



UNIVERSITÉ CATHOLIQUE DE LOUVAIN
ÉCOLE POLYTECHNIQUE DE LOUVAIN
INSTITUT ICTEAM
PÔLE D'INGÉNIERIE MATHÉMATIQUE

EXTRACTION OF INFORMATION

FROM LARGE NETWORKS

Gautier KRINGS

Thesis submitted in partial fulfillment
of the requirements for the degree of
Docteur en Sciences de l'ingénieur.

ADVISOR : Vincent BLONDEL

JURY :

Dr. Alain BARRAT, CNRS, France
Prof. Vincent BLONDEL, Université catholique de Louvain
Prof. Jean-Charles DELVENNE, Université catholique de Louvain
Dr. Jari SARAMÄKI, Aalto University, Finland
Dr. Zbigniew SMOREDA, Orange Labs, France
Prof. Paul VAN DOOREN, Chairman, Université catholique de Louvain

Acknowledgements

As Bilbo Baggins, a main character of J.R.R. Tolkien’s *The Lord of the Rings* says: “It’s a dangerous business [...] going out your door. You step onto the road, and if you don’t keep your feet, there’s no knowing where you might be swept off to”. I have started my thesis five years ago without a well-defined topic, and I would have hardly imagined at its start that it would have taken its actual aspect. Fortunately, I had the opportunity to collaborate during these five years with a number of great people who helped me to stay on track, who I would like to thank here.

First of all, I would like to thank my advisor Vincent Blondel who carefully guided me during these five years. I particularly appreciated the many opportunities he gave me to develop new collaborations and to participate in a number of exciting projects. I am grateful for the large autonomy I received to choose my topics and to decide the directions to take for my thesis, as well as for the valuable advice he gave me throughout these five years.

This thesis has been reviewed by my thesis committee, namely Alain Barrat, Jean-Charles Delvenne, Jari Saramäki, Zbigniew Smoreda and Paul Van Dooren, which I would like to thank for the numerous constructive comments they provided me. They largely improved the quality of my text, and gave me helpful advice during my research. I am particularly grateful to Jari Saramäki for his welcome in Aalto University in Helsinki and for the many hours of intense discussions we had about my research topics. I am thankful to Paul Van Dooren for his many advice and interest in my research, as well as for agreeing of chairing my jury. I thank Jean-Charles Delvenne for the diverse collaborations we have had as well as the many friendly interactions we had when both of us were staying in Louvain-la-Neuve. Finally, I also thank Zbigniew Smoreda who has given a valuable help me by providing most of the data used in this thesis.

During these five years, I enjoyed the great atmosphere in the Euler building, which made it a pleasure to work with my colleagues. In particular, I thank the members of the Large Graphs and Networks group,

who have provided a constant feedback and ideas on my work. It is impossible to dissociate my stay in Euler from Guillaume Leclercq, with whom I have shared my office in a great atmosphere of friendship. I also had the occasion to spend some time during a very instructive stay at the Senseable City Laboratory at MIT, for which I'd like to thank the members of that time, in particular Francesco Calabrese.

It would have been much more difficult to achieve this thesis without the support of the administrative assistants in the Euler building, and I am deeply thankful for their continuous help and their numerous initiatives to keep the positive atmosphere in the building. On a different note, I also thank Etienne Huens for his very precious help in the preparation of the mobile phone dataset.

This thesis was prepared at the Pole of Mathematical Engineering of the Institute for Information and Communication Technologies, Electronics and Applied Mathematics of the Université catholique de Louvain. During the preparation of my PhD, I was supported financially by research programs ARC Large Graphs and Networks and Action de recherche concertée of the Communauté Française de Belgique and IAP DYSCO (Dynamical Systems, Control and Optimization), a Belgian research network funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office.

Last but not least, I would like to deeply thank my family, and in particular Caro, for their patient support and attention.

Contents

List of Notations	1
1 General Introduction	3
1.1 Context of the thesis	5
1.2 Presentation of the Chapters	8
1.3 Publications	11
2 Materials and methods	17
2.1 Standard notations for networks	17
2.2 Information Theory	23
2.3 Community detection	26
2.4 Dynamical networks	29
2.5 Real networks	35
3 Mobile phone datasets	37
3.1 Introduction	37
3.2 Social networks	40
3.3 Adding space – Geographical networks	51
3.4 Adding time – Dynamical networks	57
3.5 Adding space and time – Mobility	60
3.6 Dynamics on mobile phone networks	69
3.7 Privacy issues	76
3.8 Conclusion and research questions	79
4 Geographic networks	83
4.1 Introduction	83
4.2 Presentation of the data	85
4.3 Gravity law	87

4.4	Scaling properties	90
4.5	External factors	97
4.6	Conclusions	99
5	Dynamical networks	103
5.1	Introduction	103
5.2	Time scales	106
5.3	A model of growing networks	121
5.4	Information theory	133
5.5	Further work	141
5.6	Conclusions	142
6	Community detection	147
6.1	Introduction	147
6.2	Multi-level community detection	150
6.3	Fast multilevel community detection	156
6.4	Results on a network of criminal offenders	168
6.5	Conclusions	187
7	General Conclusions	191
	Bibliography	194
A	Identification	215
A.1	Identification based on $m(x)$	215

List of Notations

N	Set of nodes of a network $N = \{1, \dots, n\}$
L	Set of links of a network $L \subseteq N \times N$
n	number of nodes in a network
m	number of links in a network
$G = (N, L)$	Network composed of a set of nodes N and a set of links L
t	Time (in seconds)
x	Number of events
(i, j)	Link from node i to node j
(i, j, t)	Event of a link (i, j) at time t
$\mathcal{E}(t), \mathcal{E}(x)$	Set of events observed after t time steps or x events
$L(t), L(x)$	Set of links observed after t time steps or x events
$n(t), n(x)$	Number of nodes in G after t time steps or x events
$m(t), m(x)$	Number of links in G after t time steps or x events
w_{ij}	Weight of link (i, j)
d, d_i	Degree, degree of node i
$d_{i,2}$	Distance 2 degree of node i (in a bipartite network)
C	Global clustering coefficient of the network
C_i	Local clustering coefficient of node i
O_{ij}	Overlap of link (i, j)
l_g	Number of links inside the subgraph g
$p(\cdot)$	Probability distribution
$Exp[X]$	Expectation of the variable X
$H(X)$	Entropy of the random variable X
$H_r(X)$	Entropy rate of the random variable X

S_G	Sequence of events generated by the dynamical network G
S_i	Sequence of events generated by the node i
$\mathcal{P}(x), \mathcal{Q}(x)$	Probability distribution parametrized by x
c_a	Number of customers in commune a
r_{ab}	Distance between city a and city b
	Excepted in Chapter 5, where r is a growth rate
N_i	Number of nodes in group i
$p^{conn}(r)$	connection probability of two nodes separated by a distance r
$\sigma(G_1, G_2)$	Jaccard index of similarity between G_1 and G_2
P	Partition of N
π	Community, $\pi \in P$
l_π	Number of internal links of community π
d_π	Total degree of community π
n_π	Number of nodes in community π
n'_π	Number of nodes of second class in community π (in bipartite networks)
π_i	Community of node i
Q	Modularity
Q_γ	Scaled modularity
H_γ	Constant Potts Model
$NVI(X, Y)$	Normalized variation of information of the random variables X and Y

Chapter 1

General Introduction

The study of complex networks has attracted a lot of attention in the last decade, given the increase of available computational power and storage capacities of computers. This development has enabled to collect massive datasets representing networked systems such as the internet, the world wide web, online social networks or the world airline network. Those systems can be represented by a network, *i.e.* a collection of nodes connected with each other by links. The nature of the nodes on the exact application, they may be routers in the case of the internet, web pages in the web, persons in a social network or airports in the airline network. The definition of links is then identically dependent on the application: they may be physical cables connecting two routers, hyperlinks pointing from one web page to another, a friendship status in a social network or a flight between two airports.

One particularity of several types of complex networks is that their structure results from a decentralized construction scheme, links are created between nodes based on local decisions without being planned by a central authority. This is why the study of the structural properties of complex networks has gathered a significant attention in the recent years. In particular, some non-trivial properties have been observed on a large number of datasets, representing systems that apparently have nothing in common. For example, many complex networks show heavy-tailed degree distributions, hence nodes have very different levels of activity. Complex networks are usually sparse, however they show a high local density of links, and often are structurally organized in communities.

Given the large size of the studied networks, the challenge of information retrieval in complex networks is to develop algorithms that extract significant structural information from the network in a reasonable time. A well-known example is Google's PageRank, which measures the importance of the nodes in a network. The success of the PageRank is not only due to the fact that it provides a meaningful and non-trivial measure, but also because there exists efficient algorithms to approximate the PageRank even on very large networks.

This thesis fits in this context of *information retrieval* in complex networks. The contributions in this work are centred around three topics: *geographical networks*, *dynamical networks* and *community detection*.

Geographical networks represent systems where nodes have a geographical position, such as the road network between different cities. The study of the structure of such networks can then be enriched with this geographical information and allows to observe the influence of distances on the network structure.

In many cases, networks are not static entities, but show instead dynamical changes in their structure. New nodes join the network, while others disappear, and connections are added and removed over time. The study of *dynamical networks* is a recent topic in complex network theory, most contributions so far having focused on networks as static objects.

Our last contribution concerns the *detection of communities* in networks. A community is a generic term used to describe a group of nodes that is densely interconnected. The detection of such groups in a network is a difficult problem that is still subject to new developments, with many practical applications.

The remainder of this introduction is structured as follows: in Section 1, we introduce the general context in which the different questions of this thesis have been addressed. In Section 2, we present each chapter separately, and in Section 3, we give the list of publications written or in writing process related to this thesis.

1.1 Context of the thesis

This thesis does not address one central topic, it should rather be seen as a toolbox containing a collection of methods for the extraction of information out of large networks. It is centred around the three principal topics cited above; geographic networks, dynamical networks and community detection. In the following, we present more in detail the context in which this thesis has been realised, and the questions that have directed our work.

Analysis of mobile phone networks

A large part of the work done in this thesis concerns the analysis of networks where nodes are mobile phone users and links are drawn between nodes that call each other. The data necessary for constructing such mobile phone networks are the call data records (CDRs), which are collected by mobile phone operators for billing purposes, and to which we had access for research through several research contracts with one operator. Mobile phone networks represent the interactions between the customers of the operator, and their study provides valuable information: a partition in communities unveils groups of friends, families or colleagues [148, 31, 3], with applications to marketing, the dynamics of the activity of users helps to predict if customers are about to churn [51, 164], and the study of users mobility can help to detect traffic jams or emergency situations [178, 10]. However, the study of such networks presents several challenges: they have a large size (up to millions of customers), their structure changes over time as relations raise and wane, and in some cases the geographical position of nodes is known, which adds a geographical component to the network which needs to be taken into account. Given their sociological and economical interest, the study of mobile phone datasets has attracted a lot of attention in the last decade, with over 100 research papers published on their analysis, and since recently a dedicated conference gathers the members of this emerging field annually ¹.

