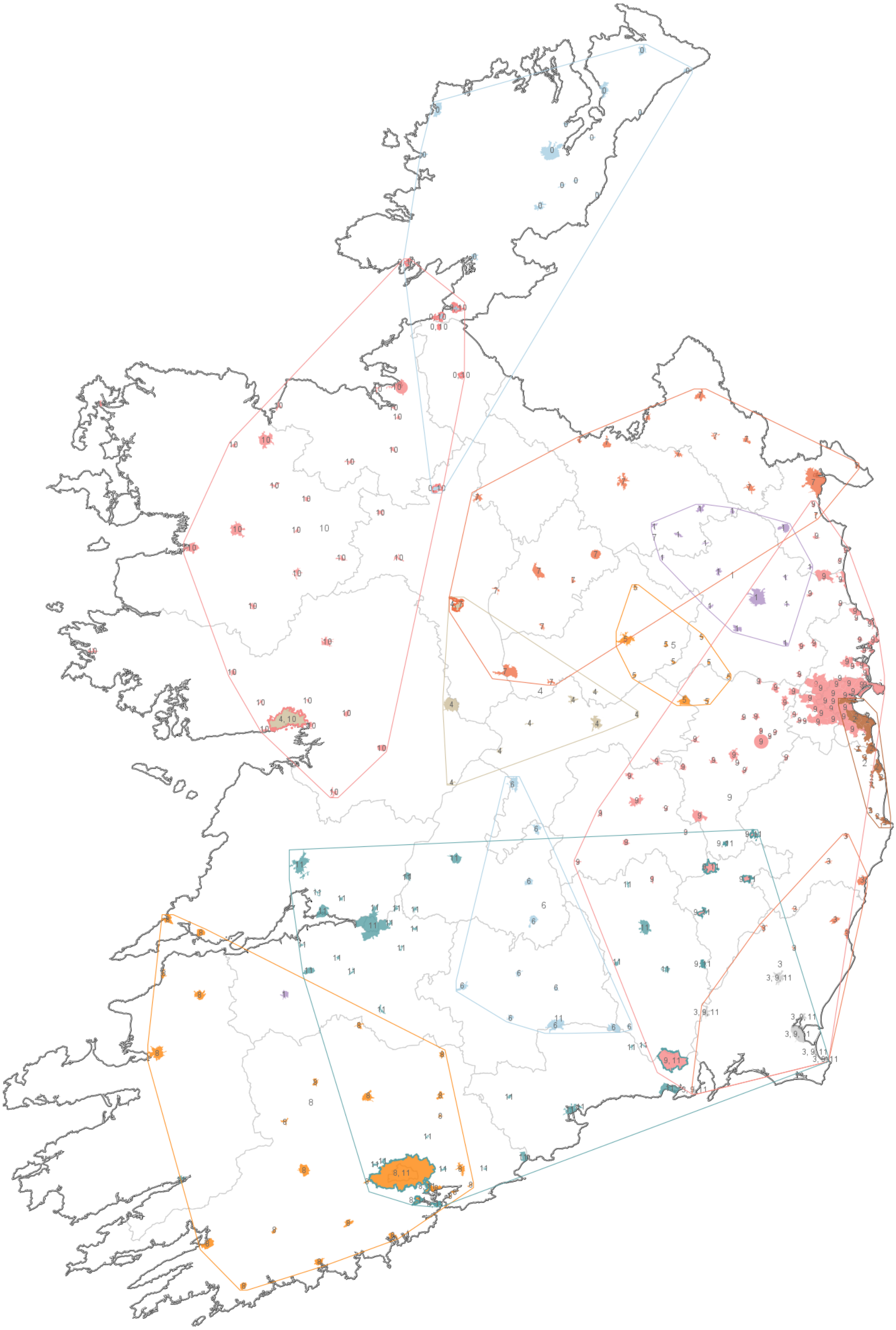


Figure 6.47: Locally optimised communities of the towns network. Hierachical level 2.

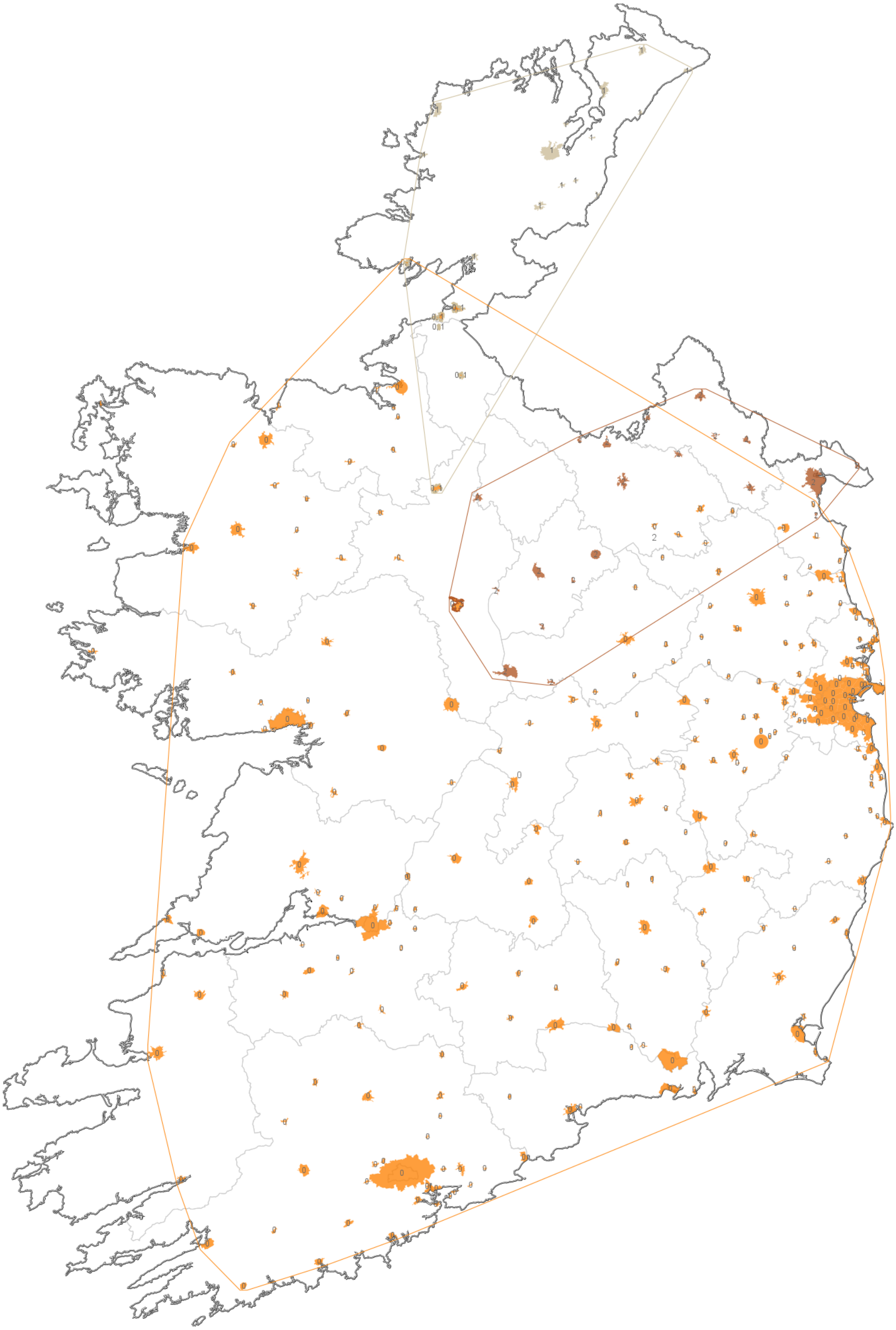


Figure 6.48: Locally optimised communities of the towns network. Hierachical level 3.

wide variance in the values at all distances.

