Universidad Politécnica de Madrid

Escuela Técnica Superior de Ingenieros de Telecomunicación

Socio geographical patterns inferred from mobile phone records

Tesis Doctoral

Carlos Herrera Yagüe
Ingeniero de Telecomunicación

2017

Departamento de Matemática Aplicada
a las Tecnologías de la Información y
las Comunicaciones (MAT)

Escuela Técnica Superior de Ingenieros de
Telecomunicación

Universidad Politécnica de Madrid

Socio geographical patterns inferred from mobile phone records

Autor:
Carlos Herrera Yagüe
Ingeniero de Telecomunicación

Tutor:
Pedro J. Zufiria Zatarain
Doctor Ingeniero de Telecomunicación

Tribunal nombrado por el Magfco. y Excmo. Sr. Rector de la Universidad Politécnica de Madrid, el día ***.

Presidente: _____

Vocal: _____

Vocal: _____

Vocal: _____

Secretario: _____

Suplente: _____

Suplente: _____

Realizado el acto de defensa y lectura de la Tesis el día ***** en la E.T.S.I.T. habiendo obtenido la calificación de _____.

EL PRESIDENTE LOS VOCALES

EL SECRETARIO

Resumen

La ubicuidad de los registros de comunicación recolectados de forma automática y la dramática reducción de los costes asociados al almacenamiento y procesamiento de información, nos permiten comenzar a estudiar el comportamiento humano de una forma completamente nueva. En lugar de limitarnos a pequeños experimentos realizados a decenas o cientos de participantes durante lapsos de tiempo relativamente cortos, como ha ocurrido con la investigación en ciencias sociales o la planificación de transportes anterior a los últimos 15 años, hoy tenemos registros detallados de ciertos comportamientos para millones de personas durante años, con el interesante matiz de que los datos se recolectan de forma pasiva, sin requerir ninguna atención ni disciplina por parte de los participantes.

Esta abundancia de información obtenida de forma sistemática para gran cantidad de sujetos, nos permite abordar la comprensión y modelización del comportamiento humano, aplicando métodos hasta ahora reservados a la física y a otras ciencias naturales, más acostumbradas a tratar con datos masivos generados de forma sistemática. En el caso de la investigación presentada en esta tesis, nos centraremos en analizar los registros de comunicación y posición asociados a unos 7 mil millones de registros de llamadas (CDRs, por sus siglas en inglés) que representan todas las realizadas por más de 25 millones de personas durante un período de seis meses. El conjunto de datos analizados incluye información de tres países distintos (Francia, Portugal y España). Esto nos ha permitido garantizar cierta robustez de nuestros resultados frente a posibles sesgos de observación asociados a las comunicaciones móviles, como pudieran ser las políticas comerciales o las cuotas de mercado del operador que facilita los datos. Además, como mostraremos, nos ha permitido también apreciar diferencias macroscópicas significativas entre las tres redes, posiblemente asociadas a la historia e idiosincrasia de cada uno de los países.

Entre las múltiples posibilidades que ofrece el análisis de CDRs, en esta tesis nos hemos centrado en los problemas de compleción de red, así como en las relaciones que se establecen entre la red social y el espacio geográfico en la que esta se enmarca.

En cuanto a la compleción de la red, nos hemos centrado en el análisis de un escenario al que nos referimos como el problema de los nodos opacos.

Este problema considera redes con dos tipos de nodos: por un lado están los nodos *transparentes* de los que se conocen todos sus enlaces y ciertos atributos específicos del nodo. Por otro lado, están los nodos *opacos*, de los que solamente se conocen sus enlaces con los nodos transparentes. El problema consiste, pues, en tratar de inferir tanto los atributos de los nodos opacos como los enlaces entre ellos. Nuestro trabajo demuestra que, aprovechando propiedades conocidas de las redes sociales y herramientas del aprendizaje estadístico, es posible conseguir predicciones sorprendentemente acertadas incluso si la población de nodos opacos supera con creces la mitad de la red. Estos resultados tienen especial relevancia en el ámbito de los operadores de telecomunicaciones, ya que demuestran que tienen una capacidad significativa para inferir información sobre usuarios que no son, ni nunca han sido, sus clientes. Asimismo, estos resultados cuestionan la idoneidad de las herramientas de gestión de la privacidad que las grandes plataformas online como Facebook, Twitter o Google han puesto a disposición de sus usuarios, que consideran que para dar a conocer un enlace social (que represente amistad, seguimiento o comunicación) entre dos usuarios, es suficiente conseguir permiso explícito de solamente uno de ellos.

A la hora de considerar la relación entre las redes sociales y el espacio geográfico que ocupan, nos hemos centrado primero en mejorar la comprensión de los resultados de uno de los experimentos más famosos del siglo XX: el experimento de Milgram o de los seis grados de separación. Tras realizar una revisión exhaustiva de los trabajos publicados sobre búsqueda descentralizada en redes sociales, tanto desde un marco de modelos teóricos como de reediciones del experimento de mundo pequeño, presentamos lo que hasta la fecha supone la simulación más grande realizada sobre datos empíricos de redes sociales. Nuestros resultados respaldan por primera vez, con datos reales, algunas de las hipótesis más relevantes sobre cuál es la estructura de la red social que permite que la búsqueda descentralizada sea efectiva. Concretamente, nuestros resultados demuestran que la cercanía geográfica es una medida muy efectiva a la hora de orientar la ruta de un mensaje en los primeros pasos; su efectividad desaparece de forma muy abrupta una vez que el mensaje llega a la ciudad del destinatario, casi independientemente del número de habitantes de esta ciudad. Sin embargo, el rutado descentralizado dentro de ciudades sigue siendo posible utilizando la estructura de comunidades de la red social.

Los resultados sobre búsqueda descentralizada nos permiten indagar más sobre una relación hasta ahora desconocida entre la red social y el espacio geográfico. Concretamente, encontramos que las comunidades detectadas algorítmicamente por optimización de modularidad pierden en gran medida la correlación espacial dentro de las ciudades. Asimismo, mostramos cómo las redes formadas por los habitantes de una parte de la ciudad geográficamente conectada, pierden su conectividad comparadas con redes del mismo tamaño en número de nodos, pero que contienen al menos un núcleo de población

completo.

Estos análisis nos han permitido, además, establecer un paralelismo entre los flujos de comunicación y de transporte: ambos decrecen con la distancia física de forma similar. Aprovechando esta similitud, presentamos dos modificaciones del modelo de radiación. En la modificación orientada a flujos de comunicación, garantizamos la simetría de las predicciones, considerando, en el denominador, la población dentro de elipses cuyos focos están en las ciudades cuyo flujo de comunicación tratamos de estimar. En el caso de flujos de transporte casa-trabajo, modelamos la capacidad de atracción de una zona como proporcional al número de negocios en el área, listados en aplicaciones como Google Places y Foursquare. Ambos modelos consiguen predicciones significativamente mejores que los modelos usados anteriormente, y tienen la ventaja adicional de no requerir de datos de entrenamiento para estimar parámetros del modelo.

Por último, nos centramos en analizar la correlación entre los patrones de movilidad urbana de personas más o menos cercanas en la red social. Encontramos correlaciones espaciales significativas, incluso entre nodos situados a distancia 3 dentro de la red social, siendo esta correlación espacial mayor cuanto mayor es la cercanía entre los nodos en el grafo social. Además, utilizando técnicas de aprendizaje no supervisado, encontramos que las relaciones entre habitantes de una misma ciudad se agrupan en 3 clases diferenciadas según la intensidad de la comunicación y los momentos en los que se producen eventos de colocalización. Finalmente, presentamos un modelo dual para la construcción de la red social y para la exploración de la ciudad, que permite reproducir buena parte de las correlaciones y distribuciones encontradas en los datos.

Abstract

The ubiquity of passively collected communication records and the dramatic cost reduction experienced by the fields of information storage and processing allow us to study human behaviour in a entirely different way. Human behaviour studies have been often limited to small experiments done with tens or hundreds of participants during relatively short time lapses, for example with social science research or transportation research prior to the last 15 years. Today, we have detailed electronic records for certain behaviours of millions of people during years. Additionally, most of this records have been passively collected, which means no especial attention or routine was followed by the participants in order to collect the data.

This abundance of systematically collected information allows us to face the understanding and modelling of human behaviour applying methods used until now only by physics and other natural sciences, more used to deal with massive amounts of systematically generated data. In the research we are presenting in the thesis, we have focused on analyzing communication and location records associated to 7 billion call detail records (CDRs) that contain all calls placed between over 25 million people during a 6 month period. The data set includes information from three different countries (France, Portugal and Spain), which allowed us to guarantee a certain level of robustness in our results against possible observations biases associated to mobile communications, such as pricing policies and market share of the carrier that facilitates the data. Also we will show how we have been able to spot macroscopic differences between the three networks, possibly related to the history and culture of each of the countries.

Among the multiple fields CDRs analysis can be useful for, in this thesis we have focused in network completion problems, and also in the relations that can be established between the social network and the geographical space where it is embedded.

Regarding network completion, we have focused in the analysis of an scenario that we refer as the opaque node problem. This problem considers networks with two different kind of nodes: on the one hand there are *transparent* nodes, about which we know all of their links and their attributes. On the other hand we have the *opaque* nodes, about which we only know how do they are connected to transparent nodes. The problem consists of trying

to infer both the attributes of opaque nodes and links between them. Our work shows that taking advantage of known properties of social networks and machine learning procedures it is possible make good predictions even if the proportion of opaque nodes is over 50% of the network. These results are specially relevant for mobile carriers, since they reveal that this companies have a significant capacity to infer information about users who are not and never have been their customers. Similarly, these results question the suitability of privacy management tools embedded into large online sites created by Facebook, Twitter or Google, that assume that to disclose to a third party the existence of a social link (whether it represents friendship or communication) it is enough to get permissions by just one of the users involved.

When it comes to the relationship between social network and the geographical space this are embedded into, we have first focused in understanding the results of one of the most famous experiments of the twentieth century: Milgram's small world experiment. First, we have exhaustively review all related work about decentralized search in social networks, both from a theoretical modelling perspective and those reproducing similar experiments in order to gather additional empirical insights. Then, we have run the largest decentralized search simulation based on real social network data published to date. Our results support, for the first time empirically, some of the most relevant hypothesis about what is the network structure that allows decentralized search to be efficient in social networks. Precisely, some of our results proof that geographical proximity is a good metric to route the messages in the firsts steps, but its effectiveness vanishes one the message reaches the target city, almost independently of the number of people living in such city. However, decentralized routing within cities is still possible leveraging the community structure of the social network.

The results about decentralized search in social networks allow us to dig deeper about the physical structure of social networks in urban environments. Precisely, we find that algorithmically detected communities, obtained through modularity optimization methods, lose almost all of their spatial correlation within cities. Additionally, we show that the networks made of the inhabitants of a certain connected area of the city have very limited connectivity compared to networks with the same number of nodes but that contain at least one complete population nucleus.

These results have also allowed us to establish a parallelism between communication and transportation fluxes: they both decrease with distance in a similar fashion. Leveraging such similarity, we present two extensions to the radiation model. In the extension oriented to communication fluxes, we ensure the symmetry of the predictions, considering in the denominator, the population that lives within certain ellipses whose foci are located in the cities whose flux we try to estimate. Regarding commuting flows, we model attraction of an area like proportional to the number of business in

the area listed by applications like Google Places and Foursquare. Both models performs remarkably better than their previous counterparts, and have the additional advantage of not requiring training data to fit model parameters.

Finally, we focus on analyzing the similarity between human mobility patterns of people depending on how close they are in the social network. We find that socially closer people have similar visitation patterns within the urban environment, and that this positive correlation holds true even up to a social distance of 3 hops. Additionally, using unsupervised learning techniques, we find that the relationship between people living within same city naturally cluster into 3 different groups, depending on what time of the week the co-locate. At last, we present a simple model for social network and city exploration that can reproduce a large portion of the behaviours found in the data.

Contents

Chapter 1

Introduction

By 2016, 4.43 billion people use a mobile phone on a daily basics[1]. According to United Nations data, that is around 85% of the worldwide population aged over 14. Every time each of those people places or receives a call or a text message, the geographic position of the user and the number dialed are recorded. The main objective of this thesis has been to learn how to extract meaningful information from such data, focusing on finding social, geographical and mobility patterns. To do so, we have used a network perspective.

In the last decade, networks have become part of our daily life, or at least we have started to be aware of them. The rising of the Internet has set up a new framework. The pieces of technology that build up the Internet (fiber optics, routers, computers) are basically useless by themselves. As many other complex systems, the Internet is better understood by trying to understand how different parts interact with each other than by carefully studying how each individual part works. Additionally, the appearance of online social networks such as Facebook or Twitter has made explicit non-trivial social structures. Those networks were there before, Facebook and Twitter simply unveiled them. In the network science community, we use the term *social network* beyond online social sites: we refer as a social network to any graph representation whose nodes represent people. In fact, the edges of a social network might represent interactions different from friendship. For example, two of the first social networks ever analyzed were mapped using co-authorship relationship in scientific papers and co-appearances of Hollywood actors in movies. In this dissertation, we will focus on networks that emerge from the aforementioned mobile phone traces, thus the links will represent communication between people.

[1] Source `http://www.statista.com/statistics/274774/forecast-of-mobile-phone-users-worldwide/`

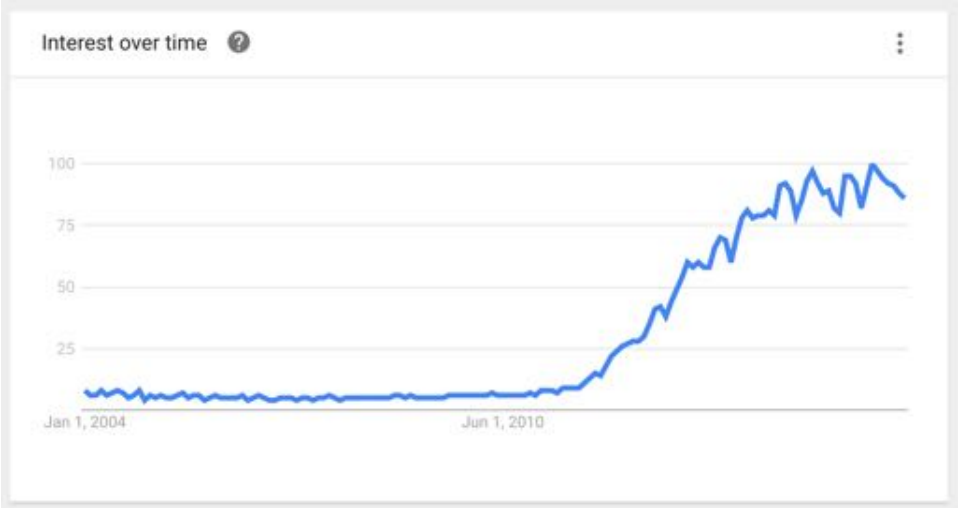


Figure 1.1: Evolution of the search volume of the keyword "Big Data" according to Google Trends.

1.1 Changes of the Big Data era

Figure 1.1 presents an snapshot of Google Trends, an application that allows to track the evolution of the number of the search queries made at Google for the specific search term. In that snapshot, it presents the time evolution of searches including the string "Big data". It is straightforward to notice that term was virtually non-existent until 2011 (roughly the beginning of this PhD) when it exploded and became mainstream.

One of the challenges around Big Data is the definition itself. Oxford English Dictionary defines it as "data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges". Similarly, Wikipedia defines Big Data as "an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using on-hand data management tools or traditional data processing applications". Another similar definition provided in a widely-quoted McKinsey study[2] reads "datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze".

All these definitions agree in highlighting that the size of the dataset is a key aspect for considering some analysis to be "Big Data". However, they fail at specifying which is the critical size so that "little data" turns into Big Data. A common definition used in the data science community refers to "datasets who do not fit into the memory of commodity computers". This is an interesting precision, because it rules out the most commonly used

[2] `http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation`

software for data analysis, like Excel spreadsheets. However, this definition is prone to change over time according to Moore's Law.

Another insight is provided by industry leaders. They have proposed the $5V$ mnemonic to highlight volume, velocity, variety, veracity and value as key ingredients of Big Data. Although the latter two are probably difficult to asses, it is measurable that the first three[3] have definitely motivated the development of new IT pieces of infrastructure:

- Dealing with datasets of high volume (for example, over 100GB, although different authors propose different figures) has motivated the rise of distributed computing, so that a presumably huge task can be split into several (in some cases, even thousands) of computers that work in parallel. In this regard, it is worth noting Google's MapReduce paradigm, with very popular open-source implementations such as Hadoop and Spark.

- By 2010, Facebook API was answering already up to 13 million queries per second. For such velocity, hard drives are simply too slow. Therefore in the last 10 years we have witnessed the rise of in-memory databases, like Memached or Redis that provide 10x to 100x improvement in throughput, at the cost of non-persistence.

- Also motivated by volume, but even more by the variety of the stored data has been the paradigm shift in databases. While from the 1970s relational databases had been the norm across all industries, lately NoSQL databases with more flexible data schemes such as MongoDB, Couch or Cassandra have gained significant market share.

Even if it is difficult to find a unique definition of Big Data, what is clear is that something has occurred in the last 5 years. By 2011, only two IT companies (Apple and Microsoft) were among the 5 most valuable companies in the New York Stock Exchange. By August 1st, 2016, the top 5 is made only of tech companies. Google, Facebook and Amazon, all three of them companies whose main competitive advantage is based on user-generated data, have made it to the top. By February 2015, President Obama appointed a Chief Data Scientist for the first time in US history.

Is it possible that data has been able to generate so much value so fast? Definitely, there is no shortage of the "raw material". According to Google's former CEO Eric Schmidt [137], every two days mankind generates more data than in the entire history up to 2003. The implications of this immense abundance have impacted beyond the tools needed to cheaply process and store large volumes of data. Precisely, it has allowed a new path for humans

[3]These 3Vs are actually the original ones proposed by Doug Laney as early as 2001, in his research note [83].

to learn about themselves. Most social science research in history has first defined a research question, then designed an experiment to collect enough data to answer the question and finally run the experiment and analyze the results. However, the abundance of data has allowed a different approach. The approach is based on assuming that for anything that we want to answer about human behaviour, the data needed to answer the question is stored somewhere already in digital form. There are billions of devices out there generating data about human behaviour. According to Albert-László Barabási in his opening keynote at the Network Science conference 2010,

> ...for the first time in history we can study humans using the same models and tools we have used to study planets, the iPhones of the world can be our telescopes.

This new form of expanding human knowledge has been named *data science*. The person who coined the term was precisely Dhanurjay "DJ" Patil, the first Chief Data Scientist appointed by President Obama. Data science practitioners are referred to as *data scientists*. In 2012, Harvard Business Review named data scientist as the sexiest job of this century[4]. However, this paradigm shift has its shortcomings, too. For example, the data we will use during this dissertation was collected to correctly bill customers and detect failing cells. We will be using it to answer questions about the structure of social relationships. This means there will always be biases and errors in the data that cannot be assessed by experimental design, as social scientists used to do. Instead, we will need to identify and eliminate misdialed numbers and marketing calls, filter out calls without enough information and perform some other tasks that can become tedious and that are typically described as "cleaning" the data. Indeed, in 2014 Harvard Business Review published another article under the title "The Sexiest Job of the 21st Century is Tedious, and that Needs to Change"[5].

1.2 A brief introduction to complex networks

A complex system is defined as one whose parts are interconnected or interrelated, and these relationships add additional information that was previously non-observable. As a result of these relationships between elements, new properties appear that cannot explained using the isolated parts. These are known as emergent properties.

This idea of a complex system as one where "the whole cannot be explained in terms of the behavior of the parts" is typically exemplified by

[4] `https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century`

[5] `https://hbr.org/2014/04/the-sexiest-job-of-the-21st-century-is-tedious-and-that-needs-to-change`

Figure 1.2: Two classic examples of complex systems, the human brain and the ants colonies.

systems such as those in Figure 1.2. The first is the community formed by the ants that build structures such as the one in the figure. These structures, additionally to their large size compared to that of the builders, present features such as very efficient temperature control through air streams in the tunnels built by the ants. This efficiency is so high that has been recently studied by experts in architecture so its design principles can be applied to our homes and offices. In any case, it is obvious that the building of such a complex system cannot be explained through an isolated ant, no matter how deep is our knowledge of the ant anatomy and metabolism. In a similar way, while neuroscience has been able to describe the process of synapses, it is still far from explaining how the interaction of very simple elements like neurons can produce high-level phenomena like thoughts or memory.

Network science tries to create new tools for the study of complex systems, because its focus is to model the relationship between the elements in the system. Since complex systems are precisely characterized by those interactions, network science is considered the architecture of the complex systems. The rise of network science is related to the finding, during the last decade, of several emergent properties that are in common for networks of very different kinds. In [118], Mark Newman compiles a number of these features that are found in social, biological, informational or biological networks, even when the only common feature among the systems is that they were *self-organized*. Self-organization is used to refer to networks that either were not designed by humans (e.g. protein interaction network) or lacked of central planning in its development. For example, we can consider the Internet a self-organized network because different Internet Service Providers

(ISP), universities and institutions added IP devices without thinking beyond their local needs. However, the Interstate Highway System in the US is not self-organized, because the government decides where to expand the system in a centralized manner.

Network science is related to a branch of mathematics named graph theory. Indeed, one of the first university courses on Network Science, taught by Prof. Hidalgo in 2011, started with the classic example of Euler's Konigsberg bridges, as many graph theory courses. It is true that network science and graph theory share a common subjects of study, although the both fields commonly refer to them in slightly different terms. Network, node and link; as opposed to graph, vertex and edge are used by network science and graph theory respectively. Also, the common curriculum of courses of network science and graph theory does not go beyond the second lecture. Born in an era of big data and complexity, network science focuses on a statistical analysis of the graph.

In this section, we will refer some of the seminal contributions of network science that we will be mentioning all along this dissertation, whereas other contributions more specific to each of the problem addressed in the chapters will be referred when describing the problem itself.

1.2.1 Small worlds

One of the first large social networks ever mapped represented co-appearances of actors in movies according to the Internet Movie Data Base (IMDB) [7]. In this network, nodes represent actors who are connected only if they performed together in at least 1 movie. For example, there is a link between Jack Nicholson and Tom Cruise, because they both appeared in *A few good men* in 1992. However, there is no link between Tom Cruise and Samuel L Jackson, because they have not acted in a single movie together, but a very short path can indeed be found between the two of them. Cruise appeared with Maureen Mendoza in *Born on the fourth of July* who also worked with Jackson in *The negotiator*. This means that within the IMDB network, Cruise and Jackson are only two links away from each other. When analyzing the entire graph, with almost 500,000 actors, each them connected on average with around 113 colleagues, one realises that very short paths exist between almost everyone in the network. In fact, the average path length between two nodes chosen uniformly at random is as little as 3.48. The reader is invited to verify this by finding the shortest paths between his favourite actors at the Oracle of Bacon[6], a gem of the early days of the web, but still updated, up and running.

The existence of very short paths despite large network sizes is one of the most common features of real world networks. In [118], Newman presents

[6] `http://oracleofbacon.org`

dozens of examples, some of them as remarkable as the network between a sample of 200 million web pages indexed by Altavista, whose average path length (typically referred to as *diameter*[7]) is only 16. Interestingly, while the mere existence short paths in social networks has caught the attention of the general public, that commonly refers to it as *the six degrees of separation* theory, it is not very remarkable by itself from a modelling point of view. Indeed, the most simple random graph, proposed by Erdos and Renyi [36], that simply connects nodes uniformly at random, presents a diameter that grows only logarithmically with the number of nodes, therefore presenting short diameters for large networks. This is straightforward to comprehend considering an Erdos network where everyone is connected with an average of 100 other nodes. In such network, from any node, it is possible to reach an average of 10,000 nodes in two steps, around a million in three steps and so on.

The role of triangles

Therefore, we now have a simple network model, the Erdos graph, that successfully reproduce short paths. However, is it a realistic modeling of social networks? We invite the reader to run the following experiment. Open your phone log, and take the last 4 people you talked to. Then, try to estimate the fraction of people you have in your phone book that at least one of them has too. Chances the are that this fraction is not negligible, and probably it is as high as 80%. This means that the previous reasoning "my 100 friends can introduce me to 100 new people each, reaching a total of 10,000 friends of friends" is simply not realistic, because empirical evidence shows that there are significant overlaps between friends and friend of friends. This overlap is typically referred to as *clustering* and can be measured as the fraction of closed triangles compared to open triangles in a network (clustering coefficient) or around a certain node (local clustering coefficient).

In 1998, Watts and Strogatz [161] focused on the relationship between clustering coefficient and diameter in networks. To do so, they studied lattices similar to the one presented in Figure 1.3. These are highly clustered networks. However, if we increase the number of nodes N in the lattice, the diameter will grow linearly with N, so paths will not be short.

To reconcile clustering and short paths, Watts and Strogatz proposed a model where, starting from the lattice, a fraction of p the links are rewired choosing nodes uniformly at random, such that if $p = 1$ an Erdos graph is built. In [161], they showed that there is a wide region of p values where

[7]In graph theory, the diameter of a graph is the maximum path length between any two nodes. However, in complex networks literature the term *diameter* is used for the average path length, mostly due to the difficulty of measuring the maximum for large data sets. In this dissertation we will use *diameter* meaning the average path length

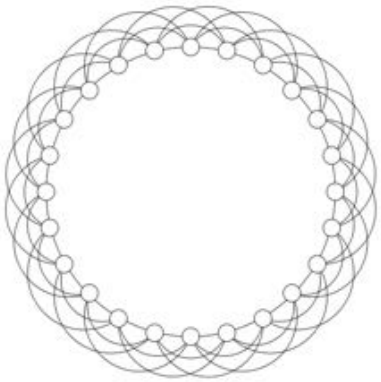


Figure 1.3: One dimensional lattice. All nodes are connected to their 4 closest neighbors.

the resulting network presents both high clustering and short paths at once. These networks are known as *small worlds*.

1.2.2 Community structure

Once we have learnt that there is an abundance of triangles in social network, a natural question arises: do these triangles tend to overlap with each other, or do they appear uniformly in the graph? One of the advantages of studying social networks is that we have been our entire lifetime living in them, so we can use our experience to formulate hypothesis that we will later validate with data. In this case, let us consider the social graph defined by the messages we exchange over Whatsapp. Whatsapp is a mobile messaging application that by the time of editing this document has over 2 billion active users. In there, users can text to each other privately or in closed groups, who users can only join if they are invited by a member of the group.

Among the 2 billion users in Whatsapp, we only talk to a few hundred of them at most, so the network is indeed very sparse, with only 1 link existing out of ten of millions possible links. However, the existence of group chats in Whatsapp highlights that there are very dense areas in the social graph. While we have not found any published results about the typical sizes of Whatsapp groups, the initial version of the app allowed for only up to 8 people in the groups, and was perceived as a major limitation by users. Even the last upgrade, from raising that limit from 100 to 256 people, was much celebrated by the users, according to specialized media[8]. This implies that the network has very dense areas with even 100 users forming a fully connected graph, also known as a *clique*.

[8] `http://www.independent.co.uk/life-style/gadgets-and-tech/news/whatsapp-group-chats-bigger-maximum-size-256-people-users-a6856491.html`

The most widely used metric to evaluate how clustered groups of nodes in a network are is referred to as *modularity*. Modularity is defined for groups of nodes that are defined so that every node belongs to only one group. It was defined by Newman and is defined as

$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - \frac{k_i k_j}{2m}) \delta(c_i, c_j) \tag{1.1}$$

where A represent the adjacency matrix, m is the total number of links, k_i is the degree of node i and $\delta(c_i, c_j) = 1$ if i and j belong to the same group, being 0 otherwise. Modularity ranges between -1 and 1, with positive value meaning there are most links within the groups than in a randomization of the graph where every link would be rewired.

Multilevel aggregation method

The multilevel aggregation method, more commonly known as the Louvain method, due to the affiliation of its authors by the time [18], tries to optimize modularity by hierarchical aggregation of small communities. While it is very difficult to find a partition of the network that optimizes modularity globally for large networks, the Louvain method is capable to provide very good communities in terms of modularity in relatively short computation times.

This method gathered a lot of recognition after its application in Belgium mobile phone network, with 2.6 million nodes. The algorithm was able to detect two large super-communities defined by the two native languages of the country: Dutch and French.

Clique percolation method

The Click Percolation Method (CPM) was presented by Palla et al in [123]. It tries to find communities by growing cliques of a certain size k. More particularly, it focuses on recursively joining k-cliques into communities if they overlap in at least $k - 1$ nodes.

The most relevant characteristic of CPM is that the resulting communities might overlap each other. This is very important in some contexts in social networks, because there are people who belong to several communities (the authors refer to them as *pivots*) and those have a important role in some phenomena, such as information diffusion.

Technically, computation requirements for CPM are higher compared to Louvain, so its application is typically finding communities in local areas of the networks, such as ego-networks, as opposed to Louvain, that is more commonly applied to the entire dataset.

1.3 Dissertation outline

This dissertation is structured as it follows. In Chapter 2 we define a specific partial information problem that we refer as the opaque node problem and we leverage emergent network properties to make predictions about missing information in a social network extracted from phone data. In Chapter 3 we address the searchability of social networks, analyzing the results of different decentralized algorithms using phone data from three different countries, finding previously unreported anomalous behaviours for urban networks. In Chapter 4, we analyze the relationship between social networks and geography at different scale that explain the behaviours exhibited in Chapter 3. In Chapter 5, we leverage the similarity between human communication and mobility to develop two new parameter-free models that improve performance at different scales. In Chapter 6, we analyze the relationship between human mobility and social networks at the urban scale, introducing a new model for individual mobility that accounts for social influence in human mobility.

Chapter 2

The network completion problem

On May 2, 2011, around 2am, two MH-160 Black Hawk helicopters packed with US Navy SEAL commandos, CIA operatives and a military dog[1], flew in a compound located a few miles out of Abbottabad, in northeastern Pakistan. The operation, codenamed Operation Neptune Spear, resulted in the killing of Osama bin Laden, founder and head of the Islamist group al-Qaeda, and the most wanted man for all American intelligence services for over 9 years, since the attacks on World Trade Center in September 2001.

Why did it take long to find the most wanted man on earth, specially considering the vast amount of resources available to US security services? How did they finally pin down the precise location after a 10-year manhunt? Not every detail has been disclosed, but American officials have stated than bin Laden did not engage in any kind of electronic communication from his hideout, in order not to disclose his whereabouts. Instead, he used off-line human couriers to deliver the messages. During the SEAL raid of the compound, thousand of email messages were found in roughly 100 flash memory drives[2], containing back-and-forth communication between bin Laden and his associates around the world. The other detail revealed by security officials was that the identification of one of those couriers, Abu Ahmed al-Kuwaiti, was crucial to the finding of the compound. According to Associated Press[3]

> The National Security Agency reportedly tracked phone calls between the courier Abu Ahmed al-Kuwaiti's relatives in the Persian Gulf to all numbers in Pakistan. And NSA surveil-

[1] http://www.nytimes.com/2011/05/03/world/asia/osama-bin-laden-dead.html?_r=0

[2] `http://www.nbcnews.com/id/43011358/ns/technology_and_science-tech_and_gadgets/t/how-bin-laden-emailed-without-being-detected/#.VVGplnVtORQ`

[3] `http://news.yahoo.com/blogs/envoy/courier-multiple-identities-man-led-u-bin-laden-133835572.html`

> lance eventually tracked Abu Ahmed al-Kuwaiti's location in Pakistan via one such phone call, the AP writes. Last August, they tracked al-Kuwaiti as he drove from Peshawar to the Abbottabad compound. And as analysts inventoried the facility's striking security features they became convinced that it housed a high-level al Qaeda figure.

Now let's look at the situation from a network perspective. Let us assume NSA had the ability to capture any electronic communication around the world, whether it is email, phone or any other kind. The other assumption would be than along those 10 years, the US intelligence services were able to identify some al-Qaeda members and its ranking within the organization. If only these two assumptions were true, the intelligence services would have been able to build a network whose nodes represent al-Qaeda operatives and whose links represent any form of electronic communication between them. In such network, of course, there would not be a node representing bin Laden itself (he did not engage any electronic communication) but all of his couriers would appear. Who would those couriers be? It is perfectly reasonable to choose just a few people for this task, and to choose them among the low-rank members of the organization, because the whole purpose of the courier strategy was not to attract any attention into the Abbottabad compound (remember that they had to travel there to pickup and deliver the flash drives with the email messages). If this would be the case, the network would have some interesting features: for instance, the *betweenness centrality*[4] of the courier nodes would appear to be very high for the rank they had in the organization. Also, the *geographic proximity* of the couriers would be revealing: many relevant messages for the organization were originated from a region were not high-rank member was known to be, therefore pointing that a high-rank *hidden node* may be present in the area, whose communications with the couriers happen off-line.

Of course these are just assumptions, because the American intelligence services have never revealed how did they find al-Kuwaiti in the first place. Maybe it had nothing to do with network science. But at least it is known that this strategy worked before. In 2003, the Pentagon released information on how they found Saddam Hussein during the Irak war[5]: a network theorist, Eric Maddox. was awarded with the Legion of Merit, for his contributions to the location of Hussein, and press reports specifically quote *betweenness*

[4] Betweenness centrality is a measure of centrality in a graph based on shortest paths. For every pair of vertices in a connected graph, there exists at least one shortest path between the vertices such that either the number of edges that the path passes through (for unweighted graphs) or the sum of the weights of the edges (for weighted graphs) is minimized. The betweenness centrality for each vertex is the number of these shortest paths that pass through the vertex.

[5] "Clan, family ties called key to army's capture of Hussein," Washington Post, 16 December, p. A27. McCallister, W.S. (2005)

centrality as a relevant tool conducting to the mission's success[6].

2.1 Network completion

The previous bin Laden case is just an example of the so called network completion problem [68]. While in the last decade the collection of large network datasets has become significantly less challenging, many times the collected network data [78] is incomplete so there are missing nodes and edges. For instance, IP networks usually contain zones where the Internet Control Message Protocol (ICMP) is disabled, therefore a fraction of the routers and connections in the network remain hidden. Similarly, networks arising from popular online social sites such as Facebook or Twitter do not completely represent all social ties of an individual due to network boundary effects, because not all the acquaintances of an individual have signed up for the site. In some biological networks, such as protein-protein interaction networks or food webs, checking the existence of a link between two entities requires a very costly experimental setting, therefore our knowledge of such networks remains limited (as an example, 80% of the molecular interactions in cells of Yeast [170] and 99.7% of human [149] were still unknown by 2008).

Previous examples include two different network completion problems, depending on whether we try to recover both links and nodes, or only links. If we consider that all networks can be presented as matrices (for example the adjacency matrix), all link recovery tasks tasks could be formalized as matrix completions [24, 62] where a data matrix with missing entries is given and the aim is to fill the entries. Node prediction is difficult to be formalized as a matrix completion problem, because it would imply identifying missing rows and columns in matrices. Even for link-prediction problems, real-world networks have very different structure and properties (e.g., heavy-tailed degree distributions) than the real-valued data matrices usually considered in the matrix completion problem.

2.1.1 Nodes and links prediction

In the previous Bin Laden example, not only the links leading to him were missing, but also the node representing bin Laden itself. This problem is significantly more difficult compared to link prediction and has only been approached more recently. In[40] authors focus on finding missing nodes and edges adjacent to them. In [68], the authors attempt to complete the network with arbitrary missing nodes and edges by assuming that the number of missing nodes is known, and then fitting a Kronecker graph [87]. In both

[6] `http://www.slate.com/articles/news_and_politics/searching_for_saddam/2010/02/searching_for_saddam.html`

works the authors only use structural information, therefore no metadata associated to nodes or edges is considered.

2.1.2 Link prediction

As in the food webs example, in this scenario all nodes are known and only some of the links between them are missing. This is by far the most studied kind of network completion problem. Although the seminal work in link prediction [92] considered the temporal component (i.e removing edges representing interactions after a certain date, and then trying to recover them) later work has abused the term *prediction*, applying it to problems where no temporal evolution is considered. Typically, link prediction algorithms are tested against well-known network datasets where edges are removed at random, and then the task consists of recovering them from the remaining graph's structure. On the other hand, while the seminal work [92] considered only structural information about the network, some later works have also used additional attributes about the nodes to improve prediction accuracy. For instance, in [158] the authors have available mobility patterns for mobile phone users, and successfully employ the similarity between those patterns as a feature for link prediction.

2.1.3 Attribute prediction

Networks datasets rarely consists of only lists of edges and nodes. They often contain also certain additional pieces of information about the nodes and the edges we will refer as attributes. For instance, an attribute we will use extensively in the following chapters is the geographical location of a node. This additional features are of course not always present for all nodes: we have found that proper location can only be found for around 50% of Twitter users, because the rest either do not fill the location field in their profile or fill it with information that cannot be geocoded due to ambiguity (e.g. "Springfield") or incorrectness (e.g. "the moon", "Narnia", ...). However, just like one can take advantage of the network structure to infer missing links, the same approach can be used to recover missing attributes, since in most networks link are more commonly found among entities with similar attributes (*homophily*), a very simple (and surprisingly effective) algorithm to infer the location for a Twitter user who did not provide a valid one could be assigning him the most common location among his followers.

2.1.4 Opaque nodes scenario

While this previous example might sound naive and obvious, it is important because it acknowledges the fact that the mere disclosure of links adjacent to a certain node provides significant insights about such node. This fact is even more interesting in nowadays society (specially when it comes to online

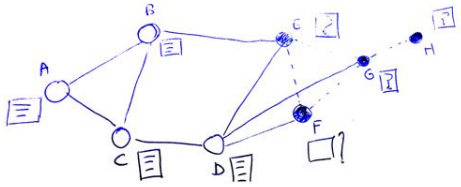


Figure 2.1: The opaque nodes problem: given a number of nodes whose attributes are known (well-known nodes, white) and some links whose events are also known, the attributes of an opaque node (black) must be estimated.

interactions) because the disclosure of links has become cheap: only the permission from one party is required to disclose a social relationship. Contrary to what happens with romantic relationships or companies merges, where a consensus is usually reached before publicly disclosing the partnership, nowadays it is enough to get a user to log in your platform via Facebook or Twitter to get his whole list of friends or the list of users they have interacted with. According to recent studies on the Facebook graph [155], this implies that it is enough to get 4 million registrations in an application to crawl relationships involving about 300 million profiles, which represents roughly one third of the total Facebook population.

These permission granting schemes generate a specific prediction scenario which we will refer to as the *opaque nodes problem* which is presented in figure 2.1 and contains two types of nodes:

White nodes whose attributes are known, and so are all the links adjacent to them.

Opaque nodes whose attributes are not known, and only the links connecting them to white nodes are known.

It is important to emphasize that the opaque nodes problem not only shows up in Facebook and Twitter applications, but it also applies, among others, to any kind of communication provider, such as mobile carriers. Generally speaking, communication providers have some data about their customers (name, address, gender, age, ...) as well as records of the interactions (such as texts or calls in the case of a mobile carrier) not only between customers but also between their customers and subscribers from other companies. This is, in fact, an opaque nodes problem, where customers would be the white nodes and subscribers from other companies would take the role of the black nodes.

In an opaque nodes scenario, there are three different network completion tasks:

1. Inferring attributes for opaque nodes. In the figure 2.1, this would answer questions such as "how old is E?" or "is F a female?".

2. Recovering links between opaque nodes. In the figure 2.1, this would answer questions such as "is E connected to F?" or "how many triangles are actually in the network?".

3. Inferring the existence of hidden opaque nodes. This would allow us to notice the existence of H in the example network, despite H be absent in every white node's adjacency list.

In this research we will focus on tasks 1 and 2. Due to the relationships between these tasks, they could be addressed iteratively feeding each step with the results of the previous one. For example, if we perform an attribute prediction step to recover locations for opaque nodes, then those inferred locations could be used to improve the link prediction, and after that, we could use the new links to improve the location prediction and so on. However, in this research we will focus on accomplishing tasks 1 and 2 using only the original information without additional steps.

In the remainder of this chapter we will present our contributions, structured as follows. First, the link prediction problem is addressed: a short review of current techniques is presented. These techniques are then applied to a standard scenario proposed as a challenge for the International Joint Conference for Neural Networks 2011. Then, the opaque nodes problem is studied from the link prediction perspective, using a mobile phone dataset and an email dataset. On the other hand, the second part of this chapter addresses the attribute prediction problem in an opaque nodes scenario. The goal will be inferring age and gender for mobile phone users.

2.2 Link prediction

As stated before, the goal of the link prediction problem is to rebuild a network when it is known or suspected that there are missing links in the data. While one could approach the dual problem (detecting spurious links in a network) with exactly the same tools, that is beyond the scope of this research.

Although link prediction was previously explored in the field of computer science to rebuild IP networks where some of the links were hidden, it can be considered that [92] was the first attempt to take advantage from complex networks properties when estimating the probability of existence of a link. As mentioned before, in this work the authors strictly *predict* links: they got access to historical citation records between scientific papers and they built a co-authorship network (two authors are connected only if they published at least one paper together). In such network they tried to predict links, which means they tried to predict if two authors who did not collaborate

before a certain date would co-author a paper later on. Interestingly, most of the tools they employed to solve the problem are of use in other scenarios where there is no explicit temporal component.

One of the most relevant applications of link prediction is to find good *link candidate*s for networks in which it is very expensive to test if the two entities interact. A clear example are protein interaction networks, because checking if two proteins interact typically requires a lot of research resources. Under these circumstances, it turns out really useful to have an algorithm capable of proposing links so that their existence probability is bigger than the one obtained from random sampling (complex networks are known to be sparse, so random sampling will always produce very poor results when trying to find a new link).

Among other applications, link prediction is used also for recommenders systems (typically using a bipartite network) [81, 140, 136], Spam detection in email [60] or even finding hidden relationships among terrorist cells [26].

In the following, a formal definition of the problem is presented, along with some metrics used to evaluate the performance of the algorithms. Then we will describe different prediction methods according to the taxonomy proposed in [98]. In this text we will focus only on algorithms based on similarity and maximum likelihood, leaving out probabilistic methods such as relational Bayesian networks [54]. The reason for this exclusion is that such probabilistic methods present some characteristics (mainly a high computational cost) that exclude them from the problem to be solved in this section. Besides, their approach is so different that it will not help the reader to understand the subsequent discussion.

2.2.1 Problem description and performance metrics

Let $G(V, E)$ be an undirected simple graph, where V represents the node set and E the edge set. Let U be the set of all possible edges of G, then

$$|U| = \frac{|V| \cdot (|V| - 1)}{2}.$$

This way, $U - E$ is the set of non-existing edges. The prediction problem consist of assuming some of the links in $U - E$ are actually links which do belong to E and trying to find them.

In general we do not know which of the links we are actually missing (otherwise we would not need the prediction). So that, to evaluate the performance of the algorithms, we will randomly divide E in two sets: a training set of existing links E^T and a validation set E^V. In order to build the training and validation set these two sets will be mixed with negative examples pulled randomly from $U - E$. The main required condition in the experiment is that after the sampling, validation set must be treated as unknown. Therefore the information contained in it is never used to improve

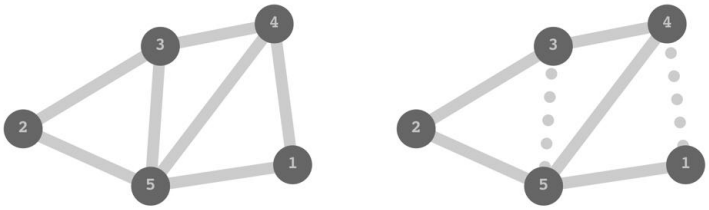


Figure 2.2: Creation of the training and validation sets,. From the original network, E^V is randomly sampled (here $E^V = \{e_{35}, e_{14}\}$). Training network $G_T = (V, E^T)$ is then defined by 5 nodes and 5 links.

the prediction. Figure 2.2 illustrates the creation of the network training set.

Regarding the sizes of the sets, although theoretically any size could be chosen, all results within this section will use a 90%-10% partition (i.e. $|E^T| = 0.9|E|$). Precisely, all results presented in this section are averaged over 10 iterations in a cross validation scheme (*K-fold* with $K = 10$), so that it is ensured all links in the original network will be in E^V exactly once. This is an standard procedure in the literature when comparing performance of different prediction techniques [98, 97, 172].

The output of a prediction algorithm could be just a sorted list of links (the most likely to exist being the first) belonging to $U - E$, although in general each edge $(u, v) \in U - E$ gets a score s_{uv} according to the probability that such edge exists. From any sorted list of links obtained as output, the receiver operating characteristic (ROC curve) [95] can be calculated by varying the decision threshold so that the curves ranges in the plane defined by the true positive rate (TPR) and the false positive rate (FPR). Figure 2.3 presents an example of three ROC curves for different classifiers. A common metric to evaluate classification performance is the Area Under the Curve (AUC), precisely referred to the ROC curve. The AUC can be interpreted as the probability that a true link has been assigned a score higher than a false link. This interpretation drives to a very simple expression to estimate the AUC. Let take n pairs of link candidates so that each pair contains a true link and a false link. Let be n_{right} the number of times the algorithm got the right link and n_{draw} the number of times the algorithm provided the same score to both links, then

$$AUC = \frac{n_{right} + 0.5 n_{draw}}{n}.$$

For example, lets assume in the experiment depicted in Figure 2.2 a prediction algorithm generates scores $\{s_{35} = 0.6,\ s_{12} = 0.3, s_{13} = 0.7, s_{24} = 0.6\}$. Since there are three possible pairs containing the true link $(3, 5)$, and

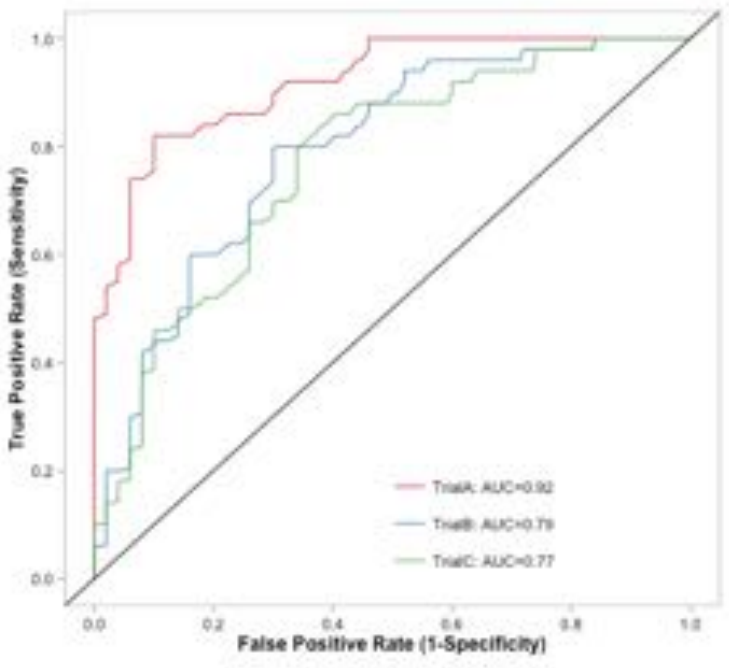


Figure 2.3: ROC curves for 3 different classifiers. Solid black line represents the performance of a random classifier. The red classifier clearly outperforms the other two.

only score (s_{12}) is smaller than s_{35}, then $n = 3$ and $n_{right} = 1$. On the other hand, since $s_{35} = s_{24}$, $n_{draw} = 1$. Therefore, $AUC = \frac{1+0.5}{3} = 0.5$, so the algorithm performs just as good a purely random prediction.

Feature degradation in link prediction

The problem described so far does not seem different from a classical supervised machine learning scheme. However, there is a key issue related to the networked nature of the data. In a classical machine learning scheme, the size of the training set is of course relevant because the emergence of patterns in the data can only occur beyond a certain size. However, features remain constant no matter the size of the training set.

When doing link prediction, many of the features (if not all) are calculated from structural information of the remaining network. As we will see below, the degree or the number of common friends are features with a lot of predicting power. However the mere existence of missing links does not allow us to know the real value of these features, which means the larger the fraction of links removed, the worse the estimation of this features. Considering the example in figure 2.2, while the real degree of node 5 is 4, in the

training set the edge e_{25} will appear in the training set as connecting nodes with degrees 2 and 3, respectively.

2.2.2 Methods based on node similarity

Methods based on node similarity (also referred as proximity methods) are the simplest prediction methods known, at least when it comes to their conception. Let u and v be nodes, then these kind of algorithms provide a higher score when u and v present similar neighborhoods.

In the following we present a number of similarity indexes, which are classified according to [98] depending on the kind of information used into local, quasi-local and global.

Local similarity indexes

1. *Common Neighbors* (CN). Lets denote N_x the set containing all neighbors of x. As is was stated in the introduction, there are many networks, including social networks, where the number of triangles formed is really high. This is called the clustering phenomenon. Hence, if two nodes share a common neighbor then is more likely they share a link. Under this principle, the score is defined as

$$s_{uv}^{CN} = |N_u \cap N_v|,$$

   or analogously $s_{uv}^{CN} = (A^2)_{uv}$ where A is the adjacency matrix of G.

2. *Salton index* (or cosine similarity), weights the previous index with the reverse of the square square root of the degrees product, so

$$s_{uv}^{Salton} = \frac{|(N_u \cap N_v|}{\sqrt{k_u k_v}}.$$

3. *Jaccard index.* One of the eldest, uses the cardinal of the union set in the denomination:

$$s_{uv}^{Jaccard} = \frac{|N_u \cap N_v|}{|N_u \cup N_v|}.$$

4. *Sorensen index.* Heuristic widely used in biological networks defined as

$$s_{uv}^{Sorensen} = \frac{2 \cdot |N_u \cap N_v|}{k_u + k_v}.$$

5. *Hub Promoted index* (HPI). Used in metabolic networks. Gives more importance to links adjacent to a hub.

$$s_{uv}^{HPI} = \frac{|N_u \cap N_v|}{\min\{k_u, k_v\}}.$$

6. *Hub Depressed Index* (HDI). Opposite of the previous one.

$$s_{uv}^{HDI} = \frac{|N_u \cap N_v|}{\max\{k_u, k_v\}}.$$

7. *Leicht-Holme-Newman Index* (LHN1). Uses the product $k_u \cdot k_v$, which is the expected number of common neighbors according to the so called configuration model [117].

$$s_{uv}^{LHN1} = \frac{|N_u \cap N_v|}{k_u \cdot k_v}.$$

8. *Adamic-Adar index* (AA). Despite that its formulation looks different from previous in a first sight, this index just reduces the contributions of each of the common neighbors according to the logarithm of its degree.

$$s_{uv}^{AA} = \sum_{z \in N_u \cap N_v} \frac{1}{\log k_z}.$$

9. *Resource Allocation index* (RA). This index, while similar to the previous one, has an an additional interpretation in transportation networks. Let u and v be two nodes which are not directly connected. Then, the node u sends resources to v through each of the neighbors. Assuming neighbors distribute each resource unit uniformly among their neighbors, the RA index is the amount of resource that v receives from u

$$S_{uv}^{RA} = \sum_{z \in N_v \cap N_v} \frac{1}{k_z}.$$

10. *Preferential Attachment index* (PA). Although the preferential attachment was first introduced by Barabási and Albert to explain the emergence of scaling in complex networks [14], later it was proven that the power law in the degree distribution can be also obtained by adding links in a network where the number of nodes is constant. The preferential attachment expression for such scenario

$$s_{uv}^{PA} = k_u \cdot k_v$$

is also useful as a proximity index.

Regarding the performance of each of the indexes, [97] includes a comparison measuring the AUC for each of the methods across several network data-sets. Another comparative, this time focused on social networks, can be found in [92]. In table 2.1 it can be observed how results for some indexes can vary a lot for certain indexes. As pointed before, most of the indexes are based

Indices	PPI	NS	Grid	PB	INT	USAir
CN	0.889	**0.933**	**0.590**	0.925	**0.559**	0.937
Salton	0.869	0.911	0.585	0.874	0.552	0.898
Jaccard	0.888	**0.933**	**0.590**	0.882	**0.559**	0.901
Sørensen	0.888	**0.933**	**0.590**	0.881	**0.559**	0.902
HPI	0.868	0.911	0.585	0.852	0.552	0.857
HDI	0.888	**0.933**	**0.590**	0.877	**0.559**	0.895
LHN1	0.866	0.911	0.585	0.772	0.552	0.758
PA	0.828	0.623	0.446	0.907	0.464	0.886
AA	0.888	0.932	**0.590**	0.922	**0.559**	0.925
RA	**0.890**	**0.933**	**0.590**	**0.931**	**0.559**	**0.955**

Table 2.1: Performance of different local similarity indexes. Best performances for each dataset are presented in bold. Network datasets: protein interaction (PIP), publications in network science (NS), power grid (Grid), political blogosphere (PB), Internet (INT) and US air traffic (US Air) . Source: [98].

on clustering, so they perform poorly on networks where such phenomena is weak (power grid and air traffic networks [118]). On the other hand, although PA index is not particularly successful, it is interesting because it requires very little information. Note also that while PA falls bellow random accuracy for INT and Grid, that means its inverted output is a classifier with $\text{AUC}_{inverted}$= (1-AUC).

Global similarity indexes

1. *Katz Index* is built of contributions from all possible paths between nodes u and v. Contributions are exponentially weighted so that long path contributions are reduced. Its definition is

$$S_{uv}^{Katz} = \sum_{l=1}^{\infty} \beta^l (A^l)_{uv},$$

   where the $(A^l)_{uv}$ term presents the number of l length paths between u and v and the $\beta < 1$ takes care of the exponential decay. Not for very small values of β, this index is proportional to common neighbors (CN).

2. *Leicht-Holme-Newman Global Index* (LHN2). Its a modification from the local LHN1 index, which considers two nodes are similar if their neighbors are so, leading to a self-consistent equation whose solution

is

$$s_{uv}^{LHN2} = \delta_{uv} + \frac{2M}{k_u k_v} \sum_{l=0}^{\infty} \phi^l \lambda^{1-l} (A^l)_{uv},$$

where δ_{uv} represents the Kronecker delta.

3. *Average Commuting Time* (ACT). Let $m(u, v)$ be the average number of hops that a random walker takes to reach v from u (note that $m(u, v) \neq m(v, u)$ in general). Then ACT index is defined as

$$s_{uv}^{ACT} = \frac{1}{m(u, v) + m(v, u)}.$$

4. *Cosine based on* L^+. Let L be the Laplacian matrix of G, and L^+ being its Moore-Penrose pseudoinverse [125], this index is defined as

$$s_{uv}^{cos^+} = \frac{l_{uv}^+}{\sqrt{l_{uu}^+ \cdot l_{vv}^+}},$$

where l_{ij}^+ terms represent elements of L^+.

5. *Random Walk with restart* (RWR). This a direct application of the PageRank algorithm (the original algorithm behind Google's search results ordering [121]). It considers a random walk from u to v in which the walker might return to u (independently of where he is at that time step) with probability $1 - c$. Denoting $q_{uv} = q_{uv}(c)$ the probability that the random walker is in v in the steady state, we have

$$s_{uv}^{RWR} = q_{uv} + q_{vu}.$$

6. *SimRank*, which is defined in a self-consistent way according to the assumption

$$s_{uv}^{SimRank} = C \cdot \frac{\sum_{z \in N_u} \sum_{z' \in N_v} s_{zz'}^{SimRank}}{k_u \cdot k_v}$$

where $s_{uu} = 1$ and $C \in [0, 1]$. The SimRank index can be understood also as random walk so that $s_{uv}^{SimRank}$ measures the average time that it takes for two random walkers who leave u and v respectively in $t = 0$ to run into each other.

7. *Matrix Forests Index* (MFI). This index is calculated as follows:

$$S = (I + L)^{-1},$$

where $s_{uv} = (S)_{uv}$, I is the identity matrix and L the Laplacian matrix. These index represents the ratio of maximal trees whose rooted in u and include v compared to all maximal trees whose rooted in u.

There are no precise results about these global indexes in a variety of real world networks such as those presented for local indexes. In [98] it is speculated that these indexes should produce better results, but no data supporting such statement is presented. The main issue regarding global similarity is the computational cost required (think for example that the adjacency matrix A for the Facebook social graph, would require nowadays about 4 exabytes, using 32 bits per element). This implies that even with the computation capability available at the moment, most of global indexes can only be applied in networks with tens of thousands of elements at most.

Quasi-local methods

To overcome the computational cost problem, there are approximations to some of the global indexes presented before. Those approximations are mostly based in considering information only from a certain neighborhood of a node, so the computational cost is then determined by the average degree of the network, rather than the number of nodes or edges in it.

1. *Local path index* LP. It is an approximation to Katz index previously defined, where only paths up to length n are considered, so that

$$S^{LP} = A^2 + cA^3 + ... + c^{n-2}A^{n-2},$$

   where $n > 2$. While computational complexity for the Katz index is $O(N^3)$, LP complexity goes with $O(N\langle k\rangle^n)$ which for small values of n turns out to be much smaller. In [97] several experiments are presented to show that the optimal n can be estimated from the diameter of the the network.

2. *Local random walk index* (LRW). It is a variation of RW in which it is considered the walk starts from a node u with a certain *a priori* probability distribution $\bar{\pi_u}(0) = \bar{e_u}$. The random walk is then modeled by a Markov process so that

$$\bar{\pi_u}(t+1) = P^T\bar{\pi}_{uu},$$

   where P is the transition matrix ($P_{uv} = \frac{A_{uv}}{k_{uv}}$). The LRW is then defined as

$$s_{uv}^{LRW}(t) = q_u\pi_{uv}(t) + q_v\pi_{vu}(t)$$

   where q_u is the initial probability, and t is the number of steps considered. In the results presented in this section q is defined as $q_u = \frac{k_u}{2|E|}$.

3. *Superposed Random Walk index* (SRW) is the sum from 1 to t from the previous index, although it can also be interpreted as a modification

AUC	CN	RA	LP	ACT	RWR	HSM	LRW	SRW
USAir	0.954	0.972	0.952	0.901	0.977	0.904	0.972(2)	**0.978**(3)
NetScience	0.978	0.983	0.986	0.934	**0.993**	0.930	0.989(4)	0.992(3)
Power	0.626	0.626	0.697	0.895	0.760	0.503	0.953(16)	**0.963**(16)
Yeast	0.915	0.916	0.970	0.900	0.978	0.672	0.974(7)	**0.980**(8)
C.elegans	0.849	0.871	0.867	0.747	0.889	0.808	0.899(3)	**0.906**(3)

Table 2.2: Performance of different similarity indexes based on random walks. Best performances for each dataset are presented in bold. Numbers between brackets represent the maximum path length considered. Source: [96]

of RWR since it also simulates the random walker restarting from the source with a certain probability. Its formulation is

$$s_{xy}^{LRW}(t) = \sum_{\tau=1}^{t} s_{xy}^{LRW}(\tau).$$

Table 2.2 presents the performance of these algorithms compared with the performance reached by local methods (CN and RA) and global methods (ACT and RWR) plus the performance reached by a maximum likelihood method (HSM) which will be presented below . Quasi-local methods are found to be quite convenient, because they perform nearly as good as global methods, and they can be up to a thousand times faster to compute for the network sizes considered in the table.

2.2.3 Maximum likelihood methods

In this section we provide an introduction, from a descriptive point of view, to maximum likelihood methods in link prediction. Precisely, we focus in the method proposed by Clauset et al. at [26], which was later on referred in the literature as the *Hierarchical Structure Model* (HSM). While there are other maximum likelihood methods which got some impact in the literature [166, 58, 5], such as the *Stochastic Block Model,* we will focus on HSM, since the goal for this section is to provide an example of a completely different approach to link prediction from the one used by the similarity indexes.

HSM first task is to extract a dendrogram (see figure 2.4) where closely related pairs have lower common ancestors in the dendrogram. This dendrogram is fitted combining maximum likelihood approach and Monte Carlo based simulations. Once the dendrogram is found, synthetic networks are generated from the dendrogram, generating links more likely between pairs with closer common ancestors. The score for candidate links is simply how often do they appear in these synthetic networks.

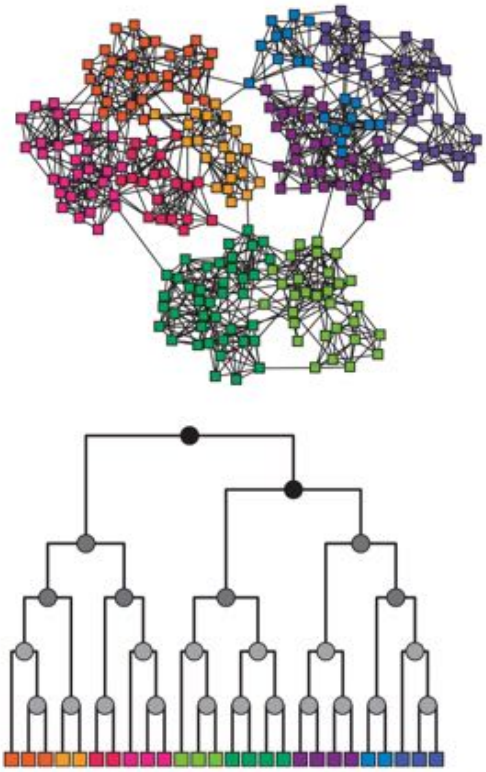


Figure 2.4: Hierarchical network and a one of its possible dendrograms.