Networks of cities

In collaboration with the Senseable City Lab from MIT, we were in-

¹NetMob: Conference on the analysis of mobile phone datasets see <http://www.inma.ucl.ac.be/~blondel/netmob/>.

interested in the influence of distance and population on communication between cities. This topic has been studied in different examples, such as road networks [94], passenger transport [196] or phone messages [53]. Here we complemented these results with an analysis of the mobile phone communication between cities using a dataset from one major operator in Belgium, where customers are grouped into cities based on their billing address [108]. Similarly to similar studies of different datasets, we observe that communication between cities can be estimated on average by a gravity law reminiscent of Newton’s law, where population sizes play the role of a mass.

Scaling behaviours in cities

Recent research has shown that different characteristics in cities grow in different ways with population size. While some characteristics are directly proportional to cities’ population sizes, other features – such as productivity or energy consumption – are not linear but exhibit instead superlinear or sublinear dependence on population size [23]. In the continuity of the previous work, we showed that communication inside cities follows also a superlinear exponent [107], which can be explained using a geographic random network model.

Communities of cities

We closed the investigation of the network of cities with an analysis of the geographical dispersal of communities of cities, showing that cultural boundaries and the influence of large cities dominates the composition of communities. Although this topic has now been studied by a number of authors [156, 69, 30], this work was one of the first studies of geographical boundaries of networks of cities.

Time scales in dynamical networks

The second part of the thesis focuses on a different topic, still with applications on the study of mobile phone networks. The initial question originated from observations that the structure of a mobile phone network varies strongly depending on the length of the period during which the data are collected. This feature has important implications which go beyond the sole study of mobile phone networks. Most studies of networks that structurally change over time define the length of their observation window based on technical constraints, such as the

format in which temporal information is available or on hypotheses of characteristic time scales of the underlying data [82, 74]. The temporal and structural inhomogeneities that characterize the dynamics of many complex systems induce a non-trivial evolution of the network structure that has to be taken into account in their study. In particular, networks extracted at different time scales show different structures.

Dynamics of network growth

In the continuation of the preceding topic, we addressed the following question: can one predict the future size of a network based on what has been observed so far? During the growth, new nodes join the network and links are dynamically created over time, does this process follow a well-defined rhythm, and could this be reproduced by a growing network model? Growing networks have already been studied in a number of contexts such as citation networks, biological networks or social network and several models that reproduce the evolution of their structure over time have been proposed, but the aspect of the evolution of the size of the network over time, removing structural considerations, has only been loosely addressed. Moreover, if a network is constantly growing, as is observed in social networks, does it make sense to study a network constructed on a small time window, or does one miss significant information given by the lack of data? We addressed these different questions, with applications on the mobile phone data.

Contract with the Belgian Federal Police

The last part of this thesis addresses a different topic, which is the result of a research contract with the Belgian Federal Police. The aim of this project is the development a multi-level community detection algorithm, to detect organized groups in their database of recorded criminal facts and offenders. Several authors proposed to rely on scaled quality functions, that use a scaling parameter to zoom in and out communities [161, 6, 116, 183], however the optimization algorithms used to optimize these quality functions haven't been adapted to take the multi-level component into account in the optimization scheme. We proposed to accelerate the search procedure with a multi-level community detection method that is tested on the network of the Police data, with success.

Collaborations

In conclusion, we would like to acknowledge that this thesis has mostly been conducted in the environment of the Large Graphs and Networks group, at Université catholique de Louvain, and partly in the Senseable City Laboratory, at the Massachusetts Institute of Technology, during a 7-months stay as visiting researcher. Many ideas and developments have been the result of discussions, exchanges of ideas and collaborations with members of these groups, and in particular, Vincent Blondel, Paul Van Dooren, Jean-Charles Delvenne, Cristobald de Kerchove, Vincent Traag, Arnaud Browet, Renaud Lambiotte and Etienne Huens in UCL and Carlo Ratti, Francesco Calabrese, Jon Reades, Daniele Quercia, Francisca Rojas and Mauro Martino in MIT. Both research groups have enabled to participate and access data from collaborations with private companies and public authorities, such as AT&T, British Telecom, Orange Labs, and the Belgian Federal Police. In addition, a large part of Chapter 5 has been the result of a 3-year long collaboration with Jari Saramäki and Márton Karsai from Aalto University.

1.2 Presentation of the Chapters

Apart from the general introduction and conclusions, this work is organized in one chapter of background material (Chapter 2), one survey chapter and three chapters of scientific contributions. A short description of the four last chapters is given hereunder.

Chapter 3: Analysis of mobile phone networks

Mobile phone datasets have attracted a significant attention in the last years. These datasets consist of Call Data Records (CDRs), *i.e.* a list of all calls passed between the customers of a mobile operator during a given period. Given their large size (typically over a million nodes) and the fact that these are real-world datasets, these networks have intensively been used to test the efficiency of new algorithms or the correctness of models. However, in these applications mobile phone datasets were mostly considered as one example of networks among others. This perception has evolved, and the study of mobile phone networks has become now a stand-alone topic with a number of dedicated studies of the structure of mobile phone networks. Moreover, CDR data have the particularity

that they also contain the base stations at which calls have been made and received, which enables to localise mobile phone users in space and time. This has led to a number of studies of human mobility, which have provided significant advances in the understanding of personal mobility and the mobility of entire populations. Having the feeling that this new community is emerging, we give a survey of the contributions made so far in the analysis of mobile phone networks. We have structured the chapter in six sections which represent each one type of information that is used: the **structure** of the social network constructed with CDRs, the **geographical dispersion** of such networks, their **temporal dynamics**, the **mobility** of users, the **dynamics of diffusion processes** on such networks and the **privacy issues** raised by the collection and study of such datasets. This survey is not intended to be an exhaustive list of contributions in this field, but rather to structure the principal advancements made so far.

Chapter 4: Geographic distribution of networks

This chapter is a collection of the results of two journal articles [108, 32] and one conference proceedings [107], in which we study the network of cities in Belgium, constructed using a mobile phone dataset from one large operator. We therefore aggregate the network of customers to a network of communes, where each commune regroups all calls passed by the customers living in this commune. We show that the intensity of communication between two cities is on average well-estimated by a gravity law, similarly to what has already been observed for different types of networks, such as car traffic. Secondly, we focus on the intensity of communication inside cities, *i.e.* the number of calls made between users of the same city. This quantity scales superlinearly with the number of users living in the city, however the observed exponent does not have a trivial explanation. In a previous study of the same dataset [115], the probability of two users to connect was found to decrease with the square of the distance that separates them. Using this probability of connection as basis of a random geographic network, we observe that it is not possible to reproduce a case where cities show similar exponents for their internal communication intensity. We explain this contradiction by showing that at short distances, the probability of connection changes depending on the city size where customers live, which tends to indicate that geographical distance might not be an appropriate measure

of distance when evaluating the probability of two persons to connect, when the population density is not uniform. This finding might seem in opposition with the gravity law found for cities, however it has mostly an effect at small distances, and hence its effect on the gravity law is weak.

The gravity law for cities represents an average prediction of the intensity of communication between cities, therefore statistical variations are observed in the real data. A part of these variations may be explained by external factors that are not taken into account in the gravity model, such as cultural barriers or different languages. We show an illustration of this on the Belgian data where the communication between cities from different linguistic origins have systematically a lower intensity as what is predicted by the gravity law.

Chapter 5: Time scales of dynamical networks

Several types of networks don't have a static structure, nodes and links join or leave the network over time, and the weights of links might be dynamic as well. The study of dynamical networks has mostly been approached by slicing the observation interval into a sequence of time windows, and by studying the obtained sequence of networks instead. The first part of this chapter focuses on how the length of the time window influences the structure of the constructed network. We study the structure of networks generated using a growing time window, and show how different characteristic parameters are influenced by its length. Using the same mobile phone dataset as in the previous chapter, which contains data over a time range of 6 months, we show that even though the network's global density increases continuously when the window length is increased, the local density reaches a maximum at an intermediate window length, suggesting that communities, or at least dense groups of nodes are observed at the start of the aggregation, while the remainder of links being added consist of more "random" links. Surprisingly, the degree and strength distributions of networks that have been constructed on windows of different sizes can be rescaled to collapse onto a single curve, as long as the length of the window is larger than two weeks. For smaller windows, this rescaling property is not observed. Hence, distributions of sufficiently large windows can be predicted from one single generative distribution. The non-scaling of small windows appears rather as a pixelation effect when too few data is used to construct

the network: the generative distribution is badly estimated due to the lack of data. In the second section, we consider again the growing mobile phone network, but having removed the time information, only keeping the order in which calls are added. We show how the network follows two well-separated growth phases: a first discovery phase followed by a steady growth phase that can be well-approximated by a power-law. We derive a network growth model that explains both of these phases. Finally, we show some exploratory work on how information-theoretical tools can be used to quantify the complexity of a growing network.

Chapter 6: Multiscale community detection

This chapter is the result of a research contract with the Belgian Federal Police. The Police stores in a large database the records on all criminal facts committed in Belgium and on the individuals known to be involved in these facts. When each individual is connected to the facts in which he/she is involved, one has a network of criminal facts and offenders that represents the collaborations between different offenders. The aim of the project is hence to detect cohesive groups of offenders, that can be assimilated to organised groups. We implement therefore a multi-level community detection method, based on the Louvain method and optimizing the Constant Potts Model, introduced in [183]. We show that the multi-level method accelerates the detection of communities at multiple levels, in particular when networks are truly organized in communities at multiple levels. The result of this algorithm is a landscape of communities at different scales, with measures on the stability of the communities at each level. This method is applied to the bipartite network of criminal facts and offenders described above. This network show typical properties of complex networks, such as broad degree distributions and high clustering coefficient. The community detection method reveals a well-defined multi-level community structure, where communities are geographically significant, and often specialized in one particular type of fact, which is comforting us of the significance of the results.

1.3 Publications

In this section, we quickly describe the list of publications published or to be submitted soon in peer-reviewed journals or conferences with published proceedings.

Publications in this thesis

The results of Chapter 4 have been published in a series of papers and conference proceedings. The results on the gravity law have been presented at NetSci'09, and were published in *Journal of Statistical Mechanics: theory and experiment* [108]. An initial analysis of the urban network has also been published in the proceedings of the *International Conference on Computational Science and Engineering* in Vancouver, 2009 [107]. Finally, the analysis of the dispersal of communities in the network of cities has been published in *Brussels Studies* [32].

The analysis of the influence of time scales on the structure of networks has been partially presented at Heeze during the 29th Benelux Meeting and at NetMob 2010. A manuscript is in preparation, and will be submitted in January 2012 to *Physical Review E*. The remaining of Chapter 5 consists of unpublished results, that will be submitted as a book chapter on dynamical networks.

The results in Chapter 6 led to a private report at the end of the contract with the Federal Police, the results of this report have partially been presented at the NetSci'11 conference in Budapest. The multi-level algorithm and the analysis of the network of criminal offenders is contained in a manuscript under preparation that is intended to be submitted in January 2012.

Other publications not related to the thesis

A first paper was published in *Physical Review E* in 2009 [54], and is entitled *The role of second trials in cascades of information over networks*. In contrast to most of the work done in this thesis, this work did not consist in the analysis of a dataset, but of the presentation of a model of information cascades over networks, where the probability of transmitting the information between two adjacent nodes depends on the number of times that the targeted node has already been attacked. In this model, the different role played by links inside communities and weak ties is shown.