We hypothesised that part of this complexity may be due to the fact that we place each person at a single spatial location rather than considering their full set of significant locations. To investigate this we identified multiple significant locations for each customer based on the frequency of visits. We found that more than 50% of customers have at least 2 locations (separate towns) which they visit on at least 15 days. When we consider multiple significant locations per person we saw that 88% of pairs share significant towns, as opposed to 51% sharing home towns. This result suggests that using a single location for each customer is likely to give a wrong impression of the effect of distance on communication. It also suggests that it is a futile exercise to attempt to model the number of contact pairs or intensity of communication between towns based purely on the attributes of those towns and their separation distance.

On the other hand these results provide some interesting insights about the nature of relationships in this dataset. We see that the majority of pairs share one or more important places in their lives. This may indicate that they live or work or attend school or university in the same place as each other and see each other on a regular basis. We cannot say for certain that a pair of individuals do actually meet with any regularity but we may surmise that their relationship is rooted in one of these shared places.

Following this we turned our attention to community structure and used our proposed measure of interaction strength to investigate if the effect of distance differs between intra-community and inter-community ties. We applied the modularity optimising Louvain method to the network at the spatial scale of towers and towns. In doing so we saw the unfortunate effect of modularity's resolution limit, demonstrating that it is a real problem in real world networks. Despite this, we found that the interaction strength between nodes in the same communities is generally higher than between nodes in different communities at the same distance. Importantly, we also showed that this is not the case for randomly generated communities or clusters found using geographical k-means clustering. This tells us that the communities that are found through modularity optimisation do capture something more than spatial relationships but we still must bear in the mind that there are many possible partitions with similar modularity scores so the results of any particular modularity optimisation procedure should not be assumed to be the only possible result or representative of the 'real' community structure. Furthermore, the resolution limit means that the communities found are not necessarily the most significant.

We tested two recent alternatives to modularity based methods, Infomap and OSLOM, at both spatial scales to assess their applicability to the dataset under study. We found that the Infomap method performed better than the Louvain method by uncovering more smaller communities in the tower network but still appeared to be affected by a resolution limit as it failed to uncover community structure within the large cities and counties outside of Dublin. This effect was further exacerbated in the towns network. The OSLOM method uncovered much more interesting structure than the other methods in the tower network, identifying 183 communities in total at the most significant level, including many inside previously unsplittable communities, with no sign of a resolution limit. This demonstrated the advantage of a local optimisation procedure over the global approach of the other methods. Additionally the distance decay tests showed clear differentiation between intra- and inter-community ties despite the smaller community sizes. On the towns network this method again performed better than Infomap but there was evidence of a resolution limit in effect due to the method's ignorance of self links. This prompted us to conclude that the level of spatial aggregation must be carefully chosen so as to avoid strong self ties. If the majority of the links from a node are self ties it is likely that the true set of lowest level communities will not be found.

In conclusion, our analysis has shown that large mobile phone call datasets offer an excellent opportunity for quantitative analysis of the relationship between social connections and space. However it has also shown that the methodologies used in other recent studies on similar datasets may not be appropriate and do not necessarily provide reliable results. We saw this in the case of the measures of distance decay that failed to account for population distributions or the limits of human social interaction. Similarly we showed that an oversimplification of peoples' spatial relationships leads to a distorted view of the effect of distance on social tie probability. Finally in the context of community detection we found that the most commonly used methods do not provide reliable results, especially when weighting schemes are used that are incompatible with the assumptions of the methods being used or the phenomena we wish to study. We have proposed alternative network representations and investigated the effects of spatial aggregation with two alternative community detection techniques. We have shown that with the correct representation and at low levels of spatial aggregation the local optimisation approach of the OSLOM method performs very well, providing a detailed interpretable view of the multi-scale spatial structure of society.

Chapter 7

Discussion

7.1 Introduction

In this thesis we set out to understand the effects of space on the formation of social ties in mobile communication networks. We wished to understand how the distance between individuals affects the likelihood of a social tie between them. We also attempted to understand if this effect is continuous across space or if it is affected by borders, be they officially defined or social constructs. The literature on large scale societal analysis from communication networks brought up two seemingly conflicting ideas. On the one hand a number of studies suggested that tie formation and communication probability decay smoothly with distance and can be accurately modelled with simple spatial interaction models. On the other hand research had shown the presence of discrete spatially contiguous communities which corresponded well with known administrative, social or linguistic regions, suggesting that interaction probability is strongly affected by these borders.

We have tackled these issues through analysis of the theory of social interactions, spatial interaction and community detection and quantitative experimental analysis of a nationwide mobile phone network. We will discuss the findings on each of these topics in the following sections.

7.2 Network representation of social ties

In the literature on societal analysis from communication networks we encountered a variety of different network representations of society. In many cases a clear distinction was not made between properties of the dataset being used and those of the society for which it was acting as a proxy. We disentangled these issues by identifying properties that are due to the technology (e.g. directionality of communications, cost of communication and spatial resolution), properties due to communication behaviour (e.g. intensity of communications) and properties that are due to the structure of society itself (e.g. the presence of reciprocate communications and repeated visits to the same locations). By disentangling these various factors we were able to separate out those that are important to incorporate into models and method inputs and those which we must take into account when interpreting results.

We also discussed the differences between societal networks derived from communication networks and various other types of networks to motivate the need for further examination of the methods being used to study them. We discussed how the properties of these networks differ from other spatially and geographically embedded networks and the impact that these differences have on the assumptions we make. Specifically, we addressed the issue of the cost of distance and argued that concept is inappropriate both for mobile communication networks and societal networks. We argued that the focus should be on understanding the opportunities for interaction rather than the minimisation of effort or cost.