For example, in the results presented in table 2.2 5000 dendrograms were considered. Regarding performance, in this table we can see HSM is in general worse than similarity based methods. In fact, some ([98]) advice against likelihood methods because of both performance and computational cost. However, as stated in the original article, there are networks where random-walk based techniques are not useful (ultimately because of lack of clustering) and then methods like HSM prove themselves useful. An example of such networks could be food-webs: while two species predating a certain third species does not make more likely that they predate each other (i.e. no clustering), food webs have a well-known pyramidal structure.

2.2.4 Statistical learning approach to link prediction

Every method presented in the previous review is focused on finding a scalar function f such that

$$s_{uv} = f(m^{uv})$$

where $m^{uv} \in \mathbb{R}^n$ contains different network features related to the possible link between nodes u and v. For example, for the Jaccard index, $m^{uv} = (|N_u \cap N_v|, |N_u \cup N_v|)$ and $f(m^{uv}) = \frac{m_1^{uv}}{m_2^{uv}}$ The goal is to find a f that completely identifies true links among the set of candidates.

This section will focus on the usage of statistical learning to address the link prediction problem. Hence it will be considered as a classification problem, since we want to find the best function when it comes to differentiate existing and non-existing links.

To illustrate the application of statistical learning to link prediction, one of the datasets released for challenges in the 2011 International Joint Conferences for Neural Networks (IJCNN) will be used. In this section different strategies to solve the problem will be presented, and later on some issues, out of the scope of the original challenge, will be presented. Precisely, the impact in the performance of the number of missing links will be evaluated, as well as how reasonable is trying to generalize the results presented here to other social networks.

Problem definition and exploratory data analysis

There are a few differences between the general link prediction problem (as presented in Section 2.2.1) and the challenge proposed for IJCNN. The dataset contains a certain neighborhood from the social site flickR, where users share pictures. In this social site a certain user A declares himself "a follower" of another user B and then every time A logs in, the site presents him the latest pictures uploaded by B. This social relation cannot be modeled as mutual (B might not follows A back) so a directed graph is employed. This is the first difference from the original link prediction problem.

IJCNN Network	
$\|V\|$	1133394
$\|E\|$	7237778 + 4480
Directed Network (reciprocity)	Yes (0.02%)
Average degree	6.386
Clustering (relative to Erdös)	0.095533 (8119.4)

Table 2.3: General properties of the IJCNN 2011 network dataset

As aforementioned, the target network comes from a mapping from the flickR site. Precisely, data include connections between 1.1 million users, providing a training set E^T with 7.2 million links. Additionally, a client set E^V with 8960 candidate links is provided, ensuring 50% of such links are actual links, and the remaining 50% have been randomly sampled for the non-existing links set $U-E$. This produces the second deviation from classic conditions, since those usually focus on scenarios where the validation set size is about 10% of the training set. In this dataset, such proportion is substantially smaller (about 0.1%).

Regarding network characteristics, the graph turns out to be a sparse directed graph. In table 2.3 some of the most relevant characteristics are presented. It is noticeable that only 0.02% of the links are reciprocated, a small ratio compared to other social networks analyzed in the literature. On the hand, while clustering coefficient might not seem too high in a first view, it is 4 order of magnitude higher than the equivalent Erdös graph. On the other hand, Figure 2.5 presents the in-degree and out-degree distribution. Although both distribution present a long tail, out-degree distribution presents an uncommon characteristic: only 3.3% of the nodes have any outgoing link. This produces the step observed for small values on the red curve in the figure.

Feature selection

Before considering any learning technique, features used to estimate scores for the candidate links need to be chosen. In this stage, 12 features are considered.

- Degree (4 per candidate link: ODS, IDS, ODT, IDT): in-degree and out-degree for both target and source nodes are considered, since the emergence of a long tail degree distribution suggests a certain preferential attachment in the network. If that would be case, edges incident in high degree nodes are more prone to exists. Besides, *assortative mixing* has been found in many real world networks, implying people tend to friend others whose degree similar to them (the opposite phenomenon is fairly common, too, and it is referred as *disassortative*

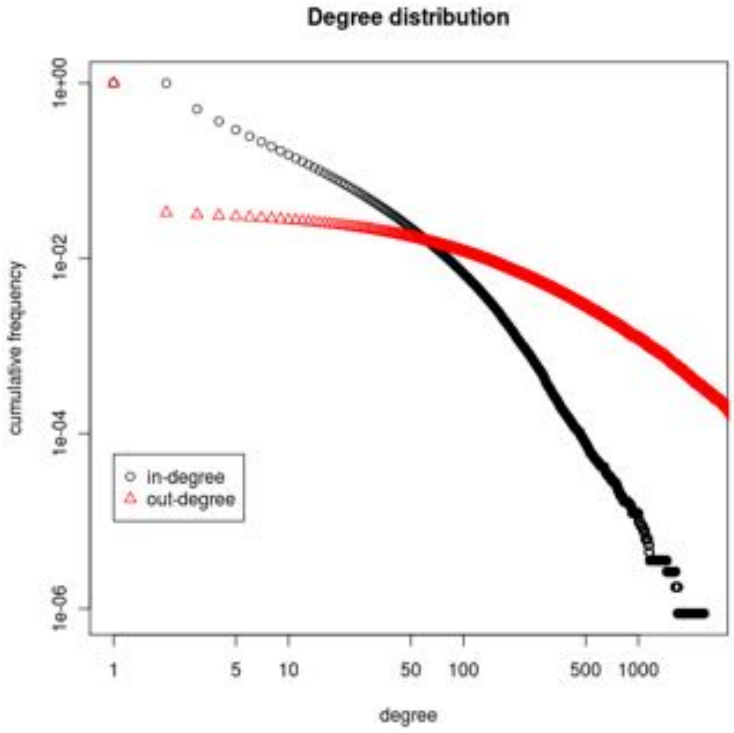


Figure 2.5: Degree distributions for the IJCNN dataset

mixing).

- Clustering (2 per link: CCS, CCT): since the clustering appears to be intense in this network, both clustering coefficients (source's and target's) are included among the features.
- Type 1 triangles (T1): number of length 2 directed paths from the source to the target. When considering a directed network, there are several kind of possible triangles, depending on edges' orientations. Considering results from [4], when it is concluded oriented triangles $A \to B$, $B \to C$, $A \to C$ are of special importance in social networks, this feature is included.
- Type 2 triangles (T2): number of common neighbors in the undirected version of the networks (where a link exists if any of its oriented equivalents exists in the original network).
- Hub and Authority Kleinberg scores [69] (4 per link: HubS, AuthS, HubT, AuthT): these metrics come from information networks. They've been included because, just like Twitter, flickR can be also considered an information network, where some of the nodes have authority when it comes to certain kind of photography, just like Kleinberg pointed out for web sites.

2.2.5 Statistical learning techniques

Linear Discriminant Analysis (LDA)

The Linear Discriminant Analysis [131] is the simplest statistical learning technique. It consists of finding the best possible hyperplane to split up different classes of data points (in our case real and false links). Formally,

$$s_{uv} = f(m^{uv}) = wm^{uv}$$

where $w \in \mathbb{R}^n$. Because in this case f is a linear function, an optimal solution can be analytically obtained [42]. The LDA is also the least computational expensive method among those presented here. Besides, the simplicity of f allows also to interpret the importance of the different features in the classification, so that it is possible to identify which features are adding relevant information to the classification.

The LDA, when applied to the network and features previously described, and using the whole training set (7.2 million links) scores 0.96 in AUC and 81% accuracy. Table 2.4 presents the confusion matrix, which shows a relatively high amount of false negatives. On the other hand, Table 2.5 presents the coefficients used by the discriminant function. Note that the degree information is mostly ignored by the classifier, which in a first look seems opposite to the hypothesis of preferential attachment as generating process.

	True links	False links
True links	2563	1437
False links	47	3953

Table 2.4: LDA confusion matrix

Feature	**LDA coefficient**
ODS	*-0.0006755337*
ODT	*-0.0001955009*
IDS	*-0.0008258207*
IDT	*-0.0057025036*
CCS	2.2029752863
CCT	-0.5307680144
T1	-0.4706423936
T2	-1.2810232432
AuthS	-2.1273195471
AuthT	-1.2385597262
HubS	3.1239834744
HubT	1.9686931287

Table 2.5: LDA coefficients

Multilayer perceptron (MLP)

The multilayer perceptron is a supervised learning technique based on a neural network. Its basic functional element, known as the *neuron,* consist of a linear combination stage (ADALINE) followed by a non-linear element. This non-linear element's transfer function is a differentiable approximation to the step function (typically a sigmoid function). Figure 2.6 presents the functional scheme of a neuron.

The multilayer perceptron combines several neurons, in order to be able to reproduce non-linear functions. The scheme employed in this text contains only one hidden layer with N neurons, in a similar manner to what is presented in Figure 2.7 (note that for the shake of simplicity, the figure presents only 4 inputs, while actually a 12 inputs MLP has been employed for this problem).

During the training process, known samples are presented to the MLP and weights are fitted using a backpropagation algorithm, which is essentially a efficient implementation of the common gradient descent optimization method[7].

One of the most relevant design decisions when using a MLP is the number of neurons and their distribution. Since a one hidden layer scheme

[7]Simulations in this text include an additional momentum term in order to mitigate the local minimum problem as much as possible.

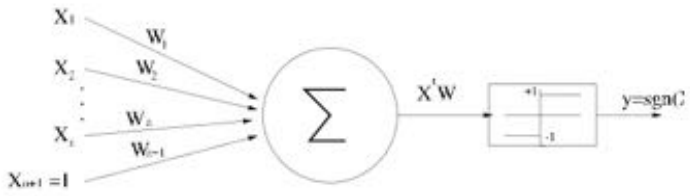


Figure 2.6: Neuron's functional scheme.

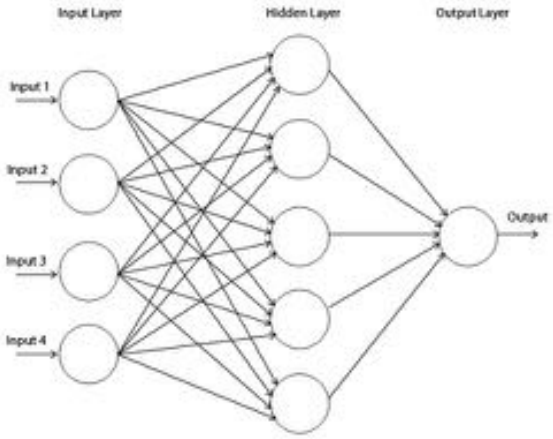


Figure 2.7: MLP using one hidden layer.

has been already chosen, only the precise number of neurons N is left to be decided. This number election involves a certain trade-off: if N is too small, the MLP will not be flexible enough to reproduce the function it needs to learn, specially if such function is strongly non-linear. On the other hand, if N is too large, the MLP will perform greatly within the training set, but it will not generalize successfully to other data points. In this link prediction scenario, $N = 4$ has been found to be an acceptable trade-off solution.

Regarding performance, the 4-neurons single-layer MLP trained using the whole E^T reaches $AUC = 0.989$ and 91% percent accuracy. This improvement compared to the previous LDA indicates that there is a non-linear function which splits the two classes better than the optimal hyperplane.

LDA using logarithmic degree in the input (LDA-log)

As stated when discussing the LDA coefficients in Table 2.5, the linear discriminant function basically ignores all degree metrics, which contradicts the idea of the network being created through preferential attachment. A descriptive analysis of the feature vector finds that none of the degree metrics follow a Gaussian distribution, but a power-law distribution instead. Since it is known that LDA works better when input data are normally distributed, logarithm are taken to the input degree data before running LDA. The effect of the transformation is depicted in figure 2.8.

By doing this simple preprocessing, the accuracy of the LDA equals MLP. Using the entire training set E^Tresults reach $AUC = 0.99665$ and 98.35% accuracy. This improvement can be observed in figure 2.9 where the two ROC curves are presented.

Feature selection and training set size impact

Once these results are obtained, next step is trying to evaluate how much is the impact in accuracy when either fewer features are included in the model or a smaller training set is used. As mentioned before, LDA is computationally much lighter than any neural network. Since it has been also shown that the performance gap between the LDA and the perceptron almost vanishes when the logarithmic preprocessing is used, we focus the following discussion solely on the LDA.

For previous examples the whole available information (7 million links) has been used to fit the coefficients. Now the performance of the predictor is evaluated when only a fraction of such information is employed. In figure 2.10 can be observed that even for really small training sets (merely $1,000$ samples) the same discriminant hyperplane W is obtained, thus the same performance is reached both in terms of accuracy and AUC.

On the other hand, the reduction in the number of features is also considered. To do so, a Wilk's lambda test is performed so that no additional

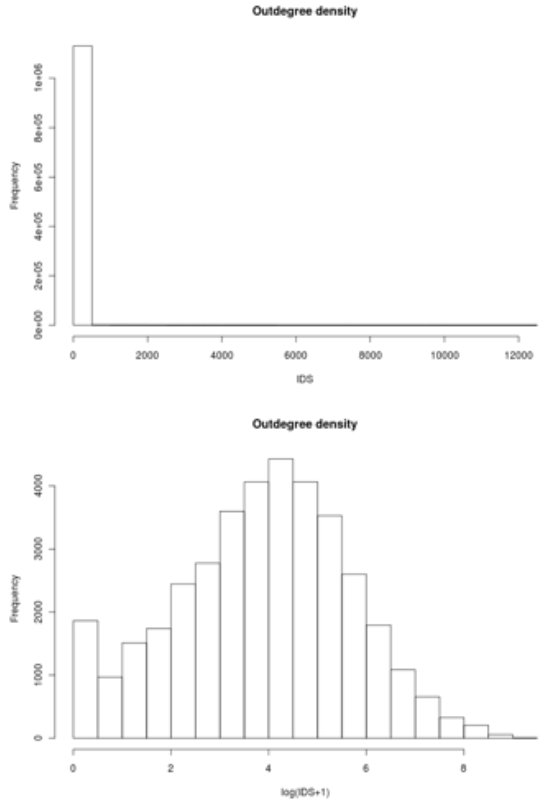


Figure 2.8: Preprocessing of out-degree data.

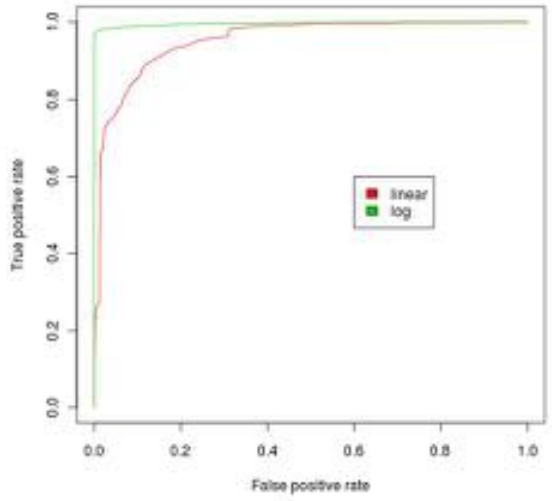


Figure 2.9: Improvement in ROC curve for LDA-log

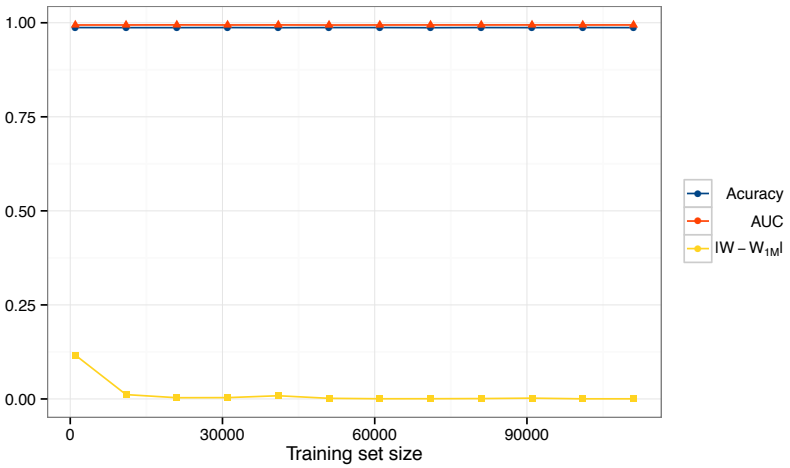


Figure 2.10: Variation of prediction performance while changing the training set size.

features are included if the performance improvement is below 0.1%. This way it is found that only two features (out degrees for the two nodes) are needed to reach a 98% accuracy.

By using any of the two options (less features, smaller training set) execution of all three stages (feature calculation, training and testing) are done significantly faster. For example, in our simulation setup, using the MLP and 5 million samples in the training set requires about 5 minutes computation, while the LDA-log with $20,000$ samples in the training sets take only 10 seconds and it keeps almost the same prediction accuracy.

Impact in performance when varying the fraction of known edges

Predictions presented so far seem to be very accurate, but one could argue that experiment conditions are really convenient: information about 7 million links is provided, while only $4,000$ links are missing, which is about 0.05% of the network total size. An interesting question is whether or not previous high precision accuracy is preserved when a larger fraction of links is removed from the network.

Note that this experiment is substantially different to the one presented right before, because in this experiment is not only the size of the training set E_T what comes into play, but also the actual features extracted for each of the links. For example, if the existence of a certain link between nodes u and v would be ignored, degrees for both nodes would be underestimated. Thus, the larger the fraction of removed edges, the worse will be the estimation of features such as common neighbors or degree.

In order to evaluate this impact in performance, simulations where a

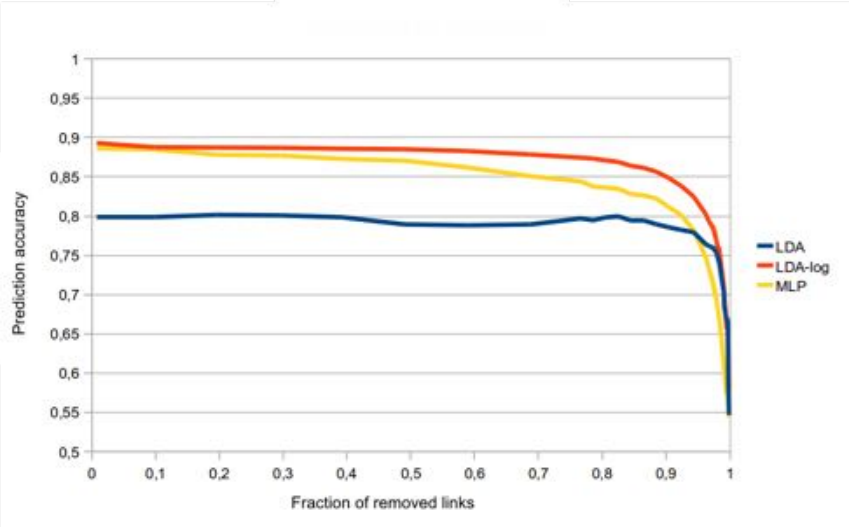


Figure 2.11: Prediction performance for several fractions of links removed.

significant fraction of edges are removed have been carried out. After the removal of such edges, train and validations sets are extracted in the same way it was done in previous sections.

Experiment results are presented in figure 2.11, showing that prediction accuracy remains almost constant even for large values of removed links (about 80%). Also, results point that MLP performance gets reduced slightly sooner that LDA based predictions. The appearance phase transitions such as those observed for LDA and LDA-log performances is common in scale-free networks, and it is ultimately produced by the sudden breakup of the giant component when a large fraction of the links are removed.

Results for other directed social networks

Finally, the same procedure has been performed in two additional directed social networks, trying to establish if previously reported results can be generalized to a broader context.

The first dataset contains a trust network extracted from the website *epinions.com*. In this dataset, a user declares if he/she trusts the reviews from another user [130]. The second case study is based on Wikipedia edits. In Wikipedia, regular users may become administrators of the site if they go through an election process. The network here analyzed links user A to user B if A voted for B [88].

After applying the exact same procedure previously described for the flickR network, the results are summarized in table 2.6. While both networks are directed and present a long tail in the degree distribution, they present significant differences when it comes to reciprocity, clustering coefficient,

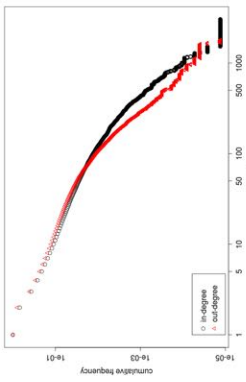
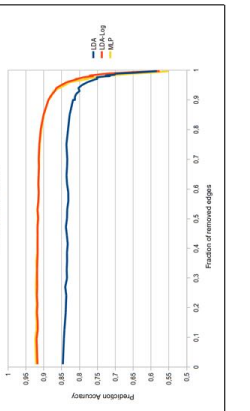
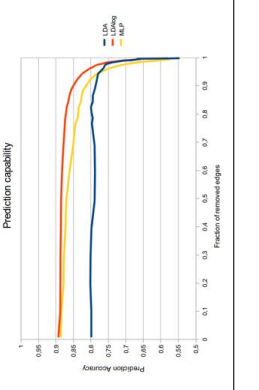
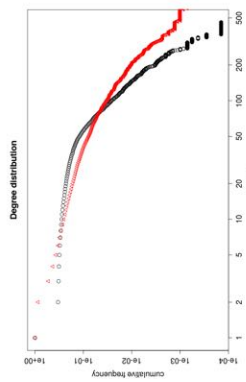
	Epinions	Wiki-vote
$\|V\|$	508836	7066
$\|E\|$	75877	103663
Directed (reciprocity)	Yes (0.2)	Yes (0.029)
$\langle k \rangle$	6.706	29.34
Clustering (rel. Erdös)	0.1377 (788.60)	0.1418 (38.48)
Degree distribution		
Prediction Accuracy		

Table 2.6: Prediction results in other directed social networks

size and average degree. Despite all these differences, prediction results in the two networks are similar to those found for flickR: very high prediction accuracy (over 90% for LDA-log) even if a large fraction of links is removed, then the prediction falls abruptly.

2.2.6 Completion of a network with opaque nodes

In the previous section, a technique has been developed which makes possible to differentiate, among a set of candidates, which of them are more likely to be links in a real network. In principle, given the high accuracy levels reached before ($AUC = 0,996$ and up to 98% accuracy), it may seem straightforward to recover the original network when some of the links are missing.

This approach, although fairly intuitive, ignores the existence of a *network effect,* which actually can significantly damage the performance of the prediction. The metrics used to evaluate prediction performance (standard in statistical learning) evaluate the accuracy on a balanced test set: this is, a high AUC value states that the prediction will do really well if there are 50% true links and 50% false links in a set.

Now think about the opaque nodes scenario presented in Section 2.1.4, where all links between certain nodes (the opaque ones) have been deleted, so that the candidate set, if we aim to reconstruct the original network, will consist of all possible pairs of opaque nodes. The problem is that this set will almost never be balanced, but on the contrary it will be severely biased towards false links.

One of the characteristics of real world networks is their sparsity: only a few links exists among all possible pairs of nodes. In fact, one the conditions assumed to explain the scale-free property of social networks, it is that the average degree remains constant [14]. This alone implies that while the number of possible links grows quadratically with the number of nodes, the number of actual links grows only linearly, so the larger a network gets, the more sparse it will be. This effect is easier to understand considering the case of the social site Facebook. Facebook has nowadays around 1.3 billion users but its use is severely limited in China, so the country is virtually absent of the social network. Consider the case that all limitations are removed, and suddenly 500 million new users from China join Facebook. Will that imply the length of friend lists of the original 1.3 billion will significantly increase? Probably not, and that is why current link probability in Facebook is about 1 per 5 million.

The sparsity associated to network growth might seem obvious nowadays, but it was not so clear by the end of the 1990s, when the Metcalf's law, named after the inventor of the Ethernet protocol, was inspiring business plans ensuring the utility of a network was proportional to the square of the number of nodes. Those business plans ultimately inflated the first *.com* bubble which burst in 2001. Thanks to examples like Facebook, nowadays

is known that while at the beginning a *local area* network might contains interactions between all of its parts, that is no longer the case when the system scales up.

Unluckily to the purpose of this section, Metcalf's law does not rule in the scenario considered, so that the training sets will be highly unbalanced. The discussion focuses in the recovery of links in an opaque nodes scenario where the networks are sparse.

Problem description

Call records from a mobile carrier are provided. These records (commonly known as Call Detail Records or CDRs) contain caller and callee phone numbers, time stamps and call durations. Records provided cover a month period. Each call or text produces a record. The carrier has access to all CDRs for calls placed among its customers, as well as communications between its customers and the rest of mobile phone users. However, the carrier lacks information of the calls placed between non-customers, thus defining a opaque nodes scenario where the non-customers play the role of opaques. The problem consists precisely in trying to infer the links between the non-customers.

The country where the carrier operates has reached mobile saturation (there are more phone lines that inhabitants) and there are approximately 50 million phone lines active in the country. The operator accounts for a 20% market share. To the extent of this text, we will consider the country has two operators: a Green operator whose records are available and a Red operator whose CDRs remain unknown.

The data provided by the carrier contains a partition of $300,000$ nodes, extracted by applying the Louvain method [18] in order to extract a social community. The network is built as an undirected graph where a link is included if there are at least 5 calls placed between the two users during the one month observation period.

An additional problem faced in this scenario is that a proper validation set is not known. To overcome this problem, the dataset is used to build two different networks.

- CG_ALL: is a network containing all provided data. The problem in this dataset is that if we remove links in order to predict them afterwards, it is likely that the algorithm will predict links between Red's nodes, whose existence is not possible to verify.

- CG_GREEN: is a subgraph induced by the Green's nodes. In this graph, all links are known and thus it would be feasible to run simulations similar to those presented in the previous section. However, since not all nodes are present, it is likely that relevant information is

	Enron	CG_ALL	CG_GREEN
$\|V\|$	33696	353206	111333
$\|E\|$	180811	1274815	287938
Directed (reciprocity)	No	No	No
$\langle k \rangle$	10.731	7.2185	5.17
Clustering (rel. Erdös)	0.085 (265.1)	0.049 (2682.6)	0.130 (2230.2)

Table 2.7: Network characteristics of the three opaque nodes scenarios.

missing. For example, consider the case of a group of 10 friends forming a clique in the complete network, but only 2 of them are Green's. Using CG_GREEN, the link between them would be really difficult to predict, because common friends have been removed.

In order to tackle the lack of a validation set for the mobile phone network, through this section auxiliary networks will be used. On the one hand, synthetically generated networks, following Barabási and Erdös models will be used to evaluate which one behaves more similar to the phone network. On the other hand, the e-mail communication network from the corporate Enron, published and analyzed in [76], will be used because of its topological and dimensional similarity to the phone network under study. Table 2.7 presents some characteristics of all three networks, showing that both are undirected networks and both present intense clustering.

Prediction performance on balanced sets

The first step when facing the link prediction in the opaque nodes scenario, is to check if the techniques developed in the previous section are still valid in the test networks considered. Due to the undirected nature of these networks, the number of features employed is reduced from 12 to 7[8].

The ROC for the Enron network is presented in 2.12, including further machine learning techniques such as decision trees or support vector machines. The specifics of each method are beyond the scope of this work, because the goal when including them was to verify that the accuracy reached by prediction came from informational limits instead of algorithmic limits. It can be observed that the results are similar to those from the previous section, with $AUC = 0.996$ for most of the methods. In CG_ALL and CG_GREEN performance is smaller, with AUC around 0.8, but this poorer performance was to be expected because of the reasons explained when describing the datasets.

[8]Note that, apart from in and out degrees, hub and authority scores are also equal for undirected networks.

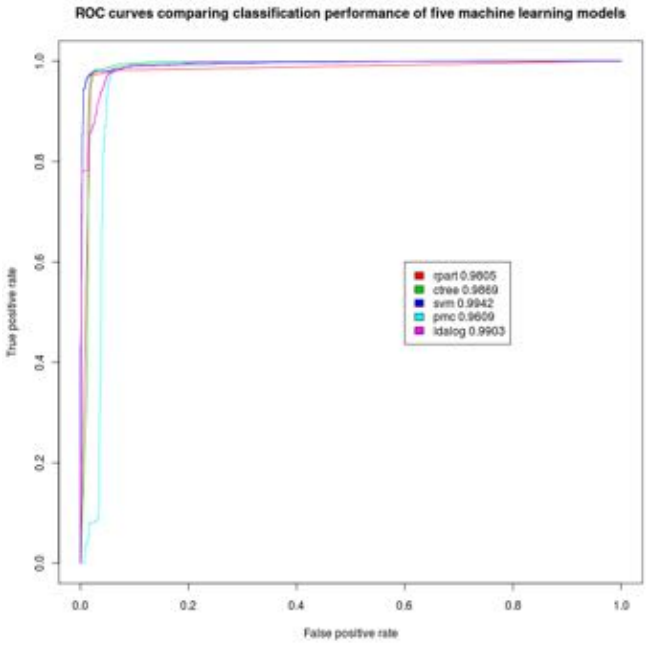


Figure 2.12: Receiver Operation Characteristic for the Enron network

Estimation of the number of observed nodes

It is important to note that in this opaque nodes scenario, if there are customers of the Red operator who never talk to any user of the Green operator, those nodes will not appear in any call graph generated by Green. Thus, the next task when trying to complete the mobile network is trying to estimate the number of nodes ignored by Green. In order to do that, the following experiment is performed.

1. Start with a certain network $G(V, E)$
2. Assign the label Green to a fraction r of the nodes, where r represents the market share for the Green operator, thus generating the V^r_{green} set, so that $|V|r = |V^r_{green}|$. All remaining nodes are assigned to the Red operator.
3. Obtain the N grade neighborhood of the nodes included in V^r_{green}, denoting it $V^{r,N}_{green}$.
4. Compute the fraction of observed nodes $t_{r,N}=\frac{|V^{N,r}_{green}|}{|V|}$.
5. Repeat steps 2-4 and compute average and standard deviation for $t_{r,N}$

It is interesting to point out that in all simulations presented here, the roles of Green and Red have been assigned purely at random (step 2), which

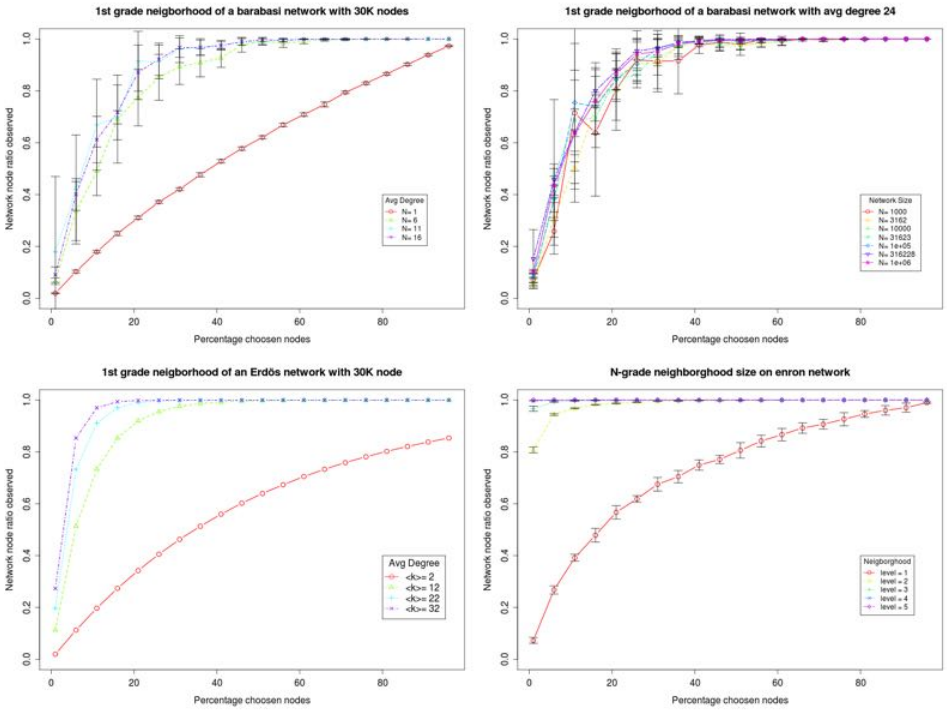


Figure 2.13: Fraction of observed nodes depending on the percentage of green nodes in the network for different network topologies. Lines present results averaged over 20 runs. Vertical bars present the standard deviation

means any kind of homophily in operator choice is neglected. While it is likely that this is not a realistic scenario, after extensive search it has not been possible to find any measure of operator homophily in the literature, mostly due to the fact that anonymization processes make very difficult to cross data from different carriers.

In order to test the node visibility an operator has, a series of simulations are performed, in order to understand how several parameters affect the visibility. Results are presented in figure 2.13. In the first graph average degree of the Barabási-Albert (BA) model is varied, concluding that adding links beyond 12 does not significantly increase visibility. In the second graph, results show how different sizes of the networks do not play a crucial role in this experiment. Both BA experiments present a high variability when only a few (0-30%) nodes are chosen. This is due to the existence of very high connectivity *hub nodes* whose degree is so high that just for choosing one of them the neighborhood size of the entire set is drastically increased. Next graph shows the simulation results for an Erdös-Rényi network where degree saturation effect also occurs. At last, Enron network results are presented, considering several grades for the neighborhood.

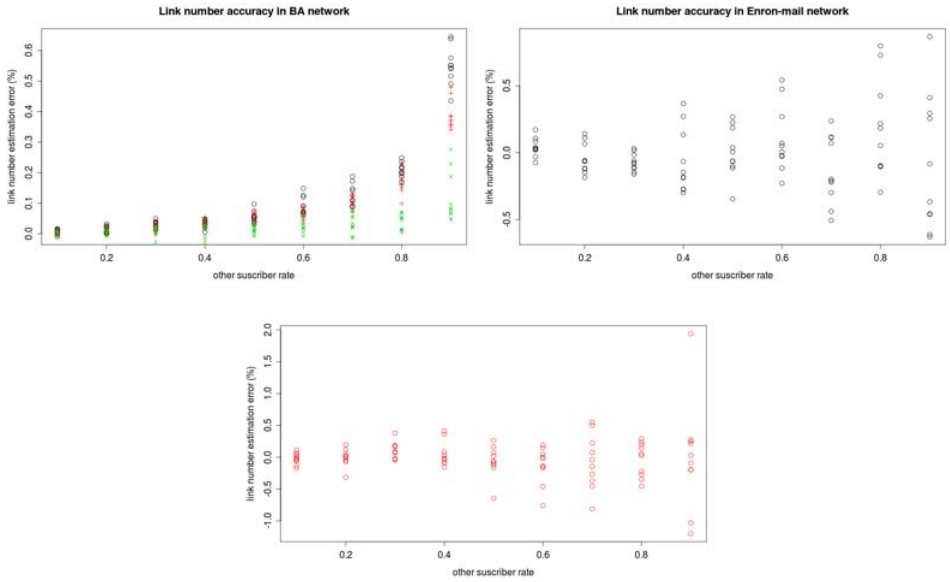


Figure 2.14: Error estimation in the number of links. Each data point presents a run.

Estimating the number of missing links

Once the number of observed nodes in the network is known, the next step consists of predicting the number of links in the entire network, so that it is possible to know how many new links should be added for later prediction stages.

To do so a simple assumption will be made: users of the Red operator have a similar average degree than Green's users. Since the exact degree of Green users is known (all calls they place or receive appear on the available CDRs), it is possible to just extrapolate that number to the whole network.

Results in figure 2.14 show that the assumption turns out to be a very accurate one, so the estimations present low error even for low values of market share for the Green operator.

Candidate filtering and final prediction accuracy

As already mentioned in the presentation of this problem, the biggest challenge is about feeding the prediction mechanism with a small number of link candidates. The trivial solution, which consists of evaluating $|V_{red}||V_{red}-1|$ candidates has to be discarded for any network exceeding the tens of thousands of nodes.

In this scenario, the proposal is to take advantage of descriptive models to generate networks. Precisely, there are models [56, 134, 38] which consider that a large fraction of the non-trivial structure commonly found in networks

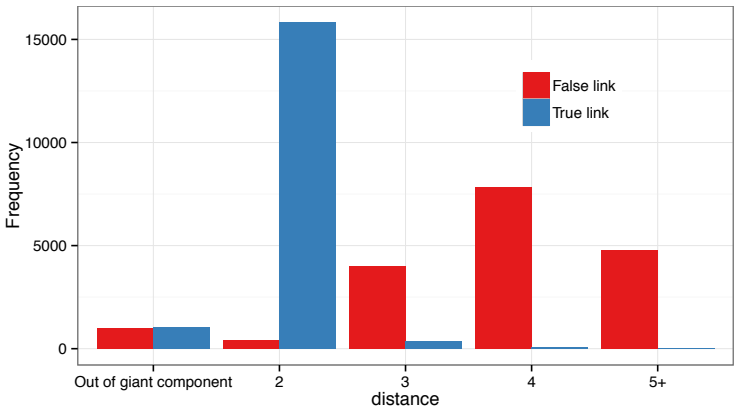


Figure 2.15: Distance distribution for true and false links after removal of links between 30% of nodes in the Enron network

can be explained just by considering that when a node joins a network, it finds its neighbors just by randomly walking the network from a randomly chosen node. Applied to social network, this implies that when someone reaches a new place, he knows one or several people, and they introduce him to his acquaintances until stable groups get configured.

Before trying if this kind of approach would work, a preliminary analysis is done. If the candidate filtering is going to be based on random walks, then most of the true links should be within short distance in the graph after the edges removal. This is tested by removing links in the Enron between 30% of the nodes (the election of such nodes being performed purely at random). As show in figure 2.15, true and false links present different distance distribution, with true links showing much shorter distance in the remaining graph. It is important to note any applied procedure will be unable to recover those links which are adjacent to a node that after the edge removal is not anymore in the giant component (they correspond to the non-observed part of the network previously discussed. Our simulations estimate in this scenario about 5% of links would become unrecoverable.

After this preliminary analysis, the next step consists of measuring the performance of random walks for candidate selection. In the scenario considered, a purely random choice of node pairs would produce 3 true links for every $10,000$ pairs of nodes, so any procedure generating more true links would prove itself useful for this candidate filtering task.

To characterize such performance, a Monte Carlo scheme has been used, where 5 million random walks of length 2 have been simulated, using as origin a node from the Red operator. In case the random walk finishes in another Red's user the pair is added into the candidate lists. As evaluation metrics, percentage of found links and ratio true vs candidates are

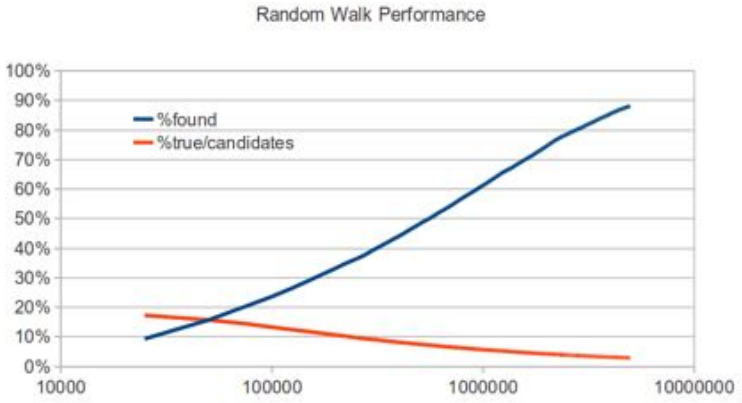


Figure 2.16: Distance distribution for true and false links after removal of links between 30% of nodes in the Enron network

considered. A pair is only added to the list if it was not added before.

Results from such Monte Carlo experiment (figure 2.16) show an interesting behavior: while the more random walks are performed the larger the number of removed links found, the ratio between found and candidates is reduced. Note that the x axis in logarithmic scale, which means both metrics evolve slowly. For example, after 1 million random walks simulated, the candidate set contains 60% of all removed edges, and the rate of true links in the candidate set is about 3.75%, which even if it seems low is still 125 times higher than the random choice. Depending of the application and the performance of the prediction machinery it might be better to choose a different size for the candidate set.

Prediction stage

The first attempted scenario is created from the Enron network by assigning 50% opaque nodes and removing all links between them. In such scenario, around 40,000 links have to be recovered from about 280 million possible links. By using the random walking technique previously described, a candidate set of 1 million is extracted, where 60% of the original links are available, and the true/candidate ratio is around 3%.

Once a reduced candidate set is available, feature extraction is done as in section 2.2.6, and several classifiers are trained to predict the sample. Unluckily, results are unsuccessful: almost independently of the prediction algorithm used, the resulting predictions consist of predicting *no link* for all candidates. Because the dataset is so unbalanced, such trivial prediction reaches an accuracy of 97% (there are only 3% of true links in the set) and because it is very difficult to increase such accuracy, all of the predictors mimic the constant function. However, this does not work for the purposes

of link recovery in an opaque nodes scenario, because basically the result would be adding no links to the network.

Interestingly, there are a number of insights previously obtained from the data that can be used overcome these problems. First, the number of links can be estimated very precisely by using the average degree hypothesis explained in section 2.2.6. This changes the problem from a classic classification to a ranking problem: for example, if our estimation is that $1,000$ true links have been removed, the problem is trying to find the $1,000$ pairs from the candidate which are more likely to be connected in the original graph. The approach just presented does not take any advantage of this information, which is arguably very valuable.

On the other hand, careful observation of figure 2.15 unveils an interesting story: the slow increase (logarithmic) of the found links (blue line), implies that some origin-destination pairs occur very often when randomly walking the network, so no new pairs are added to the candidate set. The figure also indicates that such common origin-destination pairs are actually more likely to be connected in the original network,and that is why the ratio (orange line) is a decreasing function. Thus the probability of a random walker that leaves from a node u reaching another node v emerges as relevant feature. A feasible way in this scenario to compute such probability is to use the transition matrix T which is obtained from the graph's adjacency matrix A as

$$T_{uv} = \frac{A_{uv}}{k_u}$$

where k_u represents the degree of node u. Once T is obtained, elements in T^l represent the desired probability for random walks of length l.

At last, the problem of unbalanced datasets had been widely studied in machine learning literature (see [101] for a recent survey in the topic) and several ways overcome the problem have been proposed. One of the most simple ones, commonly referred as *downsizing,* consist of training with a reduced balanced dataset which is created by randomly sampling from the most common class. Such approach turns out to be efficient in this scenario.

Compiling all this ideas, a new prediction schema is then tested, whose main steps are:

1. Computing the rows and columns corresponding to opaque nodes in T^2 and T^3. The non-zero elements out of the diagonal of these two matrix will become the candidate set.

2. Extracting features for each of the sample the candidate set, including random walks probabilities which are already available in T^2 and T^3.

3. Training using balanced sets from the candidate set got by downsizing the false links class.

4. Estimating the number of missing links M

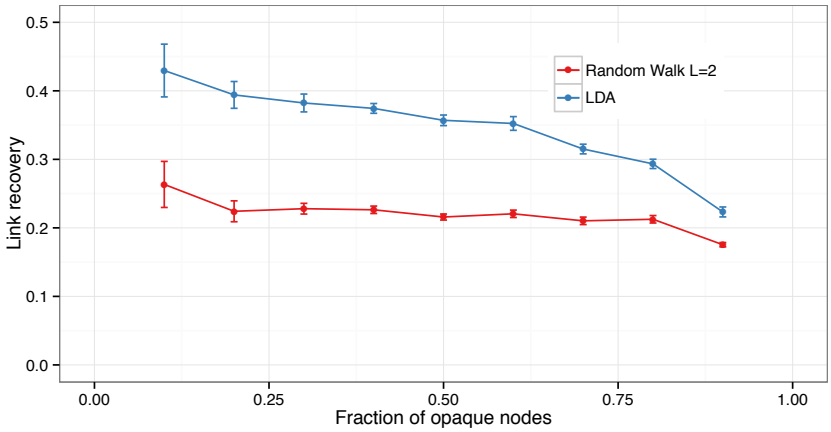


Figure 2.17: Fraction of links recovered for different amount of opaque nodes in the scenario. Error bars represent the standard error of the mean value presented after 10 runs.

5. Obtaining scores for the true links class, and gathering the top M pairs from the candidate set.

The application of this algorithm allows for up to 40% link recovery in the scenario with 50% opaque nodes. While several machine learning techniques have been tested, none of the non-linear methods significantly outperforms LDA. Interestingly, LDA coefficients indicate that most of the prediction capability comes from the probability of random walk of length 2 to happen between the pair, which is by itself a good enough predictor.

In order to generalize the results, the prediction experiment has been simulated considering scenarios with 10% to 90% opaque nodes. Each of the scenarios is run 10 times and the results are presented in figure 2.17, where error bars represent the standard error of the mean value found for link recovery. One of the interesting results shown in the experiment is that the percentage of opaque nodes (which corresponds to the complementary of operator's market share in the mobile telephony scenario) does not dramatically affect the performance of the link recovery, specially in the case with the single score from random walks.

2.3 Attribute prediction in an opaque nodes scenario

The second aspect to deal with at the opaque nodes scenario, is the ability to predict attributes for those opaque nodes. The node attribute is a paradigm flexible enough to accommodate many different aspects of the real world: if

we think of a social network, those attributes could be the age and gender of the person represented by each node (this will actually be the case in the following discussion) but could also represent the political orientation, company where he or she works, number of children, education level, native language or basically any single data that can be associated with a particular node.

The most immediate network property to take advantage of when trying to predict attributes of opaque nodes would be homophily: since people tend to create social ties with others similar to them, it would be the case that if an opaque node is known to be connected primarily to French speaking white nodes, chances are that the opaque node represents a person who is fluent in French. However, the contrary also occurs in certain networks: for example, sexual networks tend to display disassortativity not only in gender [16], but also in immune systems [164].

Beyond assortative and disassortative mixing, there is additional information in the network relevant for attribute predicted. Part of it is related to links: since in the opaque nodes the information associated to a link between the opaque node and a white node is available, it would be possible to take use such information. For example if call logs show an opaque node placing calls mainly during business hours, a fair assumption would be that the opaque node represents a professional mobile phone user. At a higher level, the entire ego-network around the opaque node may present characteristics that could be used as features in a machine learning scheme: for example, it will be shown in this section that the number of 4-star motifs in the ego-network is correlated with the age group of the ego.

The following discussion focuses on the inference of gender and age in a particular dataset. However there are other scenarios topologically equivalent where the same approach could be used. The following list describes some of them.

- Facebook applications: if an application is accepted by one user, this user becomes a well-known node, and all its neighbors become potential opaque nodes.

- Twitter private profiles: these profiles would be the opaque nodes, whose ego-networks can be built using their public profile neighbors.

- Mobile phone network: in mobile communications, carriers have information about their customers, and they also have communication records of any interaction between their customers and the rest of the users in the mobile phone network. This way it is possible to build, using records from only one carrier, ego-networks where non-subscribers are the opaque nodes, and subscribers are the well-known nodes. This scenario will be the case test for this discussion.

2.3.1 Data description and preparation

Anonymized data for this section was provided by Orange, France Telecom Spain, consisting of two data-sets:

- Call Detail Records (CDR), contain information about customers interactions via phone calls and SMSs. For each interaction, two anonymized user identifiers and a time-stamp are provided. For phone calls, duration is also available. These data include interaction between subscribers during a continuous 14 week period. According to regulator data, for the observed period, the carrier managed 20% of mobile lines in the country, where mobile phone had already reached market saturation (1.16 mobile lines per person). The data contain records for 2.2 billion interactions among 11 million users.

- User Data (UD): provide age and gender for 8 million users, identified by consistent anonymized hashes.

In order to aggregate this information in a lossless way, records were grouped by relationship (link). For phone calls, along with first and last interaction timestamps, 3 vectors per relationship were built:

- Duration vector: contains durations (in seconds) of all phone calls between the two users.

- Inter-event vector: contains inter-event time (in minutes). If the previous vector has length N, inter-event vector has length $N - 1$.

- Direction: binary vector providing caller and receiver roles for the interaction. If the relation is defined as "A-B", this vector has ones when the caller is A and zeros otherwise.

Once aggregated, 168 million different relationships were found. The next step consisted in filtering meaningful relationships. Many CDR records belong to corporate phone lines for which it does not make sense to talk about user age or gender, since usually there are several people behind each of those specific numbers. In this research, the criterion employed to filter out this kind of interactions was the criterion proposed in a seminal paper by Onella et al. [120]: only relationships with at least one call in each direction of the communication were considered. This way, 40 % of originally obtained relations were eliminated from the study.

After having chosen these mutual phone calls relationships, SMSs, inter-event and direction vectors for those relationships were built. The reason for using calls to define meaningful relationships instead of using SMS records, is that the mutual strategy does not work so well for SMS; this is so because of the increasing number of value added services which involve messaging in both directions between the user and the corresponding automatic services.

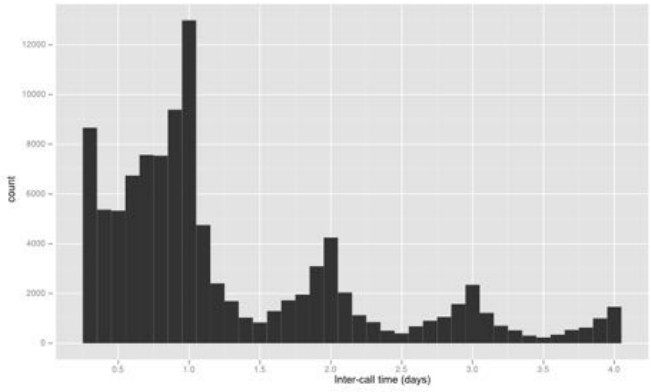


Figure 2.18: Inter-call time distribution for a 10000 links sample.

2.3.2 Exploratory analysis and learning approach

The resulting network presents characteristics similar to those found in previous works on mobile phone networks [120, 57]: high clustering coefficient, short diameter and a long-tail degree distribution.

From an exploratory perspective, an interesting behavior was found in the inter-event time distribution. Basic network engineering does usually assume that the time between two calls (within a big enough population) can be described by a exponential distribution, and that is the original assumption for many techniques used to properly dimension communication networks [139]. However, recent studies [110, 133] have proved that this exponential behavior is not present in the individual level: certain events (for example an incoming call) make the user to immediately initiate a communication burst. In our research, once the inter-event vectors for links were compiled, the distribution at this link level was studied.

Inter-call vectors for 10000 randomly chosen links were concatenated, producing 105174 inter-call time samples. Figure 2.18 shows an histogram of those samples in the time interval from 6 hours to 4 days. One can easily identify that the distribution is peaked every 24 hours. This means that if two users A and B talk to each other by phone today at 10am, it is much more likely for their next conversation to happen the day after tomorrow at 10am than tomorrow at 6pm. This fact contradicts the exponential distribution monotonic decay, proving that, at link level, there is no exponential behavior either.

After the exploratory analysis, a methodology to address the prediction problem was designed, consisting of two separate steps:

- Single link approach: given a relationship between nodes A and B, features are extracted from both A's attributes and link available data in order to predict age and gender for B.

- Ego-network: results from the previous step are used, so that for each link leading to B, there is a prediction about B's gender and age. These link-level predictions, together with ego-network features, are then used to predict attributes of B.

2.3.3 Single link approach

As previously stated, this section is aimed to perform a prediction of user attributes (gender, age) using information related to only one relationship in which the user is present. Three main sources of information are available in this approach: SMS records, Call records, and gender/age from the other user in the link.

As it will be shown later, a large part of prediction capability will come from the gender and age information about the other user, because a high level of homophily is found in the network.

Apart from taking advantage of homophily, the manner people communicate also depends on their age and gender, as it was shown by Stoica et al. in [148]. That research found out that it is possible to cluster users into a number of groups according to their communication patterns (unsupervised learning). Then it was shown that some user attributes, such as age, were correlated with the group membership of the user. Although the phenomenon leading to this research is exactly the same, our approach will be grounded on supervised learning (precisely, a classification problem). On the other hand, a larger number of features will be used for machine learning, specially those related to communication dynamics whose relevance has been recently pointed out.

To run this experiment, 9860 links were randomly chosen among those whose both users data (gender and age) were available.

SMS and call metrics

As it has been already mentioned, SMS are one of our three sources of link related data. For each relationship 3 vectors are available which contain all the information associated with direction and communication times. Seminal research on mobile social networks [120] did commonly characterize the link using only quantitative information, such as the number of interactions or total conversation time. Later is was pointed out [110] that, due to the bursty pattern of individual communication, quantitative information may not be enough to describe the nature of the link: 50 messages a week in a relationship may not be more relevant than 10 messages during a month. Using these criteria, new selected metrics from SMS have been calculated:

- Number of SMS during the observation period.
- Mean time between messages.

- Conversation length (from first message to last).
- Variation coefficient (average/standard deviation) for inter-event time.
- Fraction of calls during weekend.
- Fraction of calls during work hours.
- Peak hour of the conversation (0-23).
- Reciprocity: in a link A-B, where B is the opaque node, fraction of messages sent by A. This is the only asymmetric feature for SMS data.

For phone calls, all previous features are extracted and the feature set is completed with average call duration, so that in total each link is characterized by 7 SMS features and 8 call features.

Gender prediction

Figures 2.19 and 2.20 show the kernel density functions[9] for the metrics described above, aggregated by gender. In general, few gender differences can be found by looking at those graphs; nevertheless, there are a couple of interesting facts to be pointed out. The average call length seems to be higher if user B is female. The median[10] of call duration if B is a female is 89.67 seconds and 75.06 seconds if B is a male. On the other hand, if there is a user sending 20% or less of the messages in the relationship, it is more likely this person to be a man.

Machine learning procedure

The problem is defined as a binary classification. In order to gather an accurate idea of the data quality regarding gender prediction, several learning schemes are tested: Linear Discriminant Analysis (LDA), Decision Trees (Tree), Multilayer Perceptron (Nnet), Bagging and Support Vector Machines (SVM)[11]. In order to improve performance quality (mostly for LDA, which cannot perform non-linear transformations), long tail distributed metrics are logarithmized (similar to what was done with the degree in the link prediction problem) and, after that, all metrics are standardized (zero mean and unit variance).

[9] As available in the *density* function in R, using Gaussian kernels. See `https://stat.ethz.ch/R-manual/R-devel/library/stats/html/density.html` for details.

[10] Due to the long tail distribution on call duration, it is more robust to use the median to characterize group differences, since the mean is severely biased by samples in the tail.

[11] Implementations used in this article were the following R-packages: MASS (LDA), rpart (Tree), randomForest (Forest), nnet (Multilayer Perceptron), ipred (bagging) and e1071 (SVM).

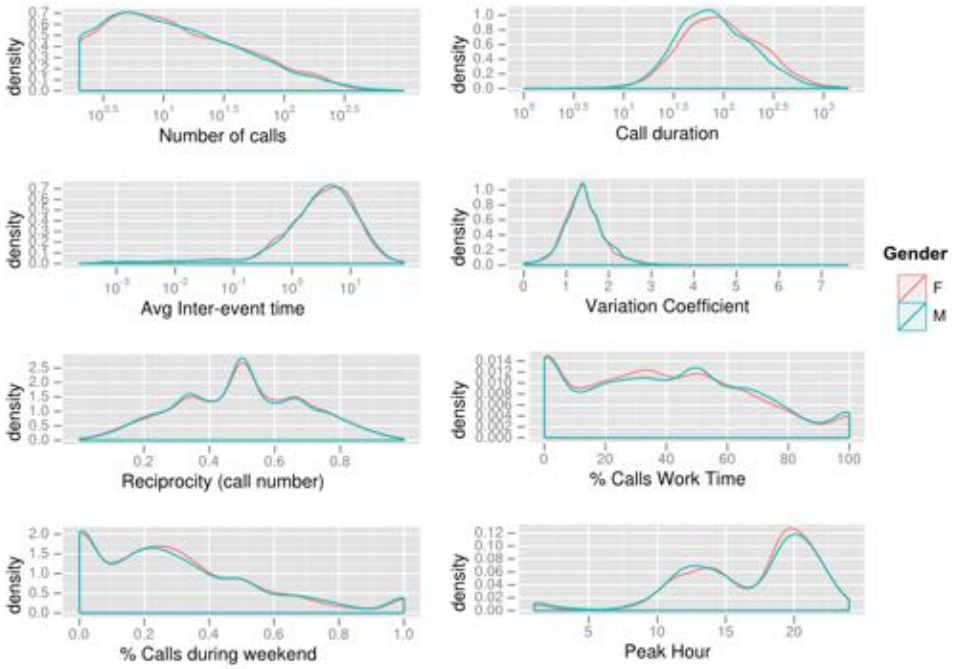


Figure 2.19: Density functions for call metrics by gender.

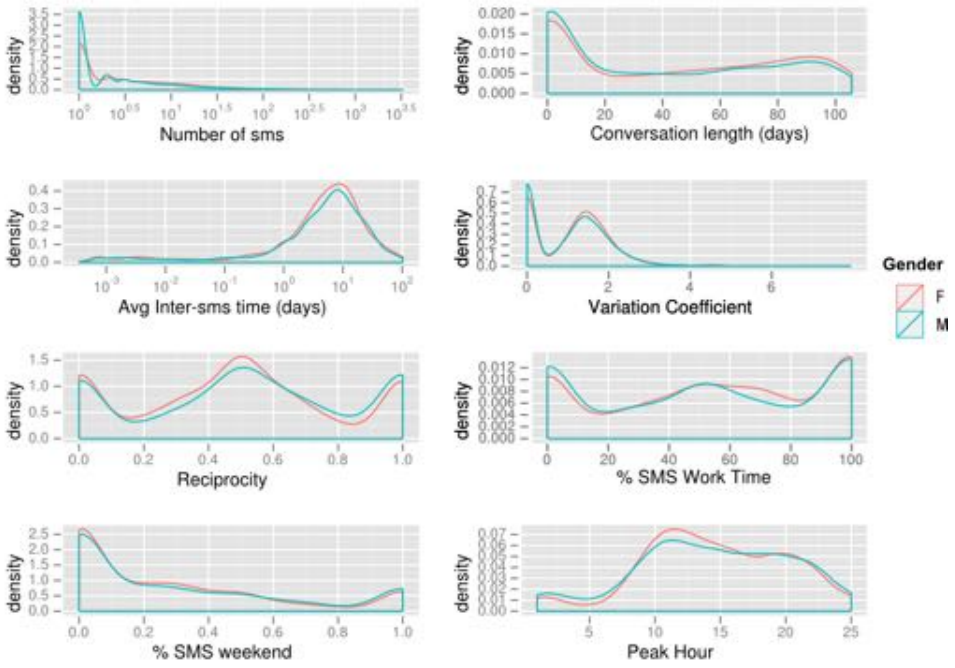


Figure 2.20: Density functions for SMS metrics by gender.

	Random	LDA	Tree	Nnet	Bagging	SVM
SMS	0.5	0.527	0.520	0.533	0.520	0.532
SMS + Calls	0.5	0.5403	0.5443	0.550	0.534	0.550
SMS + Calls + User-data	0.5	0.591	0.6044	0.595	0.579	0.605

Table 2.8: Gender prediction accuracy using isolated link information.

Once the data are ready, predictions are obtained from a 10-fold cross validation scheme, ensuring every sample belongs to both training and test sets. Data are provided to the learning machinery in 3 different steps: first only SMS data are provided, then call data are also included, and finally user data are included as well. This whole procedure is summarized in table 2.8.

The results show SMS and call metrics an small but significant prediction capability compared to the random baseline. When user data is included, prediction capability increases. This means that homophily definitely plays a role in this problem. . On the other hand, the capability of splitting non linearly separable sets does not seem to help at all, since LDA performance is almost the same as Nnet or SVM.

Age Prediction

For prediction purposes, the ages of the users were binned into 6 different age segments. These segments were chosen according to the age distribution, so that every segment has the same number of users. This way the age regression problem is transformed in a multi-class classification problem with balanced classes.