Two other publications are still in preparation, and are expected to be submitted in the coming months. The first paper is a theoretical analysis of the stability of epidemic diffusion models. We study a classical epidemic model where each city is considered as a fully-mixed system

and where agents travel from city to city, travel being hence responsible for the sickness to propagate between cities. Many epidemic diffusion models simulate the epidemics with such a model, and estimate the parameters such as travel rates and transmission probabilities. Using a branching process model, we study how the size of the epidemic varies if the parameters are not correctly estimated. In the second paper, we develop new network-based tools for the analysis of economic networks. A first series of tools use random walks, commute time and mixing rate to quantify the integration of economies into an economical union. A second measure decomposes the trade network into a series of flows of money, which allows to detect deficits accumulated by countries with debts in profit of countries that show a positive economical balance. Applied on the European trade network, from 1993 to 2007, the measures allow for example to observe the decrease of France's role in the European economies.

List of coauthors

I resume here the collaborators involved in the articles mentioned above:

- Vincent Blondel,
Professor in Applied Mathematics in the Institute for Information and Communication Technologies, Electronics and Applied Mathematics at UCL,
- Francesco Calabrese,
Advisory Research Staff Member at the IBM Research & Development - Ireland,
- Carlo Ratti,
Associate Professor of the Practice in the Department of Urban Studies and Planning at MIT and Director of the SENSEable City Laboratory at MIT,
- Isabelle Thomas,
FNRS Research Director in the Center for Operations Research and Econometrics at UCL,
- Márton Karsai,
Post-doctoral Research Fellow in the Department of Biomedical Engineering and Computational Science at Aalto University,

- Sebastian Bernhardsson,
PhD in Physics from the University of Umeå,
- Jari Saramäki,
Senior Researcher and Group Leader of the Complex Networks Group in the Department of Biomedical Engineering and Computational Science at Aalto University,
- Paul Van Dooren,
Professor in Applied Mathematics in the Institute for Information and Communication Technologies, Electronics and Applied Mathematics at UCL,
- Cristobald de Kerchove,
PhD in Applied Mathematics from UCL,
- Renaud Lambiotte,
Associate Professor in Mathematics at the University of Namur,
- Jean-Charles Delvenne,
Assistant Professor in Applied Mathematics in the Institute for Information and Communication Technologies, Electronics and Applied Mathematics at UCL,
- Jean-François Carpentier,
PhD Student in Management and Economics in the Department of Economics, Collège Dupriez at UCL,
- Sophie Hautephenne,
Research Fellow in the Department of Mathematics and Statistics at the University of Melbourne

List of publications

[108] G. Krings, F. Calabrese, C. Ratti and V.D. Blondel, *Urban gravity: : a model for inter-city telecommunication flows*, in Journal of Statistical Mechanics: Theory and Experiment, 2009.

[107] G. Krings, F. Calabrese, C. Ratti and V.D. Blondel, *Scaling behaviors in the communication network between cities*, in Proceedings of

the International Conference on Computational Science and Engineering, 2009.

[32] V.D. Blondel, G. Krings and I. Thomas, *Regions and borders of mobile telephony in Belgium and in the Brussels metropolitan zone*, in Brussels Studies, 2010

[109] G. Krings, M. Karsai, S. Bernhardsson, V.D. Blondel and J. Saramäki, *Effects of time window size and placement on the structure of aggregated networks*, submitted to EPJ Data Science, 2012

[54] C. de Kerchove, G. Krings, R. Lambiotte, P. Van Dooren and V.D. Blondel, *Role of second trials in cascades of information over networks*, in Physical Review E, 2009.

G. Krings, J.-F. Carpentier and J.-C. Delvenne, *Trade integration and deficits in the European union: a network perspective*, to be submitted to PLoS One, 2012

S. Hautephenne, G. Krings, J.-C. Delvenne and V.D. Blondel, *Sensitivity analysis in epidemic models*, to be submitted to PLoS One, 2012

Chapter 2

Materials and methods

Networks are used to represent a wide class of different systems, such as social networks, communication networks, food webs, protein interaction networks. . . In this chapter, we present the concepts and formalisms that are used to present those systems in a unified way. We also present recent results in the fields related to this thesis.

The first section introduces the common notations used to describe networks and common measures used in their analysis. In Section 2, we give a short introduction to information theory concepts, and their application to networks, necessary for Chapters 5 and 6. Next, we present the problem of community detection, which is the focus of Chapter 6. Section 4, finally, is dedicated to dynamical networks, studied in Chapter 5.

2.1 Standard notations for networks

Given their large number of practical applications, networks have been studied by scientists coming from several disciplines. There exist many textbooks on network analysis and different formalisms have been introduced in this scope. In this work, we use a general formalism, close to most available literature. For interested readers, [138, 141, 5, 59] contain extensive introductions to the field.

Network classes, neighbourhood and connectivity A network $G = (N, L)$ consists of a collection of *nodes* N and a collection of *links*

$L \subseteq N \times N$. Usually, $N = \{1, \dots, n\}$, and $|L| = m$.

If the direction of the links is important (i.e. if $(i, j) \neq (j, i)$), the network is a *directed network*, if not ($(i, j) \equiv (j, i)$), the network is an *undirected network*. If links are of varying importance, this can be translated into weights on links, we then speak of a *weighted network*. Finally, a link (i, i) is called a *self-loop*. Unless specified, we will consider G to be a *simple graph*, i.e. it does not contain self-loops or multiple links.

In an undirected network, the set of all nodes *adjacent* to node i are called the *neighbourhood* of i . The number of edges *incident* to i is the *degree* d_i of i .

For directed networks, for a given node i , we can then distinguish *inlinks* and *outlinks* of i . Then, the *indegree* d_i^- and the *outdegree* d_i^+ of i are respectively its number of inlinks and outlinks.

In the case of weighted networks, the natural extension of degree is given by the *strength* of a node i , the sum of the weights of the links incident to i . The directed counterpart of strength is obtained similarly.

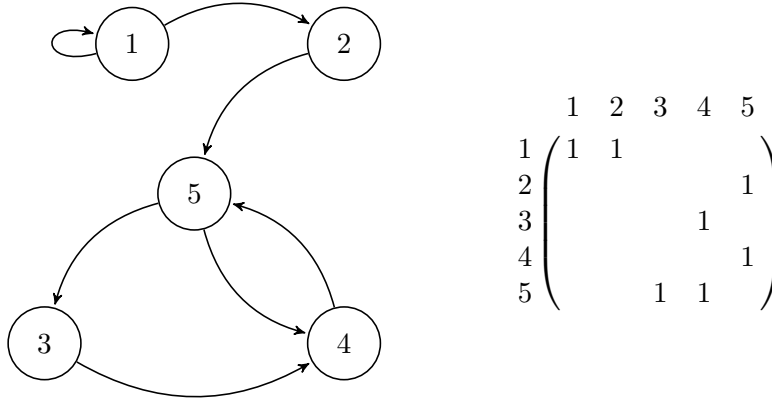


Figure 2.1: Example of a directed network of 5 nodes and 7 links and its adjacency matrix.

A *path* between two nodes i and j is a sequence of nodes $i = n_0, n_1, \dots, n_{k-1}, n_k = j$ such that each node appears exactly once in the sequence, and such that the links $(n_{l-1}, n_l), \forall l \in \{1, \dots, k\}$ are in L .

In an undirected network, a *connected component* C is a subset of N such that there exists a path between each pair of nodes $i \in C, j \in C$, and there does not exist a path between each pair of nodes $i \in C, j \in N \setminus C$. For directed networks, there exists a distinction between *strongly connected components* and *weakly connected components* which will not be addressed in this work.

Matrix representation of a network A classical representation of a network is its (square) adjacency matrix $[A_{ij}]_{i,j=1}^n$, where the entry $A_{ij} = 1$ if the link (i, j) is in the network, and 0 otherwise. The indegree of a node corresponds then to the sum of the entries of a column $d_i^- = \sum_j A_{ji}$ while the outdegree is given by the sum of the corresponding line of A : $d_i^+ = \sum_j A_{ij}$. For an undirected network, this representation leads to a symmetric matrix, hence $\sum_j A_{ij} = \sum_j A_{ji} = d_i$.

The use of the adjacency matrix allows to reformulate several network theory problems in a framework where results from linear algebra and matrix theory can be applied. In this work, we have not used these tools, the adjacency matrix has rather been used as a compact representation of the network.

Random networks, degree distribution, clustering coefficient

A *random network* is a network whose number of nodes, links and structure are defined by a set of rules which include a non-deterministic component. Equivalently, an instance of a particular random network model is a randomly chosen network in the *ensemble* of all networks having the properties defined by the model. A well-known example is the Erdős-Rényi random network $G_{n,p}$ [68], a network of n nodes where each possible pair of nodes is connected with independent probability p . Hence an instance of $G_{n,p}$ is chosen randomly in the ensemble of all possible networks of n nodes, where a network with m links appears with probability $p^m(1-p)^{\frac{n(n-1)}{2}-m}$.

An important feature of networks is the *degree distribution* $p: \mathbb{N}_0 \rightarrow [0, 1]$, which gives for a given degree d the fraction of nodes of the network that have a degree d . By construction,

$$\sum_{d>0} p(d) = 1 \quad \sum_{d>0} dp(d) = \langle d \rangle ,$$

where $\langle d \rangle$ is the average degree of the nodes of the network. Equivalently, $p(d)$ is the probability that a node chosen at random in N has degree d . A plot of the histogram of the degree distribution is then a fast visual indicator of the possible heterogeneities between the nodes of the network. For example, the degree distribution of an Erdős-Rényi network is binomial, and mostly approximated by a Poisson distribution for large networks. Several real-world networks, instead, have very different degree distributions compared to ER networks. Unlike Poisson distributions, which have an exponentially decaying tail, real-world networks show a degree distribution that ranges over several orders of magnitude, some of their nodes having a degree that is orders larger than the network average.

In some cases, the degree distribution, or at least its tail, can be approximated by a power-law distribution

$$P(d) \propto d^{-\alpha},$$

where α is generally a number between 2 and 3.

In many random networks, such as the Erdős-Rényi network, the degree of two neighbouring nodes is uncorrelated, which means that the knowledge of the degree of a node gives no information about the degree of his neighbours. However, in real networks, it might happen that nodes of similar degree tend to be connected. Such a situation is called *degree assortativity* or homophily. The simplest way to measure if such correlations exist, is to plot the average neighbour degree as a function of the degree of the node. If the obtained plot shows an increasing trend, the network is said assortative; nodes of high degree tend to be connected to other nodes of high degree. If the curve shows a decreasing trend instead, the network is said to be disassortative; nodes of high degree are connected to nodes of low degree. The concept of assortativity can be extended to other features of nodes than degree, in particular if each node has some external features (not related to the network), such as their age if nodes are persons.