In relation to the choice of nodes in network representations we discussed the issues of spatial aggregation of individual data, particularly relating to spatial heterogeneity and the modifiable areal unit problem. Choosing the appropriate level of aggregation is important for ensuring reliable and interpretable results. In many cases aggregation to officially defined population centres is most useful because the density, and hence resolution, of cell towers is much greater in areas of dense population. Where there are greenfield sites between such areas there is a lower chance of wrongly geolocating an individual because the cell tower coverage areas will be better separated. On the other hand using gridded cells or continuous regions is more problematic because of the higher chance of overlapping coverage areas and incorrect geolocation. In these cases it is very important to be aware of, and if possible account for, inter-nodal ties that are due to incorrect geolocation rather than actual inter-nodal communications and ties.

In the analysis conducted in this thesis we have taken a multi-scale approach in an attempt to understand the effects of aggregation at different spatial scales.

7.3 Distance decay in social networks

One of the key topics we wished to explore in this thesis was that of the effect of distance on interaction probabilities between individuals. In our review of the literature we identified a number of different studies looking at this distance decay effect in a variety of manners. Lambiotte et al. (2008) has been widely cited for showing that the probability of a communication tie in mobile phone networks decays with the square of the separation distance while Krings et al. (2009) and Kang et al. (2012) showed that communication flows can be modelled accurately with gravity models. We discussed the theory behind spatial interaction modelling and found such simplistic models to be lacking in theoretical basis, particularly when applied to communication networks where the concepts of the cost of distance and spatial choice do not apply.

On further analysis of the measure reported by Lambiotte et al. (2008) we found significant issues with the calculation of the interaction probabilities, due to a neglect to consider the nature and limits of human communication. We showed that the measure used is unsuitable because it is based on the flawed assumption that the number of interactions or ties per person can increase unbounded with the number of available opportunities. This is clearly impossible in real world networks and leads to distorted probabilities and an inaccurate representation of the effect of distance on communication tie formation. The results of Lambiotte et al. (2008) cannot therefore be taken as evidence of a decay in a tie probability with the square of the separation distance.

We proposed an alternative comparative measure for the interaction level between two spatially separated groups that uses the ratio of observed ties to the smaller of the two group populations. This particular ratio was chosen to reflect the fact that the number of ties between two group populations must scale with the number of people in the smaller group and is independent of the size of the larger group. It is not intended as a predictor of the number of ties between two groups, but rather as a more accurate comparable measure.

Applying both measures to real data from a nationwide mobile phone network in Ireland we showed that the distance decay is different in both cases. Furthermore we pointed out that there is much variance in interaction strength at all distances. This variance only becomes visible when appropriate visuals such as scatter and box plots are used in conjunction with more aggregated statistics. The large variance at all distances implies that models based solely on population and distance effects will poorly predict or explain these interactions.

7.4 Distance decay in dynamic settings

In our attempt to understand the factors that might explain the interaction levels between individuals and between towns we considered that perhaps an incorrect measure of distance is being used. We noted that the majority of people visited a number of different locations on a regular basis with at least 50% of the population visiting two or more towns at least once a week on average. As with other studies in the literature, in the calculation of spatial separation distances we always considered a single 'home' location for each individual but we hypothesised that perhaps contact pairs who do not share home towns may share other significant locations. We found this to be the case for nearly 90% of all pairs. If we consider pairs of people who share the same home location (tower or town) to have a separation distance of zero we must also consider those who share other significant locations to also the same separation distance of zero. Given that nearly 90% of all pairs can then be said to have an effective separation distance of zero it is a futile exercise to try to measure the effect of distance on communication ties.

This result suggests that distance affects the likelihood of communication and social ties in a very different way to commuting, migration or freight. We see that such ties generally exist between people who have opportunity to meet on a regular basis. The locations of the homes of individuals contributes to the likelihood of them having the opportunity to meet but the distance between these locations is not important if they meet at a third location. In these cases the effect of distance is more complex and the attributes of the home towns of the individuals make little contribution to the likelihood of interaction. This helps explain why we saw such poor fits with simple spatial interaction models. It also suggests that the oft-discussed 'death of distance' (Cairncross, 1997) due to communications technology is not quite as real as it may seem at first. If anything we see that it is our transport systems that facilitate the movement of people in such a manner that they can regularly visit locations separated by large distances and thus make and maintain relationships with others who may live at even greater distances from themselves.

7.5 Spatial interaction modelling of mobile communication

In our review of spatial interaction models applied to mobile communication we discovered a number of methodological and theoretical issues in the literature. Lambiotte et al. (2008), Krings et al. (2009) and Kang et al. (2012) each claim to show that a basic gravity model can accurately model the communication flows in both Belgium and China but little evidence is provided to support these claims. In the case of Lambiotte et al. (2008) and Krings et al. (2009) there is no discussion of model calibration or parameter estimation, suggesting an ignorance of how spatial interaction models have been used for more than fifty years in the geographical and transportation sciences.

Aside from the methodological issues in the application of specific models there are deeper theoretical issues. The social physics approach to modelling social phenomena in space has long since been discredited in the geographic literature for its lack of behavioural theory and the poor fit of simple models. There is now a realisation that spatial interaction models only apply when there is an element of spatial choice involved in the decision process and that spatial choice must be based on the cost of distance. We can only say that there is an element of spatial choice if the choices made by individuals are heavily influenced by the spatial configuration of opportunities or alternatives. If this is not the case then spatial interaction models are not an appropriate choice and are unlikely to give good fits and predictions, no matter what calibration methods are used. We have previously discussed how mobile communication links and the social ties they represent are embedded in space but in a different manner to other types of spatial networks such as transport, infrastructure or neural networks. When individuals make phone calls they do not make a spatial choice, instead they choose to call a specific person, usually someone they already know. The social tie itself is likely to be the result of sharing a physical location rather than due to the distance between the home locations of each individual. The shared location where the tie was rooted may not be a home location or even a current significant location of either individual. For these reasons it is quite clear that the concept of spatial choice does not apply to either mobile communication or the existence of social ties and therefore spatial interaction models are inappropriate for modelling these processes.

It must be noted that we are arguing against the use of spatial interaction models for these types of networks from a theoretical standpoint. We make no attempt to actually calibrate such models for the Irish dataset because our analysis of the underlying theory tells us that they are inappropriate.

7.6 Community detection in social networks

The second key topic we wished to explore in this thesis was the concept and detection of communities in societal networks. In the literature there are a number of studies on finding communities in both individual level social networks (Palla et al., 2007; Blondel et al., 2008; Onnela et al., 2011) and spatially aggregated networks (Blondel et al., 2010; Ratti et al., 2010; Calabrese et al., 2011a) derived from communications data. Each of these studies identifies communities that seemed to have plausible explanations and uncover interesting patterns of communication and social ties. They do not however provide *a priori* explanations or expectations of communities so it seems likely that in each case an alternative *post hoc* explanation could be provided if the results had been somewhat different. Furthermore each study uses a different network representation of the original phone call dataset.