After redefining the problem as a classification one, the same methodology discussed in section 2.3.3 can be applied. Figures 2.21 and 2.22 show the density functions for different age groups. An exploratory study shows that people over 30 years old usually call more often during work time. Concerning the amount of SMS in the relationship, it is interesting to note that density functions are sorted, meaning that the elder a person is the fewer texts he/she sends. This behavior was also observed in [148]. However, the statement just made (younger implies more SMS) has an exception in our data which does not show up in any previous study: youngest people (18-26) text a little bit less than people in the next segment. We propose an explanation for that fact: according reports available for the time when the data was collected (late 2010) [47], the youngest people (18-25) acquisition of mobile Internet flat rates is higher than in any other age segment. In the same report it is stated that the increase of mobile data plans is severely correlated with the decrease of SMS usage. Hence, we conclude that the observed "lack of messages" among youngest users is probably masked by the replacement of IP based messaging services (e.g *whatsapp*) whose traces

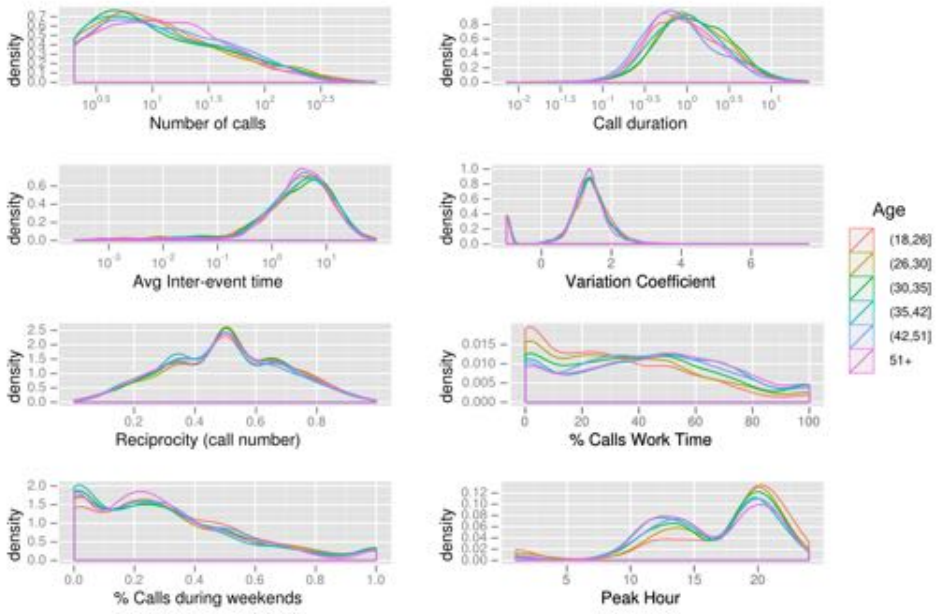


Figure 2.21: Density functions for call metrics by age group.

are not recorded by the carrier. If the same experiment would be tried again by the time of writing this document (late 2014), SMS information would probably become irrelevant: while in 2010 young people were early adopters for IP based messaging services, nowadays they are commonly used by almost every user, and in fact the overall number of SMS sent has drastically decreased during the last 3 years [12].

Table 2.9 shows the accuracy results for this classification problem. The results show that there is a prediction capability on communication metrics specially if SMS data are included. However, this prediction capability is outperformed if user data (precisely, user age) are included. Age homophily in mobile phone communications is so intense that all five classification techniques mimic the identity function on user age like the best possible classification scheme. The reason for this fact can be observed in figure 2.23, which represents a scatter density plot for ages in the same link. It is straightforward to check that the probability for a user A to be in the same ages than the user B is extremely high (dark colors in the diagonal of the plot). Interestingly, users under 30 present a second smaller maximum for a 30 years age difference, probably due parents-children relationships.

[12] "Los PCs y SMS en caída libre por el auge de tabletas y aplicaciones móviles", available at `http://www.adslzone.net/article6684-los-pcs-y-sms-en-caida-libre-por-el-auge-de-tabletas-y-aplicaciones-moviles.html`

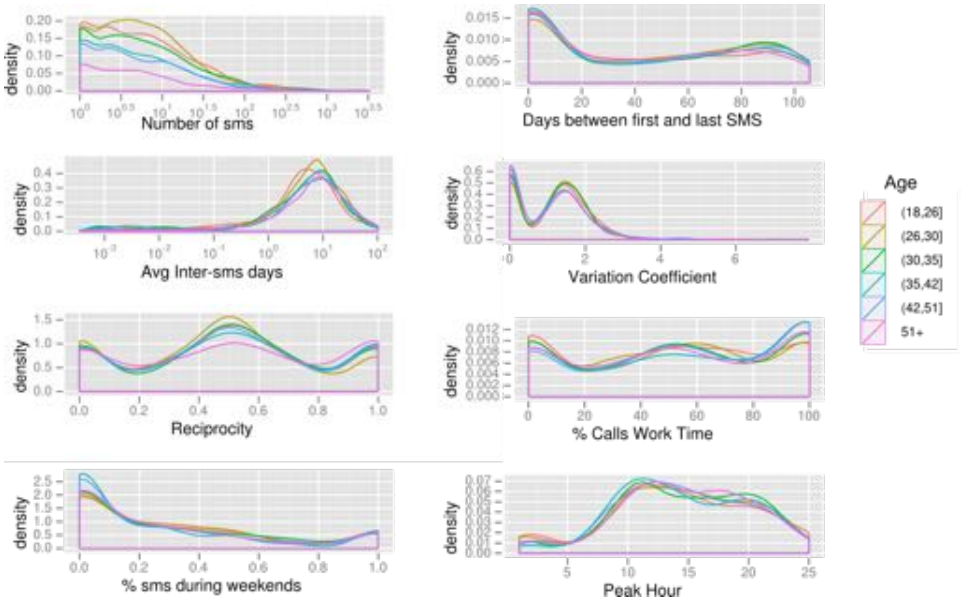


Figure 2.22: Density functions for SMS metrics by age group.

	Random	LDA	Forest[13]	Nnet	Bagging	SVM
Calls	0.1667	0.1997	0.1849	0.2194	0.2020	0.2156
SMS+Calls	0.1667	0.2340	0.2273	0.2410	0.2192	0.2395
SMS + Calls + User-data	0.1667	0.3907	0.4095	0.4027	0.3904	0.4021

Table 2.9: Age prediction accuracy in using isolated link information.

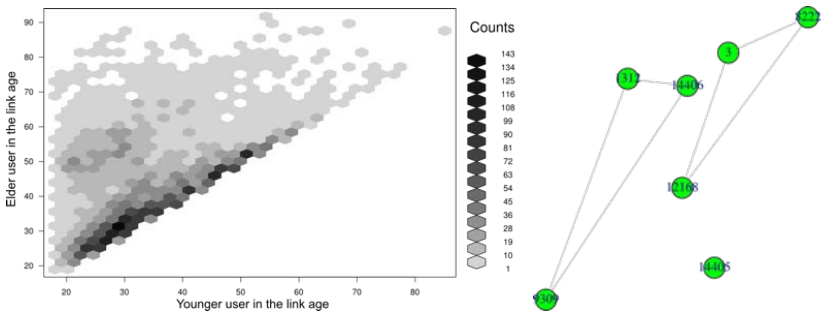


Figure 2.23: Left: Age homophily in communication links. Right: ego network: the removal of the opaque node may lead to a not connected graph.

2.3.4 Ego-network approach

In the opaque nodes scenario, not only the links between the opaque and the white nodes are available, but also all links between white nodes. Thus, there network around the opaque node is available, and therefore relevant features to feed the learning scheme can be extracted

This implies that the ego-network, formed by those nodes adjacent to the opaque node, can have triangles or other particular subgraphs, usually referred as *motifs*. On the other hand, connected components might be smaller or larger in the ego-network, and all these network features provide predictive power when it comes to gender and prediction. For this reason, this research on multi-link prediction is not only about accumulating information from the isolated stage, but also including information of the network structure around the opaque node.

Network metrics

Due to the size and inherent complexity of their analysis, real-world big networks (mobile phones, social connections online...) are not usually analyzed using global information. A very common approach to network analysis is the extraction of a certain N-grade neighborhood around a set of nodes of interest. Among these local neighborhoods, the one which has been more commonly studied is the ego-centered network. This graph includes, for a certain node, all its neighbors and the connections among them. It does not include the center node itself, neither the connections from it to the neighbors, so it is possible a ego-network not being a connected graph, as we can see in figure 2.23.

Once the subject under study is defined as the ego, a number of features are defined. Apart from quantitative information, such as the number of nodes and edges, some small structures are analyzed. It has been proved that some subgraphs show up much more often than in a purely random network. These subgraphs are called motifs, and their importance in biological networks has been already stressed: the appearance of some kind of motifs is related to some specific function within the cell. In social communications networks, the appearance of motifs has also been remarked as important for some tasks [147, 148, 171].

According to these guidelines the following metrics were used as network features:

- Number of nodes
- Number of edges
- Isolated node count.
- Number of V-motifs (unclosed triangles)

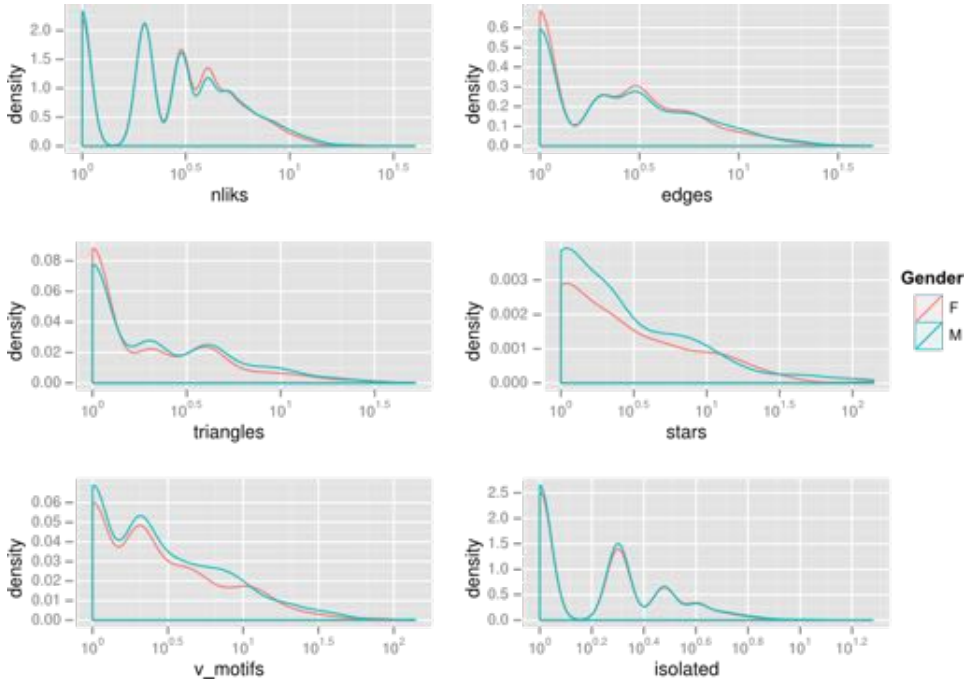


Figure 2.24: Density functions for network metrics by gender.

- Number of closed triangles
- Number of 4-star motifs (one node connected to 3).

For this experiment, data from 5670 subscribers was randomly chosen, and their neighborhood information was gathered. This way, a total number of 22098 users and 50377 relationships were analyzed for prediction purposes.

Gender Prediction

Figure 2.24 shows that the main difference between genders, regarding network features, is that women seem more likely to have triangles (less stars) in the ego-network. In order to perform final learning for gender classification, results from link level prediction are grouped in a gender score, whose value is the rate of female predictions for the node under study. For example, if there were 3 links, 2 predictions were male and 1 female, the gender score is 0.33. Therefore, a total of seven features (six network metrics plus gender score) were analyzed.

Accuracy results are shown in table 2.10, which shows that the ego-network approach increases the performance by about 5% compared to predictions using only one link. On the other hand, figure 2.25 shows the Receiver Operating Characteristic (ROC) which shows how the accuracy of

	Random	LDA	Forest	Nnet	Bagging	SVM
Network	0.5	0.532	0.529	0.536	0.519	0.537
Net + Gender Score	0.5	0.630	0.632	0.651	0.642	0.653

Table 2.10: Gender prediction accuracy using ego-network data.

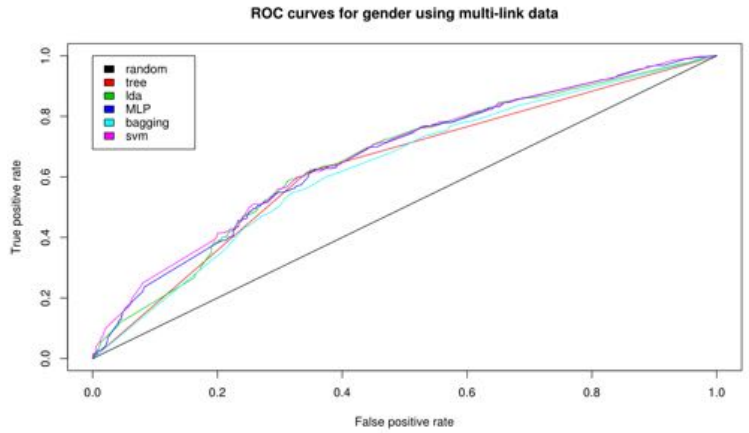


Figure 2.25: ROC curve for different machine learning techniques.

the best techniques (SVM and Multi-Layer Perceptron) is robust to small variations of the selected threshold.

Age prediction

Regarding age, there seems to be a larger diversity in ego-network structure. Figure 2.26 shows the correlation between age and the appearance of certain motifs, specially stars and triangles. For age prediction, isolated link results were included in the experiment by incorporating 6 variables which contain the number of link level predictions for each label. Prediction accuracy results using these 12 variables (6 age scores and 6 network metrics) are shown in table 2.11.

Classification results show that the use of network metrics improves link-level predictions by around 10 %, reaching a final performance three times higher than with a random predictor. In addition, prediction errors usually lead to either the following or the previous age element, as it can be

	Random	LDA	Forest[14]	Nnet	Bagging	SVM
Network	0.1666	0.2108	0.2022	0.2194	0.2030	0.2092
Net + Age Score	0.1666	0.5050	0.4904	0.5082	0.4931	0.5102

Table 2.11: Age prediction accuracy using ego-network data.

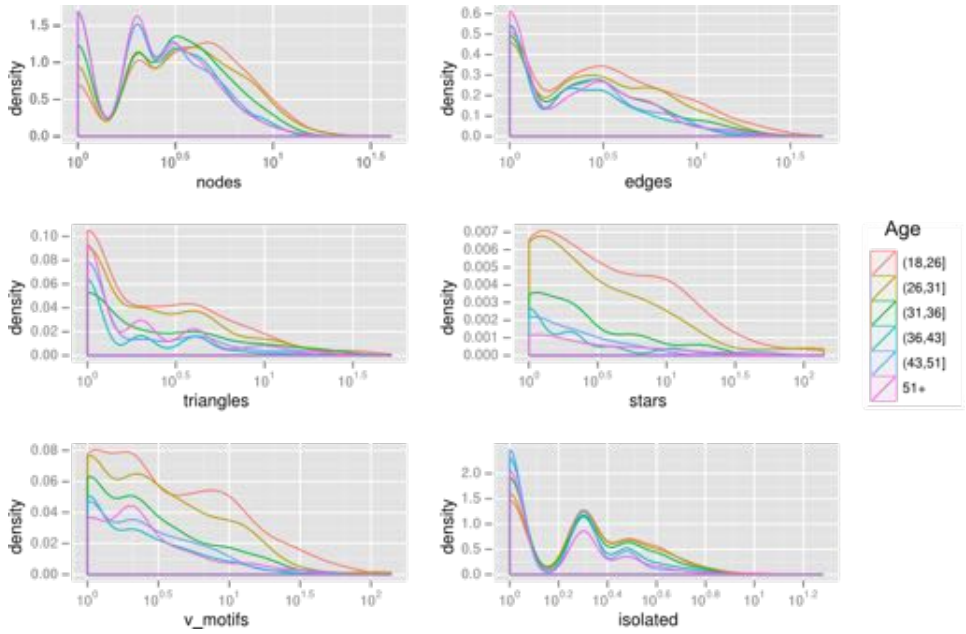


Figure 2.26: Density functions for network metrics by age.

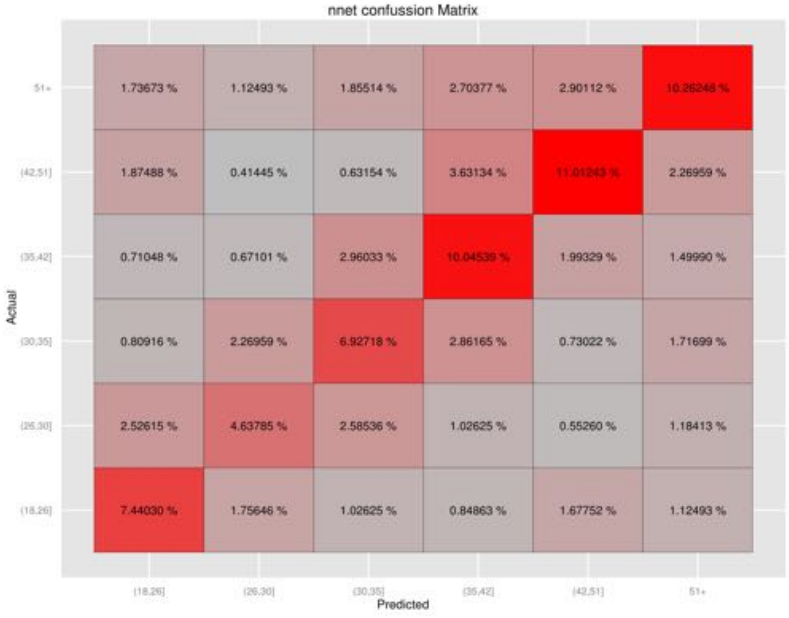


Figure 2.27: Confusion matrix for neural network multi-link classifier.

seen in the confusion matrix of figure 2.27. Note that when turning the regression problem into a classification one, the topology on the age variable was neglected; this could have propitiate that errors would lead to age segments non-contiguous with the correct one. Fortunately, the favorable results discarded this concern.

Chapter 3

Searchability in social networks

One of the main contributions of network science to date has been the concept of *small-worlds*, more popularly known as the *six degrees of separation* theory. It definitely impacted popular culture up to a level only a few discoveries have reached in the last century, such as Eintein's relativity theory or the chaos theory.

In a very personal parallelism, I like to compare the six degrees and the butterfly effect. Both come roughly from the same period of time (late 60s) and both are popular science memes the public has become familiar with. However, most people fail in both cases to understand their implications. Chaos theory is interesting well beyond the existence of sensitivity to initial conditions in dynamic systems that the butterfly example tries to exemplify. In a similar way, the very same first experiment that highlighted that we are all connected through a short social path, proved yet another more striking and interesting fact: we are able to find these paths, even if we lack a general map of the social network. This property of social networks is referred to as *searchability*.

As a telecommunications engineer, familiar with the difficulties of defining proper network topologies and algorithms so that packets can eventually find their target, I felt specially attracted to further exploring this searchability property. In the end, how is it possible that in a remarkably sparse self-organized network with hundreds of millions of nodes, actors with very limited knowledge (only their immediate neighbors) were able to reach an arbitrary target in only six hops?

In this chapter we first present key findings by previous literature, then we describe the mobile phone data available. Later we confirm that our networks exhibit the key properties previously reported in the literature and finally we present the results from simulating different decentralized search strategies in real-world networks built from mobile phone data.

3.1 Half a century of six degrees

The first research about small-worlds was carried out at MIT by Pool and Kochen (although the manuscript circulated among colleagues for over 20 years before publication in 1978 [31]). Lacking any database nearly similar to the actual Facebook or Twitter, they first ran a initial phase consisting of asking participants to write down the names of everyone they interacted with for a period of 100 days. By doing so, they were trying to estimate what nowadays would be designated as the average degree of the social network. The result was that people contacted with around 500 other people on average. According to them, if there would be no common friends, the entire American population (by that time, around 200 million) would be connected in just

$$N = \frac{\log 2 \cdot 10^8}{\log 500} \approx 3.0756$$

hops so the idea of short paths was reasonable.

Being familiar with Pool and Kochen work, Stanley Milgram, by the time a Harvard professor, considered that given the fairly common existence of triangles in the social network (which Milgram referred as *inbreeding*), further evidence of short paths was needed. So in the mid 60s, he was granted a 680 US dollars budget (nowadays equivalent to around 5,000) by the Laboratory of Social Relations at Harvard to test the idea that any two Americans were connected through just a short chain of acquaintances [108].

First, Milgram chose a sample of volunteer participants from Wichita, Kansas who would play as *sources*, and a *target* person in Cambridge, Massachusetts, whose name was Alice and who was the wife of a divinity school student. Each person who volunteered received a letter with some information about the target (name, city and occupation) and the following instructions:

- If you know the target person, please forward this letter to her.
- If you don't know the target person, do not try to contact her directly. Instead mail this folder to a personal acquittance you know on a first-name basis and who you consider that is more likely than you to know the target.
- Write your name and address in the attached document, and keep it in the folder when you forward it.

Note that the third instruction had a double aim: on the one hand it allowed Milgram and his team to analyze the chains, but also prevented participants to forward the folder to people who had it before, therefore avoiding any (potentially endless) loops.

3.1.1 First experimental results

Four days after the folders were sent out to Kansas, a theology lecturer approached Alice on street and delivered her the folder. When his student told Milgram about that, both thought it was probably a mistake and the folder never actually left Boston. When they checked the path travelled by the folder, they found themselves wrong. A source in Kansas, a wheat farmer, had forwarded the folder to an episcopalian minister in his own town, who sent it to the theology lecturer who gave the folder to Alice. The total number of intermediaries needed to connect what were assumed to be opposite poles in the American social network was only two!

After a few weeks Alice received more folders and further results were extracted. Interestingly, no further results about the Kansas experiment, apart from this anecdotal evidence, were ever published in a scientific journal. Even more, 50 years later, Professor Kleinfeld was digging in the Yale's archives and she found an undated paper by Milgram entitled *Results of Communication Project* where it was explained that only 3 of the 60 letters sent out to Kansas where able to complete the chain, and in average the path length was around 9 [75].

The Nebraska study

Probably due to inconclusive results of the Kansas experiment, Milgram carried out a second experiment in collaboration with Travers and the results were published in 1969 [154]. This second experiment is referred as the Nebraska experiment due to the location of the distant sources. Some subtle conditions were changed for this second experiment:

- The target of the experiment was a stockbroker in Boston. The sources were grouped into three different sets: random people from Nebraska recruited from marketing mailing list, blue chip stockholders also in Nebraska and responders to a newspaper advertisement in the Boston Area. The purpose of such grouping was to determine the influence of geographical or social proximity between sources and targets.
- The package sent out to participants was given the appearance of a luxury passport, probably in order to reduce the attrition rate found in the Kansas study.
- At each step, participants were asked to send a prepaid postcard back to Harvard University, therefore some data could be collected even for those chains who did not reach the target.

In total, 296 chains were started by the researchers, and 217 were forwarded by first recipients (27% attrition rate). Out of these 217, 154 originated in Nebraska and the rest in the Boston control group. 64 chains were

completed, 42 of them having origin in Nebraska. This means completion rate was not too different in the Nebraska groups (24% and 31%) than in the Boston group (35%). Average chain length was shorter among the Boston chains (4.6 vs 6.1, distribution is presented in Figure 3.1.

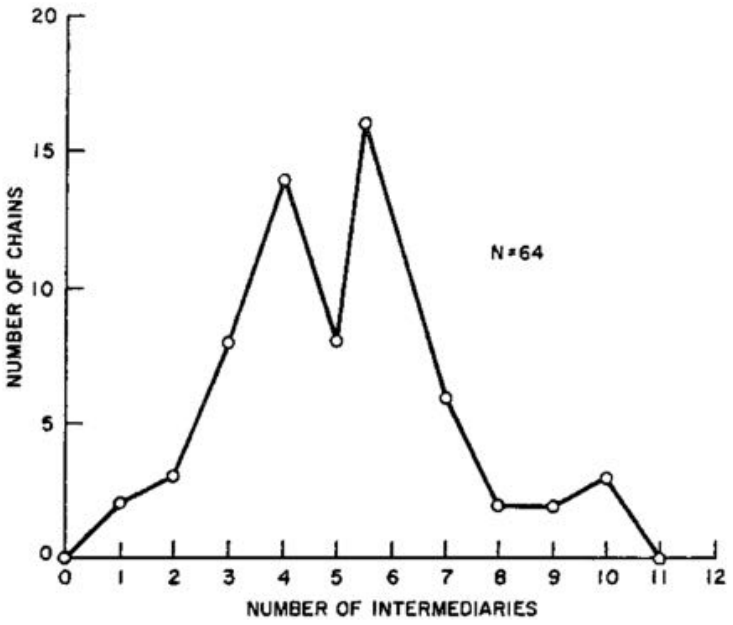


Figure 3.1: Chain length distribution for successful chain in the Nebraska study

Further results were provided in terms of attrition rates and gender homophily in the chains, but it is probably more interesting to highlight the following paragraphs from the original articles in 1967 and 1970, that summarize what the authors thought after analyzing the results:

> Qualitatively, what seems to occur is this. Chains which converged on the target principally by using geographic information reach his hometown or the surrounding areas readily, but once there, often circulate before entering the target's circle of acquaintances. There is no available information to narrow the field of potential contacts which an individual might have within the town. Such additional information as a list of local organizations of which the target is a member might have provided a natural funnel, facilitating the progress of the document from town to target person.
>
> There is a progressive closing in on the target area as each new person is added to the chain (see Figure 3.2). In some cases,

> however, a chain moves all the way from Nebraska to the very neighborhood in which the target person resides, but then goes round and round, never quite making the necessary contact to complete the chain. Some chains die only a few hundred feet from the target person's house, after a successful journey of 1000 miles.

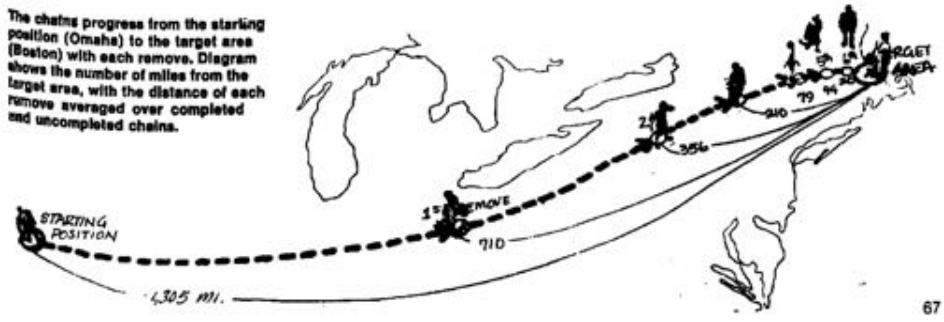


Figure 3.2: Original image from [108] explaining the geographic closing in of the chains

Reproducing the small-world experiment on *analog* means

When Milgram asked his colleagues in sociology and psychology to try to guess the average path length resulting from the experiment, answers given by those were typically in the hundreds. Actually some authors [75] have nominated Milgram's small-world results among the most counter-intuitive experimental results in social science during the 20th century.

Due to this unexpectedness, and the fact that the experiment was relatively inexpensive to run, several similar experiments were run afterwards. First, Milgram himself, this time working with Korte, [77] studied the relationships between white and black people in Los Angeles and New York, finding that there was a certain racial barrier, because completion rates substantially changed whether source and target belonged to the same racial group or not. Similar results about ethnicity effects were found in a study by Lin, Dayton, and Greenwald's restricted to a suburban community [94]. Other studies focused in more restricted scopes such as a college campus [141] and a corporate [100]. A similar study carried over telephone in Montreal reached a 85% chain completion rate over 52 initial chains, the sources being french-speaking Canadians and the target being English-speaking prominent Jew [53].

While non precisely a reproduction, it is worth noting the work by Killworth and Bernard [67]. They asked 58 participants from a small town what would be their choice in a Milgram-like experiment for 1267 different targets. They found that geography was the first election criteria and that different participants would choose the same local *hubs* when targeting certain parts of the world.

Conclusion and criticisms of the Milgram results

Overall, this seminal research carried out by Milgram, found evidence of at least three previously unconfirmed facts about social networks.

1. Short paths exist in the social network.
2. People are able to find them in a fairly efficient way even if they lack global information about the social network structure.
3. Ethnicity, gender and geography attributes are used by the participants in order to successfully choose next recipient.

Results 2 and 3 have been contested even as late as in 2002, when Judith Kleinfeld argued that the recruitment of participants through marketing mailing lists biased positively the sample towards people with higher income, and hence, higher social capital. She also argued that using only one target and choosing this target to be a prominent person positively biased the experiment as well [75].

Actually, until the Watts-Strogatz model provided in 1998 a theoretical support for the existence of highly clustered small worlds, the interest in the Milgram experiment did not explode. In 1998, 31 years after its publication, Milgram's original paper [108] gathered around 150 citations, while in the following 15 years it gathered over 6000[1].

3.1.2 Theoretical frameworks for searchability

In 1998 Watts and Strogatz proposed a simple model of a network that presents both short diameter and high clustering, by simply adding certain random links to a lattice so that those random links become *shortcuts*. Yet, even in a Watts-Strogatz network is almost impossible for participants in Milgram-like settings to find the shortcuts.

Nearly every proposed solution to the problem of routing in a small world involves the application of greedy routing. This sort of routing depends on the ability of compute a certain distance to the target so that any node in the path can choose the next node it believes is closest to the destination. That is, there must be something to be greedy about. For example, this

[1] citations counts as provided by Google Scholar

could be geographic distance, IP address prefix matching, etc. In any case, the greedy algorithm can only be fed with information from the immediate neighborhood of the current node, and it can compute some sort of distance between any of the neighbours and the destination node.

Navigating regular lattices

In the beginning of this century [70, 71, 73], Kleinberg studied what could be done to provide searchability to a 2 dimensional lattice as the one presented in Figure 3.3. He defined that a network of N nodes is *searchable* if a decentralized algorithm is able to find short paths between any two nodes, considering short as being bounded by a polynomial function of $logN$ ($O(logN)$).

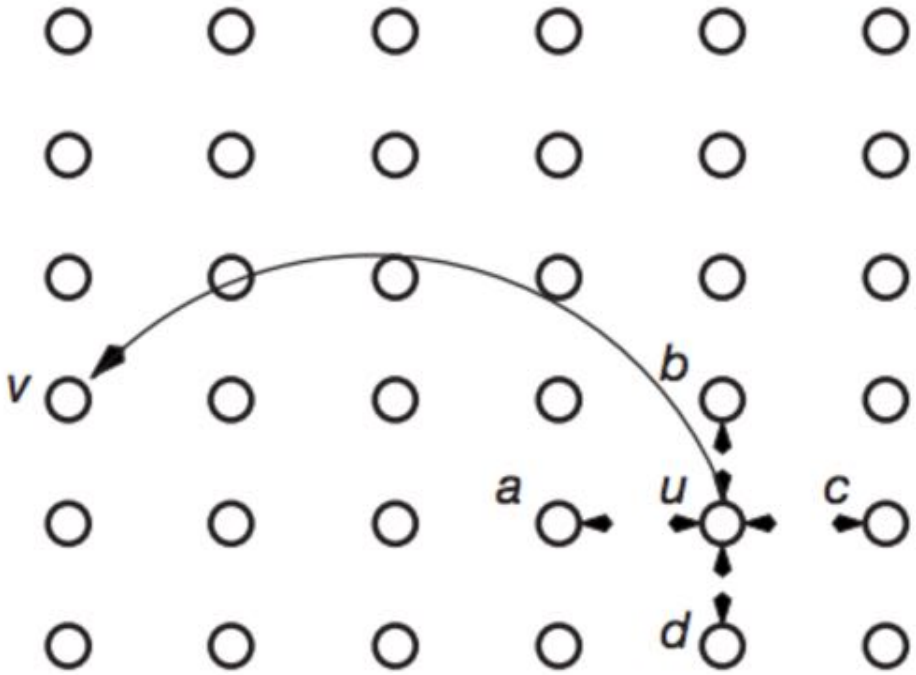


Figure 3.3: A 2D lattice becomes searchable when additional shortcuts are added depending on the lattice distance between nodes i and j.

Instead of adding shortcuts uniformly at random as proposed by Watts-Strogatz to generate small worlds, he investigated the more general case where a link is added between two nodes i and j with probability proportional to

$$\frac{1}{(d(i,j))^{\alpha}}$$

where $d(i,j)$ represents the lattice distance between i and j, and α is a *clustering exponent*. He found that in such conditions the network only becomes searchable if $\alpha = 2$. More precisely:

- For $0 \leq \alpha < 2$ decentralized search is $O(N^{\frac{2-\alpha}{3}})$.
- For $\alpha = 2$ decentralized search is $O(\log^2 N)$.
- For $\alpha > 2$ decentralized search is $O(N^{\frac{\alpha-2}{\alpha-1}})$.

Navigating hierarchical network

In 2002, Watts, Dodds and Newman [160] studied searchability in hierarchically induced networks. They proposed a model in which each node belongs to the leaves of several distinct hierarchies, reflecting the notion that participants in Milgram-like searchs were simultaneously taking into account several different notions of proximity to the target. Figure 3.4 presents one possible hierarchy. Their model constructs a random graph G as follows. We begin with H distinct complete b-ary trees whose leaves contain each of groups of size g elements. Then, every node is randomly assigned one of these elements in the leaves of each tree. As an example, for $H = 2$ the trees could represent geographical location and career field respectively. Then a particular node could be assigned to the leave representing Boston in the locations tree, and to the leave representing college professors in the careers. In each of the trees a hierarchical distance x_{ij} can be computed between any two nodes i and j, simply considering the minimum height of a common predecessor as illustrated in Figure 3.4. Links are added then into the graph G of N nodes with probability

$$p(x) = ce^{-\alpha x}$$

where α is a tunable parameter and c a normalization constant. Therefore for a large value of α, the network induced by the hierarchy would only have links between the members of the same group, while if $\alpha = 0$ the network would become a random graph.

In such networks, decentralized routing would try to reduce a *social distance*[2] which can then be computed as $y_{ij} = \min_h x_{ij}^h$. The authors define a searchable network as one where given an attrition rate p, at least a fixed rate r of chains reach the target, independently of the population size N. Using empirical data from Milgram experiment they set $p = 0.25$ and $r = 0.05$, concluding that a network is searchable if a decentralized algorithm can reach any target from any source using paths of $\langle L \rangle < 10.4$.

Having established the model to build the network and the searchability conditions, they ran simulations to explore the region of the $H - \alpha$ space

[2]Authors acknowledge this is not strictly a distance, since it violates triangle inequality

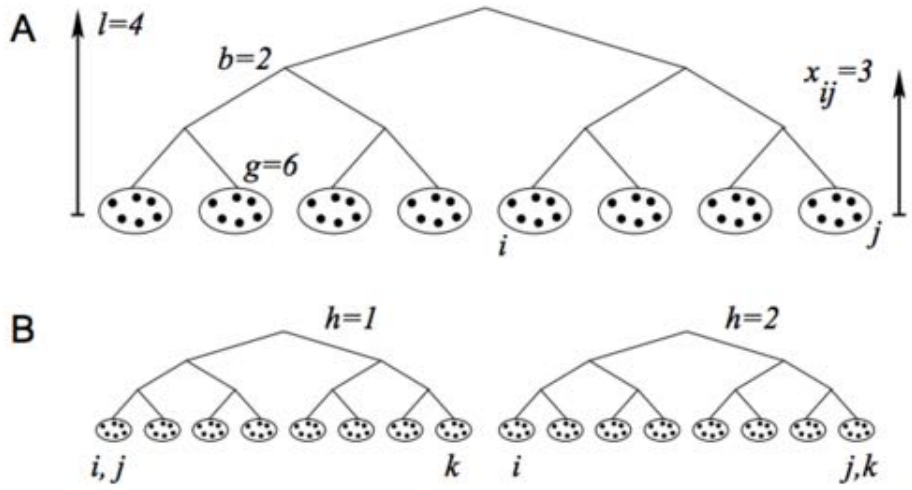


Figure 3.4: A) Figure shows a binary hierarchy ($b = 2$) where there are 6 nodes in each of the leaves ($g = 6$). Distance between i and j, defined as the the minimum height of a common predecessor, is then $x_{ij} = 3$. B) Another example where $H = 2$, $b = 2$ and $g = 6$. Note that $x_{ij}^{h=1} = 1$ and $x_{ij}^{h=2} = 4$, and given $y_{ij} = \min_h x_{ij}^h$, then $y_{ij} = 1$. Similarly, $y_{ik} = 4$ and $y_{jk} = 1$. This simple example shows that this *social distance* can violate triangle inequality since $y_{ik} = 4 > y_{ij} + y_{jk} = 2$.

where networks are searchable for different network sizes (ranging from $N = 10^5$ up to $N = 4 \cdot 10^5$) using reasonable parameters $g = 100$ and an average degree of 99. The results are presented in Figure 3.5. Additionally they were able to remarkably fit the empirical Milgram's chain length distribution (the one presented here in Figure 3.1) using a similar network size $N = 10^8$, fitting parameters $\alpha = 1$, average degree 300, $b = 10$, $H = 2$ and $g = 100$. The resulting claim, at a qualitative level, is that efficient search is facilitated by having a small number of different ways to measure proximity of nodes, and by having a small bias towards nearby nodes in the construction of random edges.

Another relevant contribution to searchability in hierarchical networks was done later by Kleinberg. In [74] he considers a network created in a very similar way that the one presented by Watts et al. In this case he focuses on $H = 1$, this is, networks induced by just one hierarchy. Also, he considers trees where there is only one node per leave ($g = 1$). Additionally, the network is built in a slightly different way: instead of connecting random pairs with probability $p(x)$, k links are created out of every node to a node chosen with probability $p(x)$ thus all the nodes have at least degree k.

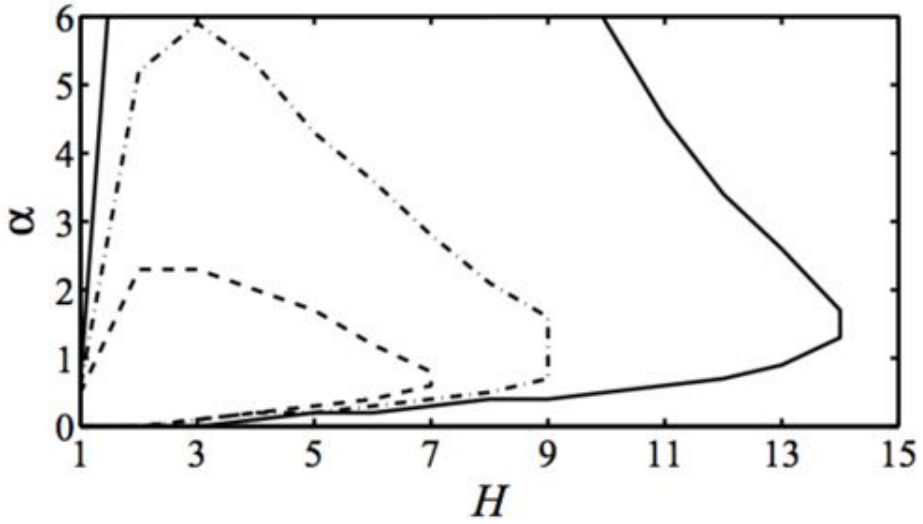


Figure 3.5: Searchable regions in the $H - \alpha$ space for networks of size $N = 102400$ (solid), $N = 204800$(dot-dash) and $N = 409600$ (dashed). Searchability shrinks with network size, but there is still a fairly big region in the $H - \alpha$ space where the networks are searchable.

In this scenario, he is able to yield analytical results similar to the ones found for the regular lattice, considering again that networks are searchable if decentralized search is $O(\log^2 N)$. Precisely, he concludes that there is just one particular value $\alpha = 1$ for which the network becomes searchable.

Searchability for group induced networks

Also in [74], Kleinberg proposed a model that generalized searchability conditions to a broader set of networks that includes both hierarchical networks and regular lattices. Qualitatively, the idea is that every node in a network belongs to a certain set of *groups*, and there is always one group every node belongs to. A social distance between two nodes can then be defined as the size of the smaller common group. Similar to previous analysis, if a network is created by connecting more likely socially closer nodes, searchability might emerge.

Formally, we start with a set of nodes V, such that $|V| = N$, and a collection of subsets $S = \{S_1, S_2, ..., S_m\}$ of V, which we will call the set of groups. Then, for constants $\lambda < 1$ and $\kappa > 1$, we impose the following three properties.

1. The full set V is one of the groups.

2. S_i is a group of size $g \geq 2$ containing a node v, then there is a group $S_j \subseteq S_i$ containing v such that $\min(\alpha g, g-1) \leq |S_j| < |S_i|$.

3. If $S_i, S_j, ...$ are groups that all have size at most g and all contain a common node v, then their union has size at most κg.

Let us now define the social distance $g(u,v)$ as the cardinal of the smallest group that contains both u and v (note that $g(u,v) \leq N$ because V is a group). We can build a network by creating $k = c \log N$ links out of every node, connecting them to other nodes with probability proportional to $g(u,v)^{-\gamma}$. In [74] the following results are proven for such network:

1. If $\gamma = 1$, the network is searchable (decentralized search is $O(\log^2 N)$) for a sufficiently large c.

2. If $\gamma < 1$, the network is not searchable.

Note that all these conditions are met by hierarchical networks, by considering that $g(u,v)$ is the number of nodes in leaves of the closest common predecessor. The regular 2 dimensional lattice also can be represented using this group model, defining $g(u,v)$ as the number of nodes that are closer to u than v, which grows with the square of the lattice distance. However, a general negative result for $\gamma > 1$ cannot be obtained because there are simple examples of networks that satisfy all the conditions and are searchable for large values of γ (see [72]).

Decentralized search with additional information

A recent trend on papers [84, 45, 102, 103] studying searchability from a analytical point of view has considered the possibility of a node to *consult* to some of his neighbors about what is the best choice for a next step. However we consider this recent trend out of the scope of this document since the subsequent discussion does not consider consultation to neighboring nodes.

3.1.3 A global searchability study

The largest small world experiment to date was carried out by Roby Muhamad while pursuing his PhD at Columbia University [114], and the most significant results were published together with Dodds and Watts in [32].

For the first time the experiment was based on e-mail, and also a supporting website was used for additional data gathering. Also the goal was more ambitious than ever before, since the authors tried to run the experiment worldwide. Due to new possibilities emerging from the use of web and e-mail, some methodological changes were introduced in the experiment. For example, a limited time was given to each recipient to continue the chain.

If the deadline was passed, the previous sender was given another chance to forward the message, thus enabling one-step backtrack to the chains. Also, for the first time in these kind of experiments, recipients had to confirm they knew their sender, trying to countermeasure participants forwarding the message to strangers they found searching online. Additional data on the tie itself was collected, such as its intensity (weak vs strong), nature (friendship, family ...) and origin (college, workplace...). The reason for choosing next recipient, as well as demographics of all participants were also collected.

In total, $24,000$ chains were originated, involving around 61000 participants in 166 countries who sent at least one message. 18 people were chosen as target, 6 decided by the authors and the other 12 at random among the pool of participants, in order to avoid bias in the social status of the targets. While intended to be a global experiment, the pool of participants was strongly biased towards the demographics of the Internet users of the time: US and UK were overrepresented and so were white, medium class highly educated participants. Also the network itself was biased towards the online community (for instance, 7% of participants reported they had met their choice online). In the end, only 384 chains arrived to their intended target (0.4% success rate)[3]. However, authors argue that the low success rate is mostly influenced by the lack of interest by participants rather than to difficulty of completing the task. This argument is supported by the constant attrition rate found for different steps in the chain, and also by the fact that, among those who did not continue the chain, a later survey found that only 0.3% declared they did not continue because they would not know whom to choose as next recipient.

Successful chains turned out to have used more often professional links over family ties and ties created during higher education. Gender homophily in the chains was an order of magnitude smaller than the ones reported by Milgram and his collaborators 40 years before. About tie strength, successful chains include significantly more often weak ties, therefore supporting Granovetter's hypothesis about diffusion of information [52].

Regarding the length of successful chains, authors reported $\langle L \rangle = 4.05$ while the acknowledged attrition affects this measure (longer chains are more prone to remain unfinished) and therefore assuming constant attrition they estimated chain average length to be around 5 for those chains where target and source live in the same country and 7 for international ones. Regarding influence of social hubs (people with high degree), results point that they do not have critical influence in decentralized search. Participants do not preferentially choose the recipient based on their degree, and targets are

[3]The reader might note that completion rate looks higher $384/24,000 = 1.6\%$. That is because the $24,000$ figure only includes those chains where at least one message was send. For an additional $74,000$ chains, the source node did not even send the first message

reached from several of their acquaintances, contrary to what was found for the target of the Nebraska study. Different success rates are found depending on the target. None of the 8000 chains towards an unemployed target in Indonesia was successful while the target who was a US professor received 3% of all the chains addressed to him, suggesting that the world is not equally small for everyone.

The most significant result for the subsequent discussion was found when analyzing the reason to choose next recipients. Authors found that while during the first steps geography was the main criterion, another social dimensions such as field of work or family origin were preferentially used in later steps (see Figure 3.6). In author's own words in [114]:

> Two factors (geographical and occupational proximities) stood out as the most-used cues for directing messages. Specifically, if we saw the reason chosen as a function of the chain length, as displayed in figure 3.6, in the early stages of the chain, it appeared that geographical reason dominated, presumably because senders were geographically distant from targets. Yet, at the later stages of the chain, occupational cue was used more than geographical cue. This finding suggested that when proximity to a target in a domain (e.g., geography) has reached a certain level of granularity at which it was too difficult to go further, senders switched to another domain (e.g., occupation) that could provide further differentiation.

3.1.4 Decentralized search on network data

The emergence of massive electronic communication records in the last 10 years allowed researchers to build large social networks upon such records. For example, email logs from a large corporation can be converted into social network data by considering that one node represents a particular employee, and links represent certain exchanges of emails between any two employees.

The new data sources motivated another trend on the analysis of small-worlds which consists of simulating how different decentralized algorithms perform when they are applied to this available real world network data. It is possible to discover also which are the structures that exist in real world networks that make them searchable.

Our study will follow this approach, but before discussing our own experimental setting and results, it is worth to reference the two previous main contributions in this field.

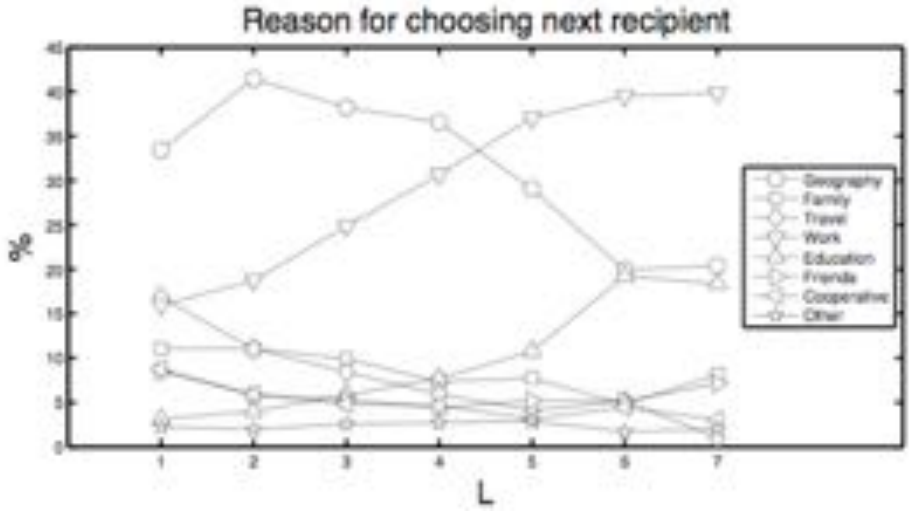


Figure 3.6: Reasons for choosing the next recipient in [32]. L is the number of steps in the chains. Geography: recipient is geographically closer; Family: recipient's family originates from target's region; Travel: recipient has traveled to target's region; Work: recipient has occupation similar to target; Education: recipient has similar educational background to target; Friends: recipient has many friends; Cooperative: recipient is considered likely to continue the chain.

Decentralized search in a corporate email network

In [2], Adamic and Adar studied a network emerging from email records within HP Labs, where they worked at the moment[4]. The network data is relatively small ($N = 430$ nodes), but additional data available, such as the hierarchical structure of the lab and the physical location of each employee, allowed the authors to provide very interesting insights regarding searchability.

The network was built considering only emails with less than 10 recipients, in order to avoid announcements and similar sort of communications that do not necessarily imply that sender and recipient know each other. With that same goal in mind, links were added to the network only if the two employees emailed each other at least once during the observation period of 6 months. The resulting network had an average degree of 12, and presented high clustering coefficient and short paths between any two nodes ($\langle L_{optimal} \rangle = 3.1$).

[4]In the same paper, they analyzed data from a pre-Facebook campus social network, but since they did not reach any interesting conclusions for our goals, this experiment is omitted from this discussion.

To test searchability of the network the authors tried three different decentralized routing algorithms, depending on whether the message was forwarded:

- to the best connected neighbor (the one with the highest degree) (DEG).
- to the neighbor closer to the target considering the departments hierarchy within the lab (COM).
- to the neighbor who seated geographically closer to the target (GEO).

The best connected neighbor strategy did not perform well. While there are theoretical results supporting decentralized routing based on degree for power-law networks, the email network did not have neither enough hubs, neither existing hubs were connected enough for the search to succeed. Median L_{DEG} was 16 while $\langle L_{DEG} \rangle = 43$. Such gap between median and average path length is explained by the difference between hubs (very easy to find using this strategy) and poorly connected nodes. This finding was consistent with empirical evidence from Muhamad experiments discussed before (degree does not play a critical role in searchability).

Before trying the COM strategy, authors study the hierarchical structure of the lab from the theoretical point of view. General inspection of the network as the one presented in Figure 3.7 shows a close relationship between organizational units and email communication. The authors also compared their data with the ideal models on hierarchical networks by Watts and Kleinberg discussed in Section 3.1.2. They found that their data could be fitted to the model proposed by Watts using $\alpha = 0.94, H = 1$, so the network was in principle in the searchable region. A similar fit to the Kleinberg model yielded a close-to-searchable $\alpha = 0.75$. COM strategy was indeed the one with the best results among the three. Median L_{COM} was 4, and $\langle L_{COM} \rangle = 5$, both remarkably close from the optimal result.

For the GEO strategy, participants building, floor and office is taken into consideration. In principle, a strong preference for short range relationships was found (87% of links were intra-floor, while employees were located in 8 different floors). However, results were significantly worse than those from COM. Median L_{GEO} was 6, and $\langle L_{GEO} \rangle = 11.7$. The authors tried to fit the network into the regular lattice model by Kleinberg discussed in Section 3.1.2 but they could not find a good fit, since in this setting the number of people within radius r from a target did not increase as fast as the square that occurs in the 2D lattice. Also, location of employees was mostly driven by hierarchy (people working together were typically closer), so it is difficult to differentiate if it is actually geography or organization the most conditioning factor.

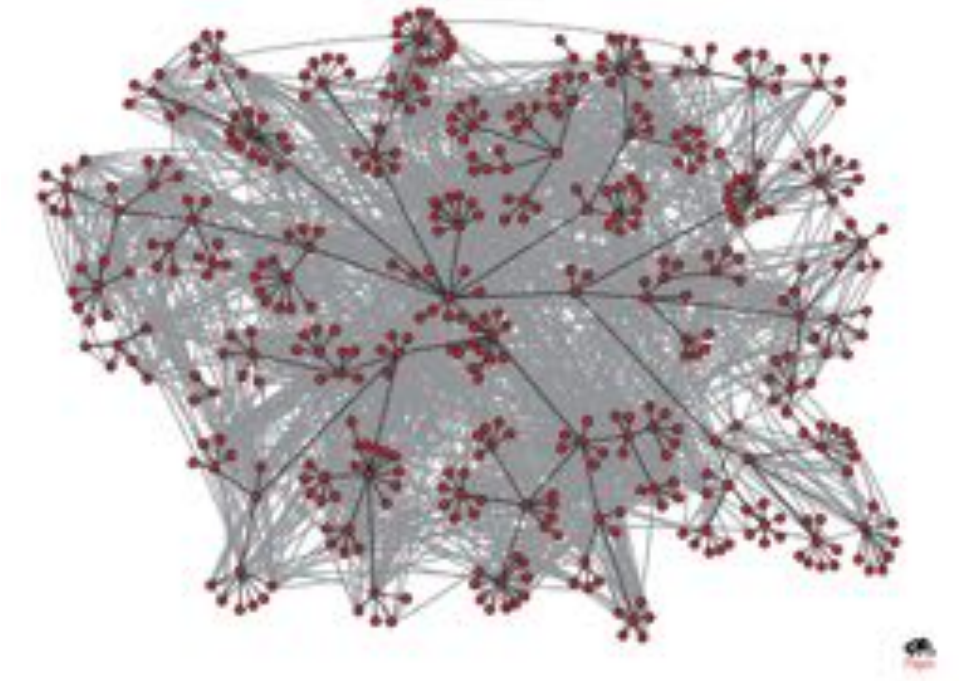


Figure 3.7: . HP Labs' email communication (light grey lines) mapped onto the organizational hierarchy (black lines). Note that communication is closely influence by the formal organizational chart.

Nationwide decentralized search using blog data

In 2005, Liben-Nowell et al. studied the role of geography in social search [93]. To do so, they gathered the connections between 500,000 blog authors from LiveJournal in the United States (by the time, one of the leading blog platforms). These bloggers also had reported the city where they live, so they could be geolocated at least at the city level, so that the simulated routing experiment is intended to reach the target city (instead of the target person). The data set contained bloggers living in around 500 different cities.

Additionally to the lack of spatial resolution, the data set has other limitations. First, messages can travel all around the US (300 million population) through only half million people (0.15%). Second, while the average degree was not reported, similar studies on LiveJournal data [10, 89] find average degree between 15 and 20, even before filtering the networks to only those nodes that can be geolocated (compared to the hundreds of acquaintances reported by participants in social search experiments).

At first, the experiment simulates a pure *geogreedy* algorithm that forwards the message to the friend geographically closer to the target. Probably

due to the limitations of the data set, this algorithm only succeeds 13% of the time and takes on average around 12 steps to reach the target city. Therefore the authors relax the conditions of the experiment so that if a message holder does not have anyone closer to the target than himself, then he would forward the message to a random person in his city. Strictly, this is not social search, because the messages do not travel then through social links (unless in the very unrealistic scenario where everyone would be connected to everyone within the city). This relaxation of the experiment increases the success rate to 80%.

Additionally to the experimental results, some interesting features about the geography of social networks where unveiled in this work. In particular, the fraction of social links between people living within distance d was found to be a decreasing function as $d^{-1.2}$ (see Figure 3.8). Authors propose, using a rank model similar to Kleinberg's group model, that this unexpected value in the exponent (a searchable geographic network should present links fraction decaying as d^{-2}) comes from the non-homogeneous population distribution.

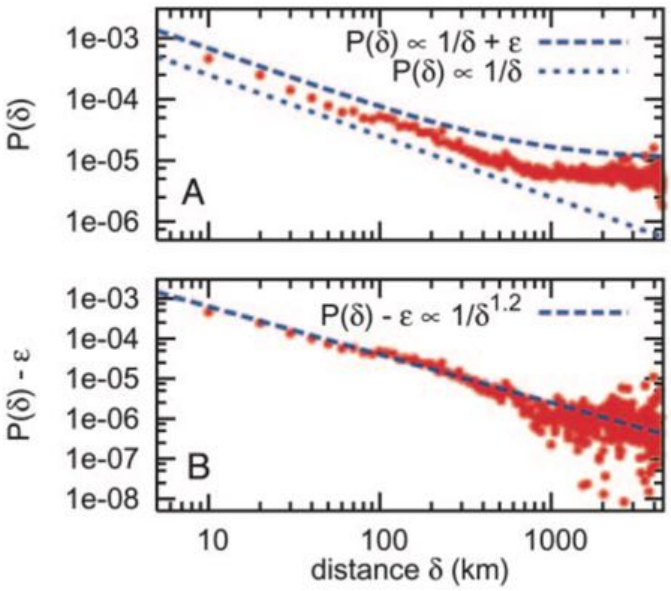


Figure 3.8: The relationship between friendship probability and geographic distance. (A) For each distance the fraction of friendships among all pairs u, v of LiveJournal users with $d(u, v) = \delta$ is shown. Distances are rounded down to multiples of 10 km. (B) The same data are plotted, correcting for the background friendship probability ϵ (See [93] for details).

Country	% GC	$N(\cdot 10^6)$	$E(\cdot 10^6)$	$\langle k \rangle$	$\langle c \rangle$	$\langle c_r \rangle$	$\langle l \rangle$	$\langle l_r \rangle$
France	99.23	18.7	81.3	8.73	0.16	$9 \cdot 10^{-7}$	8.52	7.75
Portugal	96.23	1.2	4.0	6.57	0.26	$5 \cdot 10^{-7}$	8.35	7.44
Spain	95.81	5.9	16.1	5.44	0.21	$48 \cdot 10^{-7}$	10.36	9.20

Table 3.1: Characteristic properties of the social networks in the studied countries: size of the giant component (%GC), number of users (N) and relationships (E), average degree $\langle k \rangle$, average clustering coefficient $\langle c \rangle$. Also the corresponding values $\langle c_r \rangle$ and $\langle l_r \rangle$ for random networks with the same size are provided. The observed values are typical for small-world networks.

3.2 Data description

Our dataset contains phone records for a six months period during 2011 in three countries: France, Portugal, and Spain. In total 7 billion interactions are considered.

The data was provided in the form of Call Detail Records, as Comma Separated Values (CSV) files, where a line, representing a call, presents the following format[5]:

```
timestamp, caller, receiver, tower_c, tower_r, duration.
```

In order to build the social network, only links with at least one communication per direction are included. This is a common technique in the literature [120, 119, 82] to avoid both marketing callers and misdialed numbers. However, the scale of the provided data (one or two orders of magnitude larger than the referred literature) presented a significant computational challenge, which we were able to overcome. Details of the technical solution were published as a blog post [55] and are included in the Appendix A of this thesis.

Different anonymization techniques were used by the national providers and therefore it is not possible to match a user across the three countries, so we built three different networks whose nodes are the subscribers in each of the countries. The basics metrics of such networks are shown in Table 3.1.

Regarding degree distribution, our three networks present the common heavy-tail distribution found in previous works [120, 82] with mobile phone data, which can be fitted to a high exponent power-law ($\alpha > 4$). Degree distributions for all three networks are shown in Figure 3.9. For the scope of this study, the existence of hubs (nodes with very high number of connections) in all three networks will be important.

[5]Spanish CDRs did not include tower information.

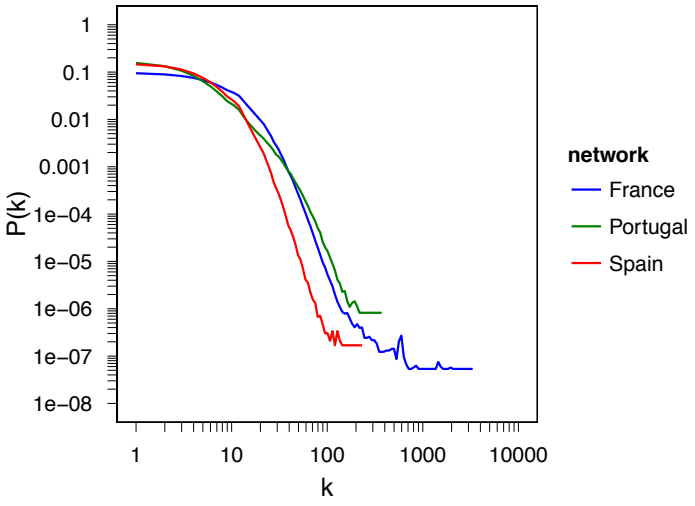


Figure 3.9: Degree distribution for each of the country level networks.

3.2.1 User location

For the subsequent discussion, we will consider that users are located in their billing zip code (Spain) or their most used tower (France and Portugal).

Spain zip codes are geolocated according to Geonames database[6], and aggregated according to latitude and longitude since some zip codes have identical coordinates. Towers coordinates were provided by the carrier, having 17475 different locations in France and 2209 in Portugal. Figure 3.10 shows the distance distribution from anly location in the country to the first, second and third closest zip code or tower in the three datasets. Although towers may provide a slightly more accurate geolocation, both are sufficient for our purposes.

On the other hand, users are not equally distributed among towers and zip codes. Figure 3.11 shows the cumulative distribution in the three data sets. Most of the towers serve between 100 and a few thousands users, while zip codes' user count is more heterogeneous (the maximum is a zip code in Madrid with 125,000 users). The explanation for these different results comes from the GSM technology used for mobile telephony: as the demand rises in an area, additional phone towers need to be installed to handle the traffic.

For simplicity, from now on we will refer both towers and zip codes as towers, unless otherwise mentioned, to explain different results among different data sets.

[6]available at `http://downloads.geonames.org/export/zip`

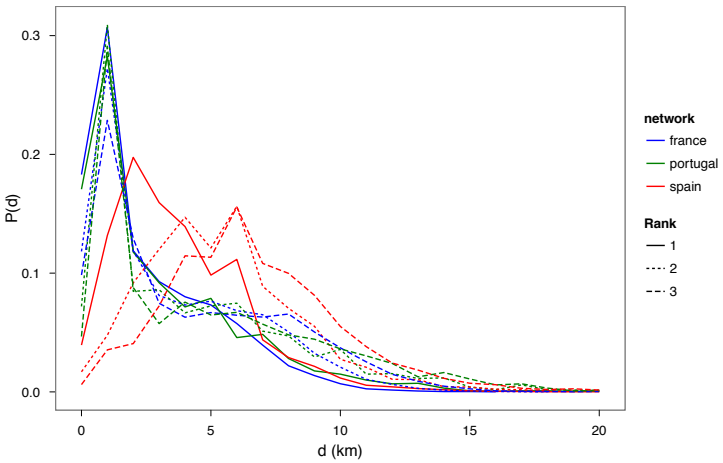


Figure 3.10: Distance distribution to the first, second and third closest tower or zip codes. Towers (France and Portugal) are slightly closer to each other than zip codes (Spain) are.