The *clustering coefficient* of a network measures the transitivity of the network, i.e. the density of triangles present in the network. It is

measured as

$$C(G) = \frac{6 \times \text{number of triangles in } G}{\text{number of paths of length 2 in } G},$$

the factor 6 appearing because each triangle counts for 6 paths of length 2 (3 possible pairs of nodes, in both directions). $C(G)$ is contained within the interval $[0, 1]$, and gives the average probability that two randomly chosen nodes such that they are neighbours of the same node are neighbours themselves. An alternative definition is provided by the average local clustering coefficient C_i :

$$C_i = \frac{2 \times \text{number of triangles connected to } i}{d_i(d_i - 1)},$$

for nodes of degree 0 or 1, $C_i = 0$ by convention. The average clustering coefficient is then taken as the average over all nodes: $\langle C \rangle = \frac{1}{n} \sum_i C_i$.

Bipartite networks and their projections A particular class of networks are bipartite networks, sometimes also called two-mode networks. Bipartite networks are composed of two classes of nodes V and V' , such that $(i, j) \in L \Rightarrow i \in V, j \in V'$ or $i \in V', j \in V$, *i.e.* nodes of the same class may not be adjacent. A different way of stating this definition is that bipartite networks cannot contain cycles of uneven length. Similarly to classical networks, we write $n = |V|$, $n' = |V'|$, and $m = |L|$.

Such networks are often encountered in real datasets, we provide here a number of examples of the two node classes that can be encountered: scientific collaborations (authors and papers), the actors network (movies and actors), biological networks (reactions and metabolites), file sharing (users and files). More examples of real-world datasets can be found at Section 5 of [119]. Several notions and measures of standard network theory can be easily extended to bipartite networks. The maximum number of links in this network is given by nn' , hence the density of a bipartite network is given by $\frac{m}{nn'}$. The degree distribution in a bipartite network is separated in two distributions, one for each class. Similarly, the *distance 2 degree* counts the number of nodes at distance 2 to which a node is connected. In a bipartite network, this measure has some interest, since it measures for a node the number of nodes of the same

class to which it is indirectly connected.

Other measures are harder to transpose onto bipartite networks. The global clustering coefficient measures in a unipartite network the density of triangles in the network, while the local clustering coefficient measures the density of triangles around a node. In bipartite networks, triangles don't exist, since these are cycles of length 3, several authors proposed to extend the notion of clustering to bipartite networks. A clustering coefficient suggested in [35] and developed in [121, 119] calculates the topological overlap between the neighbourhoods of two nodes of the same class:

$$c(i, j) = \frac{|N(i) \cap N(j)|}{|N(i) \cup N(j)|}, \quad (2.1)$$

where $N(\cdot)$ is the neighbourhood of a node. This measure, introduced as a clustering coefficient, is rather an overlap measure: it measures how many neighbours two nodes have in common. In [119] the authors acknowledge this drawback and propose to average this quantity over all distance 2 neighbours to obtain a measure on one single node:

$$c(i) = \frac{\sum_{j \in N(N(i))} c(i, j)}{|N(N(i))|}. \quad (2.2)$$

This expression also contains several drawbacks, in particular, this measure computes the density of links not only in the direct neighbourhood of node under study, but in the network containing their neighbours up to a distance 3. Other measures of clustering have also been introduced in [119] by modifying the expression of 2.1 or in [147] where the number of open and closed paths of length 4 are counted, however those measures have an increasing complexity of interpretation.

Since bipartite networks are not always easily analysable, one classical way of handling such networks is to *project* the network onto one class of nodes, and consider the created unipartite network instead. The projection consists at considering a network with only the nodes of one class, and to draw a link between two nodes if they have a common neighbour in the bipartite network. If two nodes have more than one common neighbour in the bipartite network, the link can possibly be weighted by the number of common neighbours. This procedure is often implicitly done in the construction of a network, *e.g.* the movie actors' network

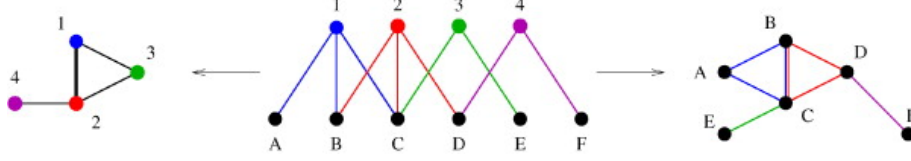


Figure 2.2: A bipartite network and his two projections. In this case, self-loops have been disregarded in the projections, in some applications self-loops are also added. Image reproduced from [119].

(studied in *e.g.* [15]) is constructed such that two actors are connected if they have played in a movie together, the links being weighted by the number of common movies in which they played. However, some authors pointed out that this construction removes a part of information from the network [122], and creates highly clustered networks. Indeed, in a projection of a bipartite network on the class of nodes V , each node in V' of degree d will create a clique (*i.e.* a complete subgraph) of d nodes, and create d^2 links in the projection.

2.2 Information Theory

Let X be a random variable with alphabet $\mathcal{X}$ and probability density function $p(x) = \Pr[X = x]$, $x \in \mathcal{X}$. The *entropy* H of X is a measure of uncertainty on one random outcome of X .

$$H(X) = - \sum_{x \in \mathcal{X}} p(x) \log_2 p(x). \quad (2.3)$$

For example, if X is the outcome of a flip of a biased coin giving heads in 90% of the flips and tails in the other 10%, the uncertainty of the outcome of X is low, since a prediction of a head will be correct in 90% of the cases. The concept of entropy has also a more practical meaning, it is the theoretical bound on the average number of characters per symbol used to describe a sequence of random outcomes of X . Since we use the logarithm in base 2, the entropy is expressed in bits. In an optimal code, each possible outcome is coded by $-\log_2 p(x)$ bits, such that frequent variables have short codewords and infrequent variables have a longer codeword. The expectation of the length of the codeword

of an outcome of X is hence given by (2.3). Entropy has been firstly introduced by Shannon [170] along with an application to English text, and has further opened a large number of applications, see [50] for a survey of the field.

On networks, the entropy has been principally used as a measure of the heterogeneity of the network. The entropy of a network can be defined by several ways, depending on the probability density function that is taken into account. The simplest measure is the uncertainty on the degree of a randomly chosen node

$$H = - \sum_{d>0} p(d) \log_2 p(d). \quad (2.4)$$

Minimal uncertainty is achieved when all nodes have the same degree, while the maximum is obtained with a uniform degree distribution. This measure of entropy has been set in relation with the resilience of networks to random attacks [186]. The same measure has already been used as classification tool of networks [42]. Other authors [57] have proposed to consider instead the entropy rate of a random walk on the network as measure of robustness, and proposed an evolutionary model for networks, in which additional nodes choose preferentially to connect to nodes such that it increases the robustness of the network. The study of the entropy rate of modified random walks on synthetic and real networks has also been studied, with potential applications in information diffusion on networks [77].

In [174], the authors use the *remaining degree distribution*

$$q(d) = \frac{(d+1)p(d+1)}{\langle d \rangle}, \quad (2.5)$$

is used to measure degree correlations. The remaining degree of a node is number of links leaving the node other than the one we arrived along in a random walk. Its entropy measures the uncertainty on the degree of the node incident to a randomly chosen link. The *information transfer* $I(\{q_d\})$ of a system is defined as

$$I(q) = H(q) - H_c(q|q'), \quad (2.6)$$

where $H_c(q|q')$ is the conditional entropy that involves a set of conditional probabilities $p(d|d')$, which give the probability to observe two nodes with remaining degree d and d' at the ends of a randomly chosen link. The *capacity* of a network is then given as the highest possible information ($\mathcal{C} = \max_{q_d} I(q)$).

Well-organized networks such as lattices have $H(q) = I(q) = 0$, while the less assortative networks are, the higher their information will be.

Information-theoretic measures have also been set in relationship with information diffusion in networks. This has been related to the difficulty of localizing a node in the network [173, 168, 167]. Let $p(i, b)$ be a path from node i to node b in the network, the probability to take that path during a random walk starting from i is

$$p[P(i, b)] = \frac{1}{d_i} \prod_{j \in P(i, b)} \frac{1}{d_j - 1}, \quad (2.7)$$

where d_j are taken with exclusion of i and b . The *search information* quantifies then the total information needed to identify one path going from i to b .

$$S(i, b) = -\log_2 \sum_{\{P(i, b)\}} p[P(i, b)], \quad (2.8)$$

where the sum is taken over all the shortest paths between i and b . Averaged over all pairs of nodes in a connected network, this measure is used to quantify how easily the network is navigable. It is also possible to use (2.8) to quantify if a node is easily accessed from others or if it has easily access to others. A longer description of these measures is also available in Chapter 8 of [49] and in the references cited therein.

A different approach consists at considering entropy from the thermodynamical point of view: the entropy of a system reflects its quantity of disorder. Suppose that one has a collection of measurements (such as clustering, number of links, etc.) on real networks $\{x_i\}$, an *ensemble* of networks is a collection of networks that all share these measurements with probabilities associated to each network. If, instead, only estimates of these measurements $\langle x_i \rangle$ are available, each possible network G is considered in the ensemble along with a probability p_G , such that the average value of the measurement on all networks of the ensemble, weighted

by the probability p_G is equal to the expectation $\langle x_i \rangle$. Several sets of probabilities p_G are consistent with these constraints, in [151, 20], the authors search for distributions having the maximal entropy. This approach allows to produce predictions on the characteristics of networks based on a set of measurements on real networks. In a different approach, network ensembles are constructed so as to match exactly some measurements, such as degree distribution. In this case, all networks of the ensemble are equiprobable, hence the entropy of the ensemble is the logarithm of its size (possibly normalized by the number of nodes of the network) [34, 25, 26, 28].

The applications of entropy on networks described in this section are intended to use entropy in order to characterize the structure of networks. For a different purpose, entropy has also been used in order to determine the maximal quantity of information that can be produced by a noisy source where the symbols produced by the source can be confused, where the possible confusions are represented in a *confusion graph*. Since this area of research differs strongly from the context of this thesis, we won't address this topic further, interested readers may refer to [171].

2.3 Community detection

The problem of *community detection* in a network is an ill-defined problem, because the exact description of a *community* is not unique. From an abstract point of view, communities represent an intermediate levels of organization in the network, between the microscopic level where nodes are taken individually and the macroscopic level where the network is studied as a whole. A general (and weak) definition of communities is that a community is a group of nodes that are relatively densely connected to each other but weakly connected to the rest of the network. This definition lacks of precision, in particular, the meaning of “relatively densely connected” and “weakly connected” are not clear. Moreover, it is not clear if every node belongs to a community or if nodes may belong to several communities. The problem of community detection consists then of finding in a given network the communities that fit the best to this definition. The ambiguity in the definition of communities is partly the reason why community detection in networks has been

a flourishing topic in the last decade, given that it opens a large number of directions for tentative solutions. The initial methods that have been proposed are related to the problem of *graph clustering*, for which a number of methods have been proposed, such as the Kernighan-Lin algorithm [99] or spectral methods [177]. In these approaches, the number of communities is often fixed in advance, and the algorithm is asked to find a partition of N so as to minimize the number of links between communities. These methods have nevertheless some drawbacks, such as that the number of communities is fixed in advance, and that there is no guarantee that the internal structure of the obtained communities fits to what the user would *a priori* expect. A number of methods have therefore been proposed in which a heuristic is used to find a partition that optimizes a *quality function*, *i.e.* a function that determines based on the structure of the network if a given partition of its nodes is good or not. The nature of the quality function varies with the methods, it can be done by comparing the number of links observed inside communities to a random null model [162, 142, 183]. Other quality functions integrate instead a dynamical component, such as the path followed by a random walker on the network [116, 166], time [134] or geographical information [69]. In opposition to these quality function methods, that optimize a function which estimates the quality of a whole partition, local methods have also been proposed, which repeatedly search for groups of nodes that show good structural properties. These methods, instead of returning a partition of the nodes in communities, produce a list of groups of nodes, without constraint on the number of communities to which a node may belong. The most prominent method in this class is the Clique Percolation method, which detects communities as groups of densely connected cliques [149]. For an extensive review of the field, see [71, 153].