Our review of the literature of community detection methods showed that there are known problems with the most widely used family of methods based on the modularity quality metric. These issues include a resolution limit and a quality function that has no clear maximum with a high number of dissimilar high quality solutions. Additionally we saw that each of the methods make certain assumptions about the networks to which they are applied. We noted that while the clique percolation method has an *a priori* local definition, it is designed for unweighted networks. Modularity based methods on the other hand can handle weights but there appears to be confusion over what types of weights are appropriate. In particular it does not differentiate between weighted edges and multiple edges in its null model. The OSLOM method on the other hand does make this differentiation and also does have an *a priori* local definition of community. It ignores self edges, however, on the assumption that they are not important in real networks. While this is generally true for individual level social networks our analysis has shown they become increasingly important in spatially aggregated networks as the size of spatial units increase.

7.7 Community detection in spatially aggregated social networks

In our review of previous studies on spatially aggregated networks we noted that in each study the choice of network representation, and weighting scheme in particular, was inappropriate for the method used or the phenomenon under analysis. The first attempt at spatial partitioning using mobile phone networks by Blondel et al. (2010) used two alternative weighting schemes, the first of which suffered from the same misspecification problem as the probability measure of Lambiotte et al. (2008). The second scheme used average communication time which is perhaps generally a more informative measure than the total number of calls or the total duration but is unsuitable from a modularity optimisation perspective. The aggregate number of calls or communication seconds used by the studies of Ratti et al. (2010) and Calabrese et al. (2011a) is equally unsuitable in the context of modularity optimisation given the lack of independence in the units of weight. These examples highlight the need to consider carefully both the phenomenon under study and the assumptions of the method used for analysis. Bearing this in mind, we determined the most appropriate choice of edge weights for spatially aggregated networks when using modularity optimisation to be a single undirected edge per pair of connected individuals where an edge is present if there has been at least a given number of reciprocated communications.

Our empirical analysis on Irish data showed that the previously discussed issues with community detection methods are not merely theoretical but have significant practical implications, especially in the case of spatially aggregated networks. The aim of this analysis was not to find the best community detection method for this type of data. Rather we wished to demonstrate the effects of the known issues with the most commonly used methods and show that there are alternatives that do not suffer from the same limitations. The best method to use in a given situation will always depend on the specifics of the dataset under study and the aims of the analyst. Our analysis shows however that there are certain methods whose use cannot be justified because they suffer from fundamental flaws that affect the quality of their output. We compare and discuss these results below to highlight these issues and demonstrate the difference in output between alternative techniques.

7.7.1 Comparison of results of community detection methods

	Louvain	InfoMap	OSLOM
Number of communities	14	37	190
Median number of towers	100	26	6
Median population	113459	25300	4979
Fraction of intra-community ties	0.80	0.77	0.55

Table 7.1: Summary of results of community detection on the towers network.

Louvain Method

The results of our analysis have shown that while the Louvain method reveals a hierarchical structure (Figures 6.11, 6.12 and 6.13) it fails to find valid communities below a certain size at any level (Figures 6.17 and 6.18). Not only does this show that the hierarchical unfolding is unreliable but also that the modularity optimised communities at the final level are amalgamations of other valid but undetected communities (Figure 6.17). This demonstrates that the resolution limit has a significant effect in spatial networks, especially in cases where there is an uneven spatial distribution of edge weights, as is the case for this network in Ireland. Smaller communities are found in the eastern part of the country where there is a higher concentration of high degree nodes. This resolution limit coupled with the previously discussed unoptimisable nature of the modularity function makes the Louvain method, and modularity optimisation methods in general, a poor choice for community detection.

Despite these limitations we have shown that the communities found by this method are not completely arbitrary and are more meaningful than some other purely spatial partitions of the country. We showed this by calculating the distribution of the median normalised interaction level over a range of distances for both intra- and inter-community ties. These results (Figure 6.21) showed that in general the intra-community ties were stronger than the inter-community ties at the same distance. The effect was shown to be greater for each of the optimised partitions than both a random partition (Figure 6.23) and one determined by geographical k-means (Figure 6.24). This tells us that people are more likely to share a social tie with someone else from the same community at a given distance than someone from a different community at the same distance. These results provide further explanation as to why there is a lot of variance in the interaction likelihood for towns separated by the same distance.

Infomap Method

Our analysis using the Infomap method has shown that the method suffers from the same resolution limit problems as the modularity optimisation approaches. While the number of detected communities was higher and the size of those communities was smaller than those found with the Louvain method (see Table 7.1), the results still show large communities in the west of the country (Figure 6.34). The communities align quite well with the county boundaries in many places and thus may seem reasonable divisions of the country. On the other hand a community is meant to be an indivisible component with random internal structure and it seems unreasonable that there is no community structure within each county. The fact that Dublin is divided into a number of communities does not necessarily show that Dublin has a more divided social structure but rather that the method is affected by the population and weight distribution in the network. This is due to the global optimisation approach of the method. When a single quality measure for the whole partition is optimised the size of communities is determined by the overall size of the network. As the network size increases so does the size of the communities that the method will naturally uncover.

OSLOM Method

The results of the OSLOM method applied to the towers network (Figure 6.40) showed a very different picture to the previous two methods. From the maps and Table 7.1 we see that the number of communities is much higher and the size of those communities is much smaller than either of the other two methods. While smaller communities are generally preferable to larger ones as they are more likely to be indivisible units, we don't just blindly look for the smallest possible communities. This of course would lead us to the trivial solution of each node being its own community. Instead we have shown that these communities genuinely capture the social structure in the country by confirming that in each case the intra-community tie strength is greater than the inter-community tie strength at the same distances. This would not be true if these communities were not in fact standalone communities but subdivisions of a larger community.

While the local optimisation approach of the OSLOM method appears to prevent the resolution limit that plagues global methods like Louvain and Infomap our analysis at the spatial scale of towns (Figure 6.46) has shown that there is another type of resolution limit with this method that must be taken into account. Self loops are often ignored or explicitly removed from networks for the purposes of community detection but in spatial networks they take on much greater importance. The likelihood of two individuals being connected is high if they share the same location and obviously as the size of that location increases, so does this probability. For this reason in the tower level network self links were less significant than they were in the town level network. The fact that OSLOM ignores self links means that a single node could never be considered an independent community, even if it is nearly completely isolated from the rest of the network. This is why we saw the City of Cork in the same community as Limerick City in the OSLOM results at the town level, which completely contradicts the results from the tower level. Our own intuition helped us to see the problem and understand which set of results where correct. This analysis shows the importance of considering the assumptions and limitations of the methods being used. As long as self links are not significant, OSLOM appears to perform well but if self links are important then it cannot uncover the true structure and will instead produce misleading results.

The downside of this finding is that higher resolution networks require more computation time for the community detection process. This is particularly relevant for OSLOM which required more than 48 hours of computation time for the modestly sized towers network in this study, compared to a matter of minutes for the fast modularity or Infomap methods. It is likely, however, that large performance gains could be achieved with an implementation that takes advantage of parallel architectures given that the method performs the same computationally expensive operations on each community found independently.