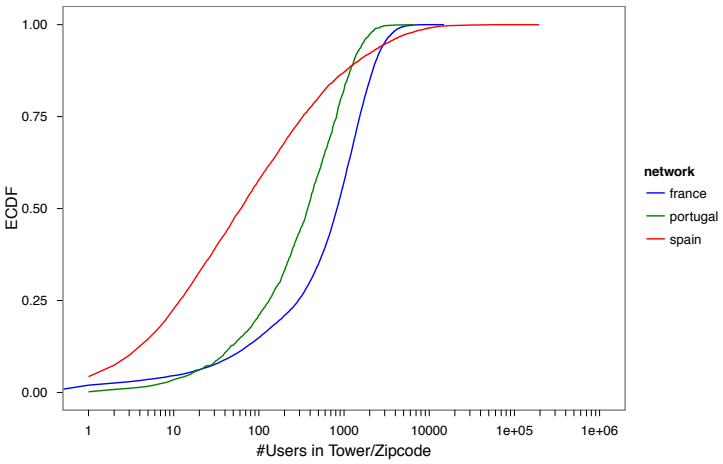


Figure 3.11: Empirical cumulative distribution function of number of users in each tower or zip code. Due to technical reasons, the range of users per location is smaller when towers are used, while zip codes distribution is more broad.

Assignation of user to cities

The subsequent discussion assigns users to an *area where they live*. For systematically delimiting this *area where the user lives* we have chosen administrative divisions over a regular spatial grid, since this way we can more exactly reproduce previous results in the literature we want to benchmark against . Specifically, we will study two levels of aggregation in each of the networks:

- We will generically refer as provinces to the following administrative divisions: *départements* in France, *provincias* in Spain and *distritos* in Portugal . This way we divide each country into 97, 50, and 20 provinces, respectively. According to the official census, the population ranges from 77 thousand (Lozère, France) to 6.4 million (Madrid, Spain). A province map for all three countries is depicted in Figure 3.12a.
- We will generically refer as municipalities to the following administrative divisions: *cantons* in France[7], *municipios* in Spain and *concelhos* in Portugal. Our users are located in 3520, 5446 and 297 different municipalities respectively. A map depicting municipalities in all three countries is presented in Figure 3.12b.

To map the user coordinates into the appropriate divisions we have used Global Administrative Divisions database[8] except for France's *cantons,* where the GEOFLA database by the French geographic institute (IGN) has been used[9].

3.2.2 Sampling effects

Users in the network are not homogeneously distributed, since in some regions there is a slightly higher concentration. This variance may come from a higher market share of the mobile phone provider or from a higher usage of mobile phone service in the area (only users who have at least one mutual relationship appear in the network). The differences between different regions are depicted in Figure 3.13.

We refer user density as the ratio $u_i = \frac{\text{Users}}{\text{Total population}}$ in a certain region i. Having different u_i seems to affect mainly the average degree of the resulting

[7] We have used this division instead of the *communes* because of the high number and high heterogeneity of the latter (over 36 thousand different *communes,* ranging from 10 people to 2 million). Most of *cantons* are composed of several *communes*, being Paris a special case: Paris city actually fills the whole department 75, and is divided into 20 *arrodisements* (districts) which are counted as cantons. In any case, when we refer the Paris city in the intracity network experiment, we mean department 75. Some other large French cities are also divided into several cantons.

[8] http://www.gadm.org/

[9] http://professionnels.ign.fr/geofla

a) b)

Figure 3.12: Provinces (a) and municipalities (b) map of the three studied countries

subnetwork. Figure 3.14 shows this relationship, which turns out to be close to linear. For a network where all inhabitants are present (i.e, $u_i = 1$), a projection of the resulting linear model would be $\langle k \rangle \simeq 16$ (although one has to be careful that extrapolation so far beyond the existing range of available data).

In any case, the number of contacts in a phone network is relatively small when compared to other social networks obtained from online social sites (whose average degrees are in the hundreds [111, 48]) or compared to different figures proposed as average degree for humans: an extrapolation from the observed correlation between social group size and neocortex volume in primates drove Dunbar to propose 150 [33], while recent statistical estimation methods based on self-reported data range between 290 [106] and 610 [107]. We will show that increasing the average degree has a positive effect on routing, which means that any result we get by studying the phone social network can be considered as a lower bound for the real world's social network. On the other hand, the phone network can be seen as the backbone of the social network, since it contains only interactions people are willing to pay for.

3.3 Data suitability for searchability simulations

Before proceeding with the simulation experiment, it is of interest to confirm that all our three networks exhibit features previously observed in the literature. Namely, we confirmed that our networks were small worlds in the Watts-Strogatz sense (i.e, they present high clustering and short diameter simultaneously). Additionally we confirmed that the network was geographically clustered as previously found by Liben-Nowell in [93].

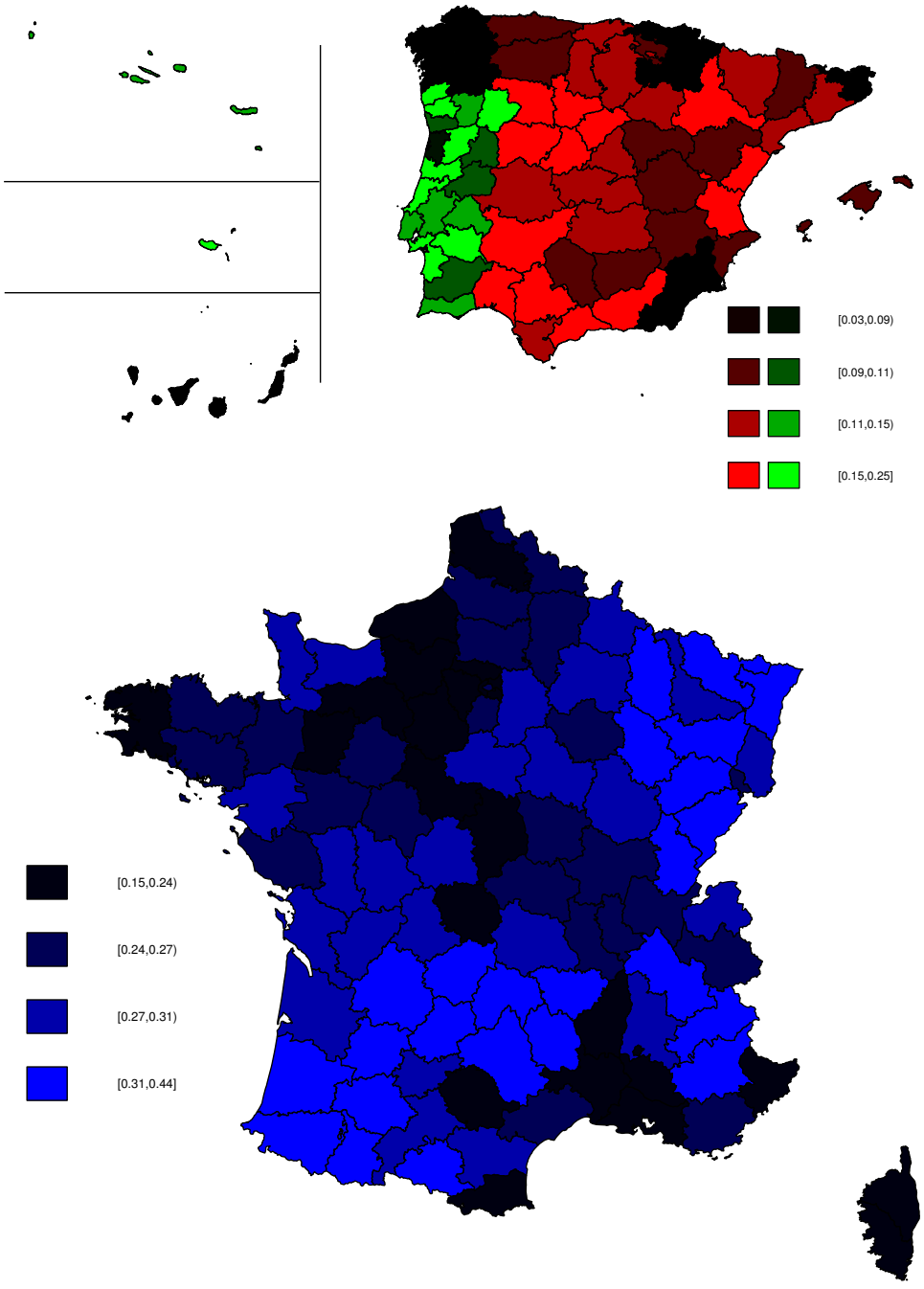


Figure 3.13: Users/Population ratio in the province level. Brighter colors represent a higher ratio.

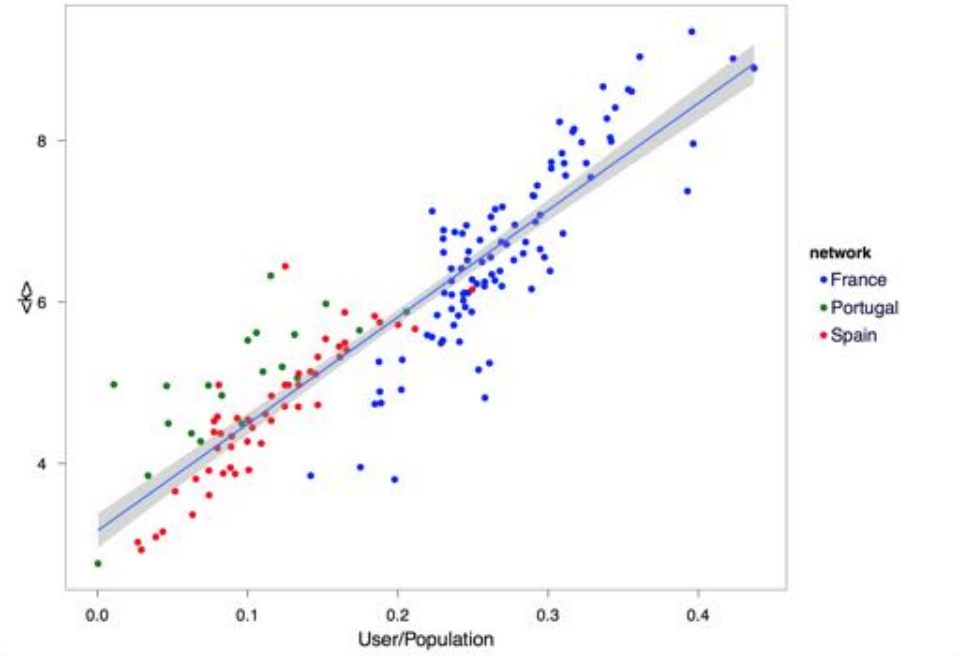


Figure 3.14: Dependence of the average degree $\langle k \rangle$ on the ratio between users and population. Each point represents a province network. It can be appreciated how closely related $\langle k \rangle$ and $u_i = \frac{\text{Users}}{\text{Population}}$ are. Blue line presents a linear fit $\langle k \rangle = 3.16 + 13.24 u_i$ where $R^2 = 0.818$.

3.3.1 Small world properties

A necessary (and obvious) condition for a network to be searchable, is that short paths actually exist between most of nodes in the network. As discussed before, while this is to be expected for a random graph, it is not necessarily true for a network with an abundance of triangles.

As reported in Table 3.1, all the three networks present around a million times more triangles than a random graph, and therefore it is necessary to confirm that short paths exist indeed.

In order to establish the diameter of the networks, 1000 random pairs of nodes were chosen as source and target, and shortest paths were found using standard Depth First Search. The resulting diameters, as reported in Table 3.1, range between 8.35 and 10.36, remarkably similar to the ones expected for random graphs of the same size. Therefore we conclude that our networks are good candidates for simulating decentralized routing algorithms.

Geographical dispersion of centrally located actors

Given the spatial resolution of the dataset, we could actually extract the spatial distribution of the most central people in the network. For every node in the network, we compute the shortest paths to any other node in the network and calculate the average number of nodes in each of the paths $\langle l \rangle$. This value is also known as the inverse of the closeness centrality and it ranges from 3.8 to 11, so everyone in the country is in average within 4 hops from the most central people and within 11 of the less central ones. In Figure 3.15 (upper left corner) we present $p(\langle l \rangle)$ distributions for each of the three networks.

In the main part of Figure 3.15, each dot represents a mobile phone tower, which is our smallest spatial resolution. In order to expose the backbone of the social network, the color intensity of each mobile phone tower represents the closeness centrality of the most central person in that tower. While main cities appear brighter, centrality is not only determined by population density: Barcelona area (Spain NE) is highly populated but it seems to be socially less central than Alicante (SE) even though the latter has half the population. Notice also, that the geographically most central city is not necessarily the socially most central one. Additionally, the links highlight the social connections only among the 50 most central people in each country, showing significant differences in the social network analyzed in the three countries. Whereas in France, Paris hosts the most central people, in Portugal two cities, Lisbon and Porto, seem to be equally important. In contrast, the most central people in Spain are spread over the entire country including even the Gran Canaria and Mallorca islands. Current spatial network models (like those we will introduce in Section 4.2) would not reproduce these differences observed in the data, but they would rather place

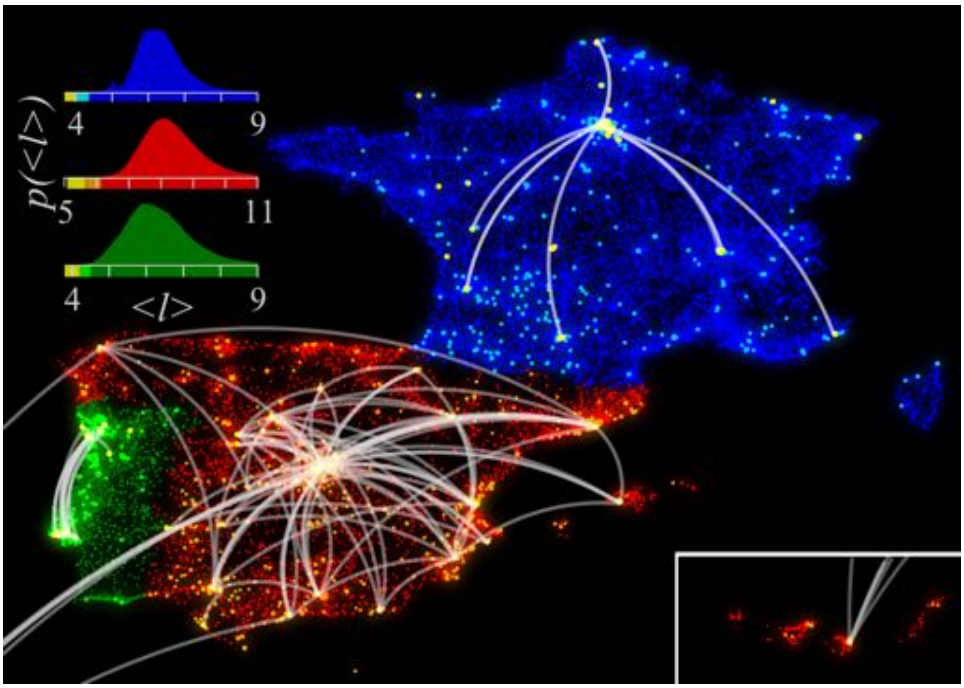


Figure 3.15: Visualization of central places in France, Spain and Portugal. Each circle represents a mobile phone tower and its color (the brighter the more central) corresponds to the inverse of closeness centrality $\langle l \rangle$ (average number of hops to any other person) of the most central persons in this tower. A person is always assigned either to his billing address or most used tower. White lines highlight the social network between the 50 most central persons of each country. In the the upper left corner, the distributions of $\langle l \rangle$ for each country (indicated by the corresponding color) are also shown.

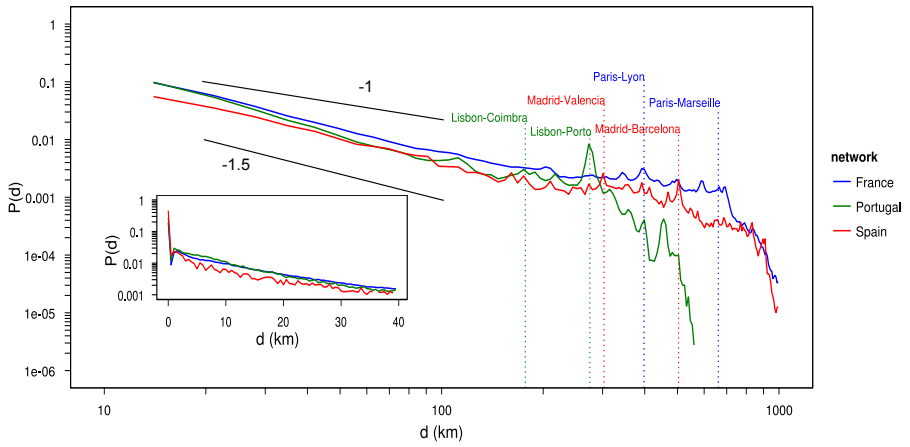


Figure 3.16: Probability of a link to have distance d in each of the networks. In the main figure distances are grouped in 7 km bins. The inset presents grouped a zoom on the low-range interval with 500 m resolution. Distributions present a power law decay (exponents between -1 and -1.5) until 100 km. There is a large number of links within the same tower ($d = 0$), reaching 40 % in Spain, 18% in France and 21 % in Portugal. This different behavior in Spain is probably an effect of the zip code population distribution discussed in Section 3.2.1.

central actors in the geographical center of the territory and/or the most populated cities. This suggests that the relationship between social networks and their underlying geography is yet to be fully understood, thus further research is needed in this topic.

3.3.2 Geographical distribution of links

The probability $P(d)$ of finding a social tie between two users decreases with geographical distance between them, regardless the proxy used to infer the social network: blogs [93], location based social networks [135, 25] or mobile phone data [82, 119, 80]. In all of them the probability $P(d)$ decreases (at least during a certain interval) as a power law, with exponents between -1 and -2. As shown in Figure 3.16, our data fits this behavior for all three networks. Moreover, due to the high number of links considered, we are able to observe long-range peaks. The reason for these peaks is the heterogeneity in the population geographical distribution (we observe the same peaks even if we randomize the links while keeping actors in the same location). To support this point, in Figure 3.16 we highlight how the peaks match the distances between main cities.

3.4 Decentralized routing simulation

In order to better understand searchability in a large scale social network, we have divided the simulations into two different kinds of experiments:

- Intercity experiment. Given two users chosen uniformly at random from the dataset, one of them will act as a source and the other one as a target. Different decentralized algorithms are used to find a social path from the source to the target. The simulation stops if a user living in the target's city is reached. This scenario is analogous to the one studied in [93].
- Intracity experiment. For the considered city, subnetworks are extracted containing users in such city. The experiment runs similarly to the intercity one, except that reaching the target is required for the simulation to be considered successful.

In both scenarios, a number of different algorithms are tested as described below.

3.4.1 Algorithms

In order to deliver the message, several strategies can be used. In the following we describe the criteria used in our experiments.

RAN We use random routing as a baseline comparison; by employing depth first search (DFS) into a routing algorithm, we effectively avoid the message to get into infinite loops. The application of DFS in the Milgram experiment is quite straightforward: when a participant receives a message, he knows the list of people who already got the message. The participant will never forward to one of these people, unless all of his friends are in the list. In this case, he will send the message back to the person who first sent the message to him. In a tree network, this would be the case of a branch which has been explored without success and the search process continues going backwards. Since our social network is far from being a tree, the ratio of rolling back events is extremely low (less than 10^{-6} in all of our simulations).

GEO This procedure consists of sending the message to the friend geographically closest to the target. In the intercity scenario, locations are considered on the municipality level. In the intracity scenario, tower locations are employed. Note that this discretization is prone to face ties (two or more friends are at the same distance from the target).

DEG In this case, the message is forwarded to the friend with the largest number of friends among the candidates.

COM In order to mimic social attributes (school, work), communities are detected in the network. To detect communities in social networks, we use the well-established Louvain method [18, 43, 119, 3, 39]. This method is a greedy optimization method that attempts to optimize the network modularity by aggregating nodes belonging to the same community and building a new network whose nodes are the communities. This method assigns to each person a set of communities at different hierarchical levels. Although the number of aggregation levels L depends on the network and it is automatically obtained from the algorithm, in all of our networks the algorithm provided between 3 and 7 aggregation levels. Note that this algorithm provides hierarchical communities. If two nodes i and j have a community of level l in common they will share as well all the communities in higher levels, formally:

$$\forall i, j \in \{1, ..N\} \mid c_{il} = c_{jl} \Rightarrow c_{ix} = c_{jx} \forall x \in \{l+1, .., L\}$$

where N is the number of people. A person will send the message to a friend with the lowest possible community level in common with the target. While it is arguable that community detection requires global information and such might not be available to participants in a Milgram-like experiment, recent research [90] has reported that people are able to relate communities detected in their network to certain social attributes and affiliations, thus making communities a reasonable proxy for those unknown attributes in our data set.

In our experiments, these criteria are combined, by using several of them to solve ties: this way, we will denote *ran-deg* to a routing scheme where first the already visited nodes are discarded from candidates (*ran*), and then those with the highest degree are chosen (*deg*). If there is still more than one possible friend after the routing logic is completed, the message is forwarded to one of these candidates at random. In our *ran-deg* example, this happens if two or more friends were not previously visited and have the highest degree.

3.4.2 Intercity experiment

For the intercity scenario we ran the experiment in the following setup: in each country we chose $60,000$ random pairs of sources and targets among all nodes in the network. Next, we attempt to deliver the messages from each source using combinations of the techniques described in Section 3.4.1. Additional to those, we have performed a pure *geogreedy* (passing to the geographically closer friend, and if no one is closer than the current user, consider the chain broken) as well as the modification proposed in [93], which we have denoted *geogreedy++*, which consists of forwarding the message to

another user in the same city even if she is not connected to the current user[10]. Only up to 1000 hops are simulated before reaching target's city.

Simulation results

First of all, in intercity routing, using provinces as target seems to make the routing process trivial (even random routing delivers a significant amount of messages), so we will present the results of the routing trying to reach the right municipality. The main conclusion is that any routing strategy other than random will deliver the messages with a high probability (as we can see in Figure 3.17). If we study small differences in error rate after 100 steps between the algorithms (see Figure 3.18) we find statistically significant differences between them. In general, *geo* methods outperform *com* methods, and solving geographical ties (two people are at the same distance from the target) using degree increases routing performance. Another relevant finding is that these distributed routings reflect the same behavior than the optimal routing: it is harder to route in Spain (due to the smallest average degree) than in France, despite the number of nodes in France being around 4 times larger.

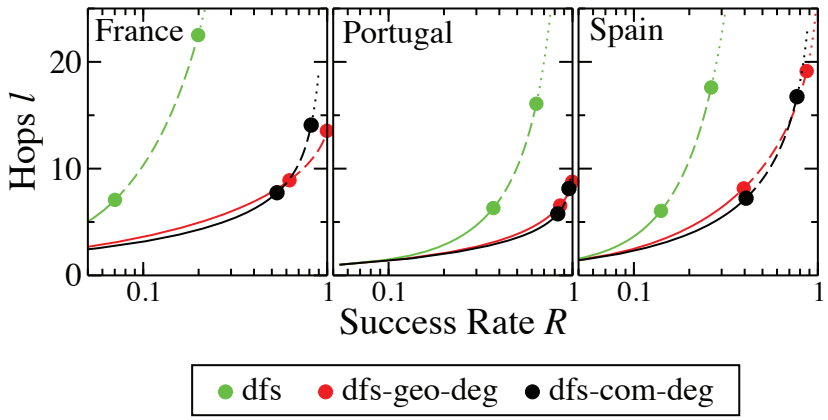


Figure 3.17: Results for different routing strategies in the intercity scenario. Dependence of the number of hops l on the success rate R for intercity routing (results for completing the delivery within 15 and 100 hops are highlighted by circles).

[10]We acknowledge this breaks the concept of routing in a network, or alternatively considers that social networks within cities are fully connected graphs. We include the algorithm in here for comparison purposes with previous work in [93].

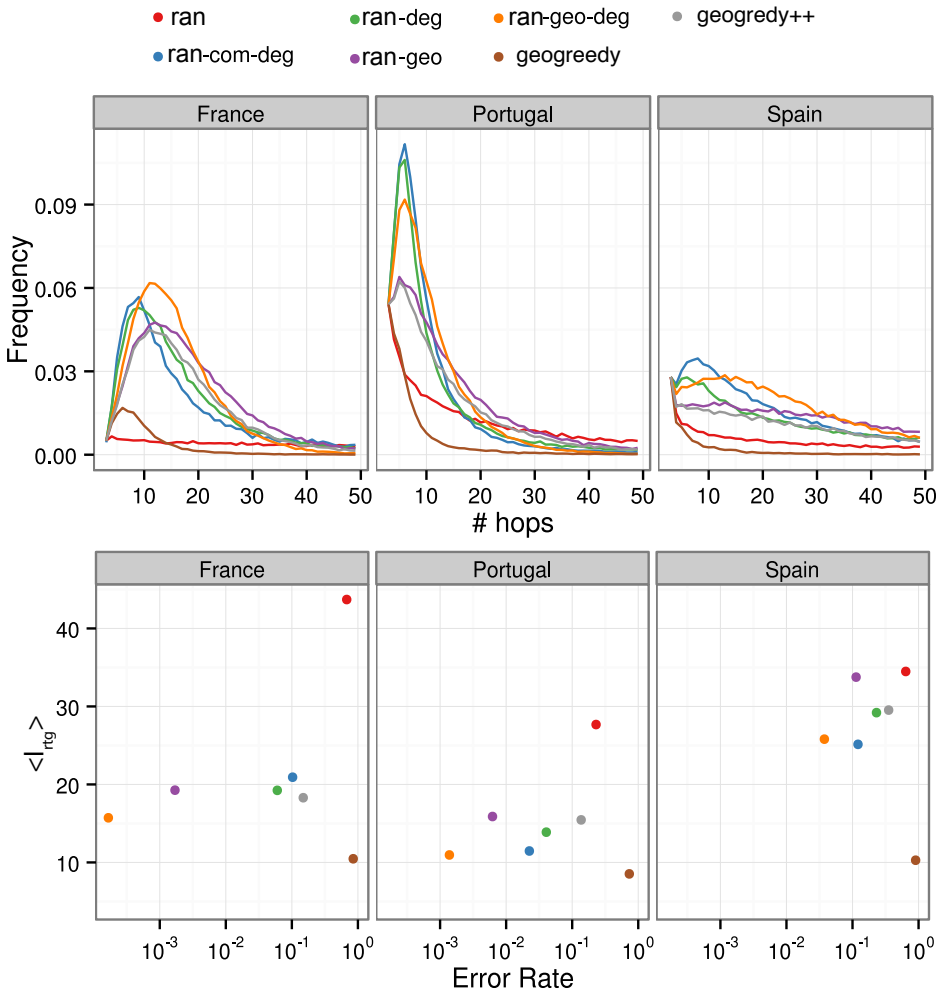


Figure 3.18: Intercity performance for different routing algorithms. Top graph shows the fraction of messages arriving at the target in the first 50 hops $f(n)$. Since the fraction of messages decreases with the number of hops, it is possible to compare the performance of different algorithms by measuring the mean and the cumulative value of this distribution after N hops, with N being large enough. The bottom graph presents the average path length of the delivered messages $\langle l_{rtg} \rangle = \sum_{n=1}^{N} n f(n)$ and the fraction of failed messages $E = 1 - \sum_{n=1}^{N} f(n)$ for $N = 100$.

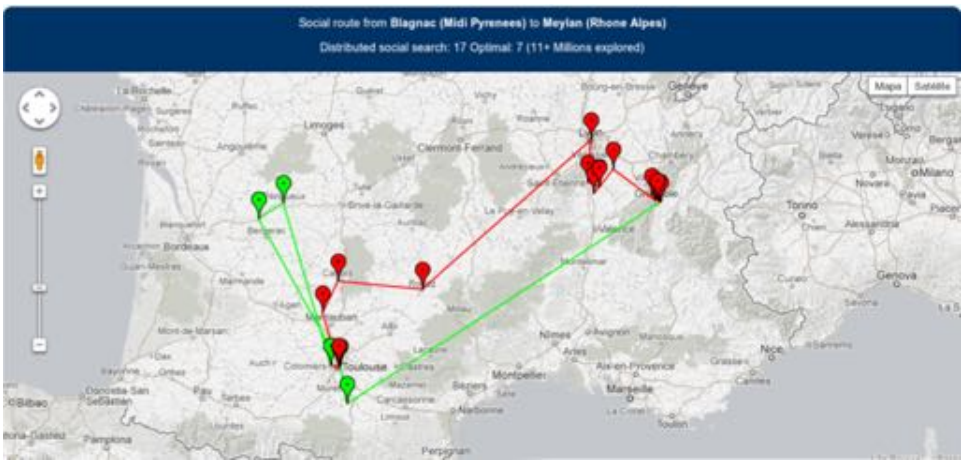


Figure 3.19: Snapshot of the app we developed for visualize our results. The red route is the result of distributed *ran-com-deg* while the green one displays the optimal route. In this example, the distributed route needs 17 steps to reach the destination city, while the optimal route uses 7. However, the distributed algorithm explores only 17 nodes, while more than 11 million nodes are checked for finding an optimal route.

Interactive webapp

In order to provide a more detailed look of this experiment, we have published the following webapp: `http://humnetlab.mit.edu/findingbacon`. In the app, the user can pick among 180 thousand routes we have simulated, choosing first the target city and then the source. To illustrate the difference between distributed an optimal routing, both optimal and best decentralized (*ran-geo-deg*) routes are plotted, and also the number of nodes explored to find the path is presented. Theoretically, a run can go "backwards" in the exploration of the network if all friends have been already visited, producing a loop in the sequence of explored nodes. However, we did not find evidence for this in our simulations (overall, over 3.2 million hops were simulated). In Figure 3.19 an snapshot of the app is presented, with one route as an example. On average, in France, the found distributed routes have 18.1 hops, while the optimal ones have 7.2. However, in order to find the optimal routes on average 8.1 million nodes have to be explored.

Effect of city size in search

Something definitely interesting and not reported in [93] is how the size of the target's city influences the length of the found distributed route. Intuitively, it is easier to reach a big town like Madrid (3.2 million inhabitants in the municipality and half million users in our network) than a small city with

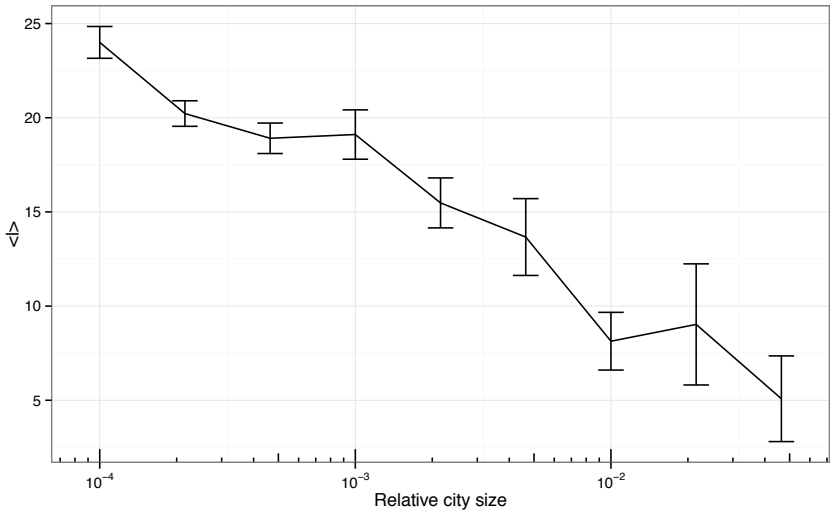


Figure 3.20: Average number of hops needed to reach each city versus city size (relative to the size of the country's network). Error bars show the standard error of the mean.

just a few hundred inhabitants. However, our simulations show how the size of the destination city affects only logarithmically to the length of the route found to reach it when using *geo* routing (see Figure 3.20).

3.4.3 Intracity experiment

For the intracity experiment, we consider networks from provinces and municipalities. All provinces and the 100 most populous municipalities in each country (300 municipalities and 168 provinces in total) have been studied. Province networks are almost connected (over 95% nodes in the giant component) and municipalities have also a quite big giant component (over 80%). The reason for this is that the classification of nodes in either provinces or municipalities is indeed a good community classification (modularity [11] scores over 0.4 and 0.5 respectively) probably due to the high spatial clustering of our networks. For the routing experiment, we take into account the nodes in the giant components (a path between any given two nodes actually exists) just as we did with the country networks.

We repeat the experiment in each of the networks with the same setup we used for the intercity experiment. In this case 100 thousand random pairs are simulated for the algorithms in each scenario.

[11] Modularity is a standard metric to evaluate the performance of a community detection method, defined in [115] as $Q = \frac{1}{2m}\sum_{ij}(A_{ij} - \frac{k_i k_j}{2m})\delta(i,j)$ where A is the adjacency matrix of the network, m is the number of edges, k_i is the degree of vertex i and $\delta(i,j) = 1$ if i and j belong to the same community and $\delta(i,j) = 0$ otherwise.

Simulation results

In Figure 3.21 we present the routing results for the three capital cities (in fact these are worst case scenarios, since the networks are the largest), showing 3.21 both $P(l_{rtg})$ distributions and their equivalence in the $(\langle l_{rtg}^{100} \rangle, E^{100})$ plane, which we will use for comparison. In Figures 3.22-3.27 we include results for the top 20 provinces and municipalities in each country. Careful observation of these graphics allows us to draw the following conclusions:

- The algorithm ranking, from best to worst, is almost constant over all studied networks.
- In *ran* methods (algorithms avoiding loops), $\langle l_{rtg} \rangle$ and E are fairly correlated. If an algorithm A outperforms another algorithm B by finding smaller $\langle l_{rtg} \rangle$ it will also provide a smaller error rate. Thus, we can compare algorithms by using only one of the two metrics. In Figure 3.28 we show the relation between these two metrics for the *ran-com-deg* algorithm.
- Contrary to what takes place in the intercity scale, using geography to route within the city does not produce efficient routing. Consistently over the studied networks, community based routing *ran-com-deg* significantly outperforms *ran-geo-deg*. Interestingly, having additional geography information besides the community structure (this means there is more information to make the routing decision) seems to be misleading, specially in large networks, as it can be observed in the performance of the *ran-com-geo-deg* routing strategy.
- Among all tested algorithms, *ran-com-deg* is the one producing the best results.

Influence of sample size

As explained in Section 3.2, all our observations are influenced by the market share of the data provider in the region, which mostly influences the average degree of the local network. Therefore, it makes sense to analyze the influence of the average degree in the routing simulation.

In Figure 3.29 (top) we show the relation between network size and error rate. Although in most networks we find that the error rate depends logarithmically on the number of nodes, we see a number of outliers. We find that these outliers have small average degree. In fact, the majority of networks that do not lie in the $O(\log N)$ behavior have average degree smaller than 4. Although to the best of our knowledge there is no previous result in the literature to explain this finding, we suggest the following explanation.

In a random graph where all nodes have the same degree k,$k \geq 3$ is needed to be able to find short paths $O(\log N)$ [20]. On the other hand,

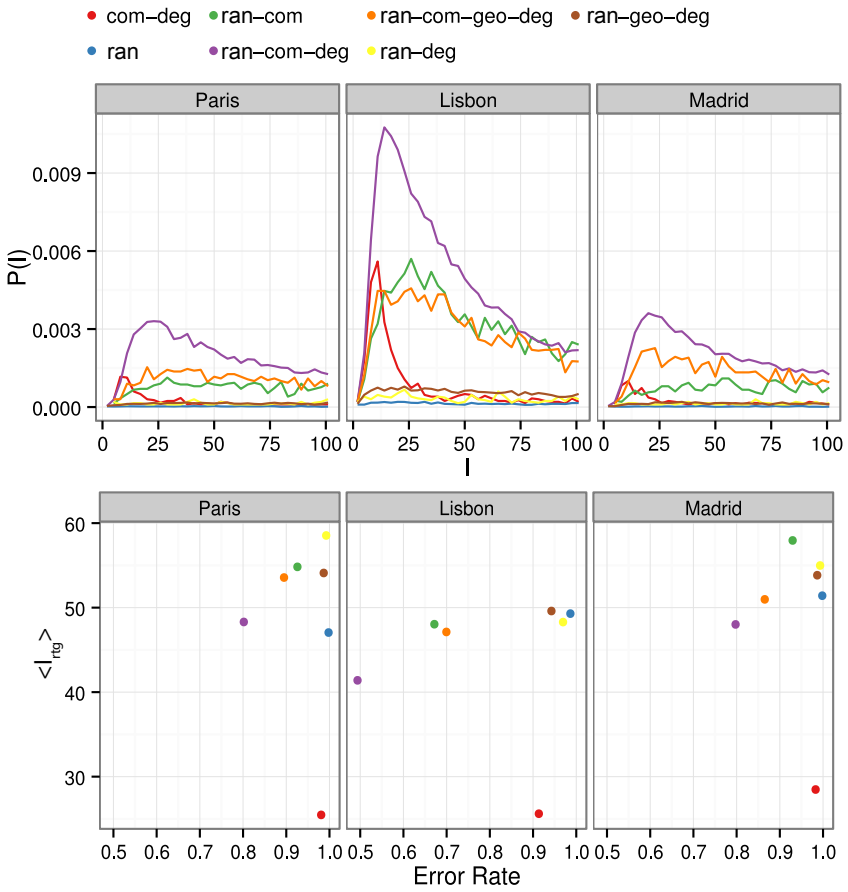


Figure 3.21: Intracity experiment results for the 3 main cities. Top graph shows the fraction of messages arriving at the target in the first 100 hops $f(n)$. Since the fraction of messages decreases with the number of hops, one could evaluate the performance by measuring the mean and the cummulative value of this distribution after N hops, with N being large enough. The bottom graph presents the average path length of the delivered messages $\langle l_{rtg} \rangle = \sum_{n=1}^{N} n f(n)$ and the fraction of failed messages $E = 1 - \sum_{n=1}^{N} f(n)$ for $N = 100$.

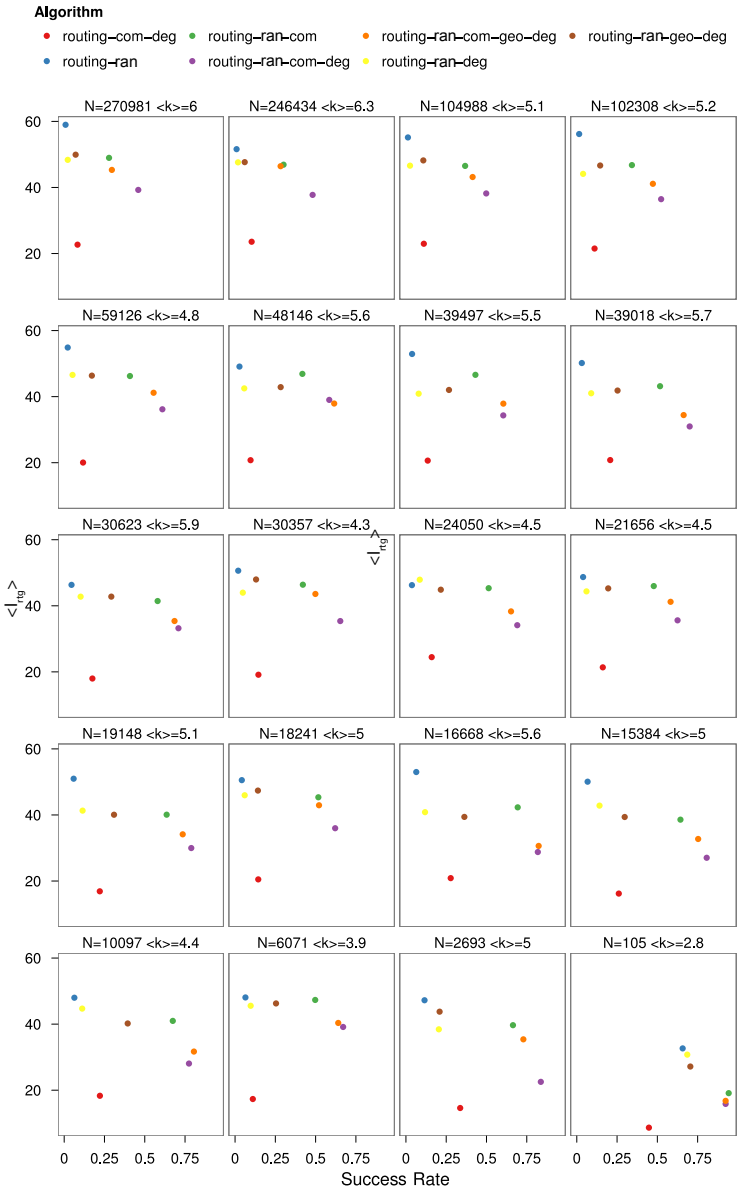


Figure 3.22: Intracity results for the 20 biggest provinces in Portugal. N denotes the number of nodes, and $\langle k \rangle$ the average degree. Success rates refer to the proportion of messages delivered after 100 steps and $\langle l_{rtg} \rangle$ to the average path length of successful chains.

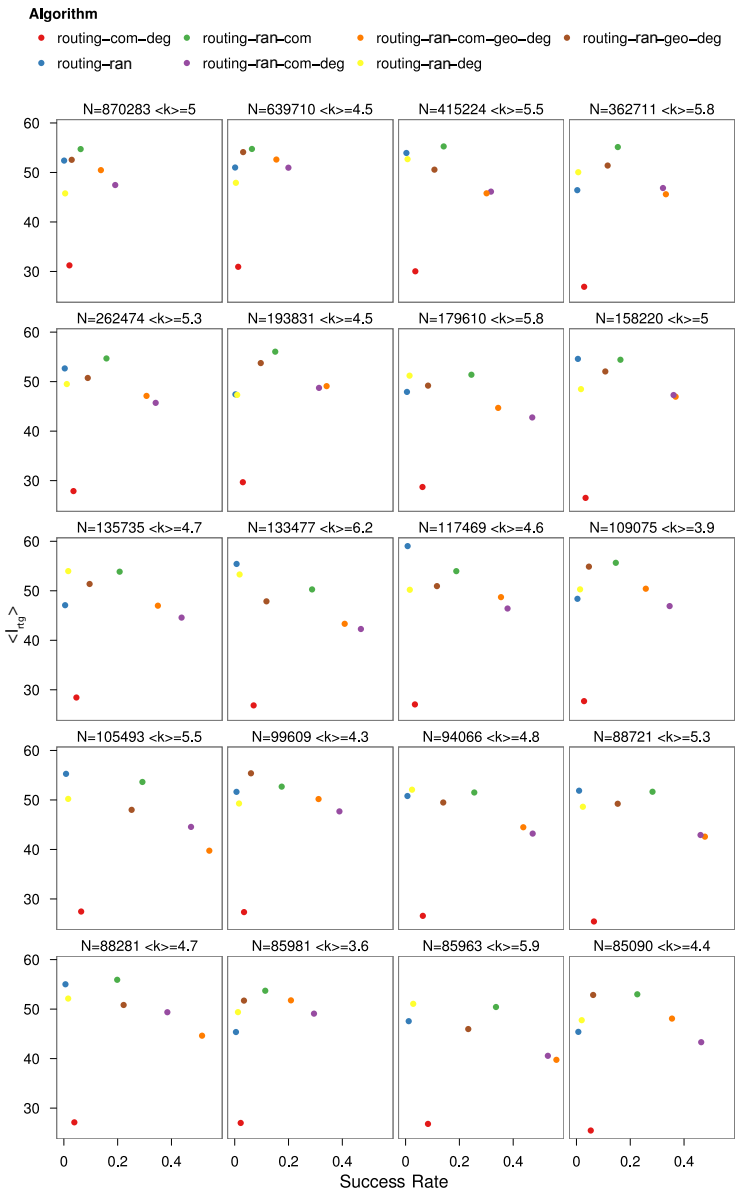


Figure 3.23: Intracity results for the 20 biggest provinces in Spain. N denotes the number of nodes, and $\langle k \rangle$ the average degree. Success rates refer to the proportion of messages delivered after 100 steps and $\langle l_{rtg} \rangle$ to the average path length of successful chains.

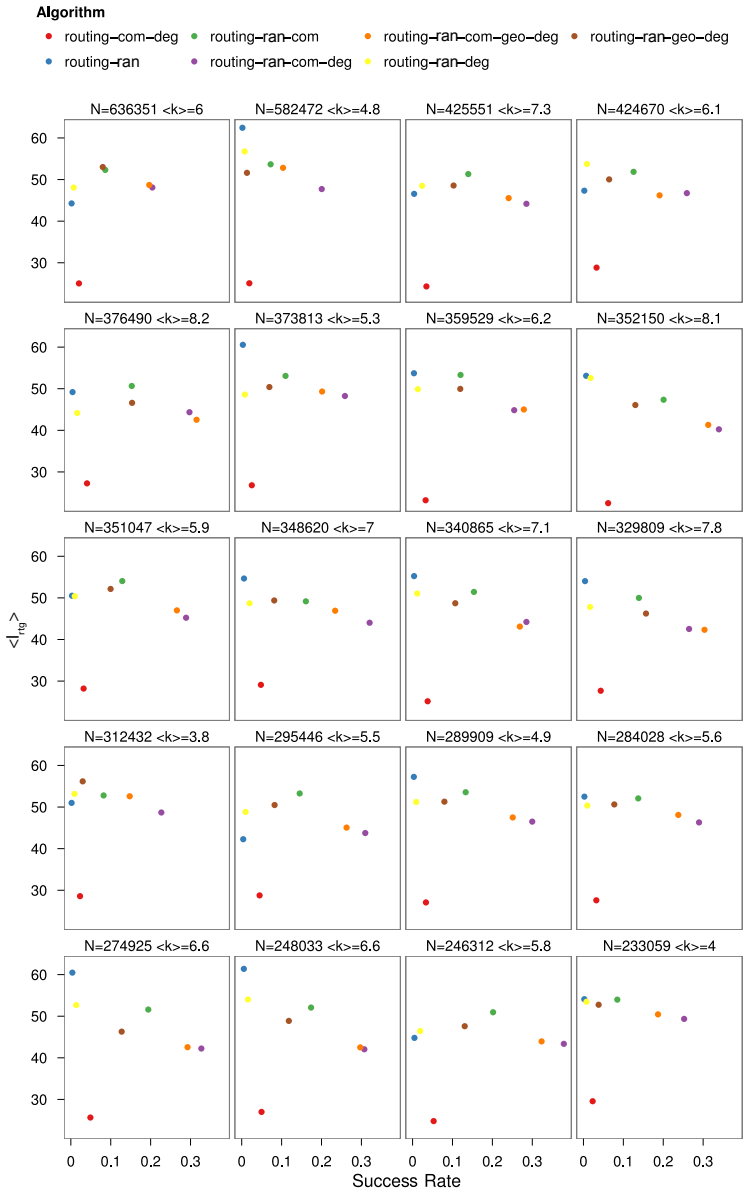


Figure 3.24: Intracity results for the 20 biggest provinces in France. N denotes the number of nodes, and $\langle k \rangle$ the average degree. Success rates refer to the proportion of messages delivered after 100 steps and $\langle l_{rtg} \rangle$ to the average path length of successful chains.

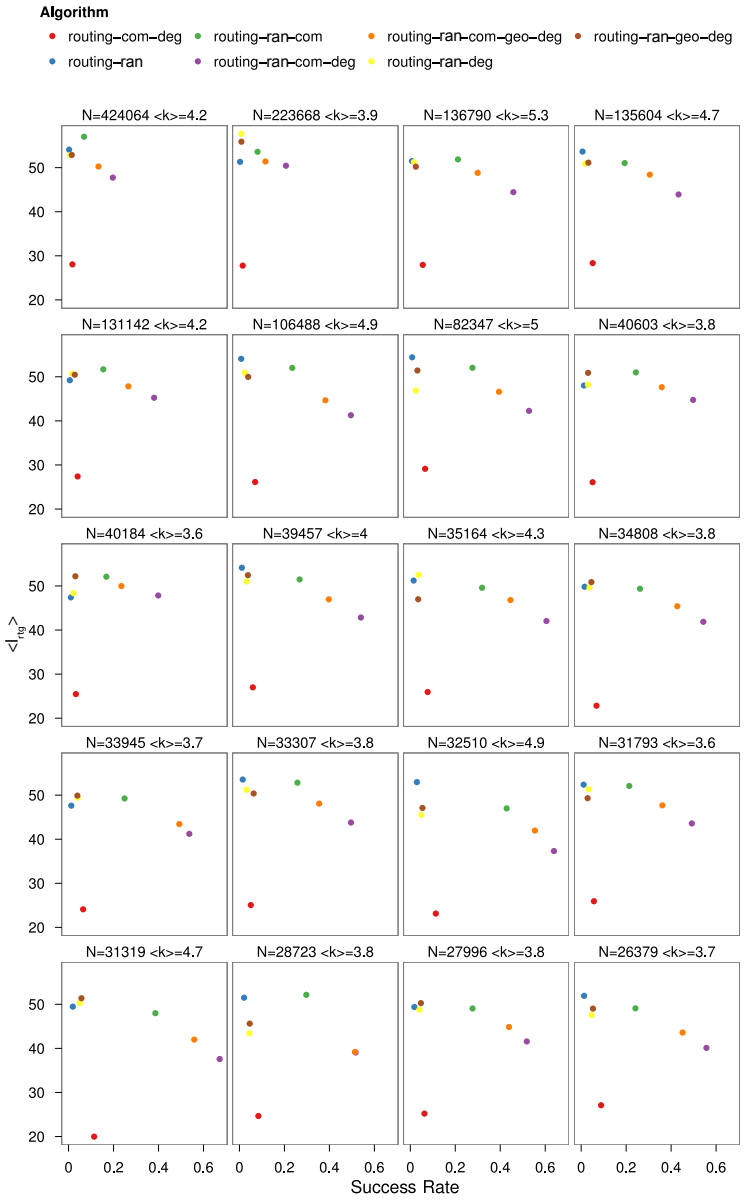


Figure 3.25: Intracity results for the 20 biggest municipalities in Portugal. N denotes the number of nodes, and $\langle k \rangle$ the average degree. Success rates refer to the proportion of messages delivered after 100 steps and $\langle l_{rtg} \rangle$ to the average path length of successful chains.

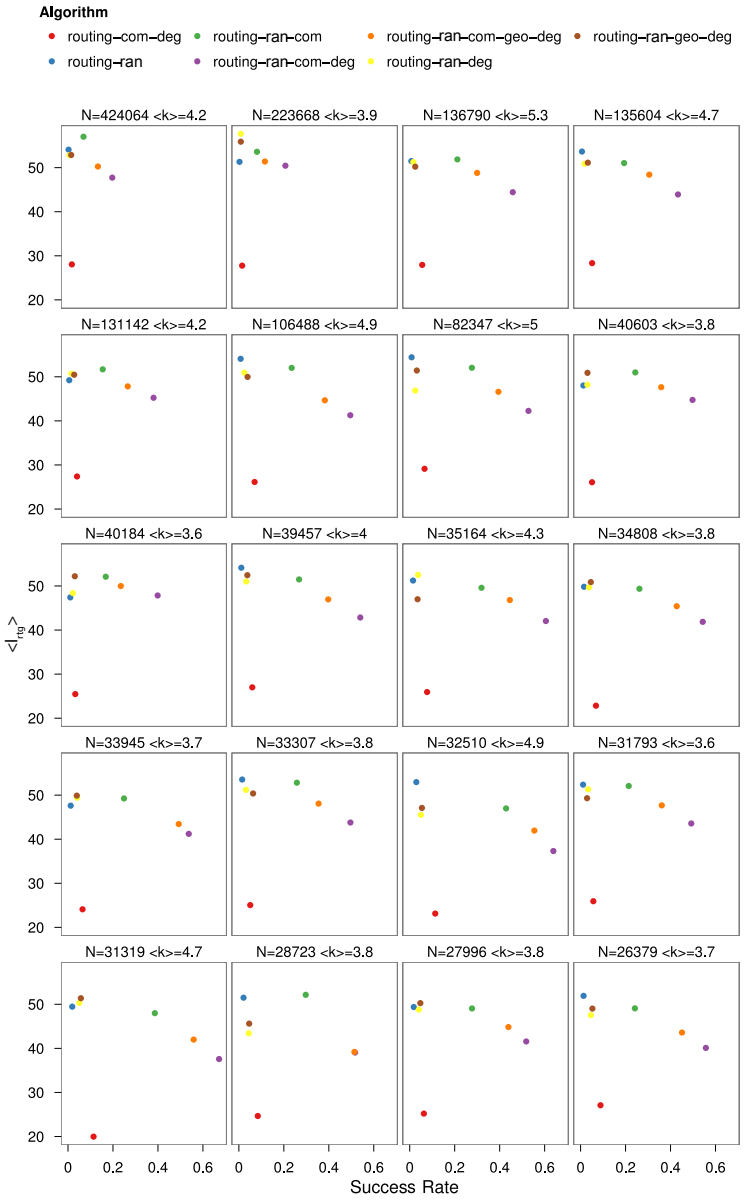


Figure 3.26: Intracity results for the 20 biggest municipalities in Spain. N denotes the number of nodes, and $\langle k \rangle$ the average degree. Success rates refer to the proportion of messages delivered after 100 steps and $\langle l_{rtg} \rangle$ to the average path length of successful chains.

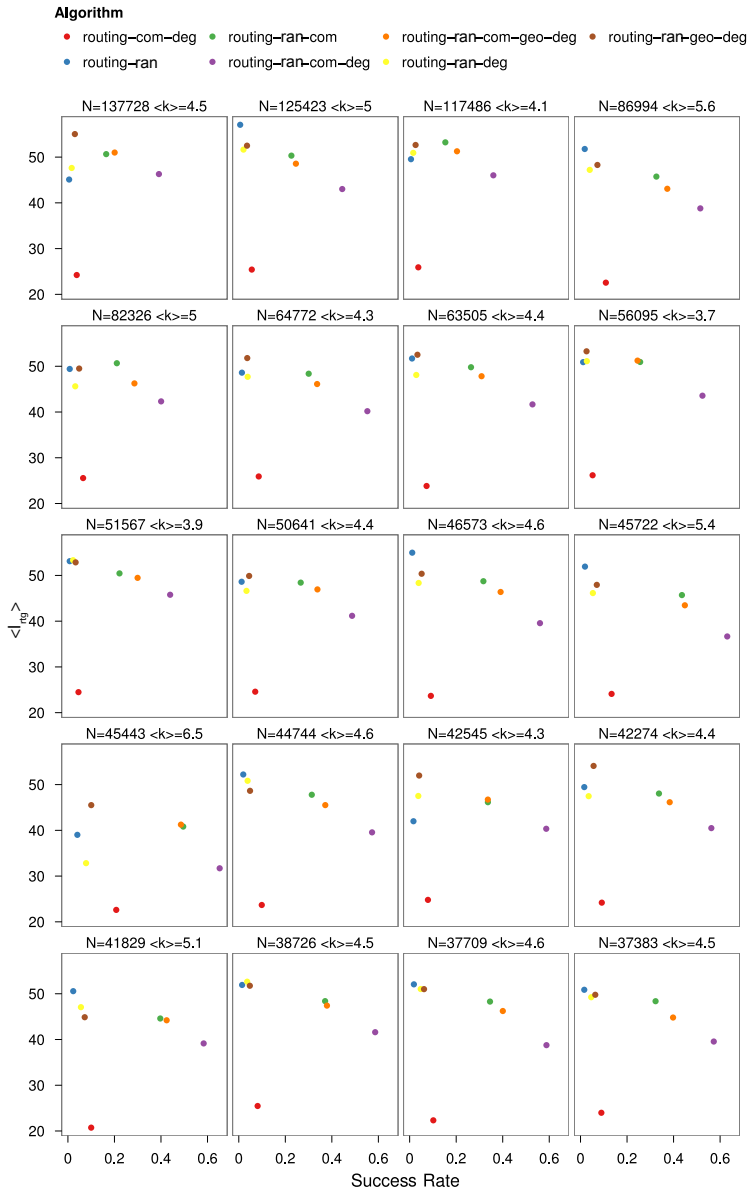


Figure 3.27: Intracity results for the 20 biggest municipalities in France. N denotes the number of nodes, and $\langle k \rangle$ the average degree. Success rates refer to the proportion of messages delivered after 100 steps and $\langle l_{rtg} \rangle$ to the average path length of successful chains.

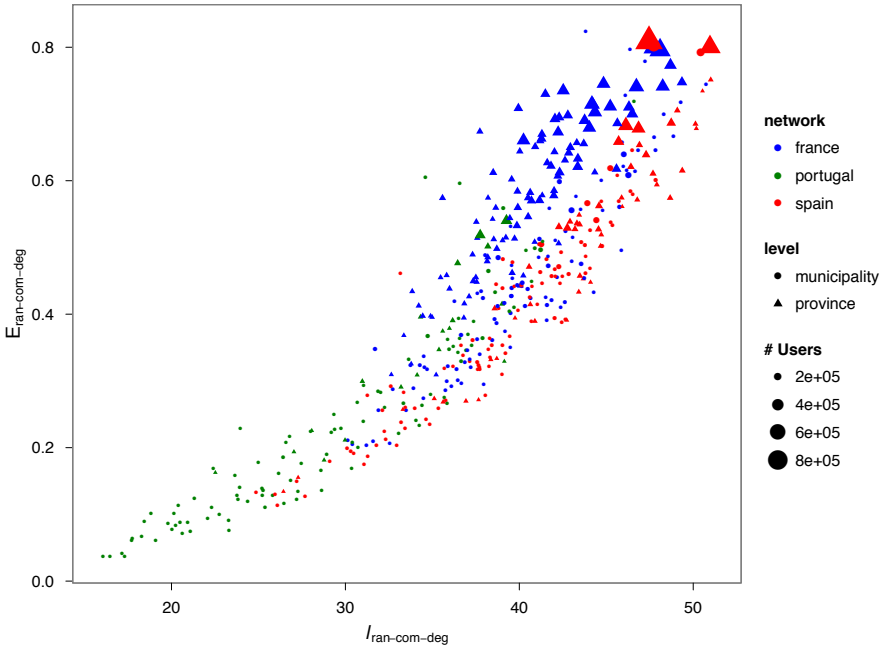


Figure 3.28: Correlation between the average shortest path length and the error rate for each province and municipality, using the *ran-com-deg* routing strategy. The size of the symbols is correlated with the corresponding population.

recent work in the effect of clustering in percolation studies show how a growing transitivity implies that a higher average degree is needed for the emergence of a giant component[116, 1, 9]. Since having a connected network is a necessary condition to route[12], we conclude that our empirical observation is consistent with previous theoretical results: is not feasible to route efficiently in networks with an average degree smaller than 4. In fact as we show in figure 3.29 (bottom), networks with low average degree actually have a significantly larger relative diameter.

As a result of these observations, for the subsequent discussion about intracity routing efficiency, only networks with $\langle k \rangle \geq 4$ have been considered.

Efficient routing within cities

Networks are considered to be *small-worlds* if they have a high clustering coefficient (ratio between closed triangles and connected triples), and at the same time the shortest path length scales with the number of nodes in the network N like $O(\log N)$ [161]. A routing algorithm is considered to be efficient if it is *polylogarithmic* [70]: i.e, it is able to find, between any two nodes, a path of length $O(\log^{\alpha} N)$ with high probability.

Thus, we analyze the three different routing strategies in 155 social networks from the large municipalities and all 150 provinces of the three countries. In contrast to intercity routing, routing inside municipalities is significantly more successful if the strategy uses community information. For different routing strategies Figure 3.30 shows the success rate for municipalities (filled circles) and provinces (open circles) in each country as a function of the population size N; an upper limit of 100 hops is employed. We find that at both scales the community based routing is efficient because of the slow decay in success rate $R \sim c - b \ln N$ ($c = 2 \pm 0.03$ and $b = 0.133 \pm 0.003$) and in contrast to the random strategy, which, as expected, decays almost reversely linear as $R \sim N^{-a}$ ($a = 0.95 \pm 0.03$). Interestingly, the geographically based routing presents a crossover behavior between municipalities (only intracity routing) and provinces (including an initial intercity stage): while within municipalities the routing success rate scales similarly to the random routing $R \sim N^{-a}$ ($a = 0.66 \pm 0.03$), the province routing success rate scales similarly to community routing $R \sim c - b \ln N$ ($c = 0.82 \pm 0.05$ and $b = 0.056 \pm 0.004$), but with a lower success rate.