In this work, communities are considered as partitions of N , hence each node belongs to one and only one community. Most community detection methods work with this assumption. One such class of community detection methods consists at optimizing a quality function named *modularity* [142].

Definition 1. Given a network $G = (N, L)$, represented by its adjacency matrix A and a partition P of N , such that the community index of node

i is π_i , the modularity of P on G is given by

$$Q = \frac{1}{2m} \left[\sum_{i,j \in N} \left(A_{ij} - \frac{d_i d_j}{2m} \right) \delta(\pi_i, \pi_j) \right], \quad (2.9)$$

with $\delta(\cdot, \cdot)$ being the Dirac delta function.

The modularity measures the difference between the number of links inside communities and what would be expected in a random network having the same degree distribution, normalized by the total number of links in the network, the higher it is, the more modular are the communities. Weak bounds on Q can be easily derived from (2.9), namely $-1 \leq Q \leq 1$, however stronger bounds can be derived depending on G .

Initially, this measure has been introduced to quantify the quality of a partition, and was used to compare partitions on a network. In a second time, methods have been developed that aim at finding the partition of N that maximizes modularity. The decision problem associated to finding the partition that maximizes modularity has been proven to be NP-hard [36], hence the available methods are heuristics that provide a good compromise between computational time and quality of the partition [46, 139, 31]. These methods have encountered a significant success, which is mostly due to the qualities of the modularity function; it is parameter-free –the number of communities and their size are endogenous to the method– and its interpretation is very natural: a “good” partition has more links inside communities than expected in a random network.

However, modularity suffers from a significant drawback. Fortunato and Barthélemy [72] have shown that modularity suffers in fact of a so-called resolution limit: the size of the optimal communities are dependent of the size of the network. They showed that for any two communities π_1 and π_2 being connected by l_{12} links and having $d_1 = \sum_{i \in \pi_1} d_i$ and $d_2 = \sum_{i \in \pi_2} d_i$, a better partition can be achieved by merging both communities if

$$2ml_{12} > d_1 d_2. \quad (2.10)$$

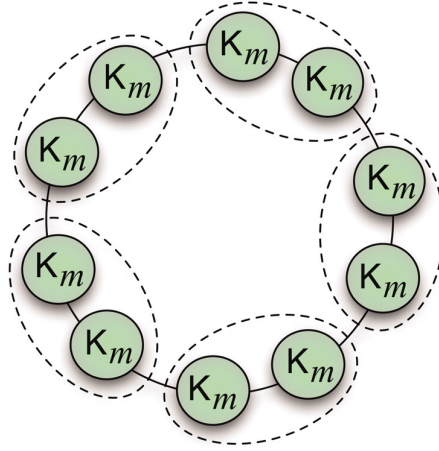


Figure 2.3: A network composed of a ring of cliques of identical size. If the number of cliques is larger than $\sqrt{m}$, modularity optimization would group the cliques in groups of two or more.

This condition imposes a non-trivial bound on the sizes of communities, in disfavour of small communities, which may lead to counter-intuitive cases. In particular, two communities connected by at least one link and whose degree is inferior to $\sqrt{2m}$ are likely to be grouped into one community. This problem has been called the *resolution limit* problem, we show an illustration of it at Figure 2.3

To circumvent this problem, alternative versions of modularity have been proposed where a scale parameter is introduced in order to lower the resolution limit. This topic will be covered more extensively in Chapter 6.

2.4 Dynamical networks

All of the preceding notions are defined on static networks, constructed by aggregating data over a given time interval. However, in practice many systems are changing over time; nodes and links are added or removed and some links may not be continuously active. The evolution of a network needs to be taken into account, both in the development of appropriate measures and in new models that could generate random

evolving networks, to be compared to real datasets.

Formally, such a *dynamical network* can be represented by a network with link weights that vary over time.

Definition 2. Given a network $G = (N, L)$, a time interval $[0, T]$, and a set of functions $\{f_{ij}\}_{(i,j) \in L}$, $f_{ij} : [0, T] \rightarrow \mathbb{R}^+$, the system $(G, \{f_{ij}\})$ is a dynamic network, where the weights on the links of G are defined by the functions f_{ij} .

A link (i, j) is *active* at time t if $f_{ij}(t) \neq 0$, and similarly a node is active at time t if at least one of its incident links is active at time t . A large number of systems may be represented in this way, e.g. social networks, airline networks or citation networks, opening a number of potential research questions.

Temporal networks Several approaches are used to study dynamic complex networks, depending on the object of the study. The most fine-grained approach consists at keeping for each link the exact sequence of timestamps at which it is active. Such an approach is necessary for information propagation studies or models of epidemic spreading, where the time order of contacts is important. When transmitting a packet of information in a network where links have an on/off dynamic, the path followed by the packet needs to respect a time constraint, such that links are used only when they are active. This has led to the notion of *time-respecting paths* [98]. A time-respecting path is an ordered sequence of links and timestamps

$$\{((i_0, i_1), t_0), ((i_1, i_2), t_1), \dots, ((i_l, i_{l+1}), t_l)\},$$

where the link (i_k, i_{k+1}) is active at time t_k and $t_k < t_{k+1}$. A number of different extensions of classical network measures and network models have been developed in this framework, for a survey of the contributions, see [85].

In this framework, it is also possible to follow the behaviour of single nodes or groups of nodes and their dynamics. Such analyses on social networks have shown that the time interval between two contacts of two nodes shows a fat tailed distribution [12, 43, 96, 133]. These *bursty* dynamics differ strongly from the classical dynamics of a Poisson process, instead of having a monotonic contact rate, such contact patterns

are made of successive bursts of interactions separated by long waiting times. These particular dynamics affect in turn the dynamics of spreading processes on the network. Moreover, the study of the correlations of those temporal dynamics and the network topology [96] provide useful insights on their impact on spreading processes on such networks.

Window-based approach A different approach consists of aggregating the temporal information in a sequence of *time windows*, each of them leading to one static network constructed as follows. Given a sequence of M time windows defined by their lower and upper bounds $\{[l_1, u_1], [l_2, u_2], \dots [l_M, u_M]\}$, the sequence of subnetworks of G , with weights being

$$w_{ij}(x) = \int_{l_x}^{u_x} f_{ij}(t) dt, \quad (2.11)$$

is the sequence of networks generated by the time windows. Links of weight 0 are considered to be removed from the network and similarly nodes that are not incident to a at least one link of nonzero weight are also not in the network. Depending on the application, the time windows may be overlapping or separated, but in any case they should be of equal length. The question of the choice of a length is not trivial and may have an influence on the structure of the network, it will be addressed in Chapter 5.

Using this construction scheme allows to consider each snapshot separately and to use classical static network theory tools to study their structure. Each measure becomes then a time series that can be studied with appropriate tools. At a first step, one may compare the evolution of variables, such as the number of nodes, of links or the clustering coefficient over the different windows. If characteristic statistical distributions such as the degree distribution or the strength distribution are similar for different windows, the underlying process is likely to be stationary, at least at macroscopical level.

Even if a dynamical network is stationary at the macroscopical level, it may present strong dynamical fluctuations at microscopical level [105, 74]. Modifications of the network structure are consequences of the addition and removal of links, the first question would then be how often links are present in the network. The *persistence* of links measures the

fraction of windows in which each link is active. This simple measure enables to differentiate long-lasting connections to furtive ones. In particular, the distribution of persistence among links and its correlation with network topology [82] or external data [130] is a good source of information about the dynamical process.

Other authors [74] have instead focused on the evolution of link weights and related this to the appearance/disappearance of links. Variations between two consecutive windows can be measured by the growth rate of link weights

$$r_{ij}(x) = \log \frac{w_{ij}(x+1)}{w_{ij}(x)}. \quad (2.12)$$

In this particular expression, the underlying assumption is that a multiplicative process is ruling the evolution of weights, but it can be adapted to other types of processes as well.

These microscopical changes have in turn an impact on mesoscale structures, such as communities. Tracking the evolution of communities has been followed in two main directions so far: detecting communities on each time window separately and then tracking communities among them, or integrating the time information into the community detection method. The first approach is simpler since methods are already available for static community detection, however one needs a fast and reliable method to compare communities between two consecutive windows. The most natural way of doing so is to calculate the overlap between communities of two consecutive windows [148, 86], but other methods have also been proposed, such as using mutual information [7] or by tracking predefined “core nodes” [190]. In the second approach, time is taken into account in the detection method, by using the result at one timestep as input for the detection on the next window [7] or in the network structure, by using e.g. sliced networks [134].

Growing networks A third approach consists in considering dynamical networks as continuously growing systems, i.e. aggregating the network through a longer and longer time window:

$$w_{ij}(t) = \int_0^t f_{ij}(s) ds. \quad (2.13)$$

A large amount of research has focused on the arrival process of new nodes in the network, a node i being said to arrive at time τ if $w_{ij}(t) = 0 \forall j, \forall t < \tau$ and $\exists j : w_{ij}(\tau) > 0$. Examples such as citation networks or the internet have been a rich source of data for measuring this arrival process and its kinetics. The relationship between the number of nodes and links over time has been intensively studied, in particular the observation that for many real-world networks, the number of links grows faster than the number of nodes of the network. This feature has been called by some authors *accelerating networks* [60, 172]. Accelerating networks have the property to show an increasing average degree and a shrinking diameter [16, 84, 123]. Finally, the study of the evolution of a series of static measures on snapshots of a growing network can also provide useful information on whether the network reaches a stationary state or not [22].

Models of dynamical networks A wide number of network growth models have been proposed to explain the different features observed in complex networks. In particular, a series of models have been developed in order to create networks having a power-law degree distribution. The most common among them being the *preferential attachment* model, introduced by Price [154] and revisited by Barabási and Albert [15]. In the preferential attachment model, new nodes connected themselves to other nodes proportionally to their degree, favouring in this way nodes of high degree. This model has been shown to generate network whose nodes degrees follow a power-law distribution.

Extensions of the basic preferential attachment model have been proposed to fit different possible cases. Non-linear preferential attachment is a model where nodes attach themselves to other nodes depending on an arbitrary function of degree $\Pi(d)$. This function is observed to vary from sublinear to linear functions depending on the real data [92]. In cases where nodes are considered not equally important, the probability to attach may even be a general function of external parameters, which allow to simulate different dynamics of the network structure [27].

Most versions of the preferential attachment model introduce links and nodes at the same rate, which does not fit to the previous finding that some networks have an accelerating behaviour. This can be simulated by adding a second process on top of the preferential attachment

model [16]: not only newly coming nodes create new links, but existing nodes also create new links at each timestep.