Chapter 8

Conclusion

8.1 Introduction

While the spatial analysis of mobile phone communication logs is a relatively new area of research there is much interest in this field because of the unprecedented level of detail that these datasets capture about human interactions. The opportunities for developing our understanding of human society from these datasets are immense. There is already a rich corpus of literature of methods developed for modelling spatial interactions and analysing the topological structure of networks so it has been a relatively straight forward task for researchers to apply these techniques to this new type of data. This has led to a number of interesting findings on the decay of communication tie probability with distance, the prediction of communication flows with gravity models and the uncovering of communities which closely match linguistic and social borders.

In this thesis we have examined the methodologies behind these studies and the theory underpinning the methods used. Our analysis has shown that there are significant issues with the methodologies employed leading to a misinterpretation of the results and a misrepresentation of the phenomena under study. Our analysis has shown that these issues stem from the assumption that the same techniques, models and methods can be used to analyse all types of networks. We have discussed in detail in this thesis the human interaction processes that generate this data and the need to carefully consider this context at all stages of analysis. The process of converting raw CDR data into interaction matrices or networks requires particularly careful consideration but has received little attention in the literature. In our discussion of community detection methods we highlighted the importance of considering the assumptions regarding edges weights in the method being used. With regard to spatial interaction modelling and distance decay analysis we saw how the particular choice of method of determining the spatial location of each individual radically affected the results. By ignoring the dynamic nature of mobile communication other authors have misinterpreted their results and come to false conclusions about the effect of space on human interactions.

8.2 Summary of contributions

The contributions of this thesis can be summarised as follows:

- The concept of distance decay in spatial communication networks has been revisited and existing measures have been shown to give incorrect results due to their failure to account for the realities of human social interaction behaviour on two fronts;

    1. The failure to consider the limit on the number of possible social ties per person has led to an incorrect definition of interaction probability and thus an misinterpretation of the distance decay. We have proposed an alternative measure which shows a significantly different decay profile.

2. The assumption that each individual can be mapped to a single location has led to an inappropriate measure of separation distance that also affects the interpretation of distance decay. We have shown that the number of inter-town ties decreases by 80% when considering multiple significant locations per individual.

- The theory behind the most commonly used spatial interaction models has been shown to be incompatible with the processes of interaction in mobile communication networks and thus the application of such models to these networks is likely to give spurious results.

- The concept of community structure and its detection in spatially aggregated communication networks has been explored, yielding the following findings:

  - The appropriate network representation, in terms of spatial aggregation and edge weighting scheme, is very much dependent on both the phenomena under study and the method used for analysis.
  - The known limitations with the most commonly used family of methods have real practical effects in spatially aggregated networks leading to incorrect results and misinterpretation.
  - Global optimisation approaches to the detection of community structure in spatially aggregated networks are likely to suffer from a resolution limit and yield incorrect results due to the uneven spatial and network distribution of edges.
  - Local optimisation approaches such as OSLOM produce reliable and interpretable results provided the appropriate network representation is employed. The choice of the level of spatial aggregation is particularly important with these methods because they do not handle self edges.

    - The intra-community interaction strength has been shown to be greater than the inter-community interaction strength at the same distances in the results of OSLOM, demonstrating that the communities found are not simply due to the spatial location of nodes.
    - The performance and usability of recent methods needs to be improved to enable their further adoption and replacement of modularity as the default method for studies on community structure.

- When the appropriate network representation and methods were used, significant hierarchical community structure was found in the Irish network. Clear discrete communities were found in each of the major cities providing further quantitative evidence of known social divides.

8.3 Future research

8.3.1 Further empirical studies

The research presented in this thesis has brought into question some of the results of previous studies and proposed alternative methodologies. These alternatives have only been empirically tested on a single dataset so there is a need for further empirical analysis, ideally on the datasets of those original studies, to validate and refine the approaches proposed here. There is also an opportunity to potentially uncover new insights in these datasets that were hidden to the unfocused lenses of the global methods that were previously used.

There is also a need for further research on the link between communities in individual level networks and communities in spatially aggregated networks. There has been an implicit assumption that the latter are simply aggregations of the former but this is something that requires detailed analysis. Unfortunately the current implementations of OSLOM make this an extremely arduous task given the size difference between individual and aggregated networks. Further work on improving the performance of these implementations is therefore of great importance (see below).

It is likely that the most important work in this area is yet to come. The idea of using mobile phone call logs as a proxy for society was an inspiration for many studies, leading to the idea of using this data to analyse the network structure of society from a spatial perspective. The work presented in this thesis is an incremental improvement on this idea, hopefully providing a more firm rooting for future work. The next step, arguably the most important, is to interpret the results of these analyses in the context of other studies and datasets to understand why we observe these patterns and how this knowledge may be used in the future.

8.3.2 Improved local community detection methods

Local community detection methods have the potential to revolutionize the field of community detection, and network science in general, in the same way Newman's modularity originally did. There is also an obvious parallel to be seen with the development of geographically local techniques in quantitative geography and the effect that they have had on the field. There is a need for further research to continue the development of this area to explore alternative approaches and to improve the performance of existing methods. These improvements may come in the form of more efficient implementations of the existing algorithms or new algorithms that make better use of heuristics. The authors of recent methods such as OSLOM and Infomap are to be commended for following the lead of Newman in publishing the source code to to the software implementations of their algorithms. It is imperative, however, that we in the research community build upon these further to improve both their performance and usability. Without such efforts it is likely that modularity based methods will remain the tools of choice for the vast majority of researchers and the discussions of limitations will continue to be acknowledged but unheeded for lack of reasonable alternatives.

8.4 Concluding remarks

In conclusion, the main finding of this thesis is that it is of utmost importance to consider the full context when conducting analysis on mobile communication networks. This includes the multi-scale spatial and temporal contexts in which the communications take place, the context in which the analysis methods were developed and the context in which the results are to be interpreted. If these are not considered carefully there is a high potential for a mismatch between the assumptions of the analyst and the theoretical assumptions underpinning the methods, leading to a misapplication of methods and misinterpretation of results. If the theory and context are carefully considered however, the tools of network science can offer fascinating insights into the structure of society.

Bibliography

Ahas, R., Aasa, A., Roose, A., Mark, Ã., & Silm, S. (2008). Evaluating passive mobile positioning data for tourism surveys: An estonian case study. *Tourism Management*, *29*(3), 469 – 486.

Barthélemy, M. (2011). Spatial networks. *Physics Reports*, *499*(1), 1–101.

Blondel, V., Guillaume, J., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*, P10008.

Blondel, V., Krings, G., & Thomas, I. (2010). Regions and borders of mobile telephony in Belgium and in the Brussels metropolitan zone. *Brussels Studies*, *42*.

Borgatti, S., Mehra, A., Brass, D., & Labianca, G. (2009). Network analysis in the social sciences. *science*, *323*(5916), 892–895.

Butts, C. (2009). Revisiting the foundations of network analysis. *Science*, *325*(5939), 414–416.