Crossover in geography-based routing

The performance of different routing strategies in the intracity scenario, considering that a delivery is successful if the message was able to reach the target in less than 100 steps, is shown in Figure 3.30. One interesting

[12]Note that all our networks in our simulations are connected, since we restrict the simulation to the giant component of each area

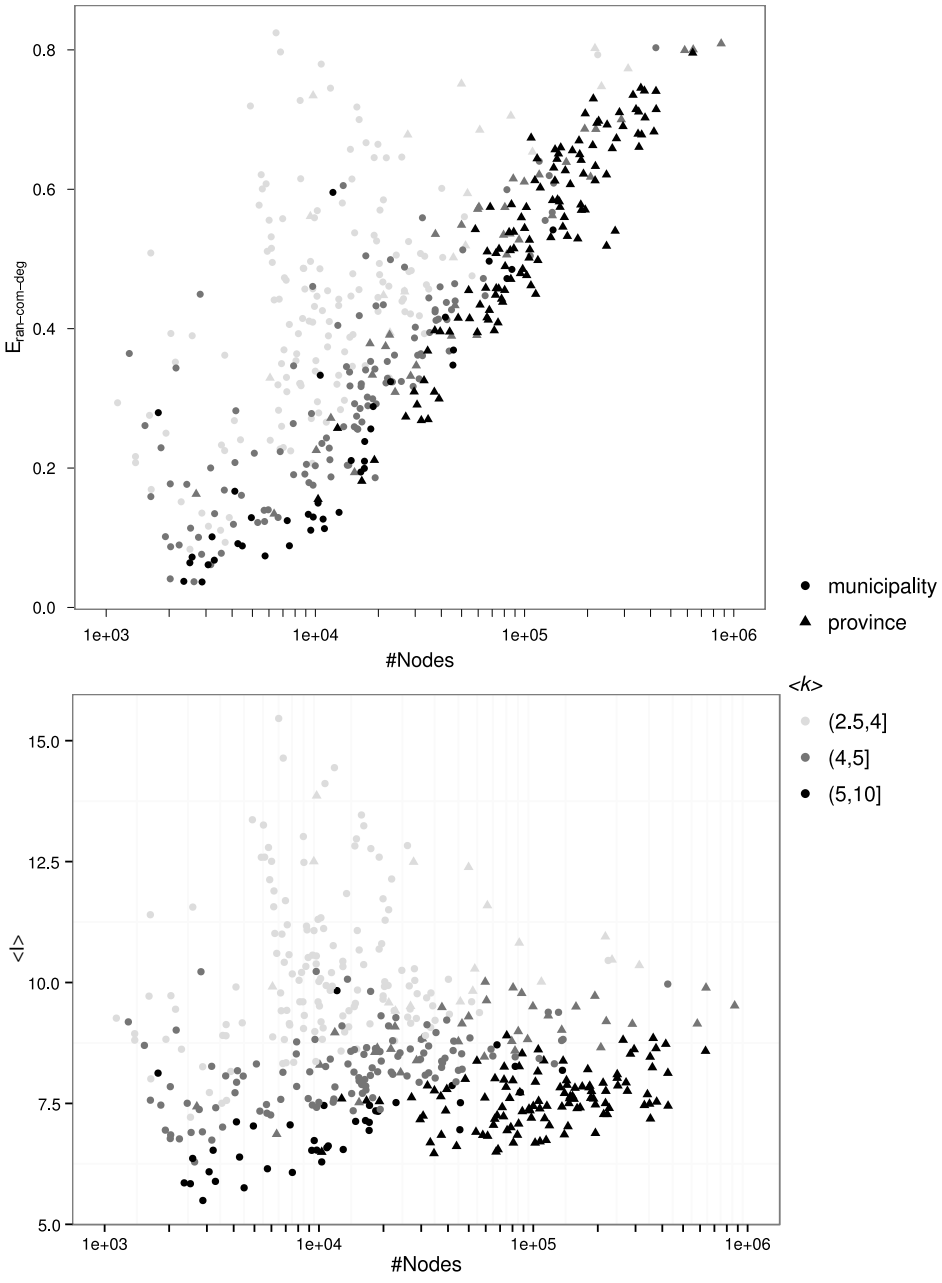


Figure 3.29: Scaling of error rate with network size for the *ran-com-deg* strategy (top). Colors represent the average degree $\langle k \rangle$. If networks are connected enough $\langle k \rangle > 4$, scaling follows a logarithmic behavior. A similar behavior emerges in the scaling of the average path length $\langle l \rangle$ (bottom), where networks with low degree have a diameter relatively large for their size.

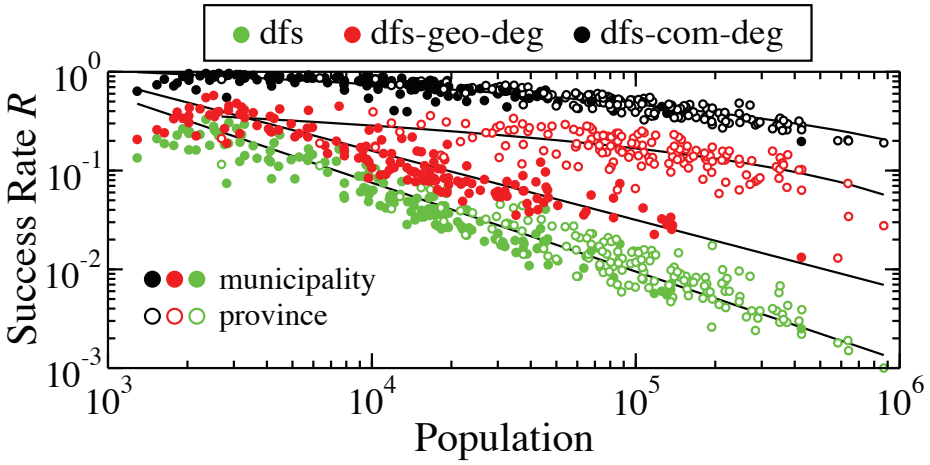


Figure 3.30: Scaling of error rate with size for the *dfs-com-deg* strategy.

aspect is the crossover behavior between municipality and provinces in the geographic based routing. In this section we explain the emergence of such behavior by using a simplified example.

The crossover cannot be linked to a critical spatial characteristic of the city. As shown in Fig. 3.31, we do not find a critical city diameter, area, or density below which the routing fails. This is a strong indication that the geography plays a different role in the social network structure between and within cities.

Figure 3.32 shows a simplified version of a province with N users and 3 cities. Let's denote $P(S)$ the probability that a message is successfully delivered. For *ran* algorithm it is straightforward to conclude the probability $P_{ran}(S) = 1/N$ being N the number of nodes in the network, no matter if the network represents a province or a city. This conclusion agrees with our results in Figure 3.30.

However, for geographic routing, we denote $P(c)$ where $c \in \{A, B, C\}$ the probability of reaching the right city c and $P(S|c)$ being the probability that the message is successfully delivered given it is already in the right city c. In the intercity experiment scheme we have proven that the *geo* approach is valid, delivering the vast majority of the messages to the right city, so we consider $P_{geo}(c) = 1$[13]. Using results from our intracity experiment we

[13]This is a fair assumption since experimental results found error rates $E < 10^{-3}$.

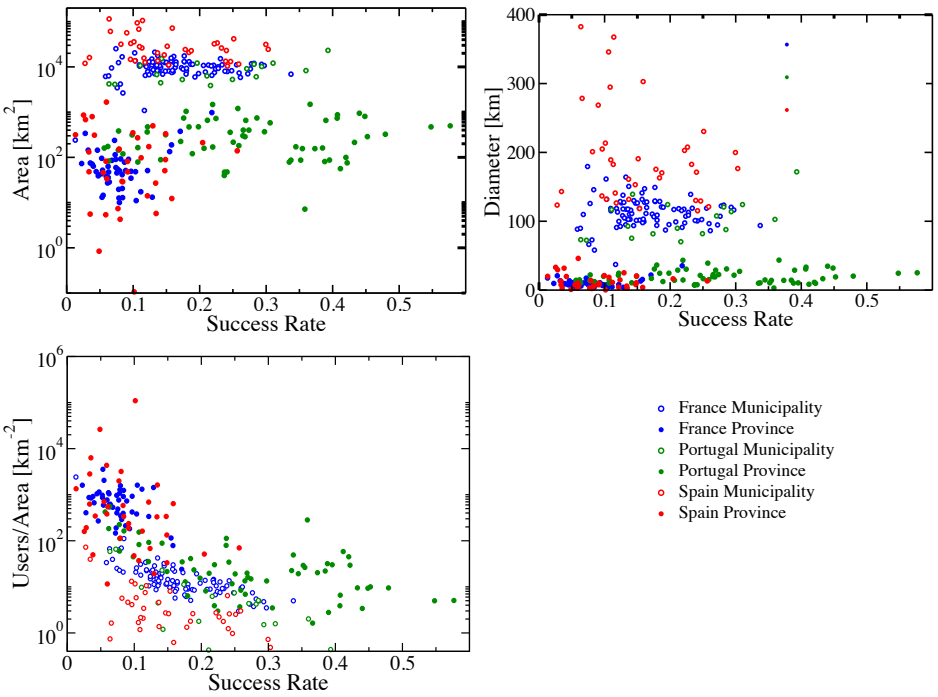


Figure 3.31: The success rate does not seems to be highly correlated with spatial characteristics of the studied region such as diameter, area, or the density of the studied region. Therefore, no critical boundary below which geographical routing fails can be identified.

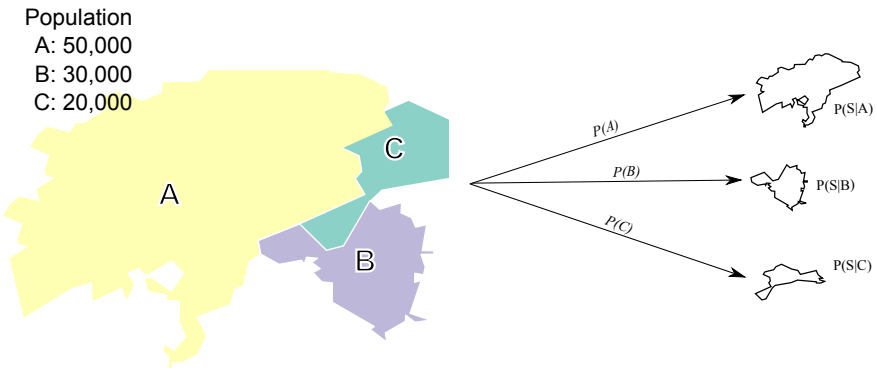


Figure 3.32: Simplified version of a province with population $N = 10^5$ and 3 municipalities. Routing process can be divided into 2 steps: reaching the right municipality and then finding the right target within that city. $P(A)$ denotes the probability that a message whose target is in city A actually reaches A. $P(S|A)$ denotes the probability that a message reaches its target given it is already in A. *geo* strategy is efficient to reach the right city so $P(A) = P(B) = P(C) = 1$ which implies the performance on the overall province is actually better than in the major city, producing the crossover behaviour observed in the results.

assume $P_{geo}(S|c) = \frac{1}{n_c^\alpha}$, with $0 < \alpha < 1$. Then

$$\begin{aligned} P_{geo}(S) &= \sum_{c\in\{A,B,C\}} \frac{n_c}{N} P_{geo}(c) P_{geo}(S|c) = \\ &= \sum_{c\in\{A,B,C\}} \frac{n_c}{N} \frac{1}{n_c^\alpha} = \frac{\sum_{c\in\{A,B,C\}} n_c^{(1-\alpha)}}{N} \geq \frac{(\sum_c n_c)^{(1-\alpha)}}{N} = \frac{1}{N^\alpha}, \end{aligned}$$

which means that using *geo* approach, a province with a certain population N and composed by several municipalities has a higher success rate than a municipality with the same size. Even if we generalize $P_{geo}(S|c) = f(n_c)$ where f is any decreasing function this result holds: if *geo* is capable to deliver all messages to the right city, then $P_{geo}(S)$ is a weighted average of the performances in the cities forming the province such that $f(n_{max}) \leq P_{geo}(S) \leq f(n_{min})$ where n_{min} and n_{max} denote the size of the smallest and biggest cities respectively.

Chapter 4

Geography of social networks

A popular misconception is that the only human-made structure which is visible from space is the Great Wall of China. Despite astronauts having repeatedly reported they cannot differentiate the Chinese structure from the International Space Station (it is basically too thin[1]), everyday one can find a new erroneous tweet on the topic. Even fewer people know that actually the greenhouses in Almería, in the southeast corner of the peninsula, are among the most noticeable human interventions when observing the Earth from outer space. Let's focus on the piece of Chinese infrastructure that is having the deepest influence in the global society during this century. If you are thinking about the wall built by the Ming dynasty, you are wrong again. I am referring to a piece of infrastructure that is neither visible nor from outer space nor from basically anywhere else. In 1997, Wired magazine labeled it as the Great Firewall of China, and it allows Chinese officials to effectively exercise border controls on information. All data packages entering or leaving China are analyzed and, if they match certain patterns, dropped.

One of those patterns is known to be applied on the origin and destination addresses of the data package. If any of those address is related to Facebook Inc., the package is systematically dropped. As a result, Chinese people cannot connect to Facebook. Not surprisingly, a government-friendly, social networking site named Qzone has reached 480 million users in mainland China[2]. However, I invite the reader to do the following thought experiment: imagine that tomorrow, by noon, the Chinese Politburo decides to immediately shutdown the Great Firewall program, and Qzone merges with Facebook, therefore allowing the nearly half billion Qzone users to connect with the 2 billion of existing Facebook users. How many of these of these new 480 million opportunities to create social links would you use?

[1] `https://twitter.com/cmdr_hadfield/status/308961809682014210` .

[2] `http://www.techcrunch.com/2009/02/24/chinas-social-network-qzone-is-big-but-is-it-really-the-biggest/`

The answer to that question would drastically change depending on your location: readers from Hong-Kong or Singapore are much more likely to use these opportunities than those in Orlando, Florida.

This implies that geographic factors drastically influence the formation of social ties. When using electronic means, one might think that the location of the friend might become irrelevant (the cost of creating a Facebook connection is the same whether she lives next door or 4,000 miles away). However, every electronically mediated social network is found to be geographically clustered, as presented in the previous chapter, with the probability $P(r)$ of finding a link within radius r being a decreasing function. The searchability experiments we just presented in the previous chapter suggest that the relation between geography an social networks goes beyond $P(r)$. In this chapter we will explore it deeply and present some consistent new findings about social networks and the geographical spaces they are embedded into. Specially, we will focus on areas where population density is relatively high when compared to that of their surroundings, more typically known as cities.

The structure of this chapter is as follows: first, we will briefly discuss the privacy concerns about geolocated social data. Then we will provide an overview on models of spatial networks. Later we will try to answer the two main questions open in the previous chapter: why geographic routing fails within cities while community routing is still effective if both perform remarkably similar in the inter-city scenario? Finally, we will try to fit our data into Kleinberg's searchability conditions presented in the previous chapter.

4.1 Privacy concerns about geolocated social networks

The relationship between geography and social networks has not been yet fully understood. One of the main reasons for this is the existence of privacy concerns regarding geolocated social data. Large scale social networks are stored by companies such as Facebook, Twitter or mobile phone carriers. Since the wake of social networks analysis in early 2000s all these entities have collaborated with researchers in complex networks to gather a better understanding of the underlying events going on in the networks, so the companies can improve user experience or profitability via targeted advertisement. As a result, a number of very relevant papers have been recently published analysing the social networks stored by these companies [120, 111, 155, 159, 27, 90].

Location data is however a bit of a different animal. While pure social network data can be easily anonymized in an effective way[3], location data

[3]For instance, by hashing the node identifiers in the network.

might enable bad actors to de-anonymize such data. For instance, I currently work at an office near Bernabéu Stadium in Madrid, and in June 2016 I was speaker in two events at Google's Campus Madrid. Both pieces of information are relatively easy to find (for instance, by combining my Linkedin profile, and the public feed of events at Campus Madrid). Consider that my phone carrier would provide someone with "anonymized" phone records from all suscribers in the Madrid area. In that data set, I am not identified by my real phone number, but by the string *dahg83radlanas*. Since it is very likely that I am the only person in the dataset that most of working days shows up near Bernabéu Stadium and was on those two specific days in the Google Campus area, a bad actor could easily conclude that *dahg83radlanas* is actually me, and from there pinpoint my location at any time during the entire time frame where phone logs were available to him.

Parallel to the work presented in this chapter, fellow researchers working at MIT Media Lab (half a mile away from our lab) with a dataset very similar to the ones we have used in this work where trying to answer precisely this question: how easy is to identify one specific person in a massive Call Detail Records (CDR) dataset? The results [30] got a lot of attention from the media, sparkling the debate about the privacy implications of metadata just weeks after Edward Snowden revelations about NSA mass surveillance systems.

The results by Media Lab were surprising, indeed. Mobile phones allow to reconstruct trajectories of those carrying them in their pockets, but with certain caveats regarding space and temporal resolutions: heavy phone users provide data points more often, and spatial resolution is higher in areas with high population density. Media Lab team considered a worst-case scenario by artificially adding certain uncertainty: calling times from the CDR would be rounded to the hour, and locations to a 1 km grid (in reality, both data can be gathered with much higher precision). They found that most of the users could be identified among one million other users by randomly choosing four of their spatiotemporal points.

4.2 Social networks as spatial networks

Social networks are not the only networks whose nodes and edges are embedded in space. All kinds of transportation networks, power grids and some sorts of neural networks are examples of complex structures in which spatial embedding plays significant role. As stated by Barthelemy in his review [15], characterizing and understanding the structure and the evolution of spatial networks is thus crucial for many different fields ranging from urban planning to epidemiology.

Much of the research about spatial networks have been focused on the cost typically associated to the length of edges which in turn has dramatic

effects on the topological structure of these networks. For example, in the world airport network, where edges represent direct regular flights between two airports, maximum edge length is constrained by commercial airplanes autonomy, which by the time of writing this document is around 14,000 kilometers [4]. In social networks emerging from electronically mediated communications, physical distance was not expected to significantly impact the structure of the networks, but it does, indeed. Moreover, next chapter will focus on how human mobility fluxes and human relationship fluxes can be accurately predicted using remarkably similar models.

Another field in which there has been a significant amount of research efforts is the modelling of spatial networks. In his review, Barthelemy distinguished five different families of spatial networks models:

- *Geometric graphs.* They are obtained for a set of nodes located in the plane and for a set of edges which are constructed according to some geometric condition.
- *Erdos-Renyi spatial generalizations.* These networks are obtained when the probability to connect two nodes depends on the distance between these nodes.
- *Spatial small-worlds.* The starting point is a d-dimensional lattice and random links are added according to a given probability distribution for their length.
- *Spatial growth models.* Spatial extensions of the original growth model proposed by Barabasi and Albert.
- *Optimal networks.* Networks that attempt to minimize a certain cost function. These networks were considered already a long time ago in different fields (mathematics, transportation science, computer science) and are now back with the explosion of studies on complex networks.

In this chapter we will refer only basic notions of the first two classes, because the third class has already been discussed for $d = 2$ in Section 3.1.2 and the last two are not relevant to the subsequent discussion.

4.2.1 Geometric graphs

A geometric graph is obtained when the points located in the plane are connected according to a given geometric rule [28]. The simplest rule is a proximity rule which states that nodes only within a certain distance r are

[4] Since 1 March 2016, the longest non-stop scheduled airline flight is Emirates flight EK448 from Dubai, United Arab Emirates to Auckland, New Zealand, and its return flight EK449, at 14,200 kilometres. The service is operated by the new Boeing 777-200LR

connected. A key property of a geometric graph is the minimum r_c distance for which a giant component emerges. This kind of graph is commonly referred as the unit disk graph and has been extensively used to model wireless networks [61].

Percolating restaurants in Boston

To illustrate a geometric graph, we will present a small experiment focused on restaurants in Boston. For the purpose of this experiment, we downloaded all the $3,634$ restaurants in the Greater Boston Area that were available by June 2012 in the Google Places API.

We also developed an small app, that given a certain radius r, produced a map with an overlay of the city with all the restaurants, connecting all those that were closer than r from each other. Figure 4.1 shows the results for $r = 433$ meters and $r = 533$ meters.

The described graph is an example of a simple geometric graph. Therefore we can study percolation in the graph, by looking at the relative sizes of the three largest connected components of the graph, for different values of r. That is precisely what we present in Figure 4.2, where there is a noticeable phase transition around $r = 500$ meters. By looking at the maps produced by the app in figure 4.1, it is easy to conclude what happened: $r = 500$ metres is roughly the minimum distance between two restaurants at different sides of the Charles river, so it is the critical radius for the subnetworks to merge.

4.2.2 Geographical generalizations of Erdös-Rényi

ER model [36] defines a simple random graph where all possible edges are equally likely to be present in the network. As such, is commonly employed as a null model to test certain hypothesis in networks datasets. One simple way to generate it is to run through all pairs among N nodes and to connect them with a probability p. The average number of links is then

$$|E| = p\frac{N(N-1)}{2}$$

giving an average degree equal to $\langle k \rangle = p(N-1)$.

A geographical random graph considers that the probability of connecting two nodes i and j is $f(r_{ij})$, where f is a decreasing function and r_{ij} is the geographical distance between the two nodes. Typically, $f(r) \propto r^{-\alpha}$, so these models are often referred as gravity models [15]. In the next chapter, we will see how gravity models have been extensively applied in the transportation field.

Given a network with N nodes and $|E|$ edges we can build a geographic randomized version by using an square matrix R containing the distances

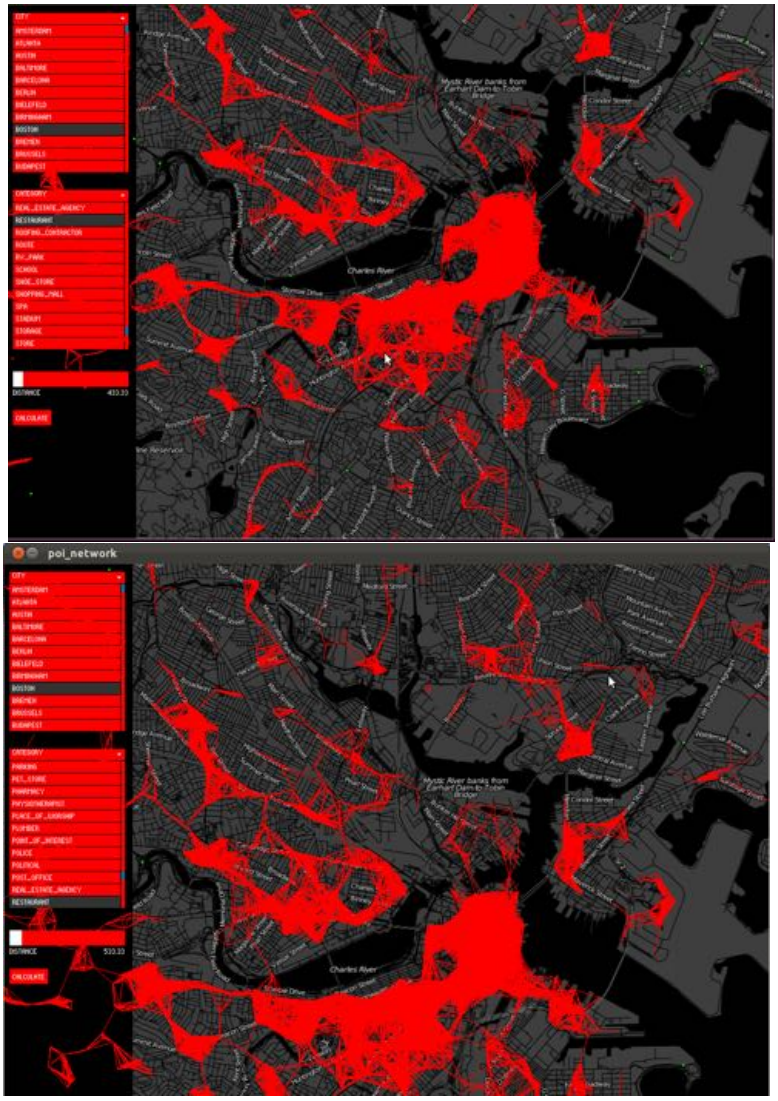


Figure 4.1: Snapshot of the app presenting a geometric graph for restaurants in Boston, for $r = 433$ meters (top) and $r = 533$ meters (bottom). Somewhere between those two radii, the components from the two sides of Charles rivers merge.

between any two nodes. Then we can build an upper triangular C matrix of the form $c_{ij} = \frac{K}{(r_{ij})^{\alpha}}$ where K is a normalization constant such that $\sum_{i=1}^{N} \sum_{j=1}^{N} c_{ij} = 1$. We can iterate through all pairs of nodes and connect them with probability proportional to $pc_{ij}N(N-1)$ where $p = \frac{2|E|}{N(N-1)}$. The resulting graph will be a geographic generalized ER graph.

Another popular geographic generalization of ER graph is the so-called

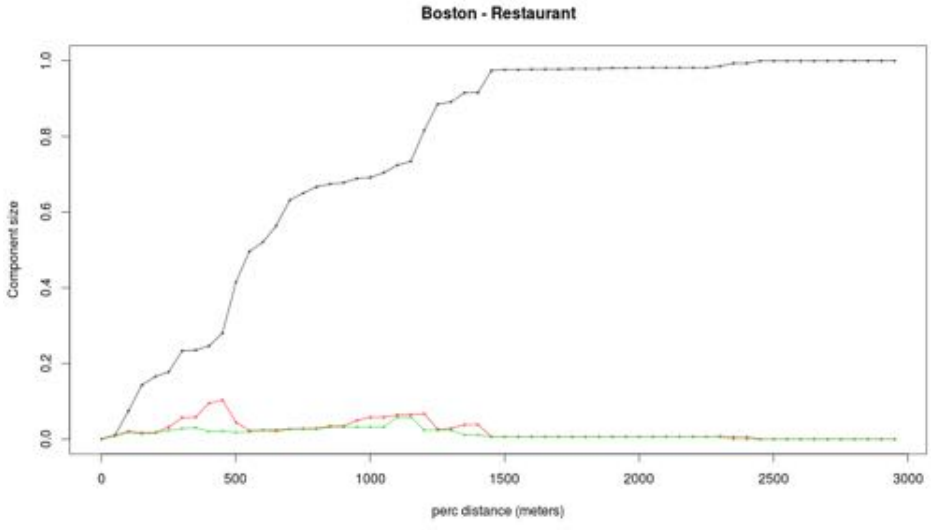


Figure 4.2: Evolution of the relative size of the largest (black), second largest (red), and third largest (green) components for the geometric graphs of restaurants in Boston defined by different values of r.

Waxman model [163], which has been used for example to model the Internet. However Waxman model considers that nodes are uniformly distributed on the euclidean space which as we will see below is not very suited for our purposes. Another interesting family of geographic network models is made of those models that consider a trade off between distance and intrinsic fitness, for example the one proposed by Masuda et al.[105].

4.3 Connectivity collapse within cities

In the previous chapter, we proved that while geographic routing was efficient to reach the right target city even among thousands of them, it performed very poorly (similar to random routing) trying to target a specific person within a city. A necessary condition for any geogreedy algorithm to succeed in a routing experiment is that the subgraph induced by the nodes located within any geographic ball of radius r must be connected for every value of r. This implies that, if a message headed to target user B has reached a user A, A and B must be in the same connected component within the subgraph induced by those nodes included in the circle whose center is in B and has radius up to A. While this is granted in a lattice, it is not necessarily true for other networks (see Figure 4.3a).

We test this structure in our data using geometric and social distances.

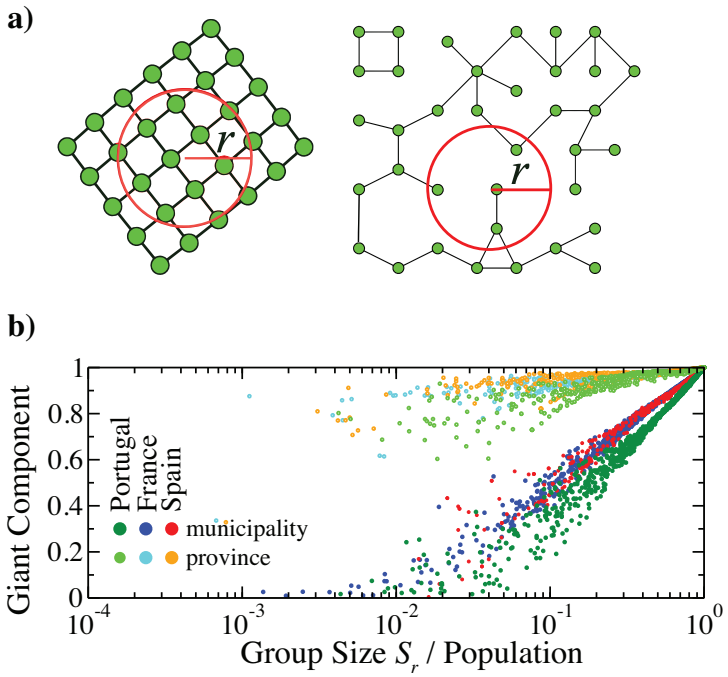


Figure 4.3: Short range connectivity a) In a 2D lattice (left), any geographic ball contains a connected network, however this is not the case for any network (right) where the path between two nodes within a geographic ball might include nodes out of the ball if the network induced by the nodes within the ball is not connected. b) Fraction of nodes in the giant component as a function of the relative size of the geographic ball for the three capitals compared to the country-wide networks. Each of the 6000 dots in the figure was calculated by selecting 2 nodes u and v at random within a city or within the country, extracting the subnetwork defined by the ball whose center is in u and radius up to v, and identifying the number individuals that belonged to the giant component of such subnetwork.

We divide the network into groups of size S_X using either geographic balls [5] of a certain radius r ($X = r$) or existing communities ($X = c$) A natural question emerging then is: which is the critical radius r_c so that geographic balls with $r > r_c$ are likely to contain a connected network? Interestingly we observe that there is not a unique r_c, but rather this radius is defined by the size of a city, so that only geographic balls containing entire cities contain a connected network.

We illustrate this fact further by calculating the size of the largest connected component within different radius and group sizes, performing this analysis centered in different locations from the capital municipality (city) or centered in a province of the three countries. Fig. 4.3b shows that the fraction of nodes in the giant component is much smaller within cities than within provinces. Surprisingly, we find that this lack of connectivity is not caused by not having enough short-distance links (actually between 18% and 40% of the links are within the same location (tower or zip-code)). When we zoom into a region of the city we find small highly clustered groups which form islands; the paths among these geographically neighboring groups exist through people living far away.

To better illustrate this finding we have studied all intra-tower networks in the capital cities and compared them to networks of the same size centered in municipalities in the countryside. Fig. 4.4a shows the average size giant component for towers and municipalities of a certain size. Municipalities with a given population have a larger giant component than a tower in a city with the same population.

Given a fixed number of nodes, a giant component emerges more likely with a higher number of links and with low clustering (a link closing a triangle does not enlarge any connected component). As shown in Figs. 4.4b and 4.4c, both effects are present at the municipality level and not within towers. This explains the different giant component sizes between municipalities and towers. However, high clustering seems to be dominant for the lack of a connected component, since in Portugal the average degree is the same in towers and municipalities. Moreover, the small average degree does not seem to be due to lack of data, since the data from France presents the highest average degree at a country scale, while it exhibits the smallest average degree on the tower scale.

Our results on geographic distance r agree with previous literature [82, 93] showing that the probability of two users within distance r to be connected follows $P(r) \sim \frac{1}{r}$. However, this sole finding does not give us any information about the number of links between people within the same location (tower/zipcocode), since in principle they are within $r = 0$ distance. In order to be able to apply pure geographical models (generating links with

[5] While in this work we only consider 2D geographic *circles* we keep the term *balls* for consistency with previous theoretical work [74]

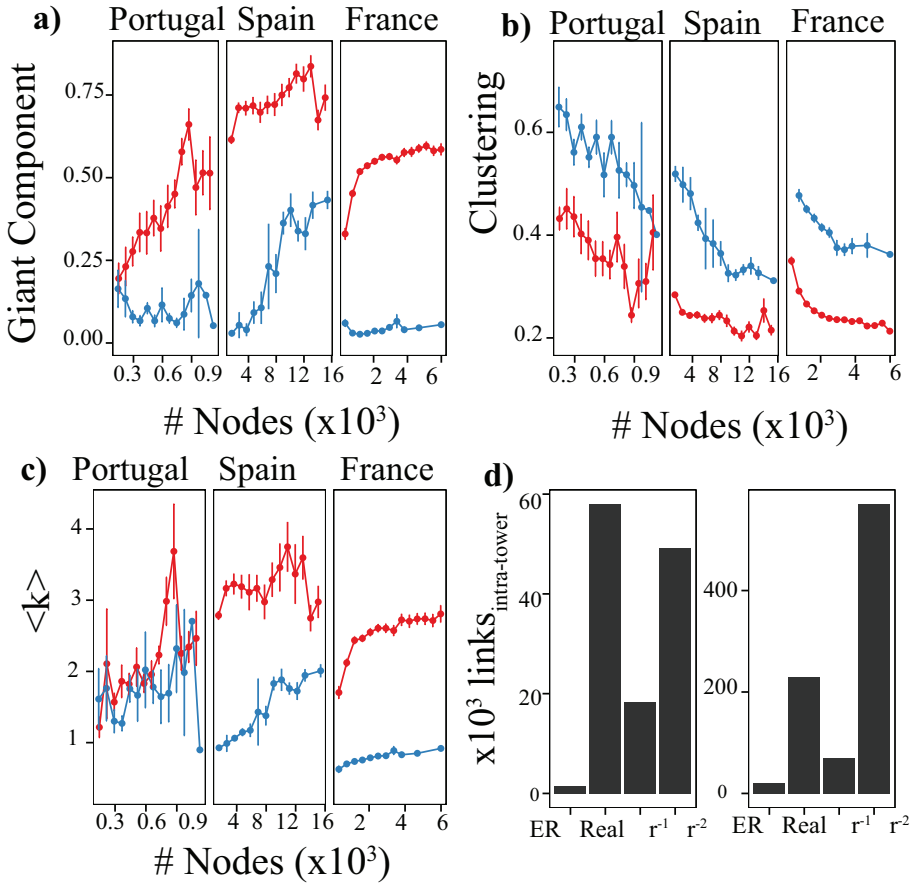


Figure 4.4: Connectivity collapse within cities. a) Relation between population size and fraction of nodes in the giant component for all towers in the capital cities (blue) and municipalities in the country within the same range of population (red). Errors bars represent the standard error of the mean $\frac{\sigma}{\sqrt{n}}$. The size of the connected components within municipalities tends to be higher than within towers of the same size. b) and c) show that towers in cities present smaller average degree and higher clustering compared to municipalities. d) Number of links within the same tower using several randomization models in Paris (right) and Lisbon (left). Results are averaged over 10 runs. The real network has a bigger number of intra-tower links than a space independent graph (ER) and a $\frac{1}{r}$ model. In the case of Lisbon, the real network has even more links than a $\frac{1}{r^2}$ model. To explain the high number of intra-tower links the geographical distance is not sufficient, thus another effect like clustering is needed.

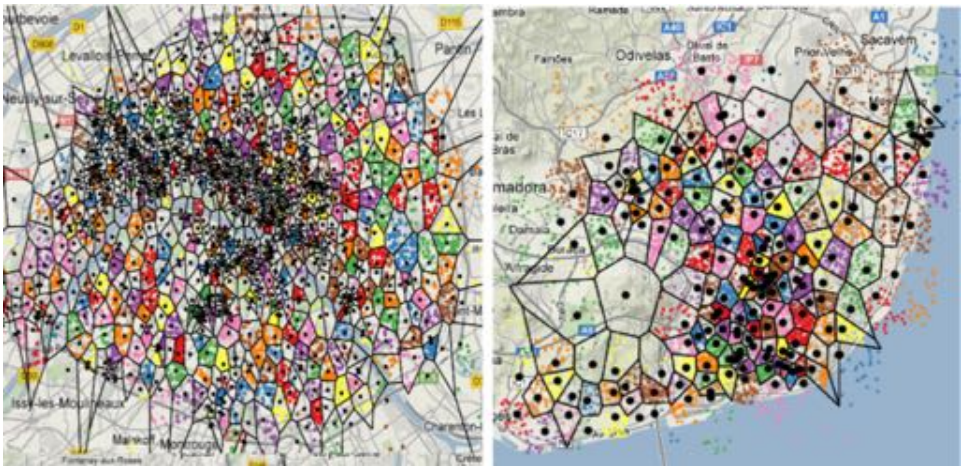


Figure 4.5: Randomization of user location within their own Voronoi cell for Lisbon and Paris. Figures displays the given locations for two thousand random users in the city. Our randomization keeps spatial distribution in the tower level (that is why small downtown cells appear to be full). The maps were created using the R packages *ggmap* and *ggplot2*.

$P(r) \sim \frac{1}{r^{\alpha}}$) to our data, we have to randomize the position of the users around the tower's location. A common assumption for mobile phone data is considering that if a call is processed by a tower, then that tower is the closest to the user's location. This assumption implies that the geographic space can be divided according to the Voronoi diagram of the towers in that region. This way our randomization assigns each user a position uniformly distributed in the Voronoi cell it belongs to. Figure 4.5 shows the randomization process in Paris and Lisbon. This simulation could not be computed for Madrid because the Voronoi assumption is not valid for zipcodes. After randomization, the distance r between any two users is greater than zero, so we can apply $\frac{1}{r^{\alpha}}$ models to compare the number of predicted and present intra-tower links for the same number of links in the whole network. In Fig. 4.4d we show that the number of observed intra-tower links in both cities is higher than what a pure geographical model $1/r$ would generate (even higher than a $1/r^2$ in the case of Lisbon). Despite this abundance of links, there is no giant component, what implies that clustering plays a major effect at this level, producing highly clustered *islands* within the same tower.

4.4 Spatial properties of social communities

Community detection has been an active field in the latest years [43, 3, 18]. Although there are many algorithms published so far, their goal is common: identify dense areas in the social graph. In our routing experiments in the

previous chapter, we used communities as a proxy for social attributes such as school attended, field of work or ethnicity. We found that, contrary to geography, the community structure of the networks still provides searchability in the urban scenario. This was an unexpected result, since previous results on geographic properties of communities always indicated a very high level of spatial correlation within algorithmically detected communities in social networks. After reviewing the most relevant results, we will show how communities do not cluster geographically for urban networks in none of the cities in our data.

4.4.1 Previous results

Geographical properties of social communities have been studied recently. Communities have been reported to be closely related to geographical distance in different scales. Lambiotte et al [82] reported triangles (which can be considered as the simplest community) in the Belgium phone network are severely geographically driven, as shown in figure 4.6.

On the other hand, Palla et. al reported that small k-clique communities in a mobile phone social graph have an unexpected number of people living in the same zip code [122], using phone data from an European country, as shown in figure 4.7.

A interesting line of research about geography of communities has been based on using communities to *redraw* the administrative borders of countries or regions. Particularly, Blondel et al. [19] considered a network similar to ours, with links representing frequency of mobile phone calls, but with nodes representing municipalities instead of individuals. Performing community detection with the Louvain method produced the map presented in Figure 4.8. Note that while the algorithm does not restrict communities to geographically adjacent regions, all detected communities were indeed made of adjacent municipalites. A similar study, named *The connected states of America* [23, 129] has been carried out by the Senseable City Lab at MIT: in this case, the network nodes represented counties for both the UK and the US, producing similar results to Blondel's, as shown in figure 4.9.

Another interesting result was found by Onnella et al [119]. They studied the geographic span of a community s defined as

$$D(s) = \frac{1}{|C_s|} \sum_{i \in C_s} \sqrt{(X_s - X_i)^2 + (Y_s - Y_i)^2}$$

where (X_s, Y_s) is the center of mass of the members of s. Particularly, they studied how D increases with the size of s and they found a significant bump in geographic spin for communities of sizes around 30 (see Figure 4.10). After having carefully tried to reproduce this result is our data, we could not find such gap in any of the three networks.

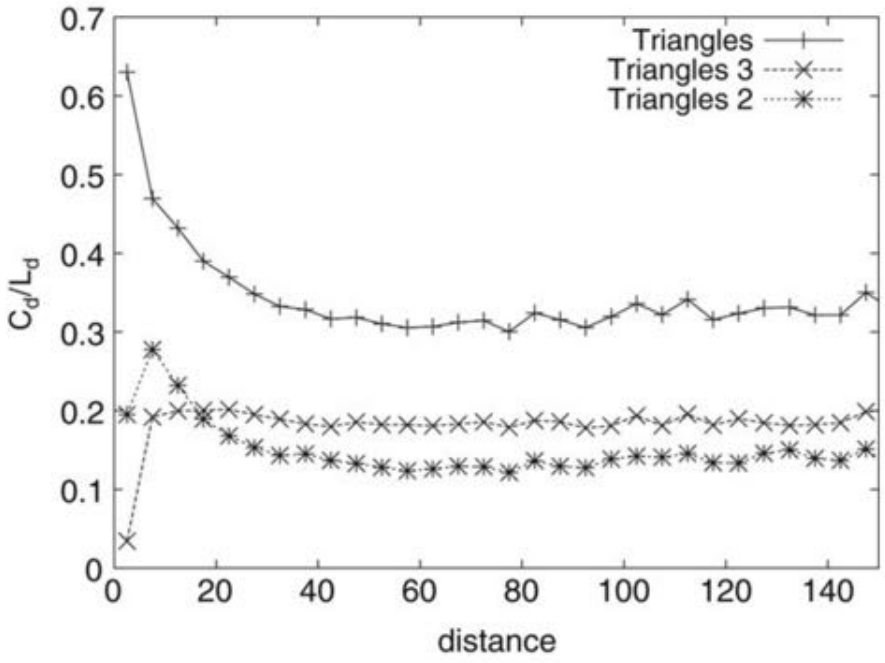


Figure 4.6: Results from [82]. Probabilities c_d, c_d^2 and c_d^3 that a link of length d belongs to a communication triangle, to a triangle of type 2 (i.e., extended over two different zip code areas) and to a triangle of type 3 (i.e., extended over three different zip code areas) respectively. The quantity c_d is seen to decrease until it reaches the plateau value 0.32.

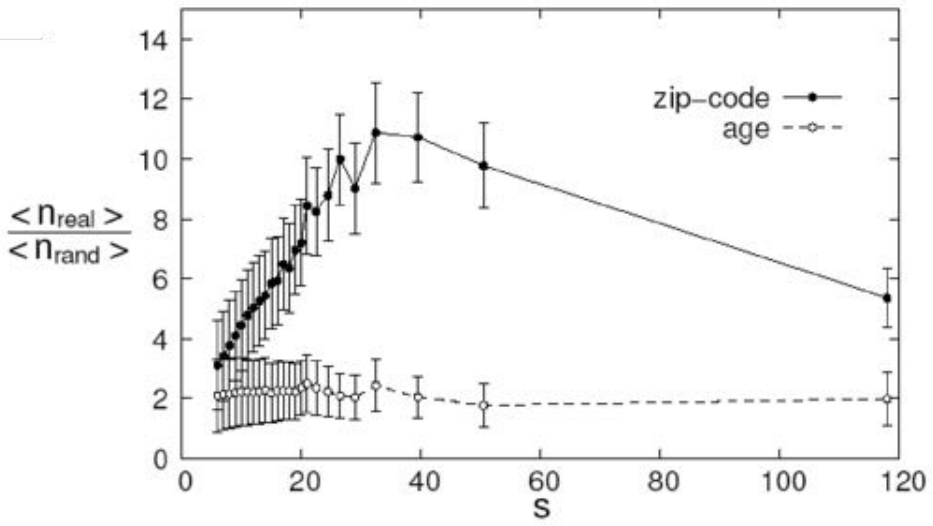


Figure 4.7: Results from [122]. The number of members of a community living in the same zipcode is much larger than when choosing groups of users uniformly at random.

The last relevant result we want to mention in this brief review is the work of Expert et al. [39]. They proposed a Louvain-based community detection that finds communities independent of the geographical space, by using a modified modularity measure that removes gravity-model induced geographic bias.

4.4.2 Methodology

To understand the relationship between communities and geographical space both qualitatively and quantitatively we have followed these steps.

- Perform a community detection on the network using the Louvain algorithm.
- Associate the tower to the most common community among that tower's users.
- Draw a map where each dot represents a tower/zip-code. Dots are coloured according to the most common community among the users living within that tower/zipcode.
- Calculate the average distance $\langle d_{com} \rangle$ between any two towers belong-

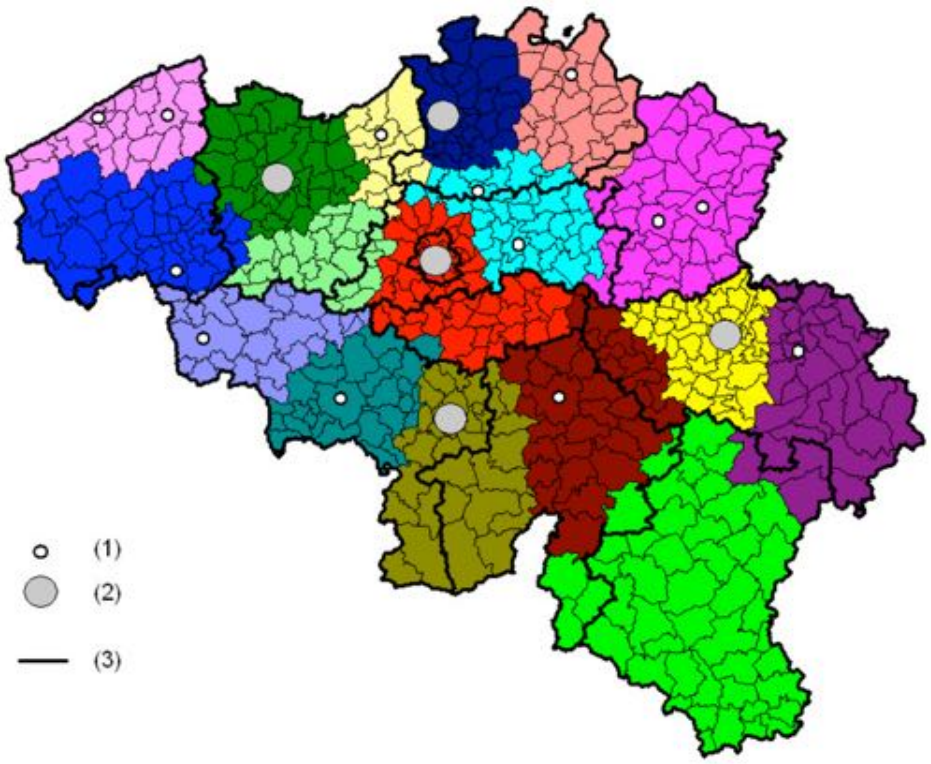


Figure 4.8: Results from [19]. Telephone areas defined based on the frequency of communications between municipalities. (1) = regional city (2) major city and (3) provincial borders.

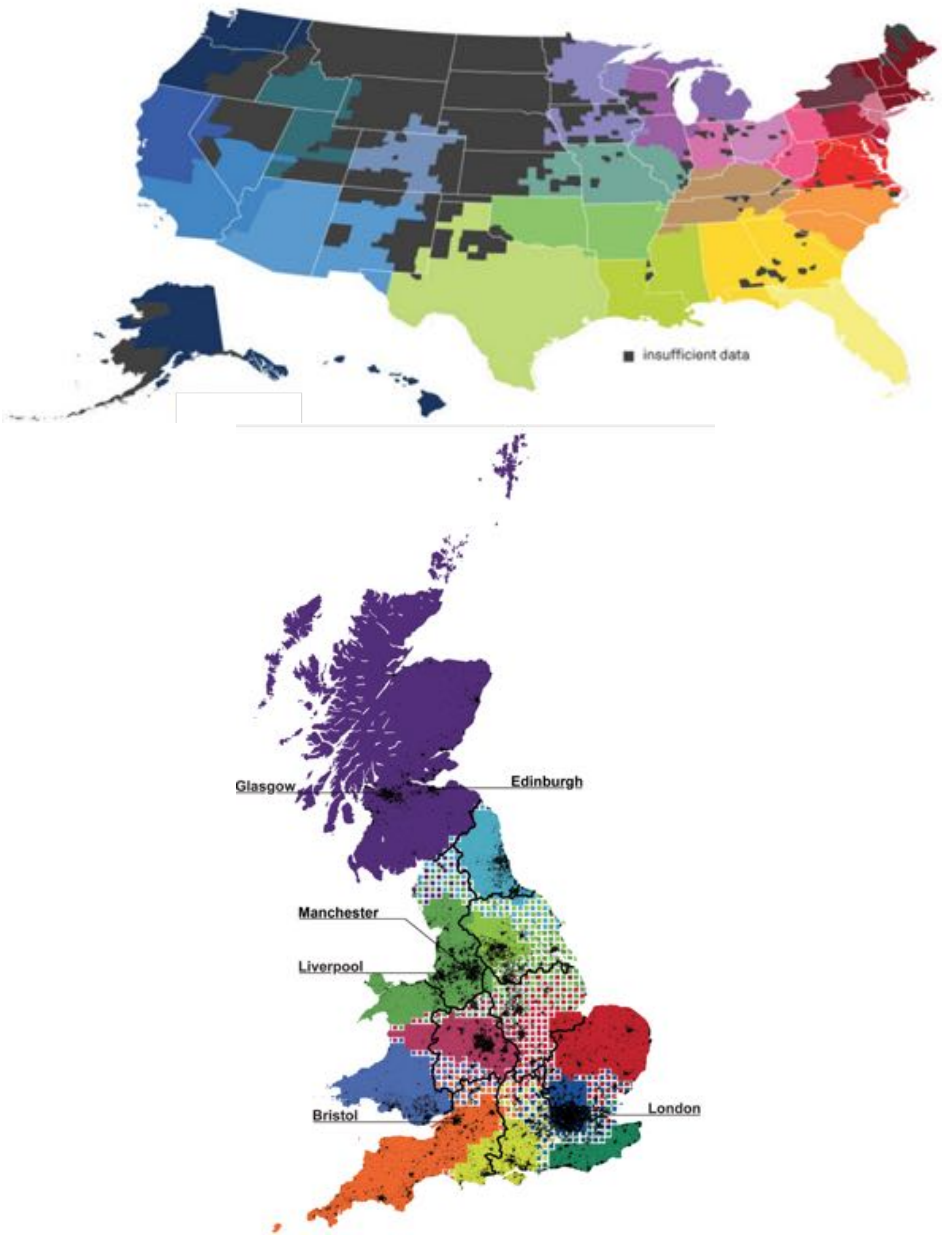


Figure 4.9: Results from [23, 129]. Community detection results from the US and the UK, with an overlay of existing administrative regions. Greys color in the US map show areas where with unstable community assignments due to lack of data.

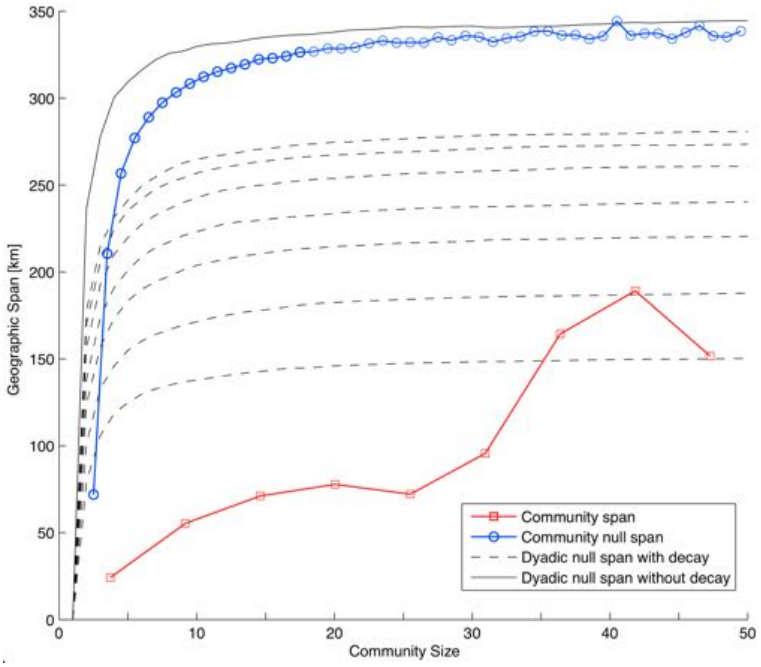


Figure 4.10: Results from [119]. Community span drastically increases when communities have more than 30 members. We could not reproduce this result in none of the three country-wide phone networks.

ing to the same community

$$\langle d_{com} \rangle = \frac{\sum_{c=1}^{C} \sum_{a=2}^{N_c} \sum_{b=1}^{a-1} d(a,b)}{\sum_{c=1}^{C} \sum_{a=2}^{N_c} (a-1)} \tag{4.1}$$

where C denotes the number of communities found, N_c the number of towers in community c and $d(a,b)$ the distance between towers a and b.

- Assign communities with the same sizes randomly to the towers and calculate the average distance as in equation (4.1) with the randomized data $\langle d_r \rangle$.

We have run this analysis with the three countries and the three capital networks

4.4.3 Results

Our analysis of country networks produces extremely spatially clustered communities you can see in Figures 4.11, 4.12 and 4.13 for Spain, France and Portugal respectively. Note that while the algorithm does not require the communities to be geographically clustered, spatial correlation is extremely high. Also a number of community borders match existing administrative borders. Exceptions to the general behaviour are also of interest. In Spain, the Ibiza island belongs to the same community to that of the capital city of Madrid. In France, a similar behaviour shows up between the French riviera and the south of Corsica and the capital city of Paris. In Portugal, the relationship between the south coast and the capital city of Lisbon is not so clear, but the south coast is definitely and area with anomalous results. All these areas have in common a warm climate and abundance of beach tourism. However, the fact that we are assigning users to towers using the most used tower during a long period (6 month) discards that we are capturing vacational effects. A more plausible explanation is that we are capturing communities of people who retired to the coastline, while keeping in touch with most of their relatives and acquaintances in the capital city.

With regard to communities in capital cities, the same analysis produces completely different results, as shown in Figure 4.14. Spatial correlation looks almost negligible in all three cities. This qualitative difference is confirmed by measurements of the $\frac{\langle d_{com} \rangle}{\langle d_r \rangle}$ ratios in the six different scenarios, as presented in Table 4.1, where we find that the ratios of the country networks are four times higher than their urban counterparts. These results show clearly that spatial clustering vanishes in urban communities, justifying how while community and geographic routing perform similarly for country networks, they do not for urban scenarios.

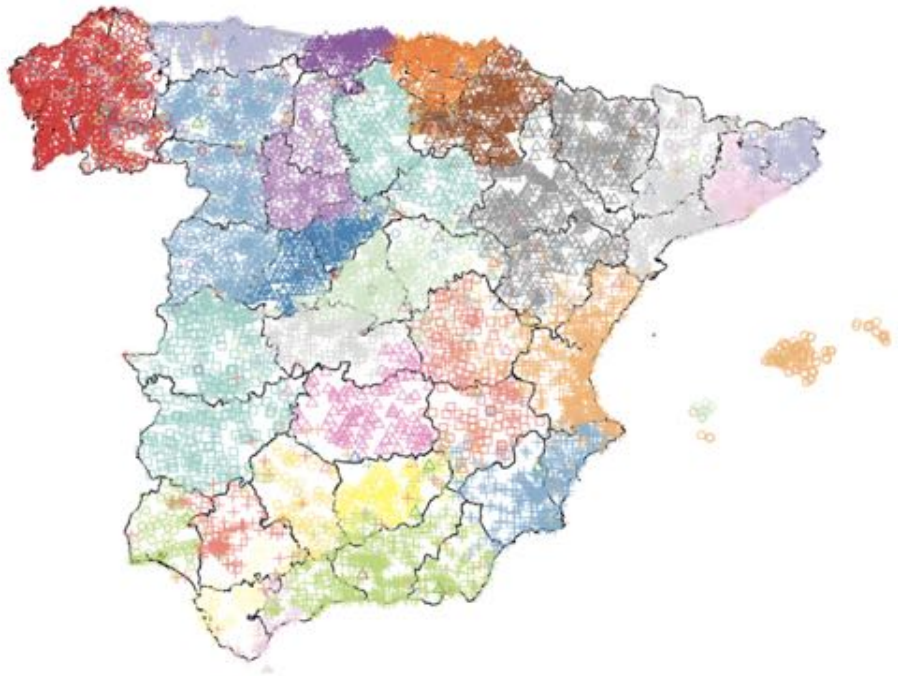


Figure 4.11: Top level communities in Spain. Note that while the algorithm does not require the communities to be geographically clustered, spatial correlation is extremely high. Also a number of community borders match existing administrative borders. A singular exception is the Ibiza island, which belongs to the same community that Madrid (center of the peninsula, green circles).

Network	$\langle d_{com} \rangle$(km)	$\langle d_r \rangle$(km)	$\langle d_r \rangle / \langle d_{com} \rangle$
Portugal	64.4	240.1	3.72
France	115.7	410.71	3.54
Spain	118.5	521.2	4.39
Lisbon (*concelho*)	3.4	4.31	1.26
Paris (*department*)	4.1	5.7	1.39
Madrid (*municipio*)	3.2	3.46	1.08

Table 4.1: Average distance between two towers belonging to the same community. The geographical effect $\frac{\langle d_r \rangle}{\langle d_{com} \rangle}$ is more pronounced in the nationwide communities.

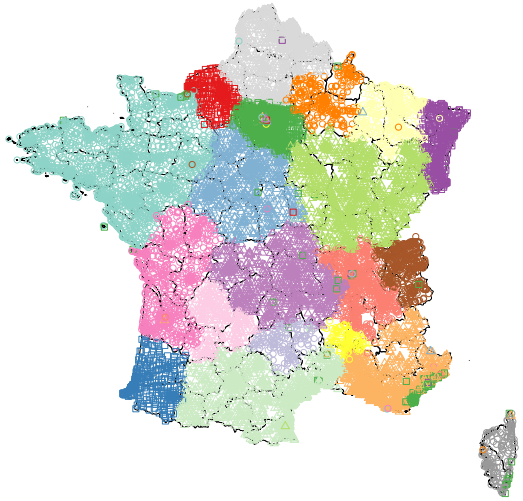


Figure 4.12: Top level communities in France. Compared to others, France is the country where we have more data, and also the one with higher spatial correlation. Remarkable exception are the French riviera in the south east, and the south of the Corsica island, because they both belong to the Paris community (green squares).

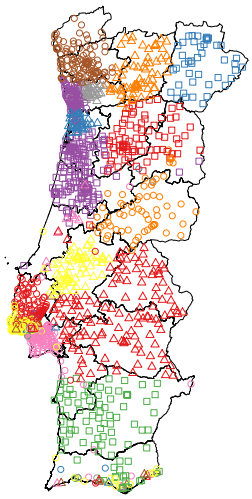


Figure 4.13: Top level communities in Portugal. Compared to others, Portugal is the country where we have less data, but spatial correlation is indeed quite high. Algarve, the southern coast, is the area where spatial correlation is lower, with some towers belonging to Lisbon (yellow circle), but also to other parts of the country.

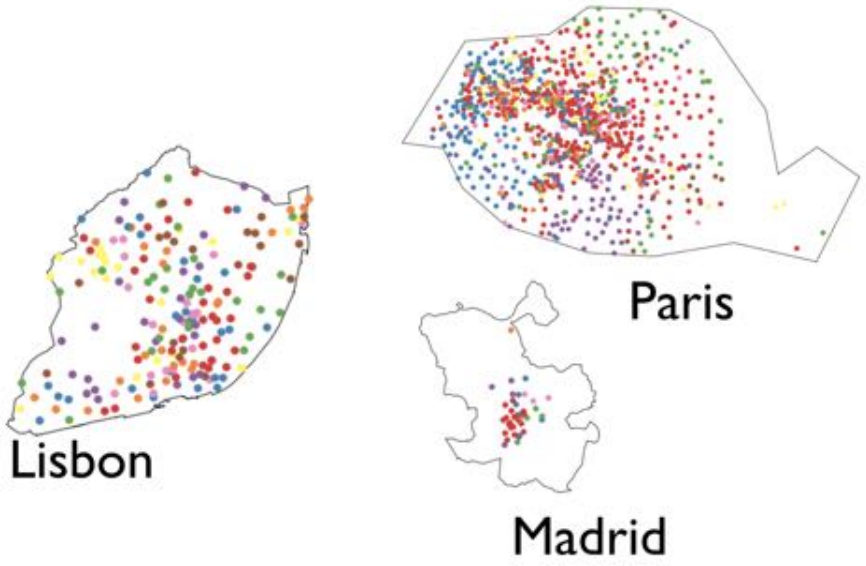


Figure 4.14: Top level communities in the three capital cities. Spatial correlation is now negligible. The communities that provide searchability in the urban scenario are clearly driven by other factors than geography.

4.5 Relation to searchability results

As mentioned in the previous chapter, a number of approaches have been employed in the literature to explain the capability of humans participating in Milgram-like experiments to find short paths: repetitions of the experiment asking the participants about routing criteria are performed [32, 114], computer simulation of decentralized search strategies are tested on real network data [93, 2], and analytic studies focusing on certain properties of networks are conducted [160, 70]. In this last category, lots of attention was attracted by Kleinblerg's work [70, 71] where it is proven that a regular two dimensional lattice can obtain a small world structure by adding randomly links between nodes. Additionally, only if these links are added with probability $\frac{1}{r^2}$[6], a decentralized algorithm is able to find these short paths. Even if this is indeed a very interesting finding, we cannot map our phone network on a two dimensional lattice with additional long-range links.