The preferential attachment-like models generate growing networks in which the produced degree distribution is similar to what has been observed in real data. However, this family of models is not the only way for networks to grow, neither the only way to create power-law distributions. Initially, the preferential attachment was motivated by analogy with citation networks, where papers tend to cite other famous publications. Similarly, a different family of network growth models has been inspired by the idea that authors writing a new paper would simply copy the whole list of references of a previous paper. The *node copying model* (the idea was proposed in [103]) is based on this idea, with slight modifications to ensure a more realistic scheme. A new node arriving in the network selects randomly a node in the network, and for each neighbour of that node, it creates a link towards the neighbour with probability γ . With probability $1 - \gamma$, it creates a link towards a different uniformly chosen node in the network. This model is particularly successful in biological networks, in which it is considered as a plausible candidate to explain the power-law structure of degree distributions.

The preferential attachment and the node copying models were inspired by simple mechanisms that would make sense in citation networks or the web graph. In social networks, these mechanisms are less interpretable, hence different models have been developed to model the growth of social networks. A common approach comes from the idea that people create new acquaintances by being introduced to them by a common friend. This has led to *triangle closure* models, in which nodes select randomly two of their neighbours, and place a link between them if they are not connected. The first idea of such a model was proposed in [52], since then several variations of this model exist [185, 129, 110], producing networks of high clustering coefficient, and depending on the model, degree-degree correlations, weight-topology correlations, and power-law degree distributions.

Several models have also been proposed to take into account the deletion of nodes and links, either based on random deletions (see [123] for an example), or based on the evolution of the link weights [74].

This list of models is not exhaustive, it contains many of the most-used models. In particular, most models focus on the structural evolution of networks, *i.e.* the creation or deletion of links, and less on the kinetics of the network growth. We will address this question in Chapter 5.

2.5 Real networks

The study of social networks has interested scientists for decades, however most of the interest has been raised only recently, when large datasets of real-world networks have been made available for scientific studies. The origins of these datasets is extremely varied; *e.g.* the movie actors network, where actors of Hollywood are connected if they have played in a movie together, online social networks, phone call networks, the internet, the world wide web, scientific paper citation networks, biological networks, food webs, etc.

From the analysis of these datasets, it appears that most of these networks share similar non-trivial properties, which is why scientists have focused on basic mechanisms that could reproduce these common properties. For instance, it appears that many real-world networks show a broadly distributed degree distribution with a fat tail [15]. An exact analytical description of such distribution is not always available, but nevertheless in all cases one can observe that most nodes have a small activity, while a small number of nodes have a large activity, and that the level of activity ranges over several orders of magnitude, with large statistical fluctuations.

Meanwhile, many studied networks appear to be small worlds, namely that the longest shortest path between two nodes in the network is small compared to the size of the network [192]. The small world phenomenon was already studied in the '60s, but only recently it could be observed on a large number of datasets.

These networks also share the property of being sparse, but locally dense, meaning that while the average degree of a node is small compared to the size of the network, the local clustering coefficient is large. In such networks, one is also likely to observe a well-defined community structure, where nodes are grouped into densely connected groups of

nodes with only few connections to the remainder of the network.

The observation of these non-trivial features in such a large range of different networks, that have apparently nothing in common led researchers to hypothesize that some simple mechanisms, non-specific to one particular application, should be underlying these structural observations. The study of real network showed that they are far from the models proposed by Erdős and Rényi.

Chapter 3

Analysis of mobile phone datasets

In this chapter, we review some of the most important advances made recently in the study of **mobile phone datasets**. This area of research has emerged a decade ago, with the increasing availability of large-scale anonymized datasets, and has grown into a stand-alone topic. We will survey the contributions made so far on the **social networks** that can be constructed with such data, the study of **personal mobility**, **geographical partitioning** and **urban planning**, as well as **security and privacy issues**.

3.1 Introduction

As the Internet has been the technological breakthrough of the '90s, mobile phones have changed our communication habits in the first decade of the twenty-first century. In 10 years, the world coverage of mobile phone subscriptions has raised from 12% of the world population in 2000 up to 78% in 2010 – 5.3 billion subscribers – while we communicate at a constantly higher rate (in 2010, 6.1 trillion text messages were exchanged, representing about 200,000 SMS per second) [89]. Mobile communication has initiated the decline of landline use – decreasing both in developing and developed world since 2005 – and allows people to be connected even in the most remote places of the world.

In short, mobile phones are *ubiquitous*. In most countries of the devel-

oped world, the coverage reaches 100% of the population, and even in remote villages of developing countries, it is not unusual to cross someone in the street talking on a mobile phone. Due to their ubiquity, mobile phones have stimulated the creativity of scientists to use them as millions of potential sensors of their environment. Mobile phones have been used as distributed seismographs, as motorway traffic sensors, as transmitters of medical imagery or as communication hubs for high-level data such as the reporting of invading species [113].

Besides these applications of voluntary reporting, where users install applications on their mobile phones in the aim to serve as sensor, the essence of mobile phones have revealed them to be a source of even much richer data. The call data records (CDRs), needed by the mobile phone operators for billing purposes, contain an enormous amount of information on how, when, and with whom we communicate.

In the past, studies on social interactions between individuals were mostly done by surveys, for which the number of participants ranges typically around 100 people, and for which the results were biased by the subjectivity of the participants' answers. Mobile phone CDRs, instead, contain the information on communications between millions of people at a time, and contain real observations of communications between them rather than self-reported information.

In addition, CDRs also contain localisation data and may be coupled to external data on customers such as age or gender. Such a combination of personal data makes of mobile phone CDRs a extremely rich and informative source of data for scientists.

Of course, there are restrictions on the availability of some types of data and on the projected applications. First, the content of communications (SMS or phone discussions) is not recorded by the operator, and thus inaccessible to any third party – exception made of cases of phone tapping, which are not part of this subject. Secondly, while mobile phone operators own the information filed by their customers and the CDRs, they may not pass all the information to a third party (such as a researchers), depending on their own privacy policies and the laws on protection of privacy that apply in the country of application. For example, in any case, names and phone numbers are not transmitted to

external parties. In some countries, location data, i.e. the base stations at which each calls are made, have to remain confidential – some operators are even not allowed to use their own data for private research. Finally, when a company transmits data to a third party, it goes along with non-disclosure agreements (NDA's) and contracts that strongly regulate the authorized research directions, in order to protect the users' privacy.

Yet, even the smallest bit of information is enough for triggering bursts of new applications, and day after day researchers discover new purposes one can get from CDRs. The first application of a study of phone logs (not mobile, though) appeared in 1949, with the seminal paper by George Zipf modelling the influence of distance on communication [196]. Since then, phone logs have been studied in order to infer relationships between the volume of communication and other parameters (see e.g. [48]), but the apparition of mobile phone data, in massive quantities, has definitely made a breakthrough in that domain. Being personal objects, mobile phones enabled to infer real social networks from their CDRs, while fixed phones are shared by users of one same geographical space (a house, an office). The communications recorded on a mobile phone are thus representative of a part of the social network of one single person, where the records of a fixed phone show a superposition of several social actors. By being mobile, a mobile phone has two additional advantages: first, his owner has almost always the possibility to pick up a call, thus the communications are reflecting the temporal patterns of communications in great detail, and second, the positioning data of a mobile phone allows to track the displacements of his owner.

Given the large amount of research related to mobile phones, we will focus in this chapter on contributions related to the analysis of massive CDR datasets.

The chapter is organized following the different types of data that may be used in related research. In Section 3.2 we will survey the contributions studying the topological properties of the static social network constructed from the calls between users. When information on the position of each node is available, such a network becomes a geographical network, and the relationship between distance and the structure of the network can be analysed. This will be addressed in Section 3.3.

Phone calls are always localized in time, and some of them might represent transient relationships while others rather long-lasting interactions. This has led researchers to study these networks as temporal networks, which will be presented in Section 3.4. In Section 3.5, we will focus on the abundant literature that has been produced on human mobility, made possible by the spatio-temporal information contained in CDR data. As mobile phone networks represent in their essence the transmission of information or more recently data between users, we will cover this topic in Section 3.6, with contributions on information diffusion and the spread of mobile phone viruses. Finally, in Section 3.7 we will consider privacy issues raised by the availability and use of personal data.

3.2 Social networks

In its simplest representation, a dataset of people making phone calls to each other is represented by a network where nodes are people and links are drawn between two nodes who call each other. In the first publications related to mobile phone datasets, the datasets were rather used as an example for demonstration of the potential applications of an algorithm [1] or model [4] rather than for a purpose of analysis. However, it quickly appeared that the so-called *mobile call graphs* (MCG) were structurally different from other complex networks, such as the web and internet, and deserved a particular attention. We will review here the different contributions on network analysis. We will address the construction of a social network from CDR data, which is not a trivial exercise, simple statistical properties of such networks and models that manage to reproduce them, more complex organizing principles, and community structure, and finally we will discuss the relevance of the analysis of mobile phone networks.

Construction While the network construction scheme mentioned above seems relatively simple, there exist many possible interpretations on how to define a link of the network, given a dataset.

The principle aim of social network analysis is to observe social interactions, but not every phone call is made in the same social purpose. Some calls might be for business purpose, some might be accidental calls, some nodes may be call centres that call a large number of people, and

all such interactions are present in CDRs. In short, CDRs are noisy datasets. “Cleaning” operations are usually needed to eliminate some of the accidental edges. For example, Lambiotte *et al.* [115] imposed as condition for a link that at least one call is made in both directions (reciprocity) and that at least 6 calls are made in total over 6 months of the dataset. This filtering operation appeared to remove a large fraction of the links of the network, but by the same time, the total weight (the total number of calls passed in by all users) was reduced by only a small fraction. The threshold of 6 calls in 6 months may be questionable, but a stability analysis around this value can comfort that the exact choice of the threshold is not crucial. Similarly, Onnela *et al.* [146] analysed the differences between the degree distribution of two versions of the same dataset, one containing all calls of the dataset, and the other containing only calls that are reciprocated. Some nodes in the complete network have up to 30,000 different neighbours, while in the reciprocated network, the maximal degree is close to 150. Clearly, the first case is hard to imagine being a single person, while the latter is a much more realistic bound. However, even if calls have been reciprocated, the question of setting a meaningful weight on each link is far from easy. It is sometimes convenient to represent a mobile call network by an undirected network, and to sum up the weights from both directions. But reciprocal calls are often strongly imbalanced, say Kovanen *et al.* [106]. In the interacting pair, one user is often initiating most of the calls, how can this be represented in an undirected network by a representative link weight? In a closely related question, most CDRs contain both information on voice calls and text messages, but so far it is not clear how to incorporate both information into one simple measure.

Besides these considerations on the treatment of noise, the way to represent social ties may vary as well: they may be binary, weighted, symmetric or directed. Different answers to such decisions lead to different network characteristics, and result in diverse possible interpretations of the same dataset. For example, Nanavati *et al.* [135] keep their network as a directed network, in order to obtain information on the strongly connected component of the network, while Onnela and colleagues [146] rather focus on an undirected network, weighted by the sum of calls going in both directions.