Cairncross, F. (1997). The death of distance. *Harvard Business School Press, Cambridge*.

Calabrese, F., Dahlem, D., Gerber, A., Paul, D., Chen, X., Rowland, J., Rath, C., & Ratti, C. (2011a). The connected states of america: Quantifying social radii of influence. In *Privacy, Security, Risk and Trust (PASSAT), 2011*

IEEE Third International Conference on and 2011 IEEE Third International Confernecee on Social Computing (SocialCom), (pp. 223–230). IEEE.

Calabrese, F., Lorenzo, G. D., Pereira, F., Liu, L., & Ratti, C. (2010). Analyzing cell-phone mobility and social events. In *Workshop on the Analysis of Mobile Phone Networks*, (pp. 59–61).

Calabrese, F., Smoreda, Z., Blondel, V. D., & Ratti, C. (2011b). Interplay between telecommunications and face-to-face interactions: A study using mobile phone data. *PLoS ONE*, *6*(7), e20814.

Carey, H. (1858). Principle of social science. *History of Economic Thought Books*.

Cariou, C., Ziemlicki, C., & Smoreda, Z. (2010). Paris by night. In *Workshop on the Analysis of Mobile Phone Networks*, (pp. 62–66).

Commission for Communications Regulation (2011). Irish communications market quarterly key data report. Report.

Coombes, M. (2000). Defining locality boundaries with synthetic data. *Environment and planning A*, *32*(8), 1499–1518.

Coombes, M., Green, A., & Openshaw, S. (1986). An efficient algorithm to generate official statistical reporting areas: the case of the 1984 travel-to-work areas revision in britain. *Journal of the operational research society*, *37*(10), 943–953.

Corcoran, M. (2004). Place re-making in dublin. In M. Peillon, & M. Corcoran (Eds.) *Place and non-place: the reconfiguration of Ireland*, vol. 4, (pp. 142–156). Institute of Public Administration.

Dodd, S. (1950). The interactance hypothesis: a gravity model fitting physical masses and human groups. *American Sociological Review*, (pp. 245–256).

Dunbar, R. (1993). Coevolution of neocortical size, group size and language in humans. *Behavioral and brain sciences*, *16*(4), 681–693.

Durkheim, E., Spaulding, J., & Simpson, G. (1997). *Suicide*. Free Press.

Eagle, N., & Pentland, A. (2006). Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, *10*(4), 268.

Eagle, N., & Pentland, A. (2009). Eigenbehaviors: Identifying structure in routine. *Behavioral ecology and sociobiology*, *63*(7), 1057–1066.

Expert, P., Evans, T., Blondel, V., & Lambiotte, R. (2011). Uncovering space-independent communities in spatial networks. *Proceedings of the National Academy of Sciences*, *108*(19), 7663.

Farmer, C. J., & Fotheringham, A. S. (2011). Network-based functional regions. *Environment and Planning A*, *43*(11), 2723–2741.

Farmer, C. J. Q. (2011). *Commuting flows & local labour markets: Spatial interaction modelling of travel-to-work*. Ph.D. thesis, National University of Ireland Maynooth.

Flórez-Revuelta, F., Casado-Díaz, J. M., & Martínez-Bernabeu, L. (2008). An evolutionary approach to the delineation of functional areas based on travel-to-work flows. *International Journal of Automation and Computing*, *5*(1), 10–21.

Flowerdew, R., & Aitkin, M. (1982). A method of fitting the gravity model based on the poisson distribution. *Journal of regional science*, *22*(2), 191.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, *486*(3-5), 75–174.

Fortunato, S., & Barthelemy, M. (2007). Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, *104*(1), 36.

Fotheringham, A. (1983). A new set of spatial-interaction models: the theory of competing destinations. *Environment and Planning A*, *15*(1), 15–36.

Fotheringham, A. (1997). Trends in quantitative methods i: stressing the local. *Progress in Human Geography*, *21*(1), 88–96.

Fotheringham, A., Charlton, M., & Brunsdon, C. (2000). *Quantitative geography: perspectives on spatial data analysis*. Sage Publications Limited.

Fotheringham, A. S., & Curtis, A. (1992). Encoding spatial information: The evidence for hierarchical processing. In *Theories and methods of spatio-temporal reasoning in geographic space*, (pp. 269–287). Springer.

Frias-Martinez, E., Williamson, G., & Frias-Martinez, V. (2011). An agent-based model of epidemic spread using human mobility and social network information. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, (pp. 57–64). IEEE.

Girvan, M., & Newman, M. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, *99*(12), 7821.

González, M., Hidalgo, C., & Barabási, A. (2008). Understanding individual human mobility patterns. *Nature*, *453*(7196), 779–782.

Good, B., de Montjoye, Y., & Clauset, A. (2010). Performance of modularity maximization in practical contexts. *Physical Review E*, *81*(4), 046106.

Greene, D., Doyle, D., & Cunningham, P. (2010). Tracking the evolution of communities in dynamic social networks. In *Proc. International Conference on Advances in Social Networks Analysis and Mining (ASONAM'10).*

Guldmann, J.-M. (2004). Spatial interaction models of international telecommunication flows. *Best Practices in Spatially Integrated Social Science*, (pp. 400–419).

Haynes, K., & Fotheringham, A. (1984). *Gravity and spatial interaction models*. vol. 2. Sage publications Beverly Hills, CA.

Huff, D. (1959). Geographical aspects of consumer behavior. *University of Washington Business Review*, *18*, 27–35.

Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Martonosi, M., Rowland, J., & Varshavsky, A. (2011). Identifying important places in people's lives from cellular network data. *Pervasive Computing*, (pp. 133–151).

Kang, C., Zhang, Y., Ma, X., & Liu, Y. (2012). Inferring properties and revealing geographical impacts of intercity mobile communication network of china using a subnet data set. *International Journal of Geographical Information Science*, *0*(0), 1–18.

Kernighan, B., & Lin, S. (1970). An eficient heuristic procedure for partitioning graphs. *Bell System Technical Journal*.

Krings, G., Calabrese, F., Ratti, C., & Blondel, V. (2009). Urban gravity: a model for inter-city telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment*, *2009*, L07003.

Lambiotte, R., Blondel, V., De Kerchove, C., Huens, E., Prieur, C., Smoreda, Z., & Van Dooren, P. (2008). Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications*, *387*(21), 5317–5325.

Lancichinetti, A., Radicchi, F., Ramasco, J., & Fortunato, S. (2011). Finding statistically significant communities in networks. *PloS one*, *6*(4), e18961.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., & Van Alstyne, M. (2009). Computational social science. *Science*, *323*(5915), 721–723.

Lloyd, C. (2010). *Local models for spatial analysis*. CRC Press.

Massen, C., & Doye, J. (2005). Identifying communities within energy landscapes. *Physical Review E*, *71*(4), 046101.

Masser, I., & Brown, P. (1975). Hierarchical aggregation procedures for interaction data. *Environment and Planning A*, *7*(5), 509–523.