However, in [74, 72] the same author proposes a generalization which in fact can be applied, which he refers as the *group model*. This generalization has already been introduced in Section 3.1.2 and focuses on the scaling parameter γ to determine if the network is searchable or not.

In principle, both our main routing strategies, communities and geography, can be mapped to groups: it is straightforward in the case of communities since the hierarchy resulting of community detection is a valid set of groups S. For geography, we can consider $g(u, v)$ as the number of people who are closer from v than u, which means S_i are the *balls* of population centered in a tower with a given radius r [7]. However, our networks do not exactly fit the theoretical model because in our data nodes have different degrees. Also several nodes have the same geographical location and therefore there is a lower limit for group sizes in the geographic approach. Despite these limitations, we find that Kleinberg's *group model* allows us to differentiate between searchable and non-searchbale networks with a remarkable degree of accuracy, as we present below.

Qualitatively, both geographically determined balls and communities seem to have the correct exponent as shown for Lisbon in Figure 4.15. However when we calculate the scaling, we find $\gamma_{geo} = 0.85$ and $\gamma_{com} = 1.07$. When we apply this fitting procedure to all cities and provinces in the 3 countries, we find γ_{geo} consistently bellow one and $\gamma_{com} > 1$ for cities while we observe no significant difference on the province level (see Figure 4.16).

To explain these result, let us consider a group S where the target belongs to (it can be a geographic ball or a community). Any decentralized algorithm will search the whole group before trying nodes in other groups. If nodes in the group do not form a giant connected component on the network,

[6] r denotes the Manhattan distance between two given nodes in the lattice

[7] A similar model was actually proposed in [93] to explain how a simple *geogreedy* technique is capable of sending messages to the right city.

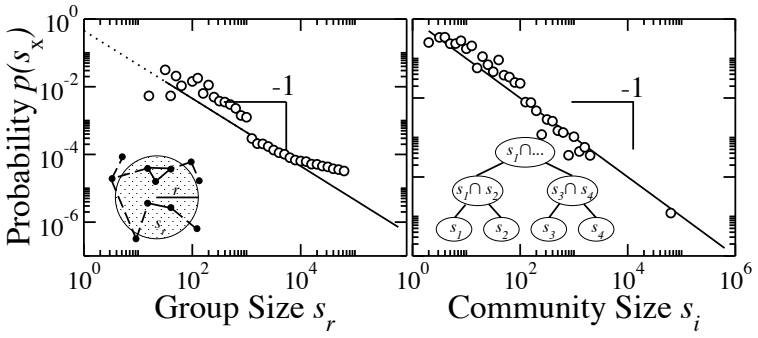


Figure 4.15: Probability of two nodes in the Lisbon urban network to be connected if they belong to a geographical or community group of size S. Both distributions are close to the theoretical S^{-1} needed for networks to be searchable with a decentralized algorithm

the decentralized search fails, because there are *islands* of users. While communities are by definition connected, geographic balls lose connectivity for small radius once within the city, as we saw in Section 4.3. However, as we saw in Figure 4.3, we cannot find such breakdown for geographic balls on the country scale (locating users in municipalities) . This finding agrees with the fact that *geo* strategies are actually efficient on the country scale, as discussed in the previous section.

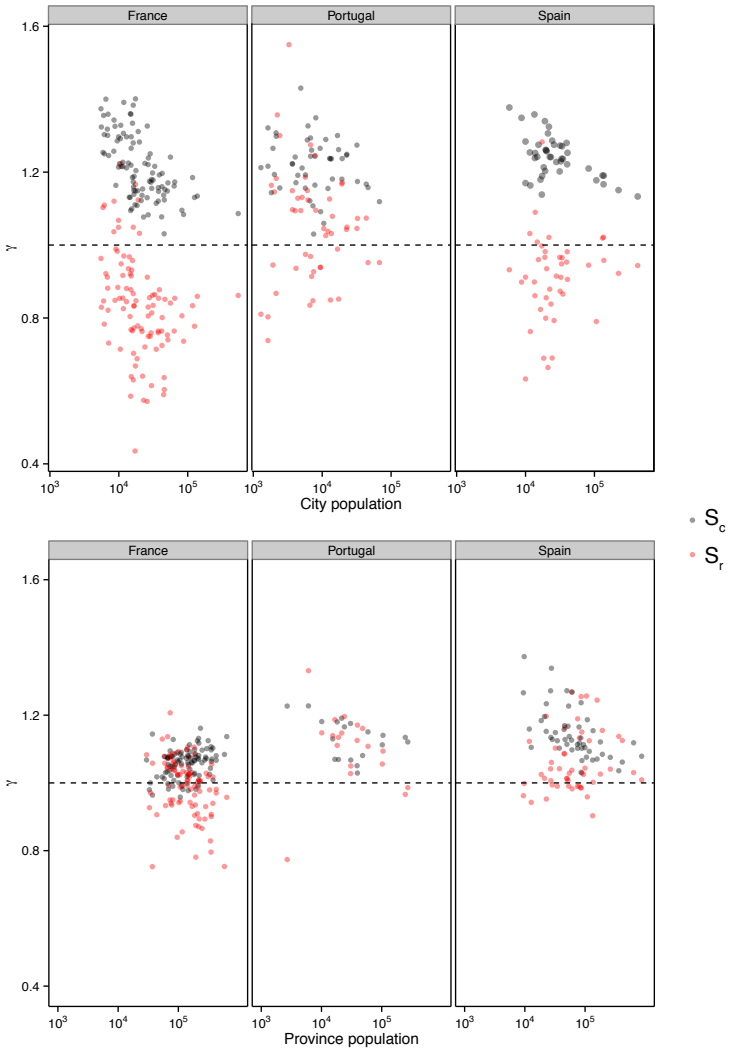


Figure 4.16: Different γ values obtained for geographical and communities groups in all provinces and cities in each country. Results confirm theoretical predictions since in those scenarios where *geo* is not efficient (i.e. cities), $\gamma_{geo} < 1$ while communities show the correct behaviour even within cities. Note some municipalities in Portugal have the right γ_{geo}, which is easy to understand when we find they belong to rural areas Portugal, where municipalities are actually a set of towns so that geographic routing is still efficient to some point because it can find the right town.

Chapter 5

Estimating transportation and communication fluxes from widely available data

In the last chapter, we have reported evidence of a strong relationship between social networks and the geographical space they are embedded into. Particularly, we have found that at any scale, the probability of finding a social tie between two users who live r kilometers far from each other decreases as $\frac{1}{r^{\alpha}}$ being α a small positive exponent.

A remarkably similar behaviour exists in human mobility. Transportation researchers have known for decades that the probability of a person to travel r kilometers decreases with r, often as a power-law like the one found for social ties. In this chapter, we will present how such relationship have been exploited for estimating trip fluxes between two locations, which is of capital importance for different fields ranging for epidemic spreading to infrastructure planning.

The structure of this chapter is a it follows: first, we will introduce the motivations and difficulties to obtain human mobility data. Then, we will briefly review a variety of models that have been used to estimate OD matrices, focusing on recent models which do not require calibration data. After, we will propose an improved methodology to estimate ODs from widely available digital data, and we will proof how it performs in different scales and locations around the world. Last, we will go back to the phone data used in the two previous chapters and propose and benchmark a similar methodology to estimate communication fluxes at different scales.

5.1 Gathering human mobility data

Good estimates of how many people frequently travel between two places are useful for several areas of study. It is not only a key ingredient in modelling

the spreading of infectious diseases [13, 157, 35, 128], but also a fundamental problem in geo-spatial economics of facility distribution. However, the biggest efforts to collect and analyze human mobility data have been done in the transportation planning.

Countries have heavily invested in transportation infrastructure. The McKinsey Global Institute estimates that by 2012 most developed and emergent economies have infrastructure stocks worth around 70% GDP[1]. Despite the huge investment, some of the new infrastructure is commonly underutilized, as in airports with no planes, high-speed trains carrying very few passengers, or subway systems moving tens of people per hour. On the other hand, traffic jams keep growing at unprecedented rate, as in the 2013 jam in Beijing, when a 50-lane highway went over capacity leaving thousands of commuters blocked for up to 12 days[2].

While the public generally perceives these dysfunctions as governance and political problems, the reality is that the engineering problem of developing the right infrastructure to fix mobility problems is still far from trivial. First, there is the problem that the transportation community refers as *induced demand*. In words of J. J. Leeming, a British road-traffic engineer and county surveyor between 1924 and 1964 [85]:

> Motorways and bypasses generate traffic, that is, produce extra traffic, partly by inducing people to travel who would not otherwise have done so by making the new route more convenient than the old, partly by people who go out of their direct route to enjoy the greater convenience of the new road, and partly by people who use the towns bypassed because they are more convenient for shopping and visits when through traffic has been removed.

Sometimes this induced demand can be so significant that it can turn a whole project viable. For example, let us consider the case of the Oresund bridge between Malmo (Sweden) and Copenhagen (Denmark), portrayed on Figure 5.1. Before the construction of the bridge, only around $1,800$ people commuted between the two cities every day. Obviously, that volume does not justify the 2.6 billion euro investment to build the bridge. However, since the opening of the bridge in 2001, the volume of commuters has been increasing year after year, reaching $20,000$ daily users by 2012 and it is expected to continue growing up to $50,000$ by 2025[3].

[1] As reported by The Economist in `http://www.economist.com/news/special-report/21586680-getting-brazil-moving-again-will-need-lots-private-investment-and-know-how-road`

[2] Source: The Atlantic. `http://www.citylab.com/commute/2015/10/chinas-50-lane-traffic-jam-is-every-commuters-worst-nightmare/409639/`

[3] `http://www.norden.org/en/news-and-events/news/more-commuters-from-malmoe-to-copenhagen`

Figure 5.1: The Oresund bridbe between Malmo and Copenhagen, a clear example of induced demand. Before the construction of the bridge, only $1,800$ commuted daily between the two cities. Today, over $20,000$ do, and the number is expected to grow up to $50,000$ by 2025.

5.1.1 Survey based data

Origin Destination matrices (ODs, also referred as trip matrices and denoted T) are the raw material that transportation planners use to decide where to build and improve transportation infrastructure. Considering N locations, the OD will be a square matrix with N^2 elements, each element representing the number of people travelling from a certain location to another.

The most common way to estimate ODs is to run travel surveys: this is, asking enough people where are they headed so that the matrix can be filled with significant data. However, this is difficult to do, since capturing enough data to fill the N^2 elements of the matrix gets very expensive. For example, cities like Boston or Chicago run their travel surveys during one specific day and spend about 10 million US dollars to significantly fill the OD matrix. Because of that, such surveys are run only once every 10 years, which means that planners usually have to deal with data that is not updated (cities significantly vary over 10 years, think of the traffic jam in Beijing mentioned before) and provides no insight about those occasions where transportation system are brought to their limits, for example Thanksgiving weekend.

If running surveys in Chicago or Boston might be problematic, it is even more challenging in developing countries. On the one hand, travel surveys relay on very accurate existing census data and previous travel surveys to decide where to run each survey in order to optimize the significance of the results for a certain budget. On the other hand, running surveys requires a large amount of trained personnel on-site. Both census data and expertise are not so commonly found in developing countries, which are precisely the ones who need the biggest investment in infrastructure.

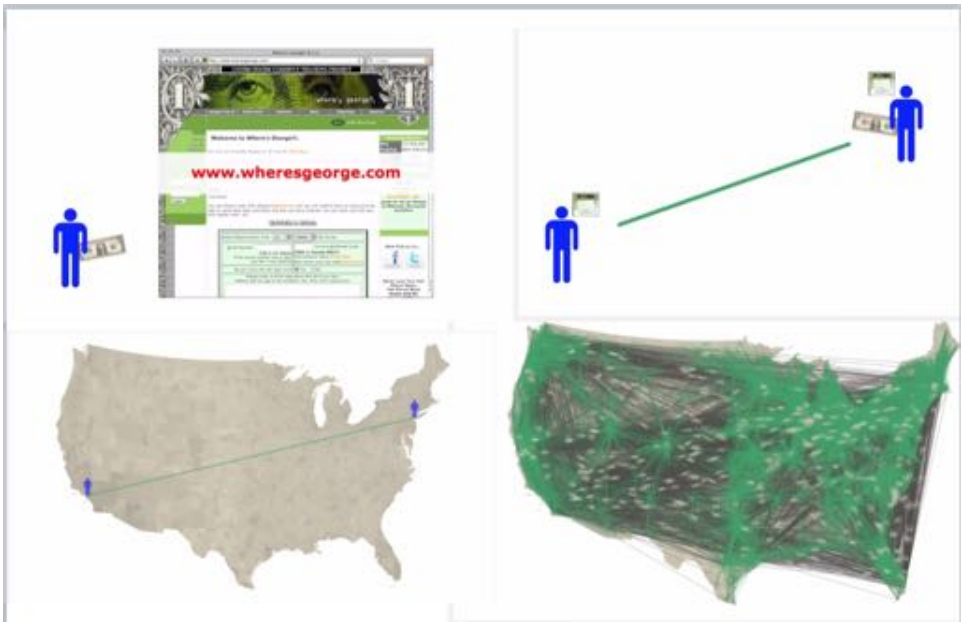


Figure 5.2: A few frames of the video explaining how the Where is George experiment worked. The full video is available at https://www.youtube.com/watch?v=kn32vavZqvg.

5.1.2 New possibilities in the big data era

Considering the limitation and challenges of survey data, the idea of an inexpensive yet comprehensive source of human mobility data has always been considered highly disruptive for the transportation research. And this idea, with the emergence of big data technologies in the last decade, is rapidly becoming a reality.

Follow the money

A very original approach was the one used by Prof Brockmann, who started to gather reports from the website *whereisgeorge.com* to compile one of the largest datasets on transportation research by the time. The website was basically a tracking system for one dollar bills. Anyone could get a dollar bill, register his location and the serial number of the note, and stamp the note to encourage other owners of the bill to register themselves on the web. While the website has been fairly successful creating a community of enthusiasts around, referring to themselves as *georgers*, it was not until Brockmann and his collaborators published their first paper [21] that they realized they had created one of the biggest datasets available about human mobility. Because dollar notes cannot travel by themselves, the fact that one specific note was

spotted in New York, and then in San Francisco, implies that there was a person (strictly, it could be a chain of people) that travelled from one city to the other. By analyzing one million of such hops, Brockmann and his team were able to clearly identify the power-law relationship in the trip length for the first time in a multiscale dataset, containing trips shorter than a mile and also others longer than 1,000 miles.

Human mobility from GSM records

The wide spread of mobile telephony in the 1990s was possible after the the adoption of the Global System for Mobile Communications (GSM) protocol. Every time a GSM compliant phone sends or receives a call, and send or receives a text message (SMS), that specific call or text is processed (and more importantly, logged) by a tower. Because of the functioning of GSM protocol, that tower will be the one that the phone is receiving more clearly, which corresponds almost every time to the tower geographically closer to the phone.

As mentioned in Section 3.2, phone towers tend to be at most 1 kilometer far from each other. Also, the adoption of mobile technology has grown to over 90% of the population even in developing countries. All these facts combined imply that the logs phone carriers were keeping during years for maintenance and billing purposes, are suddenly the most comprehensive and accurate database of human mobility ever available.

In 2008, Gonzalez and her collaborators analyzed human mobility using GSM records for the first time [49]. Their results could not be more promising: by comparing with a control group of users whose position was tracked periodically, they proved that while temporal sampling was irregular, GSM records were good enough to reconstruct people's trajectories with a very high degree of accuracy. The work has become a seminal paper of a research field interested on gathering location data from unconventional sources, having gathered over $3,000$ citations by the time of editing this thesis.

Smartphone era

One step further in the massive acquisition of human mobility data has been the adoption of smartphones, starting with the iPhone in 2007 and now reaching 2 billion devices around the world. Smartphone are location-aware, because they can get Global Positioning System (GPS) fixes, and also they can triangulate their own position based on neighboring phone towers or wi-fi hot spots.

While large scale research based on smartphone location data has yet to be published, the private industry has done a remarkable progress trying to make sense of the raw data. For example, Figure 5.3, shows my Google

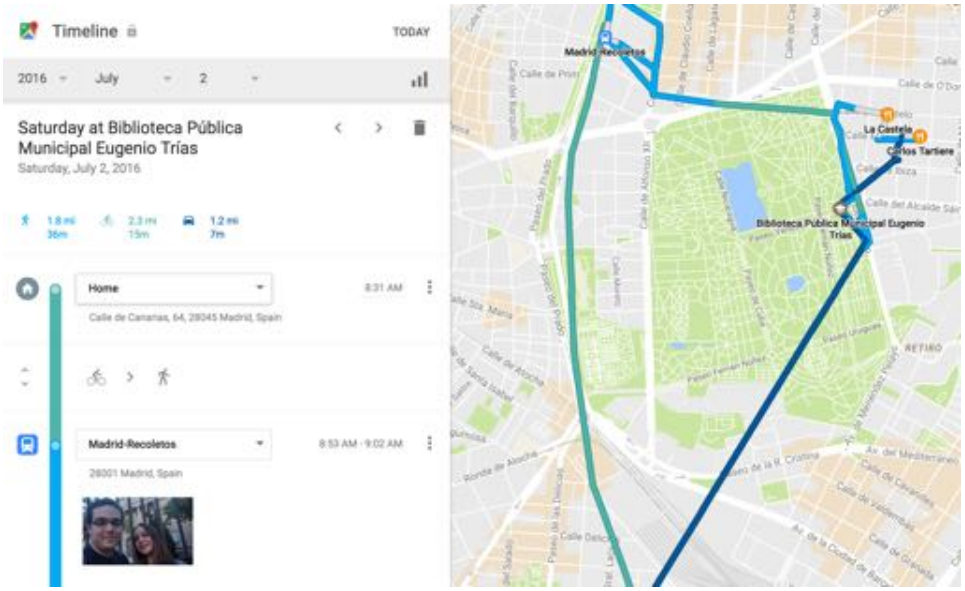


Figure 5.3: Snapshot of my own Google Timeline for July 2nd, 2016. Google was not only able to reconstruct my trajectory with a very high degree of accuracy, but also to correctly identify the two restaurants and one library I visited that day, and also the mode of transportation used (bike, walk, car).

Timeline for July 2nd, 2016. That whole report was produced by Google in a completely autonomous way. It correctly identified that I biked with my sister early in the morning to a library, then to another library until noon, then to couple restaurants for tapas (yes, it got both restaurants correct), then walked around the park and drove back home. If the reader is impressed by this, and happens to be an Android user, chances are he has his very own report available at https://www.google.com/maps/timeline.

Overall, it is reasonable to estimate at this point that not only Google and Apple as operating system owners, but also location-based apps owners such as Uber or Lyft, gather in one day more location data than transportation community has collected during its first 100 years of existence.

5.1.3 The importance of widely available data

The fact that 10 years after the launch of the first iPhone we are yet to see a peer-reviewed publication of a study based on smartphone-generated location data should probably light up some red lights. It appears that the future transportation research will not be about raising funds to grow the corpus of available human mobility data, but about negotiating to get access to massive amounts of data already stored (and probably underutilized) by

a certain corporation.

Interestingly, Google itself has made available to its users almost every piece data they produce through APIs. This way users can transfer the data to a third party that provides them with higher value. For instance, one can easily migrate his pictures or emails to a different service. However, location data has never been made available through API, and, with the exception of the Timeline mentioned before, has been removed from user's eyes.

While location data has its very special privacy issues described in Section 4.1, it is also true that, even if aggregated to preserve anonymity, it is of huge value to society even in fields outside transportation such as agriculture or society. In this regard couple projects should be mentioned as examples of trying to provide society back with value from the data. One is the *Data for development* initiative, by Orange, which made available for research Call Detail Records (CDRs) from Senegal, and asked the community about how to use them to improve the development of the country. The other project is *OpenPaths.cc*, a New York Times initiative that allows users to "donate" their phone-generated location data to different projects.

From the regulation point of view, a shared thought in the data science community is that at some point governments will regulate for holders of large amounts of data to contribute somehow back to the community. In this regard, it is specially interesting the position defended by Prof Sundararajan, who thinks governments will start regulating about accountability, not pure data sharing: for example, Uber would not be obliged to send data to government to be audited about discrimination policies, but rather being handle over the accountability of producing algorithms that helps removing discrimination [151].

Considering all these facts, one of the requirements we set to ourselves for the research presented in this chapter was to validate methodologies that can be applied using publicly available data.

5.2 Modelling OD matrices

As explained before, building significant ODs from empirical data is fairly difficult, which justifies the need of developing models that can estimate such matrices using more accessible data as an input. Additional, empirical ODs completely miss the induced demand. Coming back to the Oresund bridge example, even if Google and Apple would disclose all their data before 2001, the resulting OD matrix would say very little about the induced demand unlocked by the construction of the bridge.

Therefore, the transportation community has used models as a way to predict or estimate the flows in OD matrices using a higher level of abstraction. While gravity models has been the standard from the early days of transportation research, the recent radiation model and its derivatives,

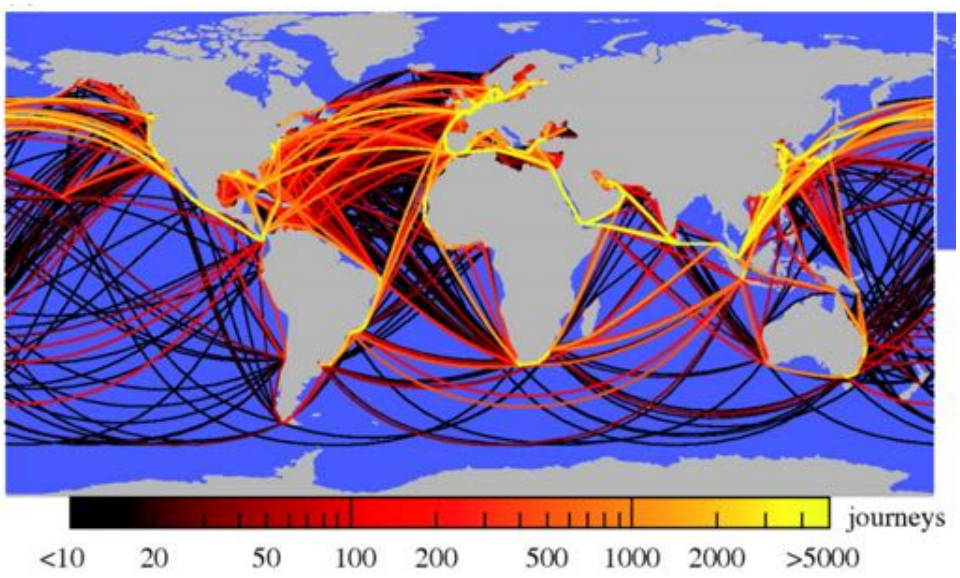


Figure 5.4: Results from [65]. Gravity model has been used to estimate cargo flows between the main ports of the world transportation network.

like the ones we will present in this chapter, have brought new interesting features very suited to a world where the cost of processing vast amount of data has become negligible.

5.2.1 Unconstrained gravity models

Formally defined by Zipf in 1946 [173], but with evidences of practical usage as old as 250 years [112], gravity models have been the standard not only for estimating population movement [15, 37, 64, 152], but also for cargo shipping volume [65] as shown in Figure 5.4, inter-city phone calls [80, 39], and trade flows between nations [127] among other applications. The simplest form of the gravity model is [13, 35, 157, 128]

$$T_{ij} = \gamma \frac{n_i^{\eta} n_j^{\beta}}{C(r_{ij})}$$

where T_{ij} is the flow between zone i and zone j, n_i and n_j are the populations of each zone, r_{ij} is the distance between them and C is an increasing function. η and β represent parameters to be fitted from data and γ is an adjustment parameter controlling the sum of all flows.

The strength of the gravity model is that n_i, n_j and r_{ij} can be computed from a map of the area and list of population of the zones, both resources

very easily available in almost every place on earth. Typically, a function of the form r_{ij}^{α} has been used for C, acquiring T_{ij} the functional form of a gravity law that the model has been named after: in this analogy, population of the two zones mimic the role of masses in a gravity system.

This simple form of gravity model is usually called the unconstrained gravity model because it does not guarantee the attainment of the desired generation and attraction marginal volumes in each zone. For instance, for the commuting scenarios we will be focusing on, the unconstrained gravity model can estimate for an area with $10,000$ population a total number of generated trips $T_i = \sum_j T_i j$ of $15,000$.

5.2.2 Constrained gravity models

In order to tackle the problem of the marginals, in transportation planning the gravity model usually takes the form [167, 37, 8]

$$T_{ij} = \gamma \frac{\eta_i \beta_j O_i D_j}{C(r_{ij})}$$

where O_i and D_j are the total trip production and attraction of zones i and j respectively. This means that for a study region of N areas, we now have to fit $2N$ additional parameters η_i and β_j.

A common approach to do so is to calculate them by iteratively applying

$$\eta_i = 1/\sum_j \beta_j D_j C(r_i j)$$

and

$$\beta_j = 1/\sum_i \eta_i O_i C(r_i j).$$

This is called the doubly constrained gravity model because it ensures consistent values of the trip production $\sum_j T_{ij} = O_i$ and trip attraction $\sum_i T_{ij} = D_j$ per zone, at the cost of needing accurate input of $2N$ additional data points, generation and attraction O_i and D_j, to calibrate the model. Unfortunately O_i and D_j are not often available, because they strongly depend of the scale considered for the entire flux matrix. For instance, a certain certain census tract i in lower east side in Manhattan, New York, will have a different production O_i when considering a T matrix representing all trips within the island of Manhattan or another matrix representing all trips within the entire city.

5.2.3 Radiation model

The radiation model [142] was first introduced in 2012, in order to address some of the issues in of gravity models. Among these issues, Simini et al. highlighted the following:

- Lacking theoretical guidance about $C(r)$, practitioners use a range of functions (power law or exponential) and need to fit too many parameters from empirical data.

- When parameters are fitted from empirical data, resulting values are very different for different T matrices, which implies is not possible to tune the model in a region and then apply it in a different one.

- Because parameter fit does not generalize to other regions, it requires previous traffic data to fit the parameters Thus, it is unable to predict mobility in regions where we lack systematic traffic data.

- The gravity law has systematic predictive discrepancies. For example, in Figure 5.5 they highlighted two pairs of counties with similar origin and destination populations and comparable distance, so according to gravity models the flux between them should be the same. Yet, the US census commuting survey documents an order of magnitude difference between the two fluxes: only 6 individuals commute between the two Alabama counties, while 44 in Utah.

To address these limitations, they proposed the radiation model, whose design principles they exemplified using the inter-county migration fluxes in the United States:

- An individual seeks job offers from all counties, including his/her home county. The number of employment opportunities in each county is proportional to the resident population, n, assuming that there is one job opening for every n_{jobs} individuals. We capture the benefits of a potential employment opportunity with a single number, z, randomly chosen from distribution $p(z)$ where z represents a combination of income, working hours, conditions, etc. Thus, each county with population n is assigned random numbers $z_1, z_2, \ldots, z_{n/n_{jobs}}$, accounting for the fact that larger a county's population, the more employment opportunities it offers.

- The individual chooses the closest job to his/her home, whose benefits z are higher than the best offer available in his/her home county. Thus lack of commuting has priority over the benefits, i.e. individuals are willing to accept lesser jobs closer to their home.

This process, applied in proportion to the resident population in each county, assigns work locations to each potential commuter, which in turn determines the daily commuting fluxes across the country. In summary, the rational behind the model is that people only travel as far as they need to satisfy their needs: a person in rural Iowa is much more likely to travel 10

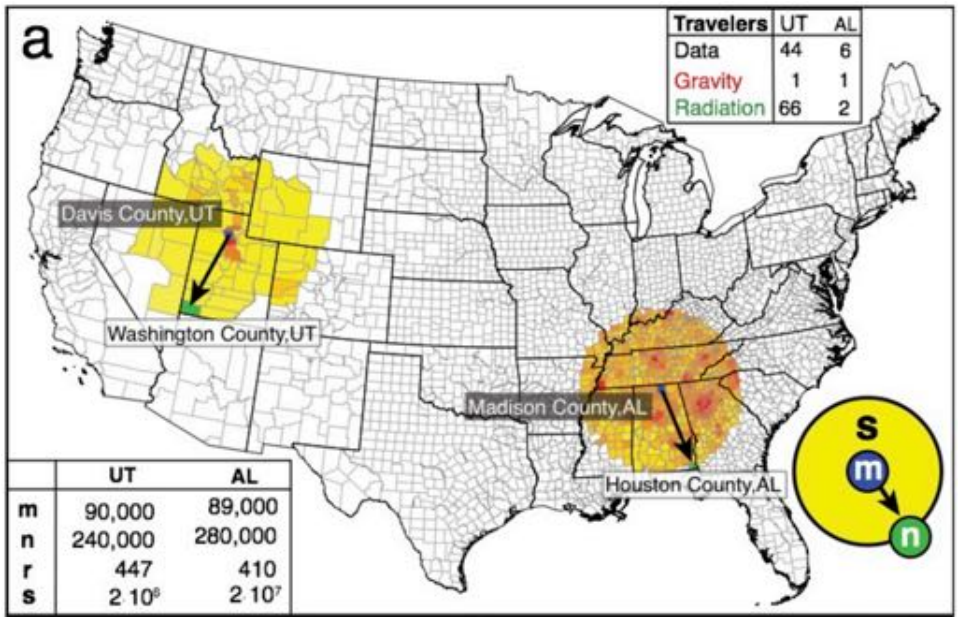


Figure 5.5: Explanation from [142]. To demonstrate the limitations of the gravity law two pairs of counties are highlighted, one in Utah (UT) and the other in Alabama (AL), with similar origin (m, blue) and destination (n, green) populations and comparable distance r between them (see bottom left table). The fluxes predicted by the gravity model are the same because the two county pairs have similar m, n, and r (top right table). Yet the US census 2000 reports a flux that is an order of magnitude greater between the Utah counties, a difference correctly captured by the radiation model.

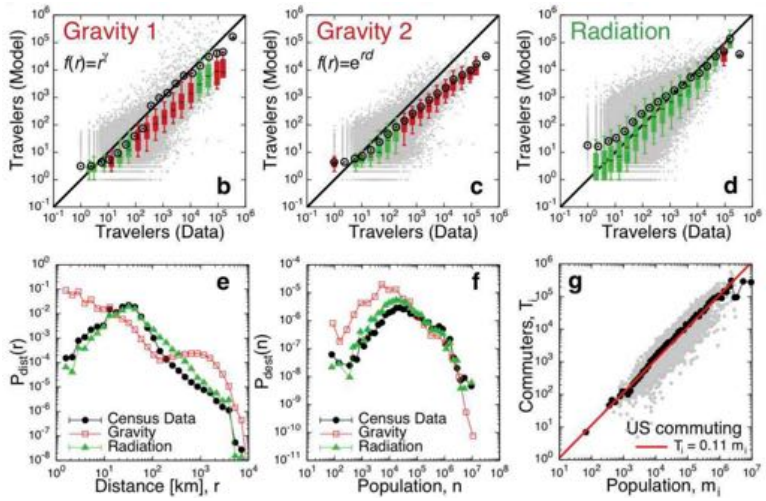


Figure 5.6: Results from [142]. Compared performance for the census data with two formulations of the gravity law, (b) and (c), and (d) with the radiation model. Gray points are scatter plot for each pair of counties. A box is colored green if the line Y=X lies between the 9th and the 91st percentiles in that bin, it is red otherwise. The black circles correspond to the mean number of predicted travelers in that bin. (e) Probability of a trip between two counties that are at distance r kms from each other, $P_{dist}(r)$. (f) Probability of a trip towards a county with population n, $P_{dest}(n)$. (g) Number of commuters in a county, T_i, is proportional to its population, m_i.

miles than a person in New York city, simply because the New Yorker can satisfy most of her needs travelling shorter distances.

The analytical formulation of the model results

$$T_{ij} = O_i \frac{n_i n_j}{(n_i + s_{ij})(n_i + n_j + s_{ij})}$$

where s_{ij} represents the total population in the circle of radius r_{ij} centered at i (excluding the source and destination population). This means the radiation model is parameter free, and can be calculated using the exact same data that the simplest radiation model (note that to compute the circle for s_{ij} only population data and coordinates are needed).

The authors reported higher accuracy in the radiation model compared to unconstrained gravity model in the inter-county migration data and some others, as shown in Figure 5.6.

5.3 Extending the radiation model for improved scaling

The model was originally developed to predict inter-county flows, which not necessarily account for daily trips. Some recent works [86, 91, 104] have shown that the parameter-free radiation model [142] does not work well at predicting intra-city trips. Two of these works [86, 91] introduced models with a calibrating parameter as a cost function of the distance to reproduce intra-city trips. However, these models introduce new parameters that cannot be estimated without trip data.

The present work seeks to answer two questions: How the value of the parameter that imposes the scale dependency can be interpreted and estimated without trip data? Under which conditions the prediction of models not calibrated with empirical trip data would work? The extended radiation model that we present here is distinct from previous formulations, in that it is a stochastic model that depends on the distributions of opportunities and population only. The one scaling parameter α depends on the region size and the heterogeneity of opportunity distribution, which makes it interpretable and estimable in many cases even without trip data. Another important ingredient for modelling trips within small scales (intra urban trips) is the separation between population density and trip attraction rates. Most models on intra-city trips use the density of population [91, 104] as a proxy for both trip generation and trip attraction rates. While this approximation is reasonable at large scales, at inner city scale trip attraction is better represented by the density of point of interests (POIs), defined as geolocated non residential establishments presented on a digital map.

5.3.1 Multi-scale benchmarking: radiation vs constrained gravity

We explore the suitability of the doubly constrained gravity model and the radiation model on predicting commuting flows at three different scales: the Western U.S., the entire San Francisco Bay area, and the city of San Francisco.

The Western U.S. is divided into 183 counties while the two smaller regions are divided into $n_{cells} = 100$ zones to calculate the ODs. Each zone is a cluster of blocks determined by applying *k-means* clustering method on the $7,348$ census blocks in San Francisco and $117,219$ blocks in the Bay area. Note that the unconstrained gravity model is not compared here because when there is empirical OD for parameter calibration, the doubly constrained gravity model performs much better. Detailed comparisons between the two gravity models can be found the SI Appendix of [169].

We apply the doubly constrained gravity model with power distance decay function $C(r) = r^k$ (which is better than the exponential decay function

in the example regions) and compare it with the radiation model. Figure 5.7a shows the commuting distance distribution $P(r)$ of different models at the three scales of study. When we compare inter-county trips in the Western U.S. both the parameter-free radiation model and the calibrated gravity model with $2N + 1$ parameters perform similarly. Adjusting the η_i, β_j and the parameter k in the distance decaying function cannot fit the model well for both short distance trips and long distance trips. This confirms the results reported in [142].

When trying to predict the commuting flows among zones within the Bay area or San Francisco, without parameters for calibration the situation is much harder because the density of population is more homogeneously distributed and commuters tend to go to various business districts across the area (see Figure 5.8). In areas where the population density is homogeneous, radiation model estimates that flows decrease similar to r^{-4} which systematically underestimates long range flows. Also, such scale of daily trips is where the calibration parameters start playing an important role and the calibrated gravity model performs better than the parameter-free radiation model.

In order to inspect further this situation, the distribution of the total number of opportunities a between trip origin and destination is calculated. Figure 5.7b shows that at the Bay area scale (zone size l is around 10 km), there is a region $a < a_{avg}$ where there is not a clear functional form on the enclosing number of opportunities a between the origin and the destination (a_{avg} is the average number of opportunities in a zone). While for $a > a_{avg}$ the probability of finding a trip start monotonically decaying. This effect of clear decaying behavior for $a > a_{avg}$ is not observed in commuting trips within San Francisco. Based on these observations we look for a way to introduce the effects of scale on the radiation model.

5.3.2 Modelling attraction with unconventional data

Let us test a usual assumption in all models used: the population density could represent both the commuting trip generation and attraction rates at different scales. We use the 2010 census LEHD Origin-Destination Employment Statistics (LODES)[4], which provides home and employment locations for the entire U.S. population at block level.

The first column in Table 5.1 shows the correlations between densities of commuting flow generation, attraction and population in the Western U.S. Both of them have high correlations, so at this scale the assumption holds. Figure 5.8 shows the commuting trip generation and attraction rates in San Francisco. Their distributions are less similar. This is straightforward to understand when we think of the classic American city. While very few

[4]https://explore.data.gov/labor-force-employment-and-earnings/lehd-origin-destination-employment-statistics-lode/zvvq-y3uj/

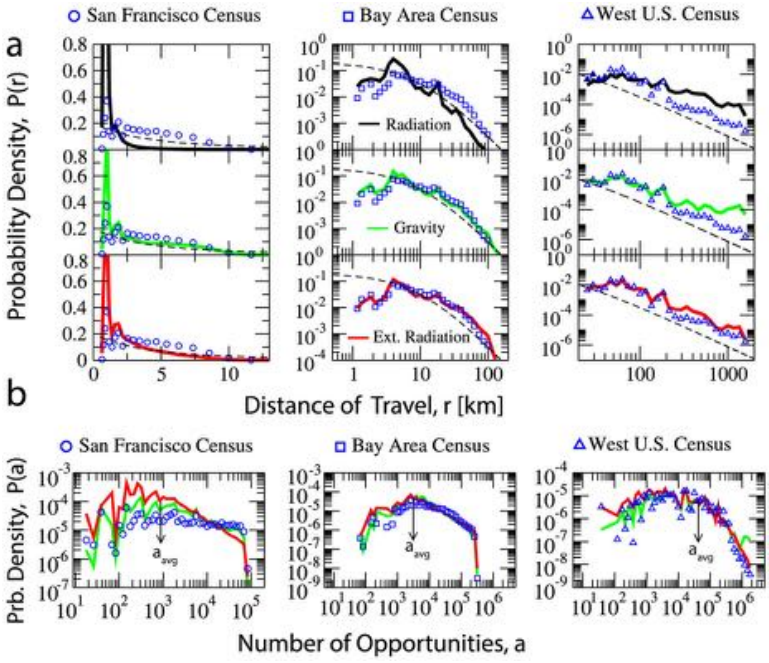


Figure 5.7: (a) The three columns represent San Francisco, the Bay area, and the Western U.S. respectively. The three rows are the results of the three different models compared with the census data. The radiation model gives relatively good prediction only at the Western U.S. scale. At the two smaller scales, the radiation model under-estimates long distance trips. The doubly constrained gravity model gives close predictions to the census data at the three scales. The extended radiation model, with one parameter α, achieves the same prediction quality. The dashed line is a guide to the eye with the distance decaying function $P(r) = 100(r + 10)^{-2.7}$. (b) The $P(a)$ distribution, is the probability of measuring a commuting trip with a opportunities between the origin and the destination. Because the radiation model is not suitable for the two smaller scales, here only the extended radiation model and the doubly constrained gravity model are compared with the census data. The flat distribution of $P(a)$ in the Census data within San Francisco differs from the other two scales, showing that the distribution of intra-city flows is influenced less by the number of opportunities between the origin and destination.

	Western US		Bay Area		San Francisco	
	Pop	POI	Pop	POI	Pop	POI
Trip Generation	0.993	0.926	0.971	0.491	0.956	0.292
Trip Attraction	0.989	0.930	0.417	0.859	0.157	0.880

Table 5.1: Spatial correlation between attraction, generation, POIs and population.

people live in financial center areas in downtown, a lot people commute there (low population, high attraction). Similar thinking can be applied to highly populated residential suburbs. Figure 5.9 presents the degradation of population-attraction correlation for increasingly smaller scenarios.

Thus, we need to find a better proxy for commuting trip attraction rates at smaller scales. Digital traces of facilities are available on-line, and they provide good estimates of commuting trip attractions [132, 150, 50]. In this study we use the density of point of interests of each zone to represent the commuting trip attraction rate. The three study regions contain 1,774,154; 319,170 and 85,230 POIs extracted from Google Places respectively. According to Table 5.1, at all the scales POI density has high correlation with the commuting trip attraction rate.

5.3.3 Formulation of the extended radiation model

In [143] the authors proposed a unified framework for different mobility models. The framework considers the probability $P_>(a)$ of not choosing the closest a opportunities. For example, in this framework a uniform selection model becomes

$$P_>(a) = 1 - a/N \tag{5.1}$$

where N represents average job openings per unit time. Using this framework, they found that for the original radiation model

$$P_>(a) = \frac{1}{1+a}. \tag{5.2}$$

In this text, we propose an extension to the radiation model of the form

$$P_>(a) = \frac{1}{1+a^\alpha} \tag{5.3}$$

such that for $\alpha = 1$ the original radiation model is recovered. In [169], we included analytical derivation of the derivation of the model using the framework of survival analysis, yielding

$$T_{ij} = \gamma m_i \frac{P(1|n_i, n_j, a_{ij}}{\sum_k P(1|n_i, n_k, a_{jk})} \tag{5.4}$$

where γ is the trips to population ratio, m_i is the population at the origin and $P(1|n_i, n_j, a)$ is the probability that a person commuting from a region with n_i opportunities to a region with n_j opportunities and with s_{ij} opportunities in between (note $a = n_i + s$) accepts one of the n_j opportunities, given that the closest n_i are not chosen. Formally,

$$P(1|n_i, n_j, a) = \frac{P_>(a) - P_>(a + n_j)}{P_>(n_j)} \tag{5.5}$$

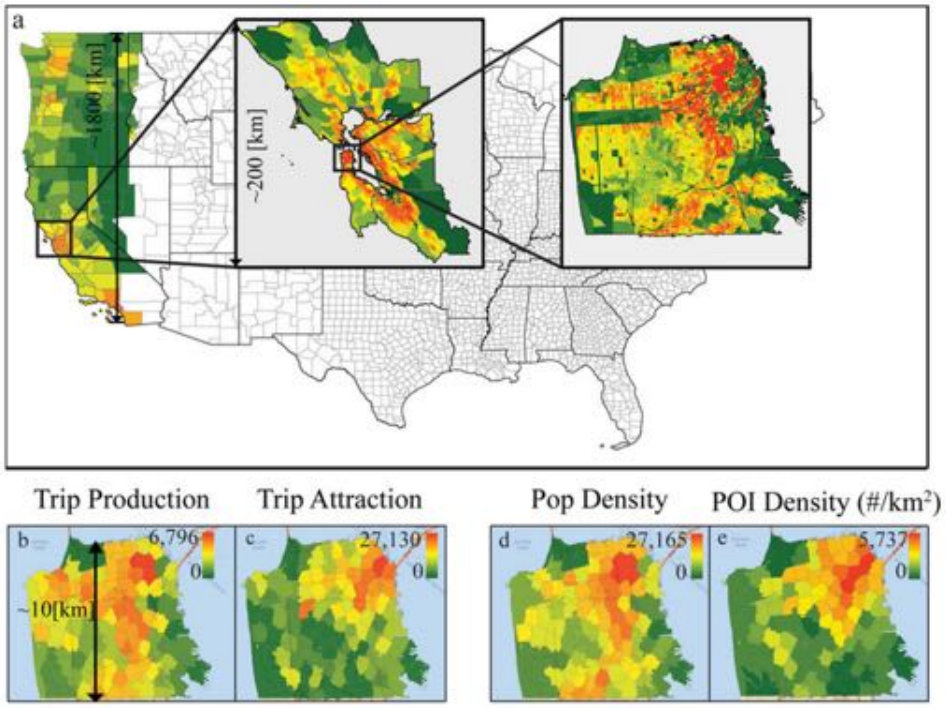


Figure 5.8: (a) A map showing the three selected regions of study representing: the Western U.S., the Bay area, and San Francisco. The color bar represents population density. (b–e) Commuting trip generation rate, trip attraction rate, the population density and the POI density in San Francisco. While the population density has high correlation with the commuting trip generation rate, the POI density has high correlation with the commuting trip attraction rate. In contrast, at the scale of the Western U.S., the population density correlates with both the trip generation rate and the attraction rate, as shown in Table 5.1

.

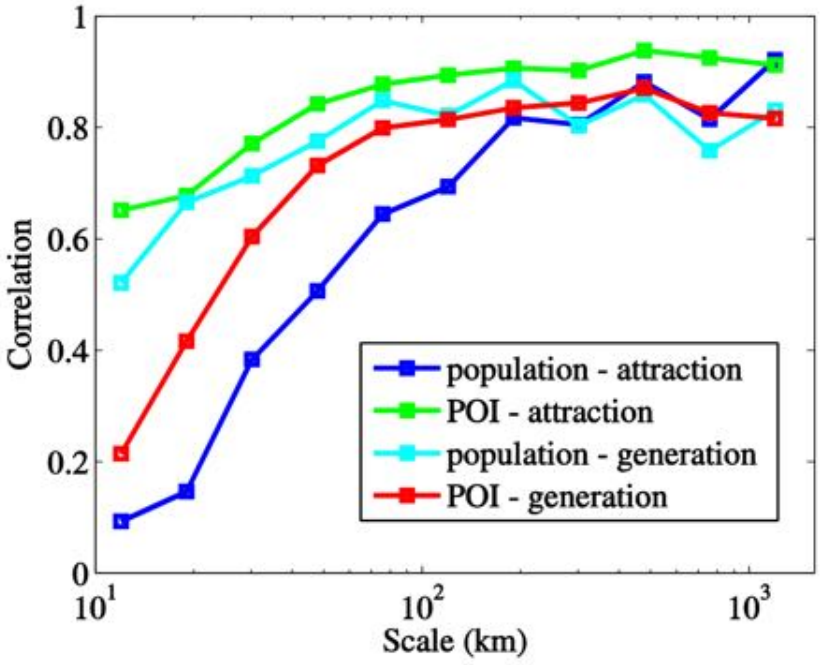


Figure 5.9: Correlation between: population – commuting generation, population – commuting attraction, POI – commuting generation, POI – commuting attraction. At small scales commuting trip attraction is better represented by POI density while at large scales these four distributions have high correlation between each other.

	Western U.S.	Bay area	San Francisco
Ext. Radiation	0.51	0.67	0.65
Gravity	0.5	0.64	0.66
Radiation	0.43	0.4	0.23

Table 5.2: CPC values for different models at different regions

and substituting from equation 5.3

$$P(1|n_i, n_j, a_{ij}) = \frac{\left[(a_{ij} + n_j)^\alpha - a_{ij}^\alpha\right](n_i^\alpha + 1)}{(a_{ij}^\alpha + 1)\left[(a_{ij} + n_j)^\alpha + 1\right]}. \tag{5.6}$$

If the empirical trip data or cell phone records (as will be shown below) are available, γ can be calculated from the total number of observed trips. If neither data source is available, the flow distribution can still be calculated because γ does not influence the relative ratio of flows between different OD pairs. The denominator in the equation is a normalization constant for finite sized area. The influence of the border effect and how this formulation can solve some of the limitations in previous models is detailed in the SI Appendix of [169].

5.3.4 Model evaluation

We evaluate the proposed extension of the radiation in the same three scenarios described in section 5.3.1: Western US, Bay Area and San Francisco. The obtained α values are 0.003, 0.05 and 1.5 respectively. We use the common part of commuters (CPC), based on the Sorensen index [146] to quantitatively measure the goodness of flow estimation.

$$CPC(T, \tilde{T}) = \frac{2NCC(T, \tilde{T})}{NC(T) + NC(\tilde{T})} \tag{5.7}$$

where T is the empirical matrix, $\tilde{T}$ the estimated one, $NC(T) = \sum_i \sum_j T_{ij}$ and $NCC(T, \tilde{T}) = \sum_i \sum_j \min(T_{ij}, \tilde{T}_{ij})$.

This index shows which part of the commuting flow is correctly estimated, 0 means no agreement found and 1 means perfect estimation. Table 5.2 shows that the extended radiation model gives estimations with similar performance to the doubly constrained gravity model at the three regions while the original radiation model's estimation power decays with the region granularity. The goodness of fit of the extended radiation model is close to other recently proposed models [86]. The difference is that the study in [86] uses actual commuting flow generation and attraction volumes as input, while in the current model we use more easily acquired population and POI density as proxies, but achieved the same level of accuracy.

5.3.5 Calibrating the model in absence of data

The $P_>(a)$ distribution plays an important role in the formulation and α parameter calibration, thus it is worth further scrutiny. When the space is infinite and the opportunities are continuous $P_>(a) = \frac{1}{1+a^\alpha}$ is a monotonically decreasing function with slope given by the α value. But if we are considering trips only within a finite sized region, this implies a finite numbers of opportunities possible, up to a_{tot}. Thus $P_>(a_{tot}) = 0$. Also, we divide a study region into a finite number of zones n_{cells}, so a can only take a set of discrete values. We define $a_{avg} = a_{tot}/n_{cells}$. a_{min} is the smallest number of opportunities in all the zones. Since within zone trips are not considered, $P_>(a)$ should start to decrease after a_{min}. This value is not known a priory but may be approximated by a_{avg} in the absence of data on trips. We correct the expression in equation 5.3 to account for these effects as

$$\langle P_>(a)\rangle = \frac{\frac{1}{1+a^\alpha} - \frac{1}{1+a_{tot}^\alpha}}{\frac{1}{1+a_{avg}^\alpha} - \frac{1}{1+a_{tot}^\alpha}}, a_{tot} \geq a \geq a_{avg} \tag{5.8}$$

Now, we explore how equation 5.8 can reproduce the $P_>(a)$ measured from data. The top panel of Figure 5.10a shows the results of $P_>(a)$ calculated from the census commuting data in San Francisco, the Bay area and the Western U.S. The solid lines show equation 5.8 with different α values, note that the two limiting values of $a_{avg} < a < a_{tot}$ determine the range of the equation. By comparing equation 5.8 with the data as seen in Figure 5.7b and Figure 5.10a, we see that in the intra-city scale the $P_>(a)$ distribution does not decay beyond a_{avg}, so we cannot use Eq 5.8 to estimate the radiation model parameter. For the Bay area and the Western U.S. scale, $a_{min} \sim a_{avg}$, equation 5.8 works well and the value of α ($0.1 \leq \alpha \leq 2$) should increase with the scale l. For a fixed scale l, if the heterogeneity of opportunity distribution increases, then a_{min} further differs from a_{avg}. In those cases $\alpha > 2$ and it is not possible to estimate α without data calibration (as shown for Las-Vegas-Seattle in Figure 5.10a).

We further explore how the parameter α systematically changes as the size l of the commuting zones changes. We evaluate the commuting within regions divided into $n_{cells} = 100$ zones of size l ranging from a few kilometers to over 100 kilometers (see Figures 5.10 and 5.11). We randomly chose 200 study regions for each scale with total population of at least $5,000 \times l$ in order to avoid unpopulated regions such as national parks. The census commuting OD data is used to calibrate the α value in each region. Figure 5.10b shows how α is close to zero for $l < 10km$ and starts to increase as a power function beyond that value. The functional relationship as a solid line is

$$\alpha = \left(\frac{l}{36[km]}\right)^{1.33} \tag{5.9}$$

The error bars show the 20th and 80th percentile of the α value at each scale. The three cases calibrated before: San Francisco, the Bay area, and the Western U.S. are marked in red squares. They are all close to the expected values calculated from equation 5.9. For trips within a city and up to metropolitan urban areas the α value is close to 0 and the error bar is narrow.

In the range $l \sim 10 \ldots 65$ km, $0.1 \leq \alpha \leq 2$ most commonly accounts for inter city trips, the model without data calibration is expected to predict trips well because of the narrow error bar. For larger scale regions enclosing trips between two or more combined statistical areas such as the ones shown in Figure 5.11, $\alpha > 2$ and has a wide range. We notice that the main driving factor influencing the α value for the same scale is the differences in the homogeneity of facility density; which are highly correlated to population density at these large scales.

In Figure 5.11, the two marked regions have the same scale: $l = 60$ km. The population distribution of the southern region has two sharp centers, Los Angeles and Las Vegas, while the rest has low population density. In the northern region, the population is more homogeneously distributed. One example OD pair is shown for each region on the right part of Figure 5.11: From Los Angeles to Lake Havasu City for the southern region and from Seattle to Wenatchee for the northern region. They have similar m_i, n_i, n_j and s_{ij} values. According to the original radiation model, they should have similar flow volumes. But in the census there are only 26 people commuting from Los Angeles to Lake Havasu City while there are 167 commuting from Seattle to Wenatchee. The reason is quite clear on the map: the distance from Seattle to Wenatchee is only 150 km while the distance from Los Angeles to Lake Havasu City is much longer because of the low population/opportunity density between the origin and the destination. To put it in another way, people have to travel further to be able to explore the same amount of opportunities. This causes the calibrated α value of the southern region to be 5, much larger than the northern one, which is 1.6. As is shown in Figure 5.11a, the more heterogeneous the distribution of population is; the larger the difference in a_{min} and a_{avg} is and the larger the α value becomes. In those cases the $P_>(a)$ from empirical trip data differs more from equation 5.8 and empirical trip data is needed for parameter calibration. More quantitative ways to measure the influence of the degree of heterogeneity as a cost function depending on the distance between the origin and the destination remains an open question for further studies.

People's location choices are not influenced by the choice of study region sizes. What the parameter α captures for the scale dependency is the granularity of aggregation. Ideally the location choice should be modeled to the smallest spatial granularity, then aggregated to the desired granularity level. But in practice such fine grained input data are usually not available, in such cases the α parameter helps the model estimation at the desired

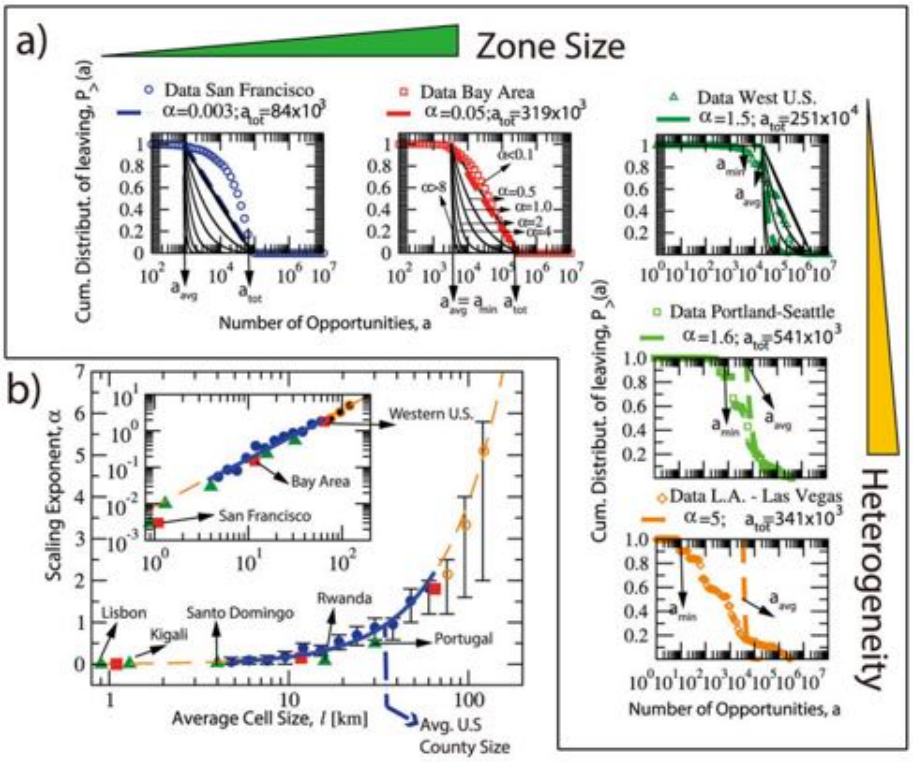


Figure 5.10: (a) Effects of the scale on the $P_>(a)$ distribution, showing as symbols the results from census data of San Francisco, the Bay area and the Western U.S. The black solid lines are evaluations of equation 5.8 for different values of α, and the dashed line is the expected α value using equation 5.9. The analytical predictions work for the Bay area and Western U.S. From top to bottom we see the $P_>(a)$ distribution for three regions with similar sizes. Given a fixed scale, the α value is influenced by the heterogeneity of the distribution of opportunities. The more heterogeneous the region is, the larger the difference between a_{avg} and a_{min}, as is shown in Las Vegas-L.A. region. In these cases the prediction of α (equation 5.8) will not resemble their calibrated values and thus calibration is needed. (b) For each zone scale l, 200 regions with random centers are selected within west U.S. In each case the corresponding α value is calibrated with census trip data. The functional relationship between α and l is $\alpha = (\frac{l}{36})^{1.33}$. The error bar shows the 20 and 80 percentile α value for each scale. The inset shows the same plot in logarithmic scale. Marked as solid blue circles is the scale range that α values can be predicted without data calibration. The calibrated results with trip data for U.S. regions are marked as red squares, while the examples from other countries are marked as green triangles. They all follow the functional approximation.

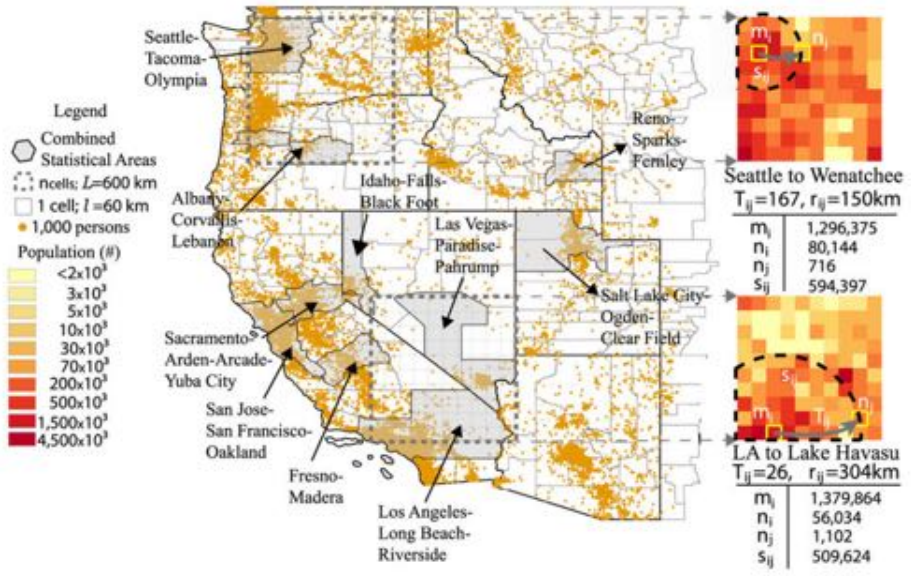


Figure 5.11: The two grey rectangles are of the same scale: 600 km. The population distribution of the southern region has only two sharp peaks: Los Angeles and Las Vegas, while the population in the northern region is more homogeneously distributed. The right section of the figure shows one example OD pair in each region with similar m_i, n_i, n_j and s_{ij} values. But the distance between Los Angeles and Lake Havasu City is much longer than the distance between Seattle and Wenetchee, which makes its commuting flow volume much smaller. This effect of distance is taken into consideration by the difference in the α value. For the southern region $\alpha = 5$ while for the northern one $\alpha = 1.6$. The grey regions are combined statistical areas which usually include one or more populated metropolitan areas.

granularity directly, without requiring finer grained data.

In summary, even without empirical trip OD data, if the commuting zones are in the range $l \sim 1 \ldots 65$ km, it can be expected for the extended radiation formula to give good commuting flow predictions. In these cases equation 5.9 gives us: $0 \leq \alpha \leq 2$. For larger zones, if the opportunity distribution do not have strong heterogeneity ($a_{min} \approx a_{avg}$), the monotonic increase of α with scale l is usually captured by equation 5.8. In other situations the model needs to be calibrated with empirical OD data.

5.3.6 Multi-regional study and the role of phone data

Not many countries in the world have detailed census data for commuting flow prediction and model calibration. Those countries with data scarcity

are often developing countries that need this kind of modeling the most. For these countries finding an alternative data source to provide guidance for their urban growth, economic planning and epidemics controlling is a pressing need. In this section we show how the extended radiation and the gravity model can be calibrated given estimated commuting trips measured from cell phone data. More importantly we can compare the phone data calibrated parameter α with the one predicted from equation 5.9 to explore the generality of that expression.

Cell phone records are increasingly showing the potential to become a data source of valuable information [99, 165, 12] since most populated areas have cell phone service coverage and the value of cell phone data in modeling human mobility has recently been highlighted in various studies [49, 145, 144, 13]. For instance, in Rwanda there is no detailed commuting census data available. Even if there were, the high migration rate of people would make the census outdated quickly. However, the country has $215,030,420$ cell phone records from one cell phone service provider in just three months. In this section cell phone records are used to extract a seed commuting OD matrix, which is expanded using iterative proportional fitting to estimate the full commuting OD matrix for the whole population under study.

We use the Bay area as an example to validate the method. Figure 5.12 shows the comparison of the results of the distribution of commuting distance, the distribution of the number of commuters between O-D pairs, and the comparison of the census commuting flow T_{ij} with the expanded cell phone user commuting flow T'_{ij}. The close fitting in all the three figures shows that we can recover the commuting patterns of the whole population from the seed matrix provided by cell phone records. For countries that do not have population density census statistics for the IPF expansion, we can use the Landscan [17] population density estimation which is available worldwide at $1km^2$ resolution.