Finally, one may raise the question of the significance of the data:

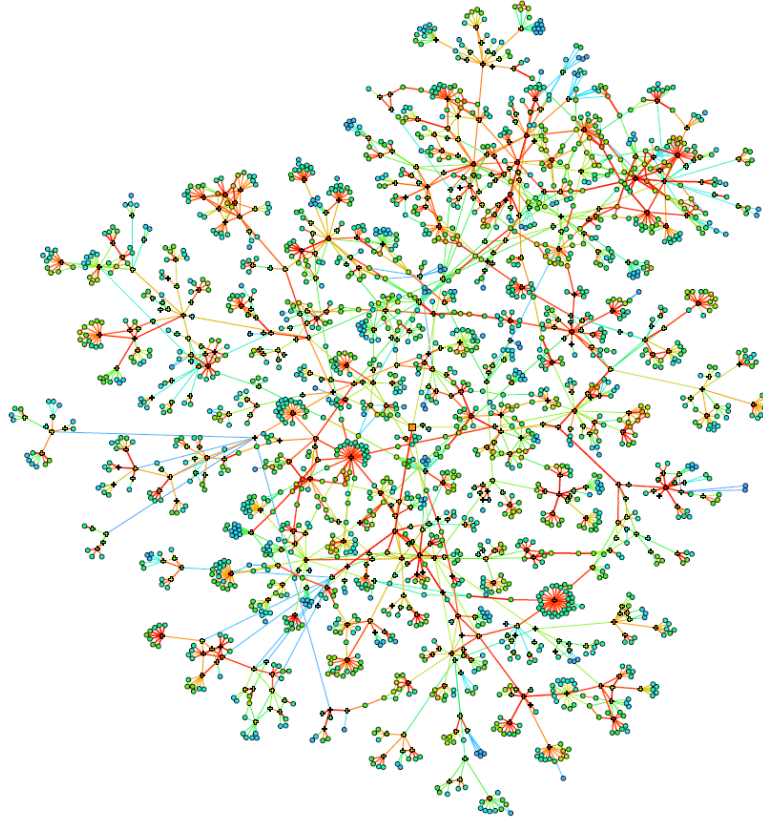


Figure 3.1: Sample of a mobile phone network, obtained with a snowball sampling. The source node is represented by a square, bulk nodes by a + sign and surface nodes by an empty circle. Figure reproduced from [145].

given that only a fraction of a country's population is reached by one operator, to which extent may the results on a dataset be generalized to larger populations? Clearly, quantitative results obtained in these studies, such as degree of nodes, cannot be taken for granted, but one may expect that as long as the population sample is not biased, qualitative observations such as the broadness of degree distribution or the organization of nodes in communities are significant information on the structure of communication networks.

Topological properties The simplest information one can get out of CDRs is statistical information on the number of acquaintances of a node, on the local density of the network or on its connectivity. Like social networks, mobile call graphs differ from random networks and lattices by their broad degree distribution [13], their small diameter and their high clustering [192].

While all analysed datasets present similar general shapes for those distributions, their fine shape, and their range differ due to differences between the datasets, the construction scheme, the size, or the time span of the collection period.

In one of the first studies involving CDR data Aiello *et al.* [4] observed a power law degree distribution, which was well explained by a massive random graph model $P(\alpha, \beta)$ described by its power-law degree distribution $p(d = x) = e^\alpha x^{-\beta}$.

Random graph models have often been used in order to model networks, and manage to reproduce some observations from real-world networks, such as the small diameter and the presence of a giant component, such as observed on mobile datasets. However, they fail to uncover more complex features, such as degree-degree correlations. Nanavati *et al.* [135] observed in the study of 4 mobile datasets that besides the power-law tail of the degree distribution, the degree of a node is strongly correlated with the degree of his neighbours.

Characterizing the exact shape of the degree distribution is not an easy task, which has been the focus of a study by Seshradi *et al.* [169]. They observed that the degree distribution of their data can be fitted with a Double Pareto Log Normal (DPLN) distribution, two power-laws joined by a hyperbolic segment – which can be related to a model of social wealth acquisition ruled by a lognormal multiplicative process. Interestingly, let us note that the time span of the three aforementioned datasets are different, Aiello *et al.* have data over one day, Nanavati *et al.* over one week, and Seshadri *et al.* over one month.

What information do we get from these distributions? They mostly reflect the heterogeneity of communication behaviours, a common feature for complex networks [13]. The fat tail of the degree distribution is responsible for large statistical fluctuations around the average, indication that there is no particular scale representative of the system. The majority of users have a small number of contacts, while a tiny fraction

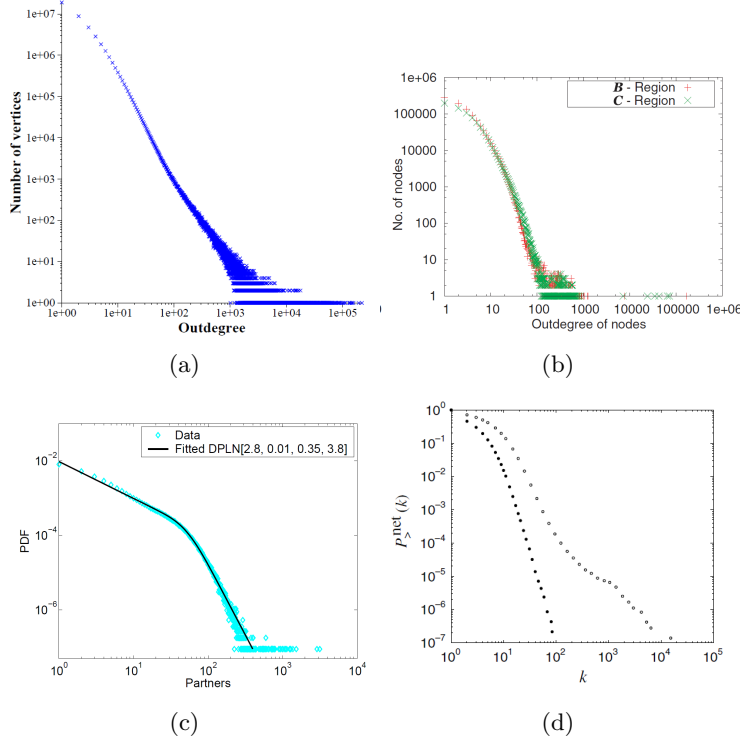


Figure 3.2: The degree distributions of several datasets have comparable features, but differences in the construction, the time range of the dataset and the size of the system lead to different shapes. Note the bump in (d), when non-reciprocal links are taken into account. Figure reproduced from (a) [4] (b) [135] (c) [169] and (d) [145].

of nodes are hubs, or super-connectors. However, it is not clear whether these hubs represent true popular users or are artefacts of noise in the data, as was observed by Onnela *et al.* [145] in their comparison of the reciprocated and non-reciprocated network.

The heterogeneity of degrees is also observed on node strengths and link weight, which is also to be expected for social networks. All studies also mention high clustering coefficient, which indicates that the nodes arrange themselves locally on well-organized structures. We will address this topic in more detail further.

Advanced network characteristics Beyond statistical distributions, more complex analyses provide a better understanding of the structure of our communication networks. The heterogeneity of link weights deserves a particular attention. Strong links represent intense relationships, hence the correlation between weight and topology is of primary interest. Recalling that mobile call graphs show high clustering coefficient, and thus are locally dense, one can differentiate links based on their position in the network.

The overlap of a link, introduced in [146], is an appropriate measure which characterizes the position of a link as the ratio of observed common neighbours n_{ij} over the maximal possible, depending on the degrees k_i and k_j of the nodes

$$O_{ij} = \frac{n_{ij}}{((d_i - 1) + (d_j - 1) - n_{ij})} \quad (3.1)$$

The authors show that link weight and topology are strongly correlated, the strongest links lying inside dense structures of the network, while weaker links act as connectors between these densely organized groups. This finding has an important consequence on processes such as link percolation or the spread of information on networks, since the weak ties act as bridges between disconnected dense parts of the network, illustrating Granovetter's hypothesis on the strength of weak ties.

The structure of the dense subparts of the network provides essential information on the self-organizing principles lying behind communication behaviours. Before moving to the analysis of communities, we will focus on properties of cliques. The structure of cliques is reflected by how weights are distributed among their links. In a group where everyone talks to everyone, is communication balanced? Or are small subgroups observable? A simple measure to analyse the balance of weights is the measure of coherence $q(g)$, introduced in [145], which is the ratio between the geometric mean of the link weights and the arithmetic mean

$$q(g) = \frac{\left(\prod_{ij \in l_g} w_{ij}\right)^{1/|l_g|}}{\frac{\sum_{ij \in l_g} w_{ij}}{|l_g|}} \quad (3.2)$$

where g is a subgraph of the network and l_g is its set of links. This measure takes values in the range $]0, 1]$, 1 corresponding to equilibrium.

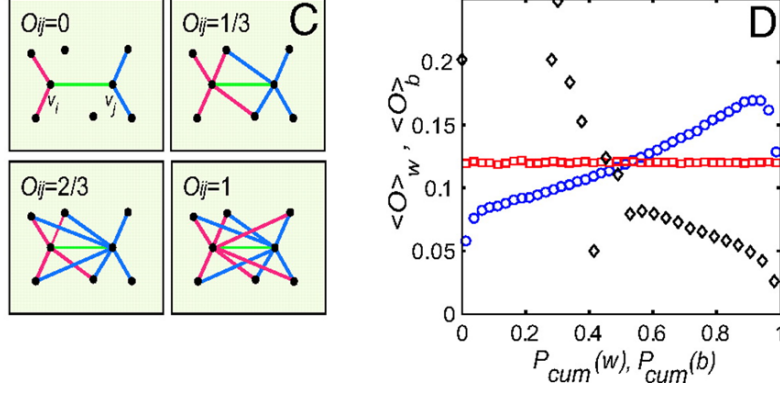


Figure 3.3: (Left) The overlap of a link is defined as the ratio between the common neighbours of both nodes and the maximum possible common neighbours. Here, the overlap is given for the green link. (Right) The average overlap increases with the cumulative weight in the real network (blue circles) and is constant in the random reference where link weights are shuffled (red squares). The overlap also decreases with the cumulative betweenness centrality $P_{cum}(b)$ (black diamonds). Figure reproduced from [146].

On average, cliques appear to be more coherent as what would be expected in the random case, in particular for triangles, which show high coherence values.

On a related topic, Du *et al.* [61] focused instead on the propensity of nodes to participate to cliques, and in particular on the balance of link weights inside triangles. Their observations differ slightly from Onnela *et al.*: on average, the weights of links in triangles can be expressed as powers of one another. The authors managed to reproduce this singular situation with a utility-driven model, where users try to maximize their return from contacts.

Communities The previous analysis of cliques and triangles opens the way for an analysis of more complex structures, such as communities in mobile phone networks. The analysis of communities provides infor-

mation on how communication networks are organized at large scale. In conjunction with external data, such as age, gender or cultural differences, it provides sociological information on how acquaintances are distributed over the population. From a corporate point of view, the knowledge of well-connected structures is of primary importance for marketing purposes. In this paragraph, we will only address simple results on community analysis, but this topic will be addressed again further in the document, when it relates to geographic dispersal of networks or dynamical networks.