Mc Fadden, D. (1973). Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics*, (pp. 105–142).

Meunier, D., Lambiotte, R., & Bullmore, E. (2010). Modular and hierarchically modular organization of brain networks. *Frontiers in neuroscience*, *4*.

Nadel, S. F. (1957). *Theory of Social Structure*. Cohen and West.

Newman, M. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, *98*(2), 404–409.

Newman, M. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, *103*(23), 8577.

Newman, M. (2012). Communities, modules and large-scale structure in networks. *Nature Physics*, *8*(1), 25–31.

Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E*, *69*, 026113.

Onnela, J., Arbesman, S., González, M., Barabási, A., & Christakis, N. (2011). Geographic constraints on social network groups. *PLoS one*, *6*(4), e16939.

Openshaw, S. (1983). *The modifiable areal unit problem.* vol. 38. Geo Books Norwich.

Openshaw, S., Charlton, M., Wymer, C., & Craft, A. (1987). A mark 1 geographical analysis machine for the automated analysis of point data sets. *International Journal of Geographical Information System*, *1*(4), 335–358.

Palla, G., Barabasi, A., & Vicsek, T. (2007). Quantifying social group evolution. *Nature*, *446*(7136), 664–667.

Palla, G., Derényi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, *435*(7043), 814–818.

Papps, K. L., & Newell, J. O. (2002). Identifying functional labour market areas in new zealand: A reconnaissance study using travel-to-work data. Tech. rep., Institute for the Study of Labor (IZA).

Porter, M., Mucha, P., Newman, M., & Friend, A. (2007). Community structure in the united states house of representatives. *Physica A: Statistical Mechanics and its Applications*, *386*(1), 414–438.

Porter, M., Onnela, J., & Mucha, P. (2009). Communities in networks. *Notices of the AMS*, *56*(9), 1082–1097.

Ratti, C., Sobolevsky, S., Calabrese, F., Andris, C., Reades, J., Martino, M., Claxton, R., & Strogatz, S. H. (2010). Redrawing the map of great britain from a network of human interactions. *PLoS ONE*, *5*(12), e14248.

Ravenstein, E. (1885). The laws of migration. *Journal of the Statistical Society of London*, *48*(2), 167–235.

Reades, J., Calabrese, F., & Ratti, C. (2009). Eigenplaces: analysing cities using the space- time structure of the mobile phone network. *Environment and Planning B: Planning and Design*, *36*(5), 824–836.

Reades, J., Calabrese, F., Sevtsuk, A., & Ratti, C. (2007). Cellular census: Explorations in urban data collection. *Pervasive Computing, IEEE*, *6*(3), 30–38.

Rosvall, M., & Bergstrom, C. (2010). Mapping change in large networks. *PLoS One*, *5*(1), e8694.

Rosvall, M., & Bergstrom, C. (2011). Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. *PloS one*, *6*(4), e18209.

Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, *105*(4), 1118–1123.

Roy, J., & Thill, J. (2003). Spatial interaction modelling. *Papers in Regional Science*, *83*(1), 339–361.

Simini, F., González, M., Maritan, A., & Barabási, A. (2012). A universal model for mobility and migration patterns. *Nature*.

Song, C., Qu, Z., Blumm, N., & Barabasi, A. (2010). Limits of predictability in human mobility. *Science*, *327*(5968), 1018.

Soto, V., & Frias-Martinez, E. (2011). Robust land use characterization of urban landscapes using cell phone data. In *Workshop on Pervasive Urban Applications*.

Soto, V., Frias-Martinez, V., Virseda, J., & Frias-Martinez, E. (2011). Prediction of socioeconomic levels using cell phone records. *User Modeling, Adaption and Personalization*, (pp. 377–388).

Sporns, O., Chialvo, D., Kaiser, M., Hilgetag, C., et al. (2004). Organization, development and function of complex brain networks. *Trends in cognitive sciences*, *8*(9), 418–425.

Stoicaa, A., Smoreda, Z., Prieur, C., & Guillaume, J.-L. (2010). Age, gender and communication networks. In *Workshop on the Analysis of Mobile Phone Networks*, (pp. 11–14).

Stouffer, S. (1940). Intervening opportunities: a theory relating mobility and distance. *American sociological review*, (pp. 845–867).

Stouffer, S. (1960). Intervening opportunities and competing migrants. *Journal of Regional Science*, *2*(1), 1–26.

Taylor, L. D. (1994). *Telecommunications demand in theory and practice*. Kluwer Academic Pub.

Tobler, W. (1970). A computer movie simulating urban growth in the detroit region. *Economic geography*, *46*, 234–240.

Traag, V., Browet, A., Calabrese, F., & Morlot, F. (2011). Social event detection in massive mobile phone data using probabilistic location inference. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, (pp. 625–628). IEEE.

Traud, A., Kelsic, E., Mucha, P., & Porter, M. (2011). Comparing community structure to characteristics in online collegiate social networks. *SIAM review*, *53*(3), 526–543.

Walsh, F., & Pozdnoukhov, A. (2011). Spatial structure and dynamics of urban communities. The First Workshop on Pervasive Urban Applications (PURBA).

Wilkinson, D., & Huberman, B. (2004). A method for finding communities of related genes. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(Suppl 1), 5241–5248.

Wilson, A. (1967). A statistical theory of spatial distribution models. *Transportation Research*, *1*(3), 253–269.

Wilson, A. (1970). *Entropy in urban and regional modelling*. Pion Ltd.

Wireless Intelligence (2010). Global mobile connections surpass 5 billion milestone. Report.

Zachary, W. (1977). An information flow model for conflict and fission in small groups. *Journal of anthropological research*, (pp. 452–473).

Zhao, Q., & Oliver, N. (2010). Communication motifs: A novel approach to characterize mobile communications. In *Workshop on the Analysis of Mobile Phone Networks*, (pp. 33–35).

Zipf, G. (1949). *Human behavior and the principle of least effort: an introduction to human ecology*. Addison-Wesley Press.