We then extended our study to three different countries: Portugal, Dominican Republic and Rwanda. We selected the capitals in the three countries: Lisbon, Santo Domingo (including the greater metropolitan area), and Kigali; and also did the analysis for the entire Rwanda and Portugal (we do not have the cell phone information available for the entire Dominican Republic). We calibrated the gravity model and the extended radiation model to test how much can they recover these regions' commuting patterns. The results are shown in Figure 5.12. The difference in the commuting distance distribution in these regions are captured by both models. The values of l and α for the extended radiation model of these regions are marked as triangles in Figure 5.10b. All of them conform to the functional form of α observed from the U.S. regions. This shows that the relationship between α and the scale l is generalizable, in this case we could have used the extended radiation formula in these countries to estimate trips, in the absence of trip data to calibrate the model.

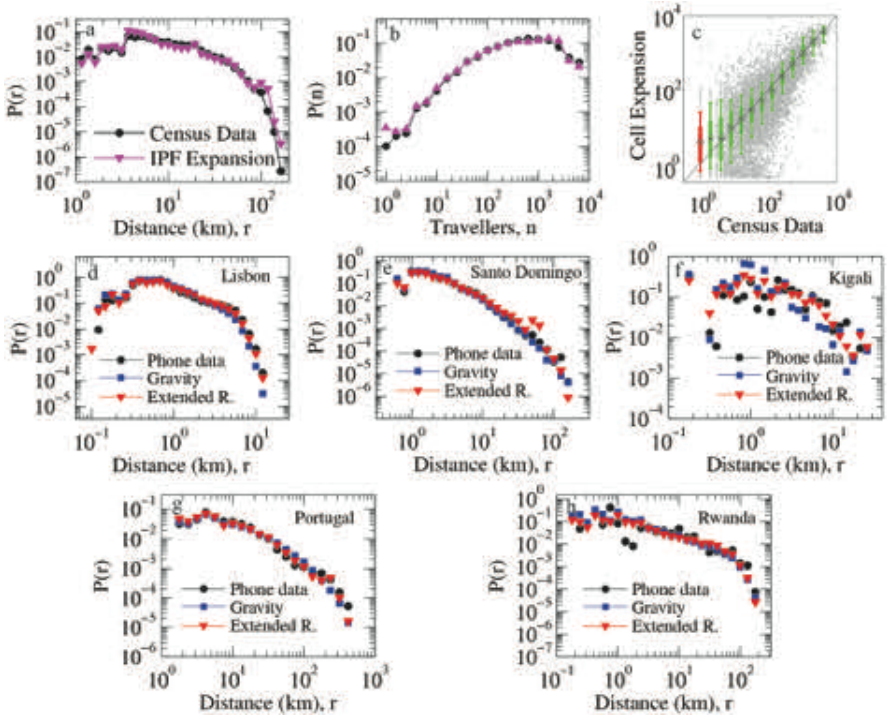


Figure 5.12: Validation of the IPF expression method and commuting patterns in different regions. (a-c) The comparison of the distribution of commuting distance, the distribution of the number of commuters between O-D pairs, and the comparison of the census commuting flow T_{ij} with the expanded cell phone users' commuting flow T'_{ij}. The close fitting in all three figures shows that we can recover the commuting patterns of the whole population from the seed matrix provided by the cell phone records. (d-h) Comparison of the distributions of commuting distance for Lisbon, Santo Domingo, Kigali, Portugal, and Rwanda. The extended radiation model can be successfully applied to all these cases and the corresponding α versus l relationship is marked in Figure 5.10.

5.4 The elliptic model for communication fluxes

While social networks have been known for years to play a key role in various human phenomena [52, 108], during decades their study was limited to certain kind of social relationships for which interaction records were available, such as authorship and cooperation in science [63, 22]. Only recently it has been possible to map large social networks representing a broader range of interactions in order to explore how their structures influence processes occurring in these networks. The required large social network data sets, usually coming from telecommunication records originated in e-mail [79], phone [120] or online communication platforms [111], have been used to explore a wide range of topics such as adoption of innovation [153], social groups discovery [168, 3, 18], epidemic spreading [124, 159, 138], social mobilization [126] or sentiment spreading [44].

Despite the publication of such studies, network data is not widely available to the community due to legal, privacy or commercial issues. In addition, even with access to the electronic records, extracting a meaningful social network may be difficult at a large scale, as we presented in the searchability chapter. For these reasons, creating models that are able to mimic different social network properties have recently attracted a fair amount of research interest [162, 14, 66, 59, 56].

While most models try to generate synthetic networks with some desired characteristics (degree distribution and clustering coefficient among others), we will focus here on reproducing a macroscopic feature of real social networks: the number of social ties between different locations, i.e. how many relationships exist between two cities, two regions or even two neighborhoods. Throughout this section, we will employ the term *location* to generically denote any of these three spatial aggregation levels. The creation of social connectivity maps between locations from widely accessible data, such as population geographic distribution (which is universally accessible for almost any region of the world through tools like Landscan [17]), will prove useful for the study of information [6] or behaviour spreading in social networks among others [41].

5.4.1 Model formulation

Similarly to what we did with the trips in the previous section, we start from the original radiation model. When applied to social relationships, the radiation model estimates the communication flux T_{ij} between two locations i and j using the population in both locations, and the population within the circle whose center is i and radius equal to the distance between i and j. Its formulation is

$$T_{i,j}^{rad} = K_i \frac{n_i n_j}{(n_i + s_{i\to j})(n_i + n_j + s_{i\to j})}$$

where n_i represents the population of location i, $s_{i\to j}$ the number of people who are not in i but closer to i than j and the normalization $K_i = n_i \frac{N_T}{N}$, where N_T is the total number of relationships to predict (which in general is considered to be available) and $N = \sum_i n_i$ the total population.

t is straightforward to verify that T^{rad} matrices are not symmetric in general. While asymmetry is a desirable feature for mobility models (commuting origin-destination matrices have strongly asymmetric suburbs-downtown flows) it is not when dealing with communication fluxes, because the number of relationships people from location i have with people from location j must be the same as the number of relationships people from j have with people from i.

A natural modification of the radiation model to deal with communication fluxes could be a simple symmetrization of the model, which we will denote *radBI* and whose formulation is

$$T_{ij}^{radBI} = \frac{1}{2}(T_{ij}^{rad} + T_{ji}^{rad}).$$

As shown below this model has a limited performance. This fact motivated us to develop the new model presented in this section. Our model, which we will refer as the elliptic model (EM), is oriented to deal with social relationships. The EM considers that the probability of someone living at location i having an acquaintance at location j is reversely proportional to the population of the area where both could meet provided their combined travel distance does not exceed certain threshold. This area forms an ellipse whose focuses are in locations i and j. Among all possible ellipses the model selects the smallest one containing the two r_{ij} radius circles whose centers are in i and j respectively (see Figure 5.13 for graphic explanation and comparison to the radiation model). Thus, the EM formulation is

$$T_{ij}^{ellip} = K \frac{n_i n_j}{e_{ij}}$$

where e_{ij} is the population within the ellipse depicted in Figure 5.13 (note that e_{ij} includes n_i and n_j) and K is a normalization parameter obtained from the total number of relationships to predict N_T ($\sum_i \sum_j T_{ij}^{ellip} = N_T$). Since $e_{ij} = e_{ji}$, $T_{ij}^{ellip} = T_{ji}^{ellip}$ and thus our model produces symmetrical matrices T.

In order to compare the quantities involved in the model, one needs to consider the sets $S_{i\to j}$ and $S_{j\to i}$ such as $|S_{i\to j}| = s_{i\to j}$ and $|S_{j\to i}| = s_{j\to i}$. Lets address the case of a very large city $C \subset S_{i\to j}$ whose population $n_C \approx e_{ij}$. While the radiation model predicts different fluxes depending on whether $C \subset (S_{i\to j} \cap S_{j\to i})$ or not (smaller when C belongs to the intersection) the EM will provide the same prediction for both cases. In fact, since $e_{ij} \geq |(S_{i\to j} \cup S_{j\to i})| + n_i + n_j$ (and usually $e_{ij} \approx |(S_{i\to j} \cup S_{j\to i})| + n_i + n_j$) the role of the union set is the main contribution of the model.

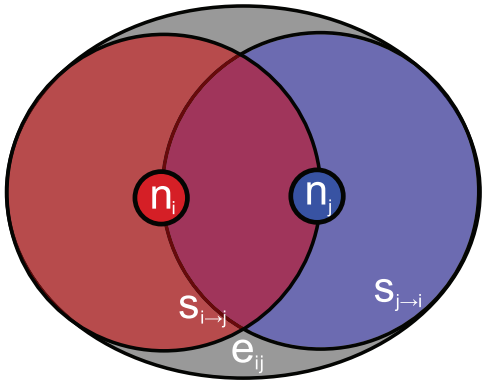


Figure 5.13: Model scenario: n_i represents the population of location i while $s_{i\to j}$ represents the population within the circle with center in i and radius up to j. As long as population is not homogeneously distributed $s_{i\to j} \neq s_{j\to i}$, the radiation model predictions will not be symmetrical. e_{ij} represents the population within the smallest ellipse whose focuses are in i and j and contains both previous circles, as well as n_i and n_j.

5.4.2 Data description

The data we have used to validate EM is the same dataset used in the searchability chapter, where an in-depth description can be found. It contains Call Detail Records (CDRs) of a six month period in 3 different countries: France, Portugal, and Spain. In total, over 7 billion calls are considered to build the social graph for each country, whose links are included only if there is at least one call in both directions during the observation period. The result is an undirected graph (this is a common technique in the literature to avoid both marketing callers and misdialed calls [120]).

In addition to the communication records, our data include a location for each user: the most used mobile phone tower in France and Portugal and the billing zip code in Spain. In order to benchmark the multi-scale performance of the EM, three aggregation levels have been used: country-wide fluxes between cities and regions and on the other hand metropolitan fluxes within cities. Table 5.3 presents the number of locations considered in each aggregation level. When applying these spatial aggregations, the center of mass of the population is used as the higher level location, instead of the centroid of the region polygon (defined by administrative boundaries), in order to avoid undesirable effects in the fairly common case of a big city located in a corner of a polygon.

Country	Towers/Zip codes	Cities	Regions
France	17475	3520	96
Portugal	2209	297	20
Spain	8928	5446	52

Table 5.3: Number of locations considered in different geographic aggregation levels for each country. At the finer level, mobile phone towers are available France and Portugal, and zip codes for Spain. Aggregation is based on administrative boundaries: cities are *cántons* in France, *concelhos* in Portugal and *municipios* in Spain while regions mean *départements* in France, and *provincias* in Portugal and Spain.

5.4.3 Communication fluxes in country scale

To validate the predictions of the EM at large scale, we consider connectivity matrices T in two aggregation levels. At the region level, matrix T has thousands of elements while at the city level there are tens of millions of fluxes to predict (see Table 5.3). Input data for the predictions only consists of the location's coordinates and populations, and the total number of relationships to predict N_T. Like the radiation model, the EM keeps the advantage of being parameter-free, so no training data is needed.

In Figure 5.14 we present a box-plot of the predictions from all the three models versus real data for fluxes between cities. The results prove consistently that the EM outperforms both the radiation model and its bilateral version. To present further evidence of the performance of the EM, we include in Table 5.4 the R^2 of the predictions in both aggregation levels. The results confirm that the EM outperforms previous models.

Overall, the accuracy of the predictions is similar to the one obtained when applying transportation models to trip prediction [142, 169]. This is an unexpected finding, since in principle, while there is an increasing cost (like time or energy) associated with distance when travelling, there is not such cost when making a phone call. While there were previous reports illustrating that social ties depend on physical distance, the capability of reproducing a significant portion of the distribution of social ties between locations just by employing a map placing them and their populations, highlights even more the importance of the geographical space when forming ties.

5.4.4 Communication fluxes within cities

While previous literature already stated that distance influences the creation of social ties between cities, our dataset allows us to study also urban relationships by using the finer spatial aggregation level available: phone towers or zip codes. Predicting all possible tower to tower relationships within the country would imply dealing with a T matrix with up to 300

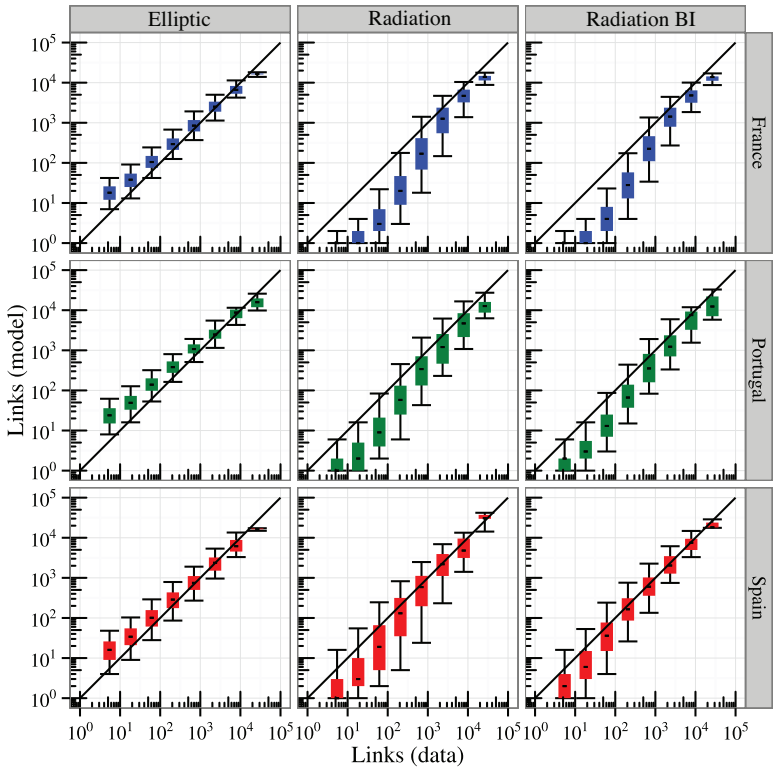


Figure 5.14: Predictions by different models versus real data. Fluxes between every city are presented, considering 297 cities in Portugal, 5446 in Spain, and 3520 in France. Error bars plot 10%, 30%, 50%, 70% and 90% quantiles. The elliptic model overcomes both radiation and bilateral radiation models in all three scenarios.

	France		Portugal		Spain	
	City	Province	City	Province	City	Province
Radiation	0.534	0.615	0.621	0.776	0.556	0.588
RadiationBI	0.626	0.723	0.730	0.847	0.676	0.668
Elliptic	0.723	0.790	0.816	0.891	0.693	0.748

Table 5.4: R^2 of the different country-wide predictions. Note that these R^2 are calculated without any logarithmic transformation on data or predictions. The number of provinces considered is 97, 20 and 52, respectively. Since the number of cities is up to two orders of magnitude larger, the flux matrix T is up to 4 orders of magnitude larger. While elliptic model is always more accurate than previous models, the improvement is specially remarkable in fluxes between cities.

	France	Portugal	Spain
Radiation	0.377	0.527	0.434
RadiationBI	0.436	0.608	0.498
Elliptic	0.653	0.658	0.501

Table 5.5: Average R^2 for urban fluxes prediction for every city in the data set where there are at least 20 different locations (towers or zip codes). The number of locations range from this 20 up to 1000 in Paris. This sums up to 40 cities in France, 29 Spain and 20 in Portugal. Although the EM again outperforms previous models, each performance is small when compared to country-wide scenarios.

million elements, with only less than 1% of them being non-zero. Thus, the prediction accuracy would be severely biased by the huge amount of zero cells. Instead, we study the short range accuracy of the model by applying it in each city where we got at least 20 different locations (the upper limit being Paris, where we have over 1000 mobile phone towers). In total, the analysis includes 40 cities in France, 29 Spain and 20 in Portugal.

Table 5.5 presents the results for the three algorithms in terms of average R^2. These results confirm that the EM still performs better, while the overall prediction accuracy is smaller compared to the country-wide experiment. The loss of accuracy within urban areas for any model purely based in distance is expected and observed in the transportation field as we saw before in this chapter. One of the main reasons for this loss of accuracy is the fact that the distance is a poorer proxy for travel time or cost in cities. People in cities tend to be within a daily radius of action and the decision of whom they communicate with depends on other metrics related to the different hierarchies that could define a social distance (e.g. ethnicity, occupation, etc.) as we presented in Section 4.4.

Correction ε as a model improvement for urban areas

When applying the EM depicted in Figure 5.13 to urban relationships one should be aware that a tower k whose distance to tower i is $r_{ik} = r_{ij} + \varepsilon$ where $\varepsilon \ll r_{ij}$ will not be taken into account for predicting T_{ij}. Since towers tend to be closer to each other in urban areas, we propose the correction in Figure 5.15 for urban environments. The variation consists of including a correction parameter ε so that the ellipse is now the smallest one containing the two circles of radius $r_{ij} + \varepsilon$ centered in i and j. After studying several values of ε, we found that the prediction accuracy peaks near $\varepsilon = 1$km for nearly all the cities (as shown in Figure 5.15).

There may be several interpretations for such a maximum: one could argue that it comes from the location error, known to be close to the average distance to neighbors from the Voronoi tessellation [156], which is around 1

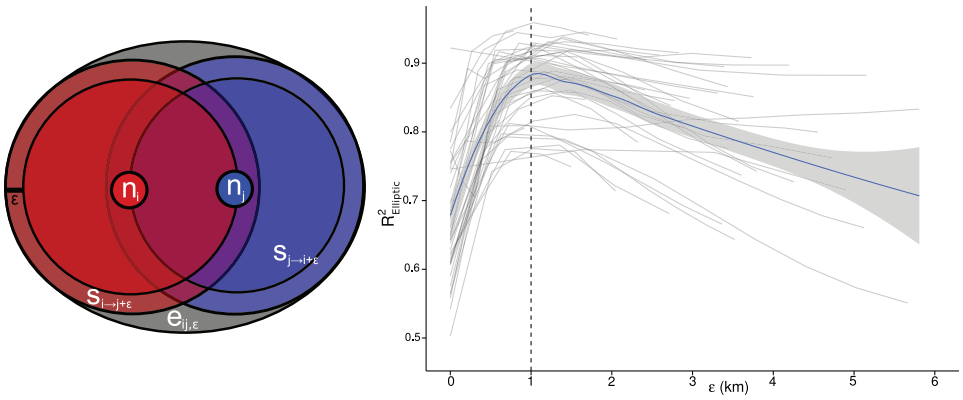


Figure 5.15: Model modified for intracity predictions, adding the correction term ε. Each grey line represents a certain city in the dataset with the blue line and the shadow representing the general trend. We find predictions improve when some correction term is included, reaching a maximum around $\varepsilon = 1$km.

	France	Portugal	Spain
Elliptic $\varepsilon = 0$	0.670	0.645	0.494
Elliptic $\varepsilon = 1\,$km	0.846	0.740	0.688

Table 5.6: Average R^2 of the predictions for the corrected model with $\varepsilon = 1$km, compared to the original (non-corrected) model.

km in average for our dataset. This agrees with the fact that the optimal ε is a fixed value and does not depend on the distance r between i and j. On the other hand, when applied back to country-wide scenarios we found the correction term does not improve the predictions and no peak emerges near $\varepsilon \backsimeq r_{Voronoi}$ or elsewhere, reinforcing the hypothesis that within cities we are reaching the boundaries of the resolution of our location data.

Another way to evaluate the performance of the different models is to compare them against empirical data in terms of the link-distance distribution $P(r)$ which represents the probability of observing a relationship between two people living r kilometres from each other. Figure 5.16 shows the improvement $P(r)$ when applying the ε =1km correction. Without the correction term, short-range relationships are over represented, while the EM with the correction fits almost perfectly with the distribution obtained from the data. Note that although radiation model predictions also improve, it still predicts shorter relationships than those observed in the data.

Table 5.6 shows results of the corrected model for urban environments in terms of average R^2, which confirm a significant performance increase when applying the corrected model with $\varepsilon = 1$ km across all cities in the data set.

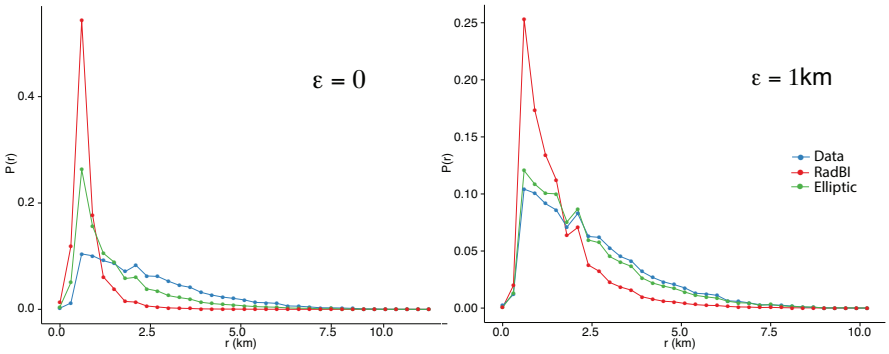


Figure 5.16: Left: fraction of relationships $P(r)$ within distance r in the real dataset compared to predictions by both elliptic and bilateral radiation models where $\varepsilon = 0$ for Porto (Portugal). Right: elliptic prediction gets very close to data when using $\varepsilon = 1$km. Although the radiation model predictions also improve, they still predict shorter relationships than those observed in reality.

Chapter 6

Coupling social ties and mobility patterns in urban enviroments

In previous chapters, the correlation between social networks and the location of social actors has been established. We are much more likely to become friends with people living 10 km from where we live than with people living 1000 km apart from us. Communities tend to be geographically clustered. These patterns are strong enough to enable us with predicting capabilities, so we can estimate communication fluxes from the geographical distribution of population. Overall, the primary interest of these previous studies was measuring and reproducing patterns of the impact of geographical distance on network topologies.

A consistent result during such research is that cities present an anomalous behaviour that requires us to modify the models that are valid for larger scale scenarios. Additionally, as explained before, mobile phone traces and other sources provide us with the ability to collect millions of spatiotemporal data points that allow us to reconstruct individual trajectories with a very high level of detail. However, in all previous analysis, we have used such data just to pin users to their most common location on a map. In this chapter, we will consider not only the most common location, but all of the locations the user has visited within the city.

Considering the whole set of locations visited by a user in a certain city and also taking into account the social network around him opens new possibilities. It is safe to assume that people who live in Madrid build social ties more often with people from a Madrid suburb than with people from a Barcelona suburb, i.e. location causes social ties. However, the list of places in Madrid I visited last week was severely influenced by my friends and acquaintances, so my social network is influencing my trajectories. With an estimated 15–30% of all trips taken for social purposes, it is not surprising

that the movement of our friends has been used to improve predictions of where we will be next [51, 25, 29].

Additionally, in cities distance is less restrictive. Residents have access to a variety of transportation options and are free to choose locations that provide the best goods and services rather than the closest location. The districts and neighbourhoods are often self-organized and make it more natural to describe mobility as movement between sets of locations, or habitats [11]. Which habitats users share with their contacts and when they share them may indicate the nature of the social relationship, e.g. a co-worker or a friend [34]. Two individuals co-located between 09.00 and 17.00 on weekdays are likely to have a different relationship than two who are found in the same area at midnight on a Saturday. In these scenarios, mobility is defined and measured based on discrete visits to places within a city that are different times; and previous work has shown that users who visit similar places are more likely to be friends in online location-based social networks [25].

Here, we describe a set of metrics to evaluate patterns of mobility and social behaviour that occur within the context of cities. Using call detail records (CDRs) produced by millions of mobile phone users, we find that individuals have far more similar visitation patterns to those of social contacts than to those of strangers and that the movement of these contacts can be used to reconstruct a considerable portion of the individuals' movements. We also find strong correlations between tie strength and mobility similarity and show that mobility similarity can be used to classify social relationships and recover semantic information about the nature of a link in the social network. Finally, we propose an extension to the mobility model described in [144] that incorporates movement based on the visitation patterns of social contacts so that it can reproduce empirical relationships found in the data. We call this model the GeoSim model and compare it against empirical data and two other mobility models. The generality of these results is demonstrated by their reproducibility in three different cities in two different countries.

6.1 Mobility models

Traditionally, human mobility has been modelled with Levy flights (LF) [46] or random walks (RW) [113], both being very simple stochastic Markov processes. These models, while simple, were good enough to reproduce the fat tailed distributions observed from empirical data both in jump size and waiting time distributions.

6.1.1 Individual mobility model

The appearance of phone-generated mobility data [49] allowed researchers to extract conclusions about human mobility way beyond the aforementioned distributions. In particular, Song and his collaborators [144] identified the following anomalies in empirical data that cannot be reproduced using traditional models:

- The number of distinct locations visited by a user increases more slowly than predicted by random walks or Levy flights.
- The frequency f of a user to visit a given location is expected to be asymptotically uniform everywhere for both LF and RW. In contrast,the visitation patterns of humans is rather uneven, so that the frequency f of the kth most visited location follows Zipf's law

  $$f(k) \sim k^{-\delta}$$

  where $\delta \approx 1.2$.
- Both RW and LF predict that the longer we follow a human trajectory, the further it will drift from its initial position. Yet, humans have a tendency to return home on a daily basis, suggesting that simple diffusive processes, that are not recurrent in two dimensions, do not offer a suitable description of human mobility.

In order to overcome these anomalies, while at the same time reproducing a fat-tailed distribution in jump size and waiting time, they proposed the following model, which we will refer to as the Individual Mobility model (IM). The IM model considers that for every jump one of the following occurs:

- Exploration: with probability ρS^{γ} the individual moves to a new location, different from the S locations visited before. The jump size is pulled from the distribution and the direction is selected at random.As the individual moves to this new position, the number of previously visited locations increases from S to $S + 1$.
- Preferential return: with the complementary probability $1 - \rho S^{\gamma}$, the individual returns to a previously visited place. In this case, the probability to visit a certain location is chosen to be proportional to the number of visits the user previously had to that location.

6.1.2 Travel-Friendship model

In [51], the authors analyzed records from location based social media services and proposed a model to create at the same time social ties and mobility patterns. We will refer to this model as the Travel-Friendship model

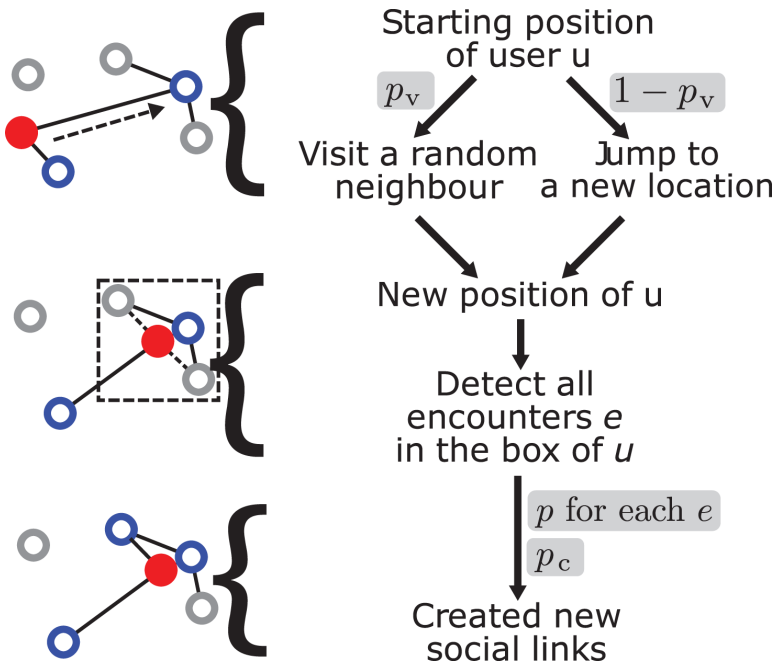


Figure 6.1: Model from [51]. The central node is the filled red circle and its neighbors are marked in blue. Directionality of links is neglected in this schematic to maintain simplicity.

(TF). TF considers that at each step, a randomly chosen agent performs actions in two stages (see Figure 6.1):

1. Travel
   (a) Visit a randomly selected friend at his current location with probability p_v.
   (b) Otherwise, travel to a new location. The distance of travel is obtained from a distribution of jump lengths, while the direction is chosen proportionally to the population density at the target distance.
2. Friendship
   (a) With probability p, create directed links to agents within a neighborhood of size $\delta \times \delta$.
   (b) With probability p_c, create a directed connection to a randomly chosen agent anywhere in the system.

An schematic of this model is presented in Figure 6.1.

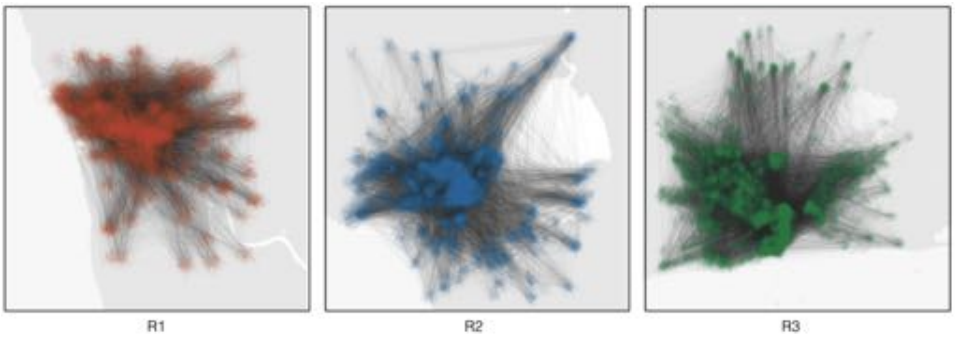


Figure 6.2: A small sample of calls between residents is shown for each of three cities. CDRs provide the location of each caller as well as the record of communication between them. A dot is drawn at the approximate location of a user calling or receiving a call and a link appears between two users calling each other. Our aim is to identify useful and reproducible patterns from this coupled tangle of social and spatial behavior.

6.2 Analysis design

In order to analyze the relationship between mobility and social networks, a number of design decisions have been made regarding data processing and the choice of metrics. Because traditionally researchers have used different approaches when analyzing social networks or mobility patterns from phone data, in here we specify clearly the kind of process and metrics of choice. The design principle has been to find a balance between mobility metrics and social network, thus our results can be compared to previous work in both fields.

6.2.1 Data description

Call detail records (CDRs) are generated when a mobile phone user performs an action that requires the provider's network, for example placing a call or sending a text message. These records generally contain the ID of the tower the phone connected through, which gives a rough estimate of the user's location. When the individual receiving a call or message is a customer of the same provider, the unique identifier of the receiver and their location may also be stored. CDRs allow us to observe mobility patterns of individuals and construct social networks containing millions of people. Figure 6.2 shows a small sample of calls between city residents during a single hour and illustrates dynamics of the urban system we wish to understand.

Our data consist of anonymized CDRs collected from three cities (R1, R2, and R3) in two different industrialized countries. Two cities (R1 and R2) were obtained from the same provider in country 1, while another provider

Table 6.1: Basic statistics on the networks and spatial extent of each region considered.

City	Nodes	Edges	$\langle k \rangle$	Towers (L)
R1	133,587	997,287	14.9	249
R2	183,486	2,487,661	27.1	447
R3	635,731	4,197,093	13.2	935

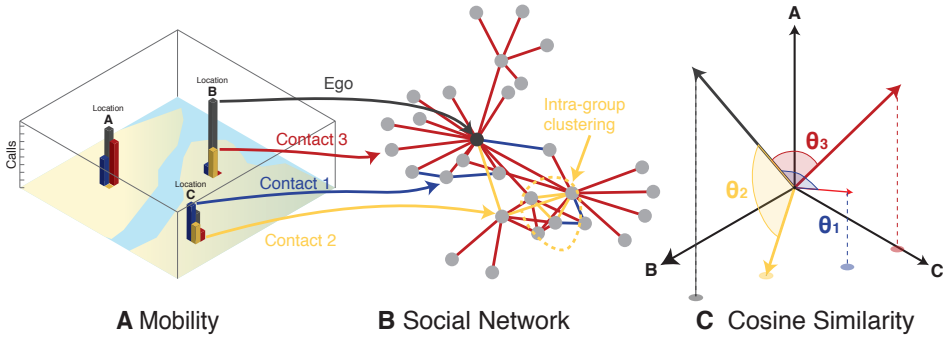


Figure 6.3: Similarity of visitation patterns between nodes in social networks. For each user, we keep track of (A) how many visits are made to locations across the city and (B) construct a social network by tracking calls to others. We can then define (C) the geographic cosine similarity between two users by computing the cosine of the angle between any two vectors in the location space.

was used for the third city (R3). The observation period covers 15 months in R1 and R2 and 5 months in R3 and contains over 1 billion events in total. Each record provides the time of the communication event, an anonymous unique ID for the caller and callee, and the ID of the tower used by at least the caller (in the case of R3) and in some cases the callee (R1 and R2).

To build the social networks for each city, we employ the following procedure. First,we consider only users that appear in over 200 communication events within each city's metro region over the course of the entire data collection period. Second, we only draw an edge between two users if they make more than two calls between them during that time. Properties of the three networks as well as the number locations (cell towers) within each metro region can be found in Table 6.1.

6.2.2 Social and mobility metrics

In each city, we construct a social network containing all users (nodes) with sufficient call volume and connect users (edges) if they have regular contact between each other. Each node is assigned a $48 \times L$ location matrix $\mathbf{L}$, where L is the number of cell towers in the city. Each row of this matrix

corresponds to an hour of a typical weekday and an hour of a typical weekend day (giving 48 hours in total) and each element $L_{t,j}$ contains the number of times that a user made a call from location j during hour t across the entire observation period (Figure 6.3A). We refer to individual rows of this matrix, $\mathbf{v(t)}$, as *location vectors*. The location matrix and location vectors can be used to compute various mobility properties of nodes (mobile phone users). Summing all elements of the location matrix gives the number of calls made and received by a user $N = \sum_{t,j} L_{t,j}$ while summing each column and dividing by N provides the frequency of visits a user made to every location in the city, $f_j = \frac{1}{N} \sum_t L_{t,j}$. Summing visits to each location at all times gives a single location vector $\mathbf{v}$ for each user and represents the total visits made to each location over the period of data collection. Similarly, the number of non-zero elements in the location vector represents the number of unique locations visited $S = \sum_j sign(v_j)$. All of these features are measures of a user's mobility behavior within the city.

We can also compare the location matrices and vectors of two mobile phone users and measure similarities between the two. While a number of metrics could be used to measure mobility similarity between nodes (Figure 6.3B), here we focus on the cosine similarity between the location vectors of two nodes i and j defined as: $\cos\theta_{i,j} = \frac{\mathbf{v_i} \cdot \mathbf{v_j}}{|\mathbf{v_i}||\mathbf{v_j}|}$. The cosine similarity measures the cosine of the angle between two vectors in our L-dimensional *location space* (Figure 6.3C). It has been shown to correlate strongly with the probability of being friends in an online social network [25] and has a number of desirable properties. It is sensitive to visit frequencies rather than set intersections alone, so two users who share frequently visited locations appear more similar than those who share less important destinations. Unlike the Pearson correlation coefficient, it does not overstate similarity when vectors contain many zero elements (as it is often the case) and finally, the cosine similarity is a measure of the angle only and is not affected by differences in the total number of calls made by two users. For the remainder of this chapter, we refer to the cosine similarity between two locations vectors as *mobility similarity*.

The mobility similarity between two users can be computed from their entire movement history or from visits during a small portion of a weekday or weekend. In the former case, we assign a single mobility similarities value to an edge in the network, while in the latter, we assign a time series of cosine similarity $\cos\theta(t) = \frac{\mathbf{v_i(t)} \cdot \mathbf{v_j(t)}}{|\mathbf{v_i(t)}||\mathbf{v_j(t)}|}$. This time series reveals how often two users visit the same places at a given time of the day and will later function as an attribute to differentiate between types of social contacts.

Within this mathematical framework, we can calculate an upper bound on how much of an individual's location vector can be reconstructed from a linear combination of the location vectors of other users. For example, a co-worker may share office space with an individual, but not live in the

same neighborhood, while the opposite may be true for a member of that individual's family. By combining the visitation patterns of the co-worker and family members, however, a complete picture of an individual's visitation patterns can be obtained. Mathematically, we define a set of users F for each individual i in the network. For example, we may choose F to be neighbors in i's ego network or a random set of nodes. The location vectors $\mathbf{v_j}$ where $j \in F$ are used as columns of an $|F| \times L$ matrix we denote as $\mathbf{A}$ and span a subspace of the L-dimensional location space. We then use QR-decomposition to find an orthonormal basis $B = q_1, \ldots, q_{|F|}$ for $\mathbf{A}$. Our target user's location vector is then projected into this vector subspace: $\hat{\mathbf{v}} = \sum_{i=1}^{|F|} \langle q_i, \mathbf{v} \rangle q_i$. This projection represents the best approximation of a user's visits based on the visits of users in F. We can quantify how it compares to a user's true visitation patterns by taking the ratio of the size of $\hat{\mathbf{v}}$ with the size of the actual location vector $\mathbf{v}$. We refer to this ratio as *predictability* and define it mathematically as $\frac{|\hat{\mathbf{v}}|}{|\mathbf{v}|}$. When predictability is 1, the visitation frequencies of a user can be completely obtained from location vectors of users in F and when it is 0, nothing about its visits can be learned. We note that for values between 0 and 1, predictability cannot be interpreted as the fraction of a user's visits that can be recovered as the vector norms are computed using the standard L2 norm. In principle, however, these two quantities should be strongly correlated because the elements of location vectors can never be negative.

6.2.3 Controlling non-uniform sampling rates

While mobile phones make excellent passive sensors of social behavior and mobility, they suffer from non-uniform sampling rates. Information is only recorded when a user makes or receives a call, leaving more observations at certain times of the day or week than others. Moreover, different users may use their devices more or less depending on habits or socio-economic variables. Because of this, we are careful to ensure that any metrics we measure in the data are not biased by different sampling rates. Figure 6.4 shows the distributions of four metrics in each region for groups of users with similar numbers of calls over the observation period. In general, we find that the calling frequency of users does not affect these distributions with the exception of the number of unique locations visited S, which increases with the calling frequency. However, even in these cases the shape and trend of the distribution remains the same for each group with only the means shifted. Finally, we note that for region R3, the number of unique locations visited takes on a slightly different shape than regions R1 and R2 due to the fact that we only obtain location information for callers in this city and not for receivers as well. Our new metrics of mobility, cosine similarity and predictability, are less affected by different sampling rates. We perform the same analysis for correlations between social behavior and mobility. For

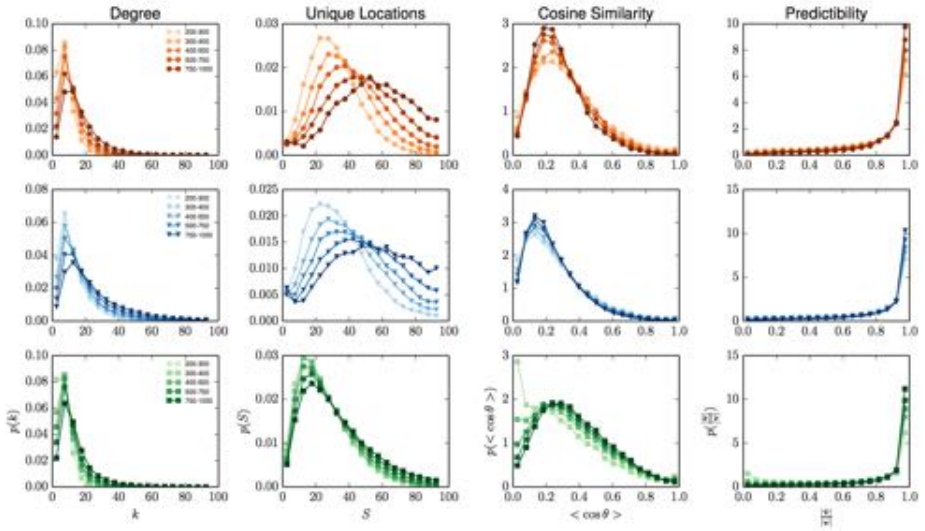


Figure 6.4: Distributions of different variables (columns) for each of the three regions (rows) for groups of users with different numbers of total calls. Note than cosine similarity is studied for all pairs in the data, while all the other metrics are computed for users. In general calling frequency does not affect these distributions with the exception of the number of unique locations visited where the mean is shifted right for users with more calls.

users with a given number of calls, we correlate their social metrics such as degree or the entropy with which they distribute calls to contacts with mobility metrics. We find that, similar to the distributions, most of these correlations do not depend on the number of calls made by a user. In cases where there is dependence, the trends hold within groups of users that make the same number of calls (Figure 6.5).

Similarly, we measure the entropy of the distribution of calls that each user makes to his or her contacts. Users with higher entropy spread their calls evenly amongst social ties, while lower entropy means most calls go to fewer. The degree of a node sets an upper bound on the entropy a user can have. Thus, any correlations we measure may be biased by differences in the degrees of each user. To control for this, we plot correlations of call entropy to other metrics for groups of users with the same degree. Figure 6.6 shows that these trends are unaffected by differences in degree.

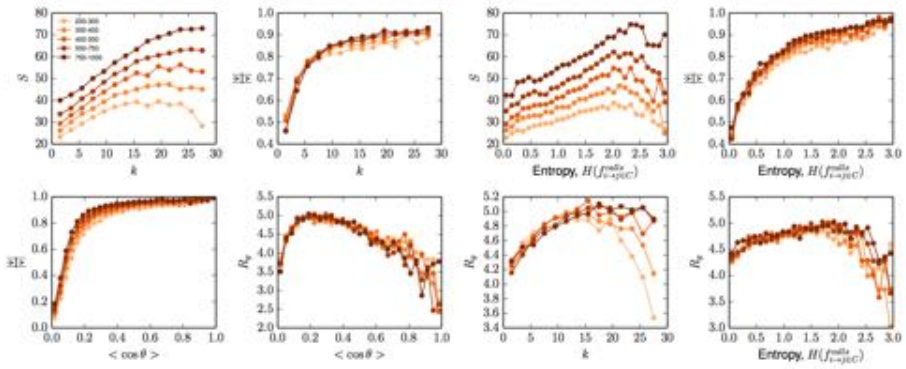


Figure 6.5: Correlation between social behavior and mobility while controlling for the number of calls made by each user. Again, we bin users based on the number of records they have in our data set and then measure correlations between social and mobility metrics. We find, as it was the case with distributions, that these correlations are unaffected by sampling frequency.

6.3 Results

6.3.1 Correlation between social networks and mobility patterns

Though similarity can be measured between any two arbitrary nodes and predictability can be computed using an arbitrary set of nodes F, we hypothesize that an individual will likely be more similar to and predictable by social contacts. To test this, we compare the mobility similarity between users that call each other regularly with the similarity between random users; and also the predictability achieved using a node's social ties with the predictability using random sets of nodes (essentially rewiring the social network, but leaving mobility intact). Figures 6.7A and 6.7B show the distribution of similarity and predictability values for the networks in each city. We find significantly more similarity and predictability in empirical networks when compared to random re-wirings. The similarity distribution is bimodal, with peaks at very low similarity near 0 and very high similarity near 1. We measure very high values of predictability when using an individual's social contacts as opposed to a random set of people in the same city. As other studies have suggested, we find that visitation patterns are strongly linked to our social relationships; our movements are far more similar to the movements of our social contacts than to those of random users.

Interestingly, we observe higher levels of mobility similarity between users separated by short network distances. We find that two connected nodes are on average 10 times more geographically similar that two randomly selected nodes. Nodes separated by two hopes, or "friends of friends",

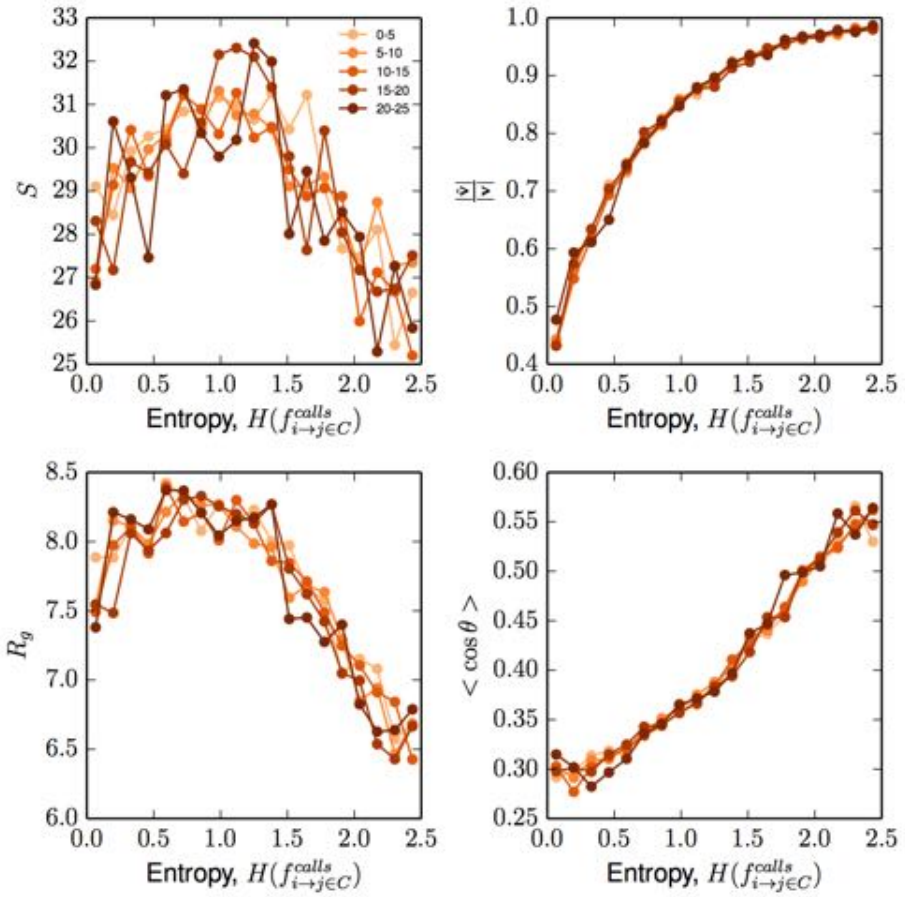


Figure 6.6: Correlation between the entropy of a node's call frequency distribution to contacts and mobility variables may be affected by the degree of each node. R_g denotes *radius of gyration* as defined in [49]. Measures such as entropy and predictability will naturally be affected by the number of contacts each user has. For example if a user has many social contacts, the maximum entropy of the distribution of call frequencies to those individuals will naturally be higher than a user who has few friends. To ensure our correlations are not artifacts of the number of contacts each user has, we plot these correlations for groups of users with the same degree and show that these relations still hold

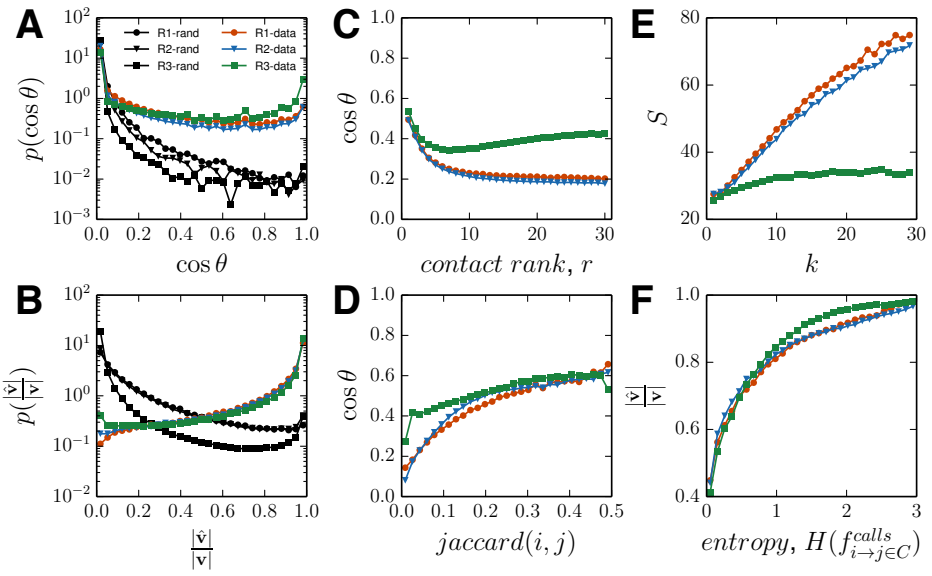


Figure 6.7: Correlations between mobility and social behavior. For each city, we compute the (A) distribution of cosine similarity and (B) predictability using observed edges (colored lines) and compare to distributions made using randomized edges. We find both mobility similarity and predictability are much higher when using actual social contacts compared to random users (C) Ranking each user's contacts by number of calls, we find that stronger ties are more geographically similar. (D) The more common contacts shared by two users, the more geographically similar those individuals tend to be. (E) Users with more unique contacts tend to visit more unique locations. (F) Users who distribute their calls to contacts more evenly (higher entropy) are more predictable than users with more uneven call distributions.

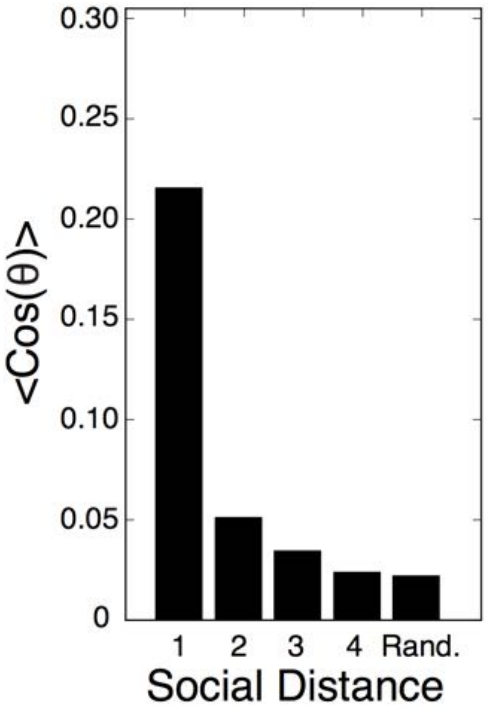


Figure 6.8: Social distance and geographic similarity. Nodes who contact each other are far more similar to each other than two randomly selected nodes. Here we compute the average mobility similarity between nodes separated by a certain number of hops. Even for nodes separated more two or three hopes, we elevated levels of similarity when compared to two randomly selected nodes in the network.

are nearly twice as similar as randomly selected nodes and this elevated similarity is observed up to three hops from an individual (see Figure 6.8 for details). This result is the location equivalent to the Fowler's result on happiness and obesity in social networks [44].

Next, we explore the relationship between tie strength and mobility similarity. We rank all contacts in each user's ego network by the number of calls shared between them (1 being contact that shares the most calls) and compute the average mobility similarity for all edges with a given rank (Figure 6.7C). Stronger contacts have higher mobility similarity on average than weaker ties, though this effect subsides for contacts below rank 10. We note that region R3 shows a slightly different trend. This is likely due to the shorter observation period in this region resulting in few individuals with more than 10 regular contacts, biasing the tail of this distribution. We also observe a positive correlation between social similarity as measured by

the Jaccard index between the neighbors of two nodes and mobility similarity (Figure 6.7D); individuals who share more social contacts share more locations.

We also find other aspects of social behavior to be correlated with mobility. Individuals with more friends tend to visit more locations, but despite this exploratory behavior, these individuals are still more predictable due to the increased information provided by their additional contacts to reconstruct their movements (Figure 6.7E). Again R3 appears as an outlier due to the shorter observation period and the absence of mobility information on the user receiving a call. We then measure the entropy of the sequence of calls made by a certain user and find that individuals who distribute their calls more evenly also visit more unique places and are more predictable (Figure 6.7F). The visitation patterns of those who spread social attention more evenly can be more easily reproduced. Finally, to ensure that these results are not an artifact of sampling frequencies, we compute these distributions and correlations controlled by the number of CDR events and by the degree of a user, finding no change in the relationships (see Figures 6.4, 6.5 and 6.6,).

6.3.2 Classifying links according to co-location events

Having demonstrated that social behavior and location choices are strongly correlated, we next use temporal variations in mobility similarity to provide context into the type of social relationship between two individuals in our networks. We measure mobility similarity $\cos\theta(t)$ over the course of a typical weekday and weekend under the hypothesis that different types of social contacts will have different levels of similarity at different times. To identify groups, we use a simple k-means unsupervised clustering algorithm on these similarity time series.

The k-means clustering algorithm must be seeded with the number of clusters to be found a priori. In order to identify a reasonable number of clusters, we run the algorithm for multiple values of k and examine the resulting clusters as well as the silhouette coefficient for each choice. The silhouette coefficient[1] decreases as the number of clusters increases indicating that there is little added benefit from additional splitting (Figure 6.9). Moreover, when examining the centroids of clustering results, the three main clusters identified break into similar groups that show small differences such as on weekends or in absolute similarity level (Figure 6.10). To ensure our results are not an artifact of the clustering algorithm chosen, we also perform clustering using a hierarchical, agglomerative clustering technique

[1]The Silhouette Coefficient is calculated using the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample. The Silhouette Coefficient for a sample is $\frac{b-a}{\max(b,a)}$. To clarify, b is the distance between a sample and the nearest cluster that the sample is not a part of.

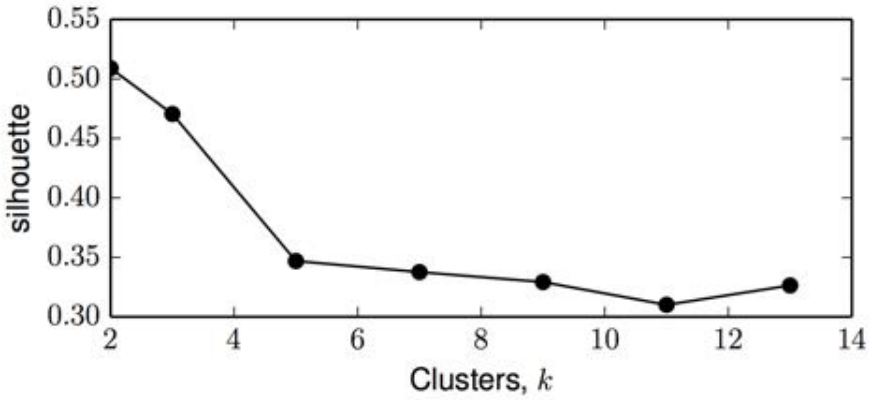


Figure 6.9: The silhouette coefficient for different numbers of clusters. The silhouette coefficient is a measure of the ratio between intra and inter-cluster variance that gives a rough measure of the quality of clustering results. It steadily from the chosen number of clusters, 3, indicating that little is gained by additional splitting

using Ward linkage. In each region, we obtain clusters that visually match those found with k-means very closely (Figure 6.11).

Overall, we find three persistent groups. While we have no ground truth data about the nature of these relationships, for clarity, we label each group according to it's qualitative signature:

- *acquaintances* with uniformly low levels of similarity,
- *co-workers* with high similarity during work hours on weekdays and low similarity on nights and weekends,
- *family/friends* with high similarity on nights and weekends.

Figure 6.12A shows the cluster centers for each group. While other interesting clusters are found for $k > 3$, they appear as subgroups of the three general archetypes we discuss here. These three groups appear in each city despite the unsupervised nature of the algorithm; cluster centers start at random locations, yet find remarkably similar final positions in each city.

Assigning each edge to a cluster based on the time series of mobility similarity effectively paints all edges in the net in a specific color as illustrated above in Figure 6.3B. Previous work has found that edges in real social networks are much more likely to be arranged in triangles, resulting in high clustering coefficients. In this case, we expect that some social groups, such as co-workers or close friends, should exhibit high degrees of intra-group clustering, while others such as acquaintances do not. For example, many of

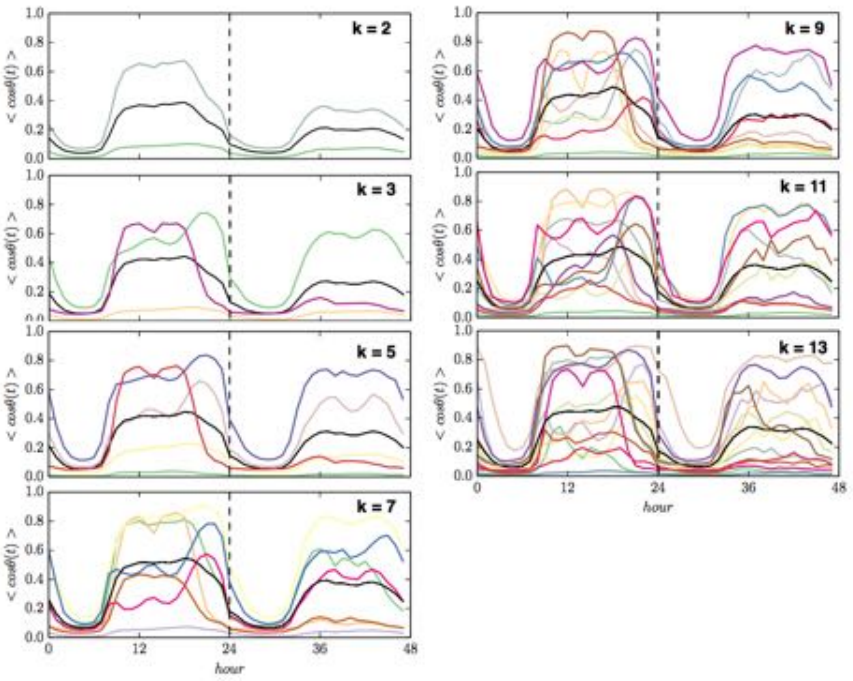


Figure 6.10: k-Means clustering results for various values of k in city R1. We perform k-means clustering for multiple values of k as a manual check that our choice of 3 clusters is appropriate. In general, additional clusters appear to be variations of three main themes used in the main text.

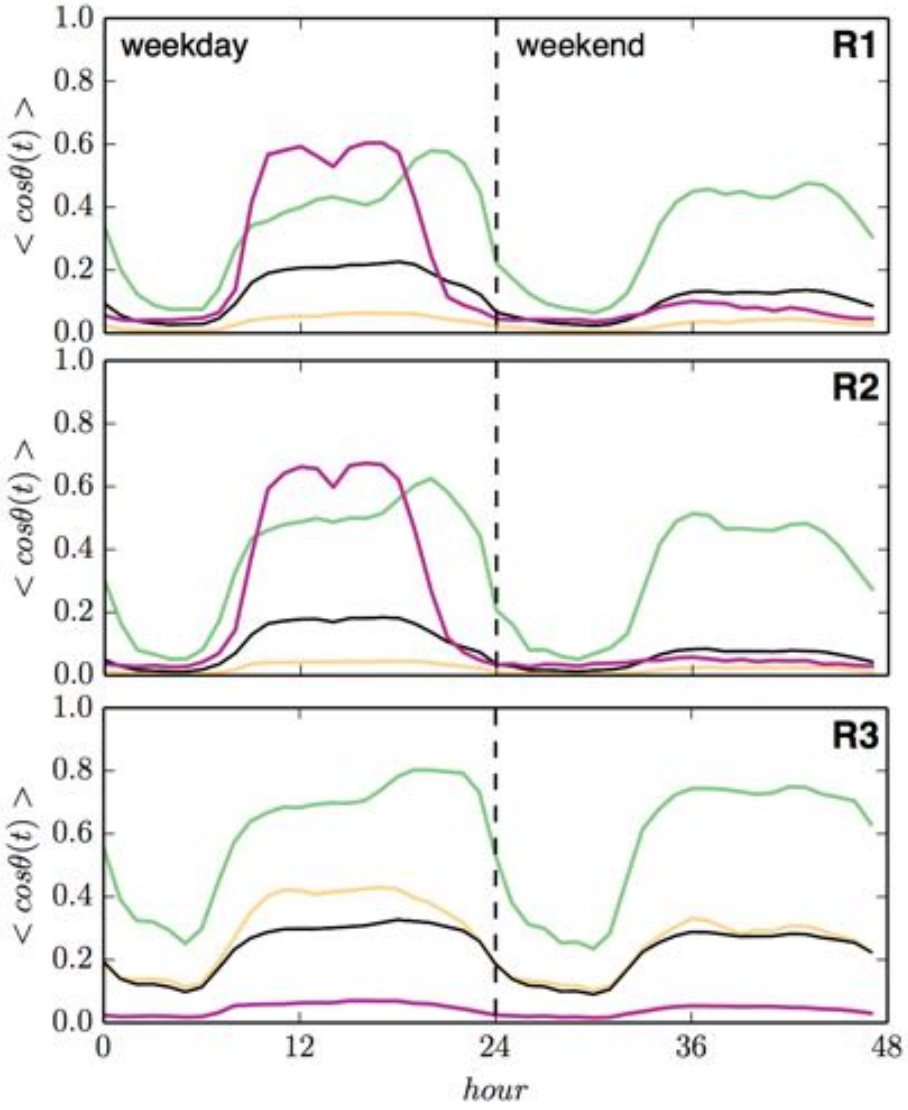


Figure 6.11: Results from a hierarchical, agglomerative clustering algorithm with Ward linkage. This clustering method clusters nodes based on connecting data points together if they are within some distance of one another and then examining connected components. The clusters in each region visually match results from k-means, suggesting that our results are robust to the exact clustering algorithm used.

an individual's *co-workers* visit similar places during work hours and tend to call each other because they are part of the same office community. We find evidence of this when measuring the clustering coefficient within subgraphs containing only edges belonging to a single mobility similarity cluster (Figure 6.12B). Interestingly, the clustering coefficient (C_g) of *acquaintances* is much lower than the *co-workers* and *family* ties despite consisting of nearly 70% of the links in the network. This provides additional evidence that we are capturing very different types of relationships with our classifications based on mobility similarity. Moreover, these results highlight mobility similarity as a property to label functional communities within social networks as well as individual edges.

Next, we consider how the composition of an individual's ego-network correlates with his mobility. Is a person with a stable job and family likely to be less exploratory and more predictable than a young college student with many acquaintances? To answer this, we bin nodes into groups based on two mobility metrics, the number of unique locations visited S and how predictable that user is $\frac{|\hat{\mathbf{v}}|}{|\mathbf{v}|}$. We then compute the fraction of edges that belong to each classification for all nodes in each mobility bin. Figure 6.12C shows that users who tend to visit more unique locations tend to have a higher fractions of *acquaintances* in their ego network, while Figure 6.12D suggests that less predictable individuals tend to have fewer contacts in this category. Conversely, less spatially explorative individuals and individuals that are easier to predict tend to have higher fraction of *co-workers* and *family/friends* labels in their ego network. These results again show the ability of mobility similarity to add contextual attributes to a network and reveal novel relationships between the structure of a user's ego network and their mobility behavior. In future works, it may be interesting to explore correlations between the mix of one's ego network and social behaviors such as their propensity to form new contacts [109].

6.3.3 Coupling social ties and mobility

Given the clear empirical relationship between social contacts and mobility, our remaining task is to identify a coupled model that captures these dynamics. While a number of models consider mobility alone [144, 142, 49], only a few have attempted to link the two [51, 25]. Those that have combined social and mobility behaviors have consistently found nearly 15-30% of trips are made for social purposes. These coupled models have had considerable success reproducing patterns of geographic distance within social network structure, but, as we show, do not always capture properties of geographic similarity.

In light of the time scales we are studying, we make the assumption that our social network is static and extend the mobility model introduced by Song et al. [144] to include movement choices based on social contacts.

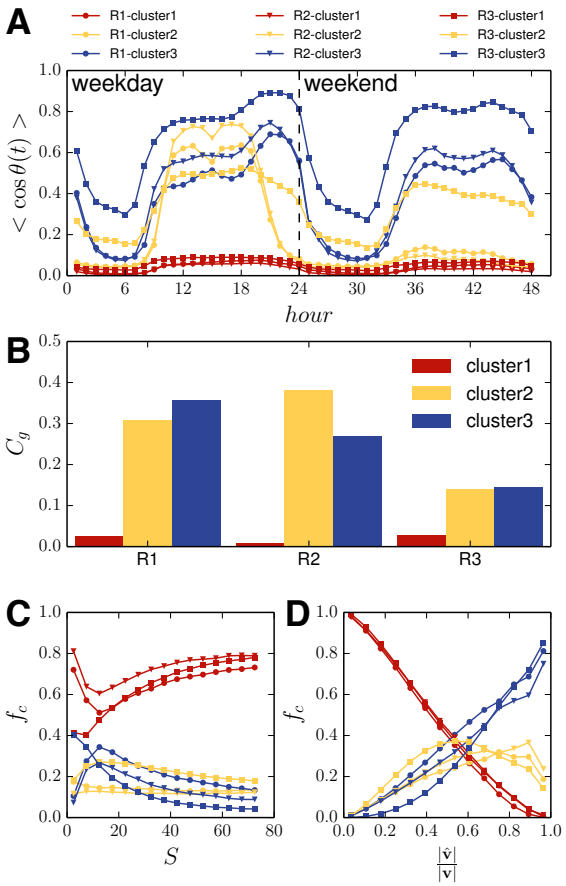


Figure 6.12: Characterizing social ties based on similarity of movement over time. (A) We perform k-means clustering on the set of similarity time series from edges in the network. We find three groups emerge in each city: (*i*) *acquaintances* who have low levels of similarity across all times, (*ii*) *co-workers* who have elevated similarity during work hours on weekdays, but lower levels on weekends, and (*iii*) *family/friends* who have high similarity on nights and weekends. (B) For each city we construct subgraphs containing only edges in a single cluster. We find that these subgraphs retain high clustering coefficient (C_g) within the co-worker and family/friend group while acquaintances are far less likely to have ties among each other. Finally, we explore how an user's behavior correlates with the mobility characteristics of their immediate social network. (C-D) We group nodes based on their mobility characteristics (unique locations visited S and predictability $\frac{|\hat{\mathbf{v}}|}{|\mathbf{v}|}$) then compute the fraction of edges that belong to each of the identified clusters for each node in the group. Individuals that are more exploratory (visit more unique places) tend to have higher fractions of *acquaintances* ties than individuals with lower mobility while the reverse trend is observed for the most predictable individuals.

We call our extension the *GeoSim* model[2]. We compare our model to the original individual-mobility model (IM model) by Song et al. and the Travel-Friendship model (TF model) described by Grabowicz et al (both of them have been briefly presented in Section 6.1).

The GeoSim model works as follows: first, a population of N agents are initialized and connected to replicate the undirected social network constructed from the CDR data in R1. Each edge that exists in the call data, exists in the model, but all weights and similarities are set to 0. Agents are randomly assigned to a location at the start and their location vectors are initialized to reflect this single visit. They are allowed to move in a discrete space of L locations replicating the towers from CDRs.

Each time step corresponds to a single hour of the day. At each time step, individuals decide whether or not to change locations according the waiting time distribution measured in [144], a power-law with an exponential cutoff $p(\Delta t) = \Delta t^{-1-\beta} \exp(\Delta t/\tau)$ where $\beta = 0.8$ and $\tau = 17$ hours. If an individual moves, they must decide to either return to a previously visited location with probability $1 - \rho S^{\gamma}$ or explore and visit a new one with probability ρS^{γ}, where S is the number of unique locations they have visited thus far and $\rho = 0.6$ and $\gamma = 0.6$ are parameters chosen by procedures outlined in [144]. In the original model, an individual u *preferentially returns* to a location l with probability proportional to the frequency of previous visits, $P(l) \propto f_l^u$ and new locations to explore are chosen uniformly at random (note that in our version of the model distance is irrelevant).

In our extension of this model, we choose some locations based on social influence. When picking a return location, our agent has two possibilities. With probability $1 - \alpha$, they select a return location with the preference for locations they have visited in the past as in the original model. With probability α a social contact v is chosen. The probability a given contact is chosen is directly proportional to the current mobility similarity between the two, $P(v) \propto \cos(\theta_{u,v})$ and a location to visit is chosen based on a preference to visit locations frequented by the selected contact, $P(l) \propto f_l^v$ (note that the location choice is repeated until an agent finds a location he has visited before). In the social case, this amounts to preferential return based on a contact's visit frequency as opposed to the ego's visits. In the event that an agent is exploring a new location, the same weighted social coin is flipped. This time, though, with probability $1 - \alpha$ a random, previously unvisited location is selected and with probability α the agent again chooses a contact based on mobility similarity and chooses a new place to visit based on the visit frequencies of that contact. The cosine similarity across all edges is computed and updated over as the model progresses and changes dynamically during the simulation. A schematic of this process can be found in

[2]We have released code and data required to run this model online at `http://humnetlab.mit.edu/wordpress/downloads`.

Figure 6.13.

In this variant of the mobility model, the parameter α controls the influence of social contacts on the visitation patterns of individuals. When $\alpha = 0$, we recover the original mobility model of [144], while when $\alpha = 1$ all location choices are influenced by social ties. In reality, each user may have an inherent value of α that we cannot observe. To incorporate this heterogeneity, we simulate this model for a number of distributions of the parameter α. We find an exponentially distributed α with a mean of $\langle\alpha\rangle = 0.2$ produces a close fit to distributions of mobility similarity and predictability observed in the population. This value is consistent with the results of both Cho et al. [25] and Grabowicz et al. [51] who find that roughly 15-30% of trips were motivated by social intentions.

Having found an appropriate distribution for α, we next compare simulation results with this distribution to results from the IM model (equivalent to the GeoSim model with $\alpha = 0$) and the TF model all run for the same 1 year duration and populations size. Like the IM model it extends, the GeoSim model is able to reproduce elements of individual mobility such as the rate of exploration of new locations $S(t)$ over time (Figure 6.14A) as well as frequency at which users visit their locations f_k (Figure 6.14B). Here the TF model adequately reproduces exploration rates, but produces a flatter visit frequency distribution. In the case of mobility similarity and predictability, however, only the GeoSim model reproduces observed behavior (Figure 6.14C-D). Interestingly, the TF model results in relatively high predictability of users, despite similarity values orders of magnitude lower than those observed in the data or with the IM model. This is likely due to the flattened frequency distribution which the cosine similarity is highly sensitive to. Even if two users share a few locations due the friendship component of the TF model, the preferential dynamics of revisiting a place will continually bring those two users back to that place, increasing cosine similarity. On the other hand, this flat frequency distribution makes very likely that users will share at least some locations in common with each other, making it possible to reproduce location vectors based on social contacts. Despite its inability to recover these distributions, the TF model is the only model tested that builds a social network endogenously. For this reason, we hope future work will find variants on this model capable of dynamically reproducing empirical data of both social and mobility behavior.

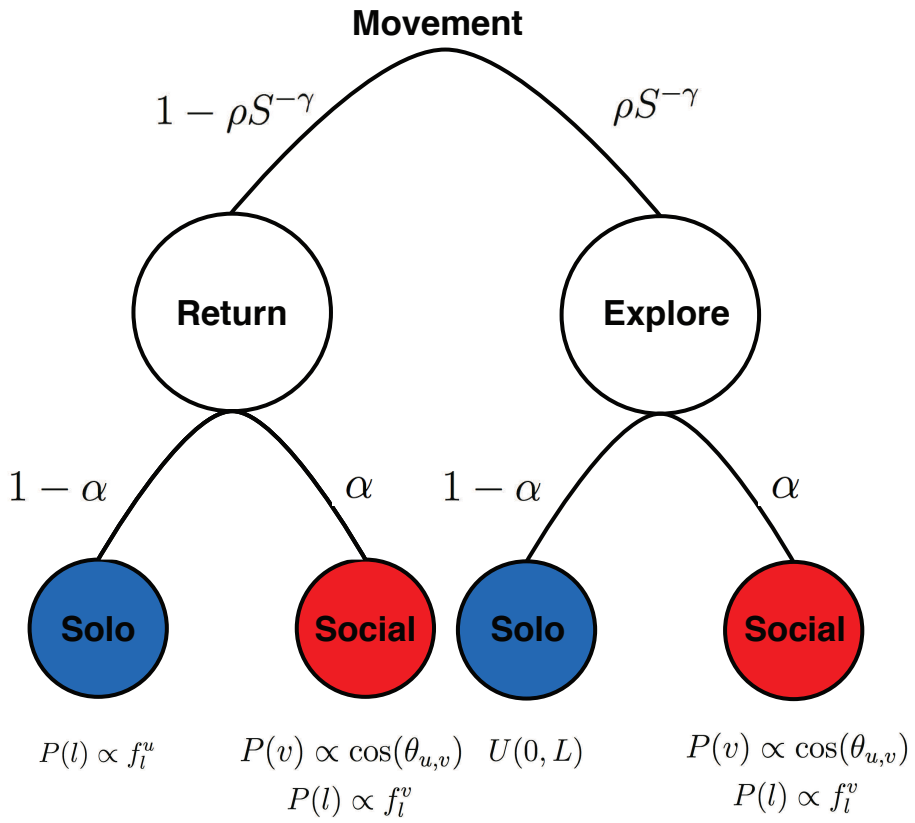


Figure 6.13: A schematic description of the GeoSim model. As in the IM model presented by Song et al., individuals first decide whether to return to a previously visited location or explore a new location. The actual choice of location to visit, new or returning, is made based on either a social influence with probability α or individual preference with probability $1 - \alpha$.

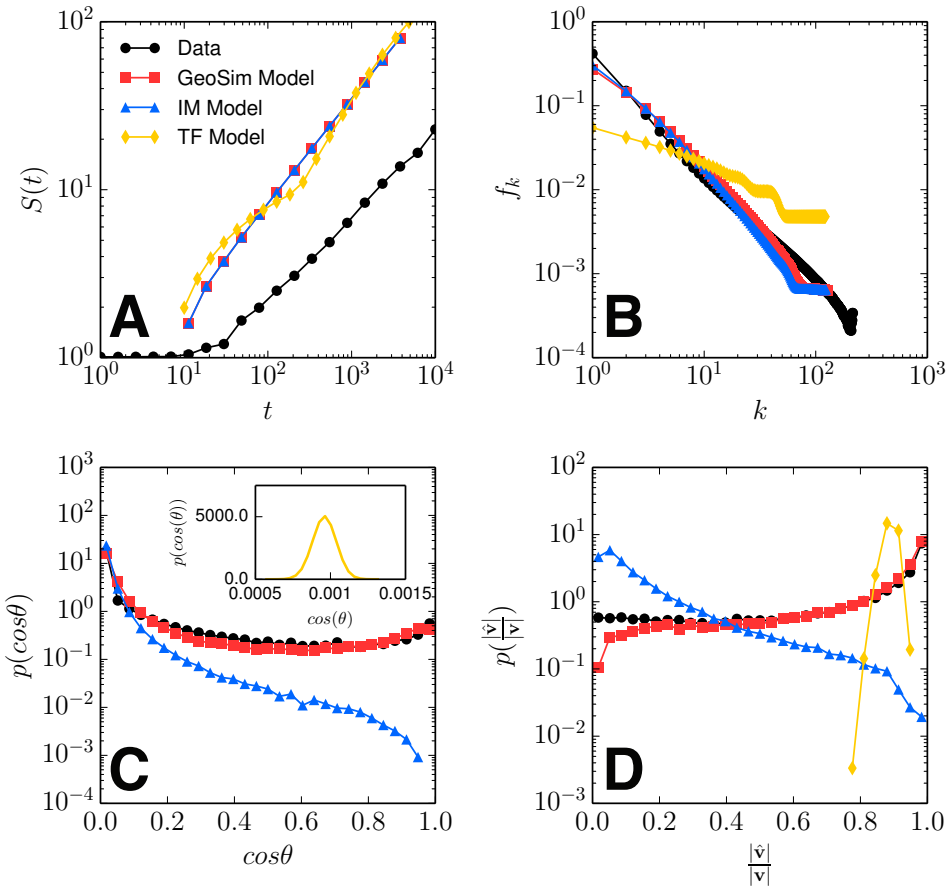


Figure 6.14: Comparing social mobility models. A) We compare model results simulating the rate of exploration $S(t)$ compared to empirical data. While all three models appear to estimate more absolute locations visited, the rate of this growth is consistent between them and in-line with data. B) For each user, we sort locations based on the number visits and compute the frequency that a user visits a location of rank k. We find that the IM models and our extension to it reproduce this distribution well, while the TF model is much flatter, distributing visits more evenly over all locations. C) Only the GeoSim model is able to reproduce patterns of mobility similarity and D) predictability. The TF model results shown in the inset in C shows similarity values orders of magnitude below the observed data. As the similarity is heavily influenced by the frequency distribution of visits, this deviation is likely due to the flatter distribution of f_k produced by the TF model.

Chapter 7

Conclusions

The main objective of this dissertation has been trying to answer the following question: given large scale mobile phone data, what can we learn about the social network and its relationship with the geographical space? In that regard, each of the chapters in this dissertation provides different pieces of the acquired knowledge by analyzing the data in a different way.

First, we have proven that it is possible to infer a lot of information about an opaque node just by looking at how it connects to other nodes. We demonstrated, by combining machine learning and network analysis techniques, that it is possible to infer, with significant accuracy, how opaque nodes connect with each other, as well as predicting certain attributes of the opaque nodes such as the age or the gender. More interestingly, we have found that it is possible to do so, even when 80% of the nodes are opaque. Theses results have huge implications regarding privacy. On the one hand, they show how any corporation that serves 20% of individuals in a certain market, can make accurate predictions about the remaining 80% of the people they have never interacted with. On the other hand, these results may change our perception of the intimacy of social links. They show that, although uploading my digital phone books to a certain service looks like a personal decision, this may allow the service to learn a lot about my contacts, even if some of them actively avoided their own data being gathered by such service.

Regarding the geographical patterns of social networks, we have demonstrated that cities (as defined conventionally by their administrative borders and population size) change the structure of social networks. Interestingly, these findings could be related to urban growth and the economic function of cities.

Taken together, our results lead to the following discoveries: (i) Communities within cities follow a hierarchical structure that favors social distance over geographic distance. (ii) While people living within a geographic radius including several cities form a connected network, the same radius within

cities leads to highly clustered components only connected through people in distant parts of the city. This behavior occurs across different cities and regions sizes, highlighting cities as functional entities of the social networks. (iii) The structure of communities (here related to social proximity) and not geographic distance is what makes social networks searchable within cities. This finding is consistent with experimental results that suggest that people do use the profession or name of the target in the final steps of the search, to make inferences about his/her education or ethnicity, as a hint to help routing within cities.

Our work uncovers an unknown feature of social networks: while at the national level descriptions of social networks consist of highly connected and geographically close communities, we find that geography plays only a minor role when forming communities within cities. Urban networks consist of geographically dispersed communities. This structure explains why people are able to successfully route in Milgram-like experiments, provided they correctly identify the community of the target. Our results support the theoretical hypothesis of Kleinberg: the likelihood to find friendships within communities decays as a power-law with increasing community size, confirming that among all possible network configurations, humans have favored those such that a message can reach anyone even if delivered using only local information. This is a remarkable example of a self-organized structure that allows a small group of individuals to solve a complex problem by cooperating to take advantage of collective knowledge.

Leveraging all this newly-acquired knowledge on the relationship between geographical and social spaces in the micro level, we have proposed two models that produce accurate estimations for transportation and communications fluxes between areas, based only on widely available data such as the population spatial distributions.

For transportation fluxes, we propose an extension to the radiation model that can be calibrated with one scale parameter to predict commuting flows at different spatial scales. The scale parameter α modulates the influence of the opportunity distribution heterogeneity and the spatial scale l of the commuting zones. The main advantage of the proposed modelling framework is that it can still be applied to predict the number of commuting trips when lacking data for calibration. The α parameter depends on the scale of the study region and the homogeneity of the population distribution. The presented results provide the first building blocks for a multi-scale generator of human mobility expressed as a function of the distributions of population and job facilities. We tested the model in different scales at different countries and discussed its range of applicability. We have shared the sample of the U.S. county level commuting flow prediction[1] to help in this direction.

For communication fluxes, the presented elliptic model successfully takes

[1] `http://humnetlab.mit.edu/extendedradiation/`

into account the symmetry of such fluxes, in order to predict the number of social ties between geographical locations at different scales, ranging from neighborhoods to regions. Interestingly, we have shown that geolocated population data is as useful to predict communication fluxes as it is to predict trip fluxes, even if digitally mediated communications are not affected by the same cost penalization that affects long trips. The proposed model is readily available to be used by researchers in different social sciences[2] studying different phenomena where human ties are known to be crucial, such as information propagation or disease spreading. Overall, our model implies that social ties are to a large extent driven by geographical factors. While there may be other factors influencing very long distance relationships (e.g. time zones, or natural, national and language borders, etc.) the available data did not allow to check them, so that further research would be needed along this line.

Finally, we came back to the micro scale to study the relationship between social and mobility patters within cities. We have offered new metrics and empirical findings that relate social behaviours to mobility similarity and predictability. Our results show that our mobility is far more similar to the mobility of our social contacts than to the one of strangers, and that this similarity can be used to reconstruct our own mobility patterns. We find strong, positive correlations between tie strength and mobility similarity. Moreover, temporal variations in this similarity reveal three distinct groups of social ties that hint at semantic types of relationships such as co-worker or family member. These subgraphs often have high levels of intra-group clustering, suggesting functional groups of individuals within the network. The mix of these groups among the edges of an individual's ego network is correlated with their mobility behaviour; users with many dissimilar contacts tend to explore more locations. Speaking to their generalizability, these results persist across three different cities in two countries. Finally, we extended an established mobility model to include choices based on social behaviour, thus replicating the empirical findings described here as well as from other works. We call this model the GeoSim model and we have compared its results to those of two similar models with the hope that it may be a useful tool for future work in the area.

7.1 Future research

The findings presented have a number of implications for researchers interested in social networks or mobility applications extracted from ICTs. Additional contextual information about the relationships may help to predict

[2]Implementations of the elliptic model written in the most widely used programming languages by the data science community have been open sourced and made available at `https://github.com/humnetlab/elliptic-model/`

missing links or to provide critical details for a more accurate modelling of the flows of information or diseases. Urban planners or those people needing good estimates of travel demand can incorporate social mechanisms like the ones described here to improve on their models and to capture movements previously unaccounted for. Robust findings that classify social contacts from passive data alone may influence future studies and help with data informed policies through city science. In the new data-rich reality of cities, deeper insights into the connections between us will help make the places we live more sustainable, efficient, productive and fun.

Overall, there is another conclusion emerging from this work. When we look at figures like 3.15 or 4.12, we realize that the kind of analysis we have performed during this research can provide valuable insights about the underlying idiosyncrasy of countries and regions. It would be amazing to explore, for example, the evolution of network modularity between the United Kingdom and the European Union during 10 years before and after Brexit, or to do similar analysis related to the rise or decline of separatists movements in Northern Ireland, Catalonia or Scotland. By analyzing the social networks, we can add important information to the public discussion, in a world that is starting to resent globalization. Measuring how each individual interacts with each other should probably be more relevant than the number of tones of iron exchanged between two areas, when it comes to select the right level of aggregation in order to decide how we rule ourselves in democratic societies.

Bibliography

[1] Dimitris Achlioptas, Raissa M D'Souza, and Joel Spencer. Explosive percolation in random networks. *Science*, 323(5920):1453–1455, 2009.

[2] Lada Adamic and Eytan Adar. How to search a social network. *Social networks*, 27(3):187–203, 2005.

[3] Yong-Yeol Ahn, James P Bagrow, and Sune Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466(7307):761–764, 2010.

[4] S E Ahnert and T M A Fink. Clustering signatures classify directed networks. *Physical Review E*, 78(3):36112, 2008.

[5] E M Airoldi, D M Blei, S E Fienberg, and E P Xing. Mixed membership stochastic blockmodels. *The Journal of Machine Learning Research*, 9:1981–2014, 2008.

[6] Thomas J Allen. Managing the flow of technology: Technology transfer and the dissemination of technological information within the R&D organization. *MIT Press Books*, 1, 1984.

[7] Luis A Nunes Amaral, Antonio Scala, Marc Barthelemy, and H Eugene Stanley. Classes of small-world networks. *Proceedings of the national academy of sciences*, 97(21):11149–11152, 2000.

[8] Alex Anas. Discrete choice theory, information theory and the multinomial logit and gravity models. *Transportation Research Part B: Methodological*, 17(1):13–23, 1983.

[9] Nuno A M Araujo and Hans J Herrmann. Explosive percolation via control of the largest cluster. *Physical Review Letters*, 105(3):35701, 2010.

[10] Lars Backstrom, Dan Huttenlocher, Jon M Kleinberg, and Xiangyang Lan. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 44–54. ACM, 2006.

[11] James P Bagrow and Yu-Ru Lin. Mesoscopic structure and social aspects of human mobility. *PloS one*, 7(5):e37676, 2012.

[12] James P Bagrow, Dashun Wang, and Albert-László Barabási. Collective response of human populations to large-scale emergencies. *PloS one*, 6(3):e17680, 2011.

[13] Duygu Balcan, Vittoria Colizza, Bruno Gonçalves, Hao Hu, José J Ramasco, and Alessandro Vespignani. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences*, 106(51):21484–21489, 2009.

[14] Albert-László Barabási and R Albert. Emergence of scaling in random networks. *Science*, 286(5439):509, 1999.

[15] Marc Barthélemy. Spatial networks. *Physics Reports*, 499(1):1–101, 2011.

[16] Peter S Bearman, James Moody, and Katherine Stovel. Chains of Affection: The Structure of Adolescent Romantic and Sexual Networks1. *American Journal of Sociology*, 110(1):44–91, 2004.

[17] Budhendra Bhaduri, Edward Bright, Phillip Coleman, and Marie L Urban. LandScan USA: a high-resolution geospatial and temporal modeling approach for population distribution and dynamics. *GeoJournal*, 69(1-2):103–117, 2007.

[18] Vincent D Blondel, J L Guillaume, R Lambiotte, and E Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[19] Vincent D Blondel, G Krings, and I Thomas. Regions and borders of mobile telephony in Belgium and in the Brussels metropolitan zone. *Brussels Studies*, 42(4), 2010.

[20] Bela Bollobás and Fan R K Chung. The diameter of a cycle plus a random matching. *SIAM Journal on discrete mathematics*, 1(3):328–333, 1988.

[21] Dirk Brockmann, Lars Hufnagel, and Theo Geisel. The scaling laws of human travel. *Nature*, 439(7075):462–465, 2006.

[22] Frances Cairncross. *The death of distance: How the communications revolution will change our lives*. Harvard Business Press, 2001.

[23] F Calabrese, D Dahlem, A Gerber, D Paul, X Chen, J Rowland, C Rath, and C Ratti. 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 conference on social computing (socialcom), pages 223–230. IEEE, 2011.

[24] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

[25] E Cho, S A Myers, and Jure Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1082–1090. ACM, 2011.

[26] A Clauset, C Moore, and Mark E J Newman. Hierarchical structure and the prediction of missing links in networks, 2008. *Nature*, 453:98, 2008.

[27] Colleen Cuddy and Nancy R Glassman. Location-based services: FourSquare and Gowalla, should libraries play? *Journal of Electronic Resources in Medical Libraries*, 7(4):336–343, 2010.

[28] Jesper Dall and Michael Christensen. Random geometric graphs. *Physical Review E*, 66(1):16121, 2002.

[29] Manlio De Domenico, Antonio Lima, and Mirco Musolesi. Interdependence and predictability of human mobility and social interactions. *Pervasive and Mobile Computing*, 9(6):798–807, 2013.

[30] Yves-Alexandre de Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3, 2013.

[31] Ithiel de Sola Pool and Manfred Kochen. Contacts and influence. *Social networks*, 1(1):5–51, 1979.

[32] Peter Sheridan Dodds, Roby Muhamad, and Duncan J Watts. An experimental study of search in global social networks. *science*, 301(5634):827–829, 2003.

[33] Robin I M Dunbar. Neocortex size as a constraint on group size in primates. *Journal of Human Evolution*, 22(6):469–493, 1992.

[34] Nathan Eagle, Alex Pentland, and David Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36):15274–15278, 2009.

[35] Rosalind M Eggo, Simon Cauchemez, and Neil M Ferguson. Spatial dynamics of the 1918 influenza pandemic in England, Wales and the

United States. *Journal of The Royal Society Interface*, 8(55):233–243, 2011.

[36] P Erdos and A Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60, 1960.

[37] Sven Erlander and Neil F Stewart. *The gravity model in transportation analysis: theory and extensions*, volume 3. Vsp, 1990.

[38] TS Evans and JP Saramäki. Scale-free networks from self-organization. *Physical Review E*, 72(2):026138, 2005.

[39] P Expert, T S Evans, Vincent D Blondel, and R Lambiotte. Uncovering space-independent communities in spatial networks. *Proceedings of the National Academy of Sciences*, 108(19):7663–7668, 2011.

[40] Ron Eyal, Avi Rosenfeld, Sigal Sina, and Sarit Kraus. Predicting and identifying missing node information in social networks. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 8(3):14, 2014.

[41] Juan Fernández-Gracia, Krzysztof Suchecki, José J Ramasco, Maxi San Miguel, and Víctor M Eguíluz. Is the voter model a model for voters? *Physical review letters*, 112(15):158701, 2014.

[42] R A Fisher. The use of multiple measurements in taxonomic problems. *Annals of Human Genetics*, 7(2):179–188, 1936.

[43] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010.

[44] James H Fowler and Nicholas A Christakis. Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham Heart Study. *BMJ: British medical journal*, 337, 2008.

[45] Pierre Fraigniaud, Cyril Gavoille, and Christophe Paul. Eclecticism shrinks even small worlds. *Distributed Computing*, 18(4):279–291, 2006.

[46] Joseph E Gillis and George H Weiss. Expected number of distinct sites visited by a random walk with an infinite variance. *Journal of Mathematical Physics*, 11(4):1307–1312, 1970.

[47] M Gimeno, B Villamia, and V Suarez. *eEspa{ñ}a 2011, Informe anual sobre el desarrollo de la sociedad de la informaci{ó}n en Espa{ñ}a.* Fundaci{ó}n Orange, 2011.

[48] B Goncalves, N Perra, and Alessandro Vespignani. Modeling users' activity on Twitter networks: validation of Dunbar's number. *PLoS One*, 6(8):e22656, 2011.

[49] Marta C González, César A Hidalgo, and Albert-László Barabási. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[50] Michael Gould, Max Craglia, Michael F Goodchild, Alessandro Annoni, Gilberto Camara, Werner Kuhn, David Mark, Ian Masser, David Maguire, Steve Liang, and Others. Next-generation digital earth: A position paper from the vespucci initiative for the advancement of geographic information science. *International Journal of Spatial Data Infrastructures Research*, 2008.

[51] Przemyslaw A Grabowicz, José J Ramasco, Bruno Gonçalves, and V\'\ictor M Egu\'\iluz. Entangling mobility and interactions in social media. *PloS one*, 9(3):e92196, 2014.

[52] Mark S Granovetter. The strength of weak ties. *American journal of sociology*, pages 1360–1380, 1973.

[53] Jean M Guiot. A modification of Milgram's small world method. *European journal of social psychology*, 6(4):503–507, 1976.

[54] D Heckerman, D Geiger, and D M Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3):197–243, 1995.

[55] Carlos Herrera-Yagüe. Extracting social networks from interactions logs. `http://toobigtobecomplex.blogspot.com.es/2012/06/extracting-social-network-from.html`, 2012.

[56] Carlos Herrera-Yagüe and Pedro J Zufiria. Generating scale-free networks with adjustable clustering coefficient via random walks. In *Network Science Workshop (NSW), 2011 IEEE*, pages 167–172. IEEE, 2011.

[57] César A Hidalgo and C Rodriguez-Sickert. The dynamics of a mobile phone network. *Physica A: Statistical Mechanics and its Applications*, 387(12):3017–3024, 2008.

[58] P W Holland, K B Laskey, and S Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.

[59] Petter Holme and Beom Jun Kim. Growing scale-free networks with tunable clustering. *Physical Review E*, 65(2):26107, 2002.

[60] Z Huang and D D Zeng. A link prediction approach to anomalous email detection. In *Systems, Man and Cybernetics, 2006. SMC'06. IEEE International Conference on*, volume 2, pages 1131–1136. IEEE, 2006.

[61] Mark L Huson and Arunabha Sen. Broadcast scheduling algorithms for radio networks. In *Military Communications Conference, 1995. MILCOM'95, Conference Record, IEEE*, volume 2, pages 647–651. IEEE, 1995.

[62] C R Johnson. Matrix completion problems: a survey. In *Matrix Theory and Applications*, volume 40, pages 171–198. Amer Mathematical Society, 1990.

[63] Benjamin F Jones, Stefan Wuchty, and Brian Uzzi. Multi-university research teams: shifting impact, geography, and stratification in science. *Science*, 322(5905):1259–1262, 2008.

[64] Woo-Sung Jung, Fengzhong Wang, and H Eugene Stanley. Gravity model in the Korean highway. *EPL (Europhysics Letters)*, 81(4):48005, 2008.

[65] Pablo Kaluza, Andrea Kölzsch, Michael T Gastner, and Bernd Blasius. The complex network of global cargo ship movements. *Journal of the Royal Society Interface*, 7(48):1093–1103, 2010.

[66] Brian Karrer and Mark E J Newman. Random graphs containing arbitrary distributions of subgraphs. *Physical Review E*, 82(6):66118, 2010.

[67] Peter D Killworth and H Russell Bernard. The reversal small-world experiment. *Social networks*, 1(2):159–192, 1979.

[68] Myunghwan Kim and Jure Leskovec. The Network Completion Problem: Inferring Missing Nodes and Edges in Networks. In *SDM*, pages 47–58. SIAM, 2011.

[69] Jon M Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[70] Jon M Kleinberg. Navigation in a small world. *Nature*, 406(6798):845, 2000.

[71] Jon M Kleinberg. The small-world phenomenon: An algorithmic perspective. In *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pages 163–170. ACM, 2000.

[72] Jon M Kleinberg. Small-world phenomena and the dynamics of information. *Advances in neural information processing systems*, 1:431–438, 2002.

[73] Jon M Kleinberg. The small-world phenomenon and decentralized search. *SiAM News*, 37(3):1–2, 2004.

[74] Jon M Kleinberg. Complex networks and decentralized search algorithms. In *Proceedings of the International Congress of Mathematicians (ICM)*, volume 3, pages 1019–1044, 2006.

[75] Judith Kleinfeld. Could it be a big world after all? The six degrees of separation myth. *Society, April*, 12:2–5, 2002.

[76] B Klimt and Yingxiang Yang. Introducing the Enron corpus. In *First conference on email and anti-spam (CEAS)*, 2004.

[77] Charles Korte and Stanley Milgram. Acquaintance networks between racial groups: Application of the small world method. *Journal of Personality and Social Psychology*, 15(2):101, 1970.

[78] Gueorgi Kossinets. Effects of missing data in social networks. *Social networks*, 28(3):247–268, 2006.

[79] Gueorgi Kossinets and Duncan J Watts. Empirical analysis of an evolving social network. *Science*, 311(5757):88–90, 2006.

[80] G Krings, F Calabrese, C Ratti, and Vincent D Blondel. Urban gravity: a model for inter-city telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(07):L07003, 2009.

[81] J Kunegis, E De Luca, and S Albayrak. The link prediction problem in bipartite networks. *Computational Intelligence for Knowledge-Based Systems Design*, pages 380–389, 2010.

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

[83] Doug Laney. 3d data management: Controlling data volume, velocity and variety. *META Group Research Note*, 6:70, 2001.

[84] Emmanuelle Lebhar and Nicolas Schabanel. Almost optimal decentralized routing in long-range contact networks. In *Automata, Languages and Programming*, pages 894–905. Springer, 2004.

[85] J J Leeming and G M Mackay. *Road accidents: prevent or punish?* Cassell, 1969.

[86] Maxime Lenormand, Sylvie Huet, Floriana Gargiulo, and Guillaume Deffuant. A universal model of commuting networks. *PloS one*, 7(10):e45985, 2012.

[87] Jure Leskovec, Deepayan Chakrabarti, Jon M Kleinberg, Christos Faloutsos, and Zoubin Ghahramani. Kronecker graphs: An approach to modeling networks. *The Journal of Machine Learning Research*, 11:985–1042, 2010.

[88] Jure Leskovec, Dan Huttenlocher, and Jon M Kleinberg. Signed networks in social media. In *Proceedings of the 28th international conference on Human factors in computing systems*, pages 1361–1370. ACM, 2010.

[89] Jure Leskovec, Kevin J Lang, Anirban Dasgupta, and Michael W Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics*, 6(1):29–123, 2009.

[90] Jure Leskovec and Julian J Mcauley. Learning to discover social circles in ego networks. In *Advances in neural information processing systems*, pages 539–547, 2012.

[91] Xiao Liang, Jichang Zhao, Li Dong, and Ke Xu. Unraveling the origin of exponential law in intra-urban human mobility. *Scientific reports*, 3, 2013.

[92] David Liben-Nowell and Jon M Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031, 2007.

[93] David Liben-Nowell, Jasmine Novak, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. Geographic routing in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33):11623–11628, 2005.

[94] Nan Lin, Paul W Dayton, and Peter Greenwald. Analyzing the instrumental use of relations in the context of social structure. *Sociological Methods & Research*, 7(2):149–166, 1978.

[95] C Ling, J Huang, and H Zhang. AUC: a better measure than accuracy in comparing learning algorithms. *Advances in Artificial Intelligence*, page 991, 2003.

[96] W Liu and L Lü. Link prediction based on local random walk. *EPL (Europhysics Letters)*, 89:58007, 2010.

[97] L Lü, C H Jin, and T Zhou. Similarity index based on local paths for link prediction of complex networks. *Physical Review E*, 80(4):46122, 2009.

[98] Linyuan Lü and Tao Zhou. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, 390(6):1150–1170, 2011.

[99] Xin Lu, Linus Bengtsson, and Petter Holme. Predictability of population displacement after the 2010 Haiti earthquake. *Proceedings of the National Academy of Sciences*, 109(29):11576–11581, 2012.

[100] Craig C Lundberg. Patterns of acquaintanceship in society and complex organization: A comparative study of the small world problem. *Pacific Sociological Review*, pages 206–222, 1975.

[101] Oded Z Maimon and Lior Rokach. *Data mining and knowledge discovery handbook*, volume 1. Springer, 2005.

[102] Gurmeet Singh Manku, Moni Naor, and Udi Wieder. Know thy neighbor's neighbor: the power of lookahead in randomized P2P networks. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 54–63. ACM, 2004.

[103] Chip Martel and Van Nguyen. Analyzing Kleinberg's (and other) small-world models. In *Proceedings of the twenty-third annual ACM symposium on Principles of distributed computing*, pages 179–188. ACM, 2004.

[104] A Paolo Masucci, Joan Serras, Anders Johansson, and Michael Batty. Gravity versus radiation models: On the importance of scale and heterogeneity in commuting flows. *Physical Review E*, 88(2):22812, 2013.

[105] Naoki Masuda, Hiroyoshi Miwa, and Norio Konno. Geographical threshold graphs with small-world and scale-free properties. *Physical Review E*, 71(3):36108, 2005.

[106] Christopher McCarty, Peter D Killworth, H Russell Bernard, Eugene C Johnsen, and Gene A Shelley. Comparing two methods for estimating network size. *Human Organization*, 60(1):28–39, 2001.

[107] Tyler H McCormick, Matthew J Salganik, and Tian Zheng. How many people do you know?: Efficiently estimating personal network size. *Journal of the American Statistical Association*, 105(489):59–70, 2010.

[108] Stanley Milgram. The small world problem. *Psychology today*, 2(1):60–67, 1967.

[109] Giovanna Miritello, Rubén Lara, Manuel Cebrian, and Esteban Moro. Limited communication capacity unveils strategies for human interaction. *Scientific reports*, 3, 2013.

[110] Giovanna Miritello, Esteban Moro, and Rubén Lara. Dynamical strength of social ties in information spreading. *Physical Review E*, 83(4):45102, 2011.

[111] A Mislove, M Marcon, K P Gummadi, P Druschel, and B Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42. ACM, 2007.

[112] Gaspard Monge. *M{é}moire sur la th{é}orie des d{é}blais et des remblais*. De l'Imprimerie Royale, 1781.

[113] Elliott W Montroll and George H Weiss. Random walks on lattices. II. *Journal of Mathematical Physics*, 6(2):167–181, 1965.

[114] R Muhamad. *Search in Social Networks*. Columbia University, 2010.

[115] Mark E J Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.

[116] Mark E J Newman. Random graphs with clustering. *Physical Review Letters*, 103(5):58701, 2009.

[117] Mark E J Newman, S H Strogatz, and Duncan J Watts. Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64(2):26118, 2001.

[118] Mark EJ Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[119] J P Onnela, S Arbesman, Marta C González, Albert-László Barabási, and Nicholas A Christakis. Geographic constraints on social network groups. *PLoS One*, 6(4):e16939, 2011.

[120] J P Onnela, Jari Saramaki, J Hyvönen, G Szabó, David Lazer, K Kaski, J Kertész, and Albert-László Barabási. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18):7332–7336, 2007.

[121] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report 1999-66, Stanford InfoLab, nov 1999.

[122] G Palla, Albert-László Barabási, and T Vicsek. Quantifying social group evolution. *Nature*, 446(7136):664–667, 2007.

[123] Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.

[124] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Physical review letters*, 86(14):3200, 2001.

[125] Roger Penrose. A generalized inverse for matrices. In *Mathematical proceedings of the Cambridge philosophical society*, volume 51, pages 406–413. Cambridge Univ Press, 1955.

[126] Galen Pickard, Wei Pan, Iyad Rahwan, Manuel Cebrian, Riley Crane, Anmol Madan, and Alex Pentland. Time-critical social mobilization. *Science*, 334(6055):509–512, 2011.

[127] Pentti Pöyhönen. A tentative model for the volume of trade between countries. *Weltwirtschaftliches Archiv*, pages 93–100, 1963.

[128] José J Ramasco and Alessandro Vespignani. Commuting and pandemic prediction. *Proc. Natl. Acad. Sci*, 106:21459–21460, 2009.

[129] C Ratti, S Sobolevsky, F Calabrese, C Andris, J Reades, M Martino, R Claxton, and S H Strogatz. Redrawing the map of Great Britain from a network of human interactions. *PLoS One*, 5(12):e14248, 2010.

[130] M Richardson, R Agrawal, and P Domingos. Trust management for the semantic web. *The SemanticWeb-ISWC 2003*, pages 351–368, 2003.

[131] B D Ripley. *Pattern recognition and neural networks*. Cambridge Univ Pr, 2008.

[132] Filipe Rodrigues, ACdCO Alves, Evgheni Polisciuc, Shan Jiang, Joseph Ferreira, and F Pereira. Estimating disaggregated employment size from points-of-interest and census data: From mining the web to model implementation and visualization. *International Journal on Advances in Intelligent Systems*, 6(1):41–52, 2013.

[133] Diego Rybski, Sergey V Buldyrev, Shlomo Havlin, Fredrik Liljeros, and Hernán A Makse. Scaling laws of human interaction activity. *Proceedings of the National Academy of Sciences*, 106(31):12640–12645, 2009.

[134] Jari Saramäki and Kimmo Kaski. Scale-free networks generated by random walkers. *Physica A: Statistical Mechanics and its Applications*, 341:80–86, 2004.

[135] S Scellato, A Noulas, and C Mascolo. Exploiting place features in link prediction on location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1046–1054. ACM, 2011.

[136] J B Schafer, J A Konstan, and J Riedl. E-commerce recommendation applications. *Data mining and knowledge discovery*, 5(1):115–153, 2001.

[137] Eric Schmidt and Jonathan Rosenberg. *How google works*. Hachette UK, 2014.

[138] Christian M Schneider, Tamara Mihaljev, Shlomo Havlin, and Hans J Herrmann. Suppressing epidemics with a limited amount of immunization units. *Physical Review E*, 84(6):61911, 2011.

[139] M Schwartz. *Telecommunication networks: protocols, modeling and analysis*. Addison-Wesley Longman Publishing Co., Inc., 1986.

[140] M S Shang, L Lü, Y C Zhang, and T Zhou. Empirical analysis of web-based user-object bipartite networks. *EPL (Europhysics Letters)*, 90:48006, 2010.

[141] R Lance Shotland and Margret K Straw. Bystander response to an assault: When a man attacks a woman. *Journal of Personality and Social Psychology*, 34(5):990, 1976.

[142] Filippo Simini, Marta C González, A Maritan, and Albert-László Barabási. A universal model for mobility and migration patterns. *Nature*, 484(7392):96–100, 2012.

[143] Filippo Simini, Amos Maritan, and Zoltán Néda. Human mobility in a continuum approach. *PloS one*, 8(3):e60069, 2013.

[144] Chaoming Song, Tal Koren, Pu Wang, and Albert-László Barabási. Modelling the scaling properties of human mobility. *Nature Physics*, 6(10):818–823, 2010.

[145] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[146] Thorvald Sørensen. ${$A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons$}$. *Biol. Skr.*, 5:1–34, 1948.

[147] A Stoica, Thomas Couronne, and J S Beuscart. To be a star is not only metaphoric: from popularity to social linkage. In *Proc. ICWSM10 4th. Intl. Conf. Weblogs & Social Media*, 2010.

[148] A Stoica, Zbigniew Smoreda, C Prieurb, and J L Guillaume. Age, gender and communication networks. In *Proceedings of the Workshop on the Analysis of Mobile Phone Networks, satellite workshop to NetSci 2010*, 2010.

[149] Michael P H Stumpf, Thomas Thorne, Eric de Silva, Ronald Stewart, Hyeong Jun An, Michael Lappe, and Carsten Wiuf. Estimating the size of the human interactome. *Proceedings of the National Academy of Sciences*, 105(19):6959–6964, 2008.

[150] Daniel Sui, Sarah Elwood, and Michael Goodchild. *Crowdsourcing geographic knowledge: volunteered geographic information (VGI) in theory and practice*. Springer Science & Business Media, 2012.

[151] Arun Sundararajan. *The Sharing Economy: The End of Employment and the Rise of Crowd-Based Capitalism*. MIT Press, 2016.

[152] Christian Thiemann, Fabian Theis, Daniel Grady, Rafael Brune, and Dirk Brockmann. The structure of borders in a small world. *PloS one*, 5(11):e15422, 2010.

[153] Jameson L Toole, Meeyoung Cha, and Marta C González. Modeling the adoption of innovations in the presence of geographic and media influences. *PloS one*, 7(1):e29528, 2012.

[154] Jeffrey Travers and Stanley Milgram. An experimental study of the small world problem. *Sociometry*, pages 425–443, 1969.

[155] Johan Ugander, Brian Karrer, Lars Backstrom, and Cameron Marlow. The anatomy of the facebook social graph. *arXiv preprint arXiv:1111.4503*, 2011.

[156] M Ulm and P Widhalm. Properties of the positioning error of cell phone trajectories. *NetMob workshop, Boston (MA), USA*, 2013.

[157] Cécile Viboud, Ottar N Bjørnstad, David L Smith, Lone Simonsen, Mark A Miller, and Bryan T Grenfell. Synchrony, waves, and spatial hierarchies in the spread of influenza. *science*, 312(5772):447–451, 2006.

[158] Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, and Albert-László Barabási. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1100–1108. ACM, 2011.

[159] Pu Wang, Marta C González, César A Hidalgo, and Albert-László Barabási. Understanding the spreading patterns of mobile phone viruses. *Science*, 324(5930):1071–1076, 2009.

[160] Duncan J Watts, Peter Sheridan Dodds, and Mark E J Newman. Identity and search in social networks. *science*, 296(5571):1302–1305, 2002.

[161] Duncan J Watts and S Strogatz. The small world problem. *Collective Dynamics of Small-World Networks*, 393:440–442, 1998.

[162] Duncan J Watts and S H Strogatz. Collective dynamics of small-world networks. *Nature*, 393(6684):440–442, 1998.

[163] Bernard M Waxman. Routing of multipoint connections. *IEEE journal on selected areas in communications*, 6(9):1617–1622, 1988.

[164] Claus Wedekind and Sandra Füri. Body odour preferences in men and women: do they aim for specific MHC combinations or simply heterozygosity? *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 264(1387):1471–1479, 1997.

[165] Amy Wesolowski, Nathan Eagle, Andrew J Tatem, David L Smith, Abdisalan M Noor, Robert W Snow, and Caroline O Buckee. Quantifying the impact of human mobility on malaria. *Science*, 338(6104):267–270, 2012.

[166] H C White, S A Boorman, and R L Breiger. Social structure from multiple networks. I. Blockmodels of roles and positions. *American journal of sociology*, pages 730–780, 1976.

[167] Alan Geoffrey Wilson and M L Senior. Some relationships between entropy maximizing models, mathematical programming models, and their duals. *Journal of Regional Science*, 14(2):207–215, 1974.

[168] Jaewon Yang and Jure Leskovec. Community-affiliation graph model for overlapping network community detection. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 1170–1175. IEEE, 2012.

[169] Yingxiang Yang, Carlos Herrera-Yagüe, Nathan Eagle, and Marta C González. Limits of predictability in commuting flows in the absence of data for calibration. *Scientific reports*, 4, 2014.

[170] Haiyuan Yu, Pascal Braun, Muhammed A Y\ild\ir\im, Irma Lemmens, Kavitha Venkatesan, Julie Sahalie, Tomoko Hirozane-Kishikawa, Fana Gebreab, Na Li, Nicolas Simonis, and Others. High-quality binary protein interaction map of the yeast interactome network. *Science*, 322(5898):104–110, 2008.

[171] Q Zhao, Y Tian, Q He, N Oliver, R Jin, and W C Lee. Communication motifs: a tool to characterize social communications. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1645–1648. ACM, 2010.

[172] T Zhou, Lü L., and Y C Zhang. Predicting missing links via local information. *The European Physical Journal B-Condensed Matter and Complex Systems*, 71(4):623–630, 2009.

[173] George Kingsley Zipf. The P 1 P 2/D hypothesis: on the intercity movement of persons. *American sociological review*, 11(6):677–686, 1946.

Appendix A

Extracting social network from interactions log

The task of extracting a social networks from interaction logs (in this example, email logs) seems trivial, and it is trivial indeed for small data, but it gets much more complicated when the data size increases. In general, it is good to use as many resources as we have available. It is not the same if you have to run the code in your mono-core netbook, your 4-core laptop or in a 32-core research server. Although this appendix will be focused in a certain architecture, I hope we provide comments to give you clues on how to adapt it to your system

Problem specification

The problem sounds like a really easy task. We are provided with a set of text files containing following info per line

```
emiter;receiver;number_of_emails.
```

which means than the *emiter* sent the *receiver* exactly $number_of_mails$ during a certain observation period.

Our goal is to get an output with this this format

```
userA;userB;#emailsAtoB;#emailsBtoA
```

So, for up to some thousand lines input, this problem is really straightforward to solve. As you may have already thought, any solution developed for this can be used to process any massive communication log, such as CDRs, just by adding proper line process.

However, this my particular case, we were provided with 25GB in text files, containing almost a billion lines. So it was a bit more complicated.

Available resources

In this case, we had access to a 16-core, 50GB RAM linux server, which although shared with other researchers, anyone can usually occupy 80% resources without no problem. Also it was a pleasure to check that disk storage was pretty fast, reaching 500-600MB/s. My limited experience makes me think this is a common setting in nowadays research labs.

Now, let me introduce you to an amazing Unix tool we use intensively to measure performance, called *pv*. It just get the data rate crossing any Unix pipe, so you can use to test either disk o network connection (just place *cat* o *nc* before). Usage example:

```
cat whatever_file_in_disk | pv -r > /dev/null
```

First attempt: mono-core python dict based solution

Our first approach was to try some previous written code to do exactly this, but in a smaller scale. Code takes advantage on python dictionary which basically allows to retrieve a value from the key in $O(1)$ time, compared with regular array search (trying to find the key by iterating on a list, which is $O(n)$).

Since, as you can see in the code bellow, it is necessary to search for a relation once per line, the advantage of using *dicts* (in java would be HashTable, in php associative arrays, although in my experience php implementation was worse in both memory and time) is big, as big I could say that without the $O(1)$ trick, the machine would probably would not finish by the end of my PhD period, and that was really a design request.

```
links={}
i=0
f=file("/dev/stdin")

for line in f:
    i+= 1
    emitter, receiver, contacts = map(int, line.split(";"))
    rel_key = tuple(sorted([emitter,receiver]))
    direction = 0 if caller < receiver else 1
    if not rel_key in links:
        links[rel_key]=[0,0]
    links[rel_key][direction] += contacts

f.close()
f= open("network-mutual-graph","w")
f2= open("network-directed-graph","w")

def output(caller, receiver, contacts):
    return "%s %s %i \n" % (caller,receiver,contacts)

for link,contacs in links.iteritems():
    link=link.split('-')
```

```
if contacts[0]==0:
    f2.write(output(link[1],link[0],contacts[1]))
elif contacts[1]==0:
    f2.write(output(link[0],link[1],contacts[0]))
else:
    f.write(output(link[0],link[1],contacts[0],contacts[1]))
```

Memory, memory, memory

The first time I used a research server having 10 times more memory my laptop does, I thought I would never use that much. Obviously I was wrong. In this case, the 50G were enough for barely 15 minutes, until the dictionary reached 100 million keys. I tried to overcome it by using a 2 integer tuple instance of previous string 'userA-userB'. Although it did not work out (memory consumption was reduced by 15-20% but not enough) I happened to learn some things about python basic types memory consumption[1].

So what to do?

Once it was clear we did not have enough memory to keep al possible relationships in a hashTable-like structure, the next thing was to start considering alternatives:

1. Write code to manage swapping to disk, keeping only a fraction of relationships in memory. However doing this efficiently requires a non-easy task like keeping in the cache last (so more likely to appear in future) relationships, and store the data in fixed width sorted files, so that in the event of having to come to disk for a relationship, it doesn't take ages (again the O(log(n)) trick).

2. Doing basically the thing described in point 1 but via a database. It could be non-relational like the all-fancy NoSQL ones. Since we're working in python ZODB looks like the best option, but MySQL will do the job. However if we're no experienced with DB, it will take a while to install, configure and managing the interaction between our code and the DB. And database performance....is an entire field.

If we think a bit out-of-the-box...why do we need such a big memory?...The answer is cause since data is not sorted by relationships in line million we may find a record of a relationship we did not see since line 14. If lines would be grouped by relationship, it would not be necessary to keep in memory more than one output line, the one aggregating the contacts of the relationship we are currently reading on input.

[1] http://stackoverflow.com/questions/449560/how-do-i-determine-the-size-of-an-object-in-python, http://stackoverflow.com/questions/1331471/in-memory-size-of-a-python-structure

So, how long it takes to sort those files? And even more important, is there a way to use several cores to accelerate the sorting process?

Answer to the second option is yes, and it was right there all the time. As I get more and more experience, I appreciate old-style unix tools more and more. In this case is the *sort* command. By default you can just send to sort via an unix pipe and it will output the data sorted. A very interesting option is $-S$ which allows you to set how much memory process may use, if you give too much data, it will just dump to file automatically sorted segments. But the amazing thing is in the last version it supports $--parallel N$ where N is the number of cores to use (We had to download source and compile, as we could not find it in our distribution standard repository, but it was just the typical *./configure ; make ; make install*).

So the only remaining thing is, before sending to sort was to add a prefix to each line such that lines like "emailA;emailB;5" and "emailB;emailA;3" end up together after the sort process. We achieved this by adding the prefix $sort(emiter, reciever)$ using a simple python script.

Appendix B

Crawling spatial databases with adaptive resolution

The goal of this text is to explain the design decisions and procedures taken to gather the POI database used in chapter 5. The aim of the project was to acquire and leverage existing publicly available geolocated data in order to better understand the structure and function of cities. The project was developed in collaboration with Prof. Cesar Hidalgo, from the Macro Connections group at MIT Media Lab.

B.1 Foursquare vs Google Places

By the time (summer 2012), there were two main databases ubiquitous and comprehensive enough to satisfy our purposes:

- Foursquare (4sq): is geolocated social media launched in 2009. In its form by the time of our study, provided users with the ability to proactively share with their acquaintances their visits to places, named *venues* in the 4sq jargon. Venues contained structured data, such as type of venue (e.g. restaurant, airport), pictures, latitude and longitude. The creation of venues was entirely crowdsourced to users. The app went viral by leveraging *gamification*: a user could become *major* of a venue, if she was the user with the most visits.
- Google Places: is the geolocated data layer one can find on top on Google Maps. It contains the type of place, latitude and longitude among others. The creation of places is only half crowdsourced, since the Google Places team reviews requests from users to create places, and most places have been created by incorporating existing databases.

Our first attempt was to gather Foursquare venues, capturing public tweets mentioning them using Twitter's Streaming API. In one month listen-

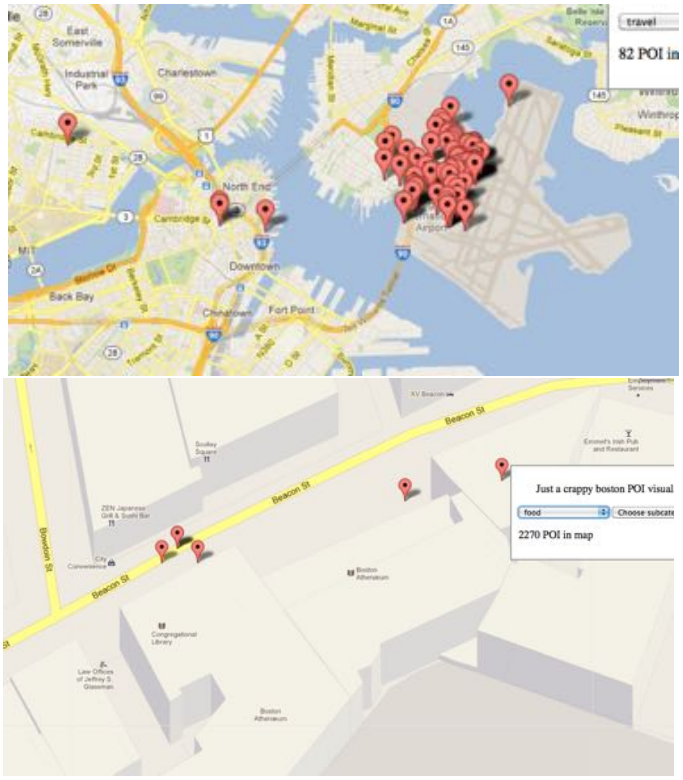


Figure B.1: Snapshot of the web set up to validate the POIs from Foursquare. Many venues are not accurately geolocated, and popular places appear as more than one venue.

ing to these feed, we captured 9000 different POIs for the Boston metropolitan area. By the third week, about 90% of the new venues captured had been created recently, implying that we had already gathered a large subset of all existing venues at the time.

However, the gathered dataset presented some problems, mostly related to the crowdsourcing of venues. First, geolocation did not appear to be highly accurate, as shown in Figure B.1. Also, the gamification of the app produced a large amount of duplicates (we were able to detect over 10 different "Fenways Park") and meaningless venues (e.g "Gate K56" in Logan Airport, categorized as an airport). Additionally, venues were more representative of leisure activities compare to other uses of the city (most bars in the city are present in the dataset, while less than 5% registered doctors appear).

These weaknesses were not present in the Google Places API, due to the review and curation processes in place creation. However, the acquisition of data from the API presented some technical challeges

B.2 Adaptive querying the Google Places API

The Google Places API was originally created to allow mobile app developers to augment their apps using neighboring geolocated data. Thus, the basic query consist of defining a circle in the map (using center and radius), and the kind of place or some search query. If within the defined circle there are more than 20 places, only 20 will be returned.

This presented a significant challenge: it did not make sense to choose a certain granularity in the crawling, because spatial density was precisely among the goals of the study. Also, systematically crawling with small radius would take very long, specially considering that because of performance reason the API is rate limited to 100K requests per day.

The solution was to develop a smart adaptive crawler, than started requesting the whole city, using the smallest circle containing the defined bounding box. If the response contained 20 places, the bounding box was divided into two by the longest edge, and queries will be placed in each half using again the smallest circle containing the requested region. If there are less than 20 results, we store results (removing duplicates due to circle overlapping) and consider the area crawled.

Figure B.2 shows the progress of the crawler requesting Madrid. This strategy turned out to be very productive. We manage to crawled the city of Madrid in only 35K requests, acquiring 78K POIs. In some areas of downtown, radius as little as 7 meters were needed. This means that for a non-adaptive crawling, over 8 million requests to the API would have been needed.

Appendix B. Crawling spatial databases with adaptive resolution

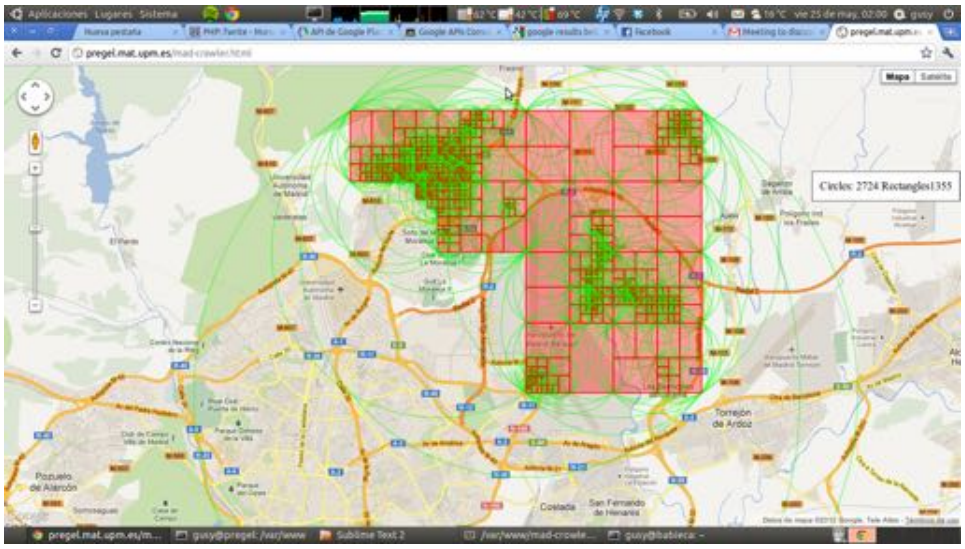


Figure B.2: Adaptive crawling process taking place on the north east corner of Madrid. Green circles represents requests, red areas are already fully crawled. The crawler queries with more resolution areas with higher density, as opposed to others with low density, for example the runways of the airport.