Appendix A

Reference maps of Ireland

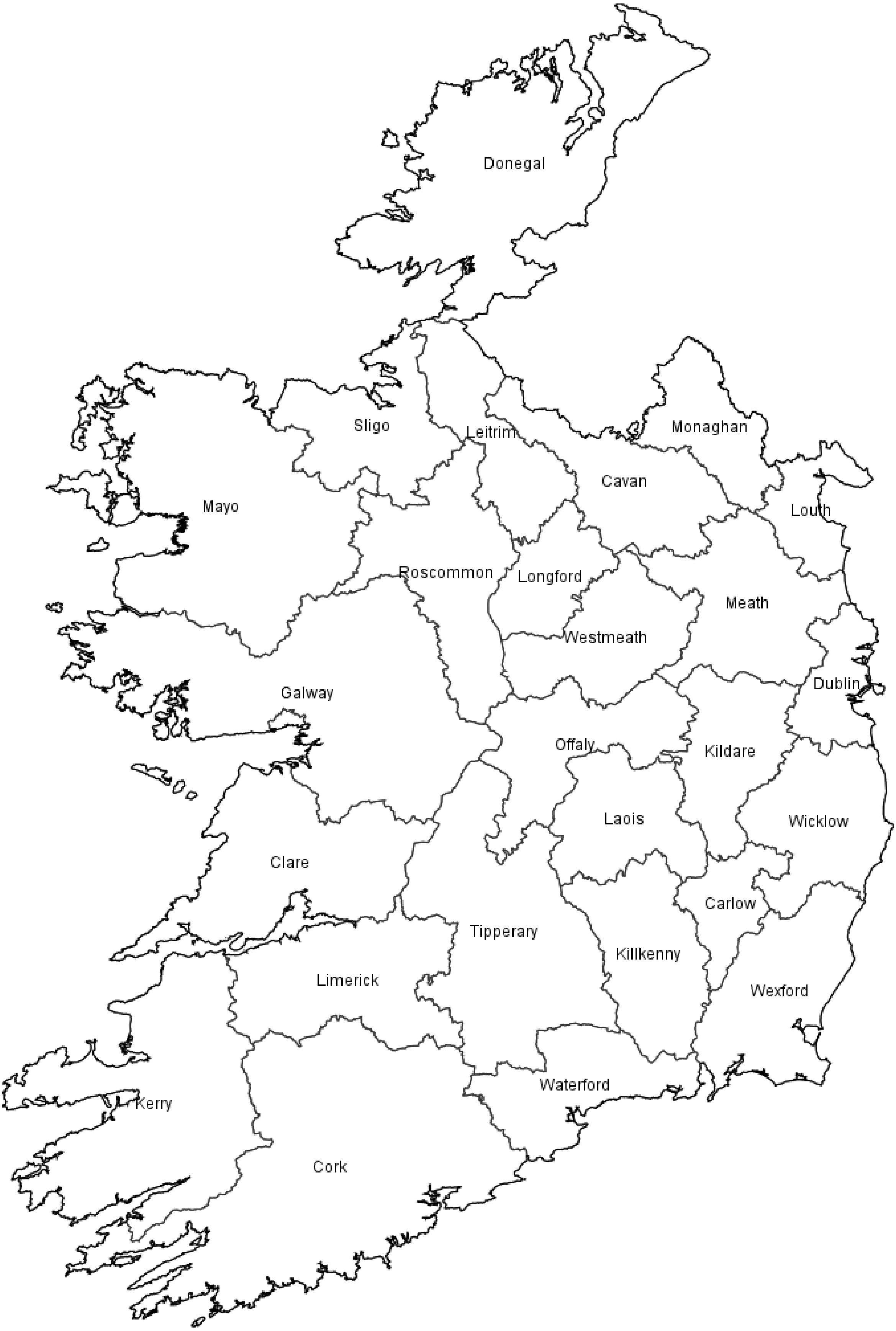


Figure A.1: The traditional 26 counties of the Republic of Ireland.

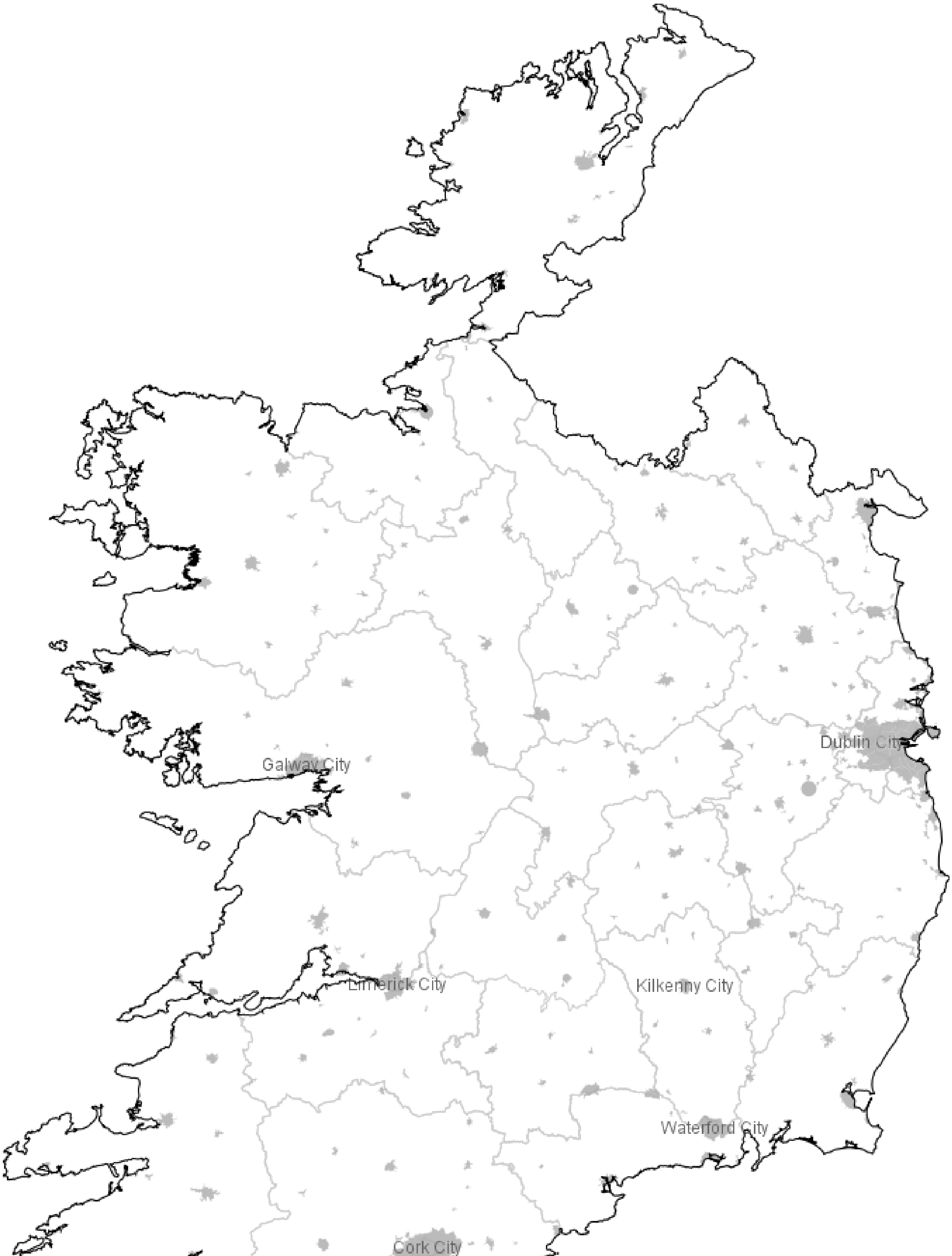


Figure A.2: Towns and cities with populations greater than 1000 people.