At small scale, traditional clustering techniques may be applied, see [100] and [195] for examples of applications on small datasets. However, on large mobile call graphs involving millions of users, such clustering techniques are outplayed by community detection algorithms.

Uncovering the community structure in a mobile phone network is highly dependent on the used definition of communities and detection method. One could argue that there exist as many plausible analyses as there are community detection methods. Moreover, the particular structure of mobile call graphs induces some issues for traditional community detection methods. Tibely *et al.* [181] show that even though some community detection methods perform well on benchmark networks, they do not produce clear community structures on mobile call graphs. Mobile call graphs contain many small tree-like structures, which are badly handled by most community detection methods. The comparison of three well-known methods: the Louvain method [31], Infomap [166] and the Clique Percolation method [148] produce different results on mobile call graphs. The Louvain method and Infomap both build a partition of the nodes of the network, so that every node belongs to exactly one community. In contrast Clique Percolation only keeps as community dense subparts of the network (see Figure 3.4).

As observed in Tibely *et al.* the small tree-like structures are often considered as communities, although their structure is sparse. Such a result is counter-intuitive given the intrinsic meaning of communities, is community detection hence unusable on mobile call graphs? The results have probably to be considered with caution, but as this is always the case for community detection methods, whatever network is used, this speciality of communities in mobile call graphs appears rather as a particularity than a problem. Although they might have singular

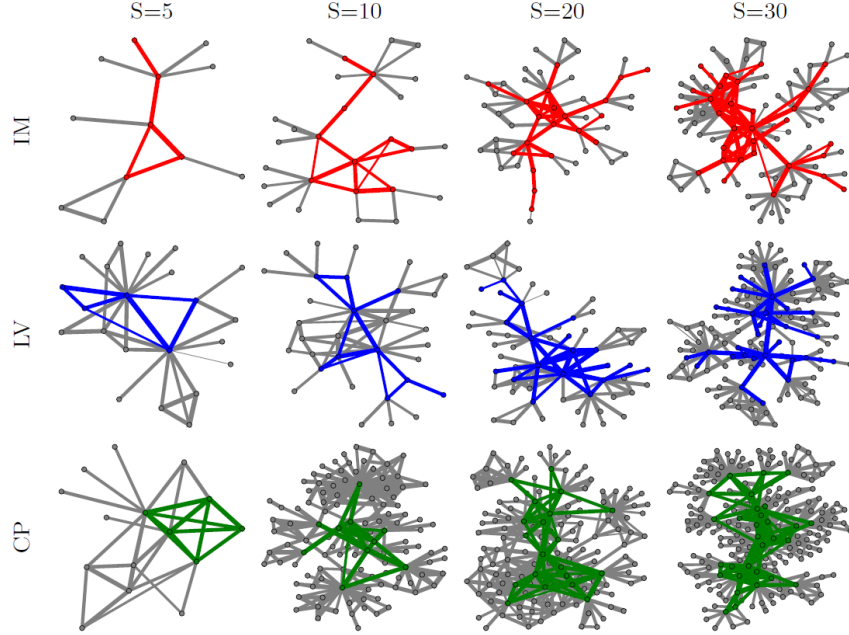


Figure 3.4: Examples of communities detected with the InfoMap method (IM, red), Louvain method (LV, blue) and Clique percolation method (CP, green). For each method, four examples are shown, with 5, 10, 20 and 30 nodes. The coloured links are part of the community, the grey nodes are the neighbours of the represented community. While IM and LV find almost tree-like structures, CP finds dense communities. Figure reproduced from [181].

shapes, communities can provide significant information, when usefully combined with external information. Proof is made by the study of the linguistic distribution of communities in a Belgian mobile call graph [31], where the communities returned by the Louvain method strikingly show a well-known linguistic split.

The notion of communities in social networks, such as rendered by mobile phone networks, has raised a debate on the exact vision one has of what a community is and what it is not. In particular, several au-

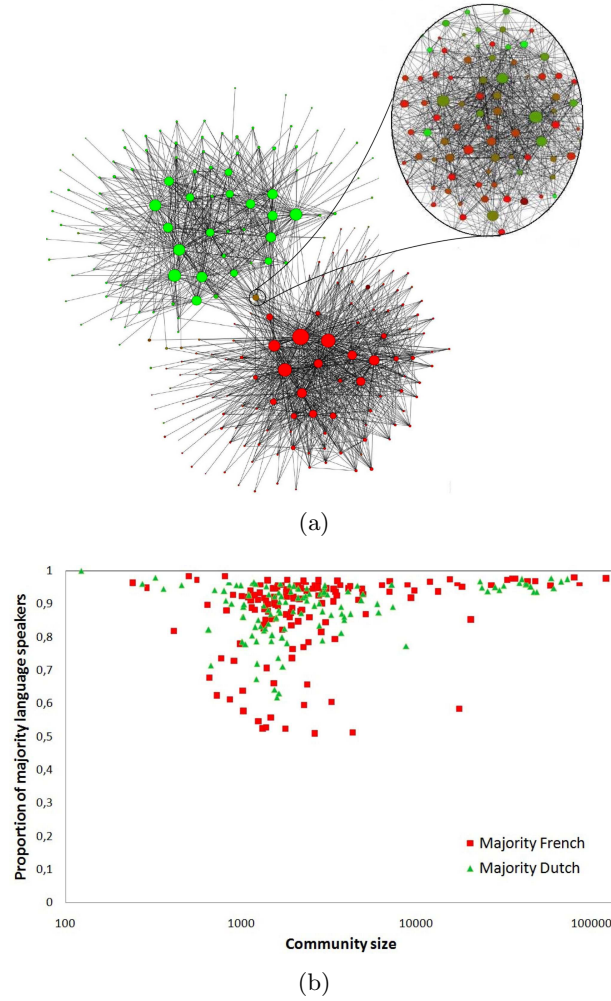


Figure 3.5: (a) The communities of the Belgian network are coloured based on their linguistic composition: green for Flemish, red for French. Communities having a mixed composition are coloured with a mixed color, based on the proportion of each language. (b) Most communities are almost monolingual. Figures reproduced from [31].

thors have favoured the idea of overlapping communities, such that one node may belong to several communities, in opposition with the classical

vision that communities are a partition of the nodes of a network. An argument in favour of this vision is that one mostly makes part of several groups of acquaintances who do not share common interests, such as family, work, sports activities. In [3], Ahn *et al.* show how overlapping communities can be detected by partitioning edges rather than nodes, and illustrated their methods with a mobile phone dataset. For each node, they had an additional information about its center of activities, with which they showed that communities were geographically consistent.

Social analysis The use of mobile call data in the purpose of analysis of social relationships raises two questions. First, how faithful is such a dataset of real interactions? Second, can we extract information on the users themselves from their calling behaviour?

It has often been claimed that mobile phone data analysis is a significant advance for social sciences, since it allowed scientists to use massive datasets containing the activity of entire populations. The study of mobile phone datasets is part of an emerging field known as computational social science [120]. These massive datasets, it is said, are free from the bias of self-reporting, which is that the answers to a survey are usually biased by the own perception of the subject, which is not objective. Still, the question remains: how much does self-reporting differ from our real behaviour, what is the exact added value of having location data? This has been studied by Eagle *et al.* [66] in the well-known Reality Mining project. By studying the behaviour of about 100 persons both by recording their movements and encounters using GSM and Bluetooth technology and with the use of surveys, they managed to quantify the difference between self-reported behaviour and what could be observed. It appears that observed behaviour differs strongly from what has been self-reported, confirming that the subjectivity of the subjects own perception produces a significant bias in surveys. In contrast, collected data allows to reduce this bias significantly. While most studies use external data as validation tool to confirm the validity of results, Blumenstock *et al.* shortly addressed a different question, namely if it was possible to infer information on people's social class based on their communication behaviour. Apparently, this task is hard to perform, even if significant differences appear in calling behaviour between different classes of the

population [33].

3.3 Adding space – Geographical networks

Besides basic CDR data, it happens that geographic information is available about the nodes, such as the home location (available for billing purposes) or the most often used antenna. This allows then to assign each node to one geographic point, and to study the interplay between geography and mobile phone usage. Studies on geographical networks have already been performed on a range of different types of networks [19].

Relationship space-communication This idea has been followed by Lambiotte *et al.* [115], who assigned each of the 2.5 million users from a Belgian mobile phone operator to the ZIP code location where they were billed. By approximating the position of the users to the centre of each ZIP code area, they showed that the probability of two users to be connected decreases with the distance r separating them, following a power law of exponent -2 . The probability of a link to be part of a triangle decreases with distance, until a threshold distance of 40 km, after which the probability is constant. Interestingly, this threshold of 40 km is also a saturation point for the average duration of a call (see Figure 3.6). A different study on the same dataset also showed that total communication duration between communes in Belgium was well fitted by a gravity law, showing positive linear contribution of the number of users in each commune and negative quadratic influence of distance [108, 107]:

$$l_{ab} = \frac{c_a c_b}{r_{ab}^2} \quad (3.3)$$

where l_{ab} represents the total communication between communes a and b , c_a and c_b the number of customers in each commune and r_{ab} the distance that separates them.

While it seems sure that distance has a negative impact on communication, its exact influence is not unique. Onnela *et al.* [144] observed in a different dataset a probability of connection decreasing as $r^{-1.5}$ rather

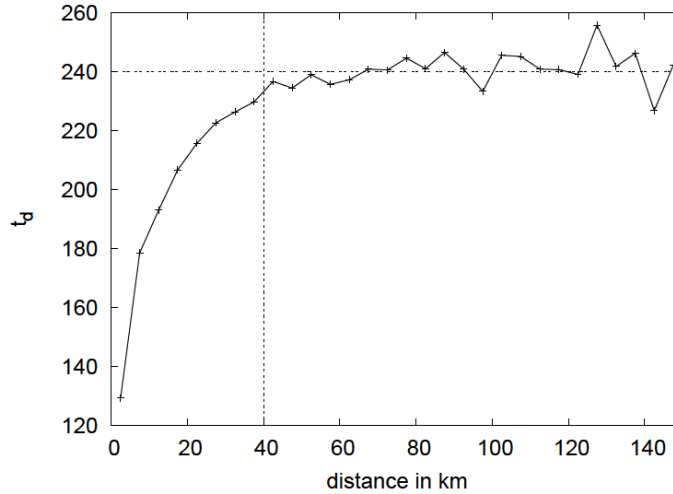


Figure 3.6: The average duration of a call in function of distance between the callers. A saturation point is observed at 40 km. Figure reproduced from [115].

than the gravity model observed by Lambiotte *et al.* However, these differences might be explained by the differences that exist between both countries, such as the distribution of the population density. A different study on mobility data from the location-based service Foursquare [143] levelled those variations using a rank-based distance [124], which could also be helpful in this case.

Further, Onnela *et al.* also studied the geographic structure of communities, and showed on the one hand that nodes that are topologically central inside a community may not be central from a geographical point of view, and on the other hand that the geographical shape of communities varies with their size. Communities smaller than 30 individuals show a smooth increase of geographical span with size, but bounces suddenly at the size of 30, which could not be clearly explained by the authors.

While communities appear to be geographically tightly bound, Eagle *et al.* investigated if some relationship could be found between the structure of a user's social network and the type of environment in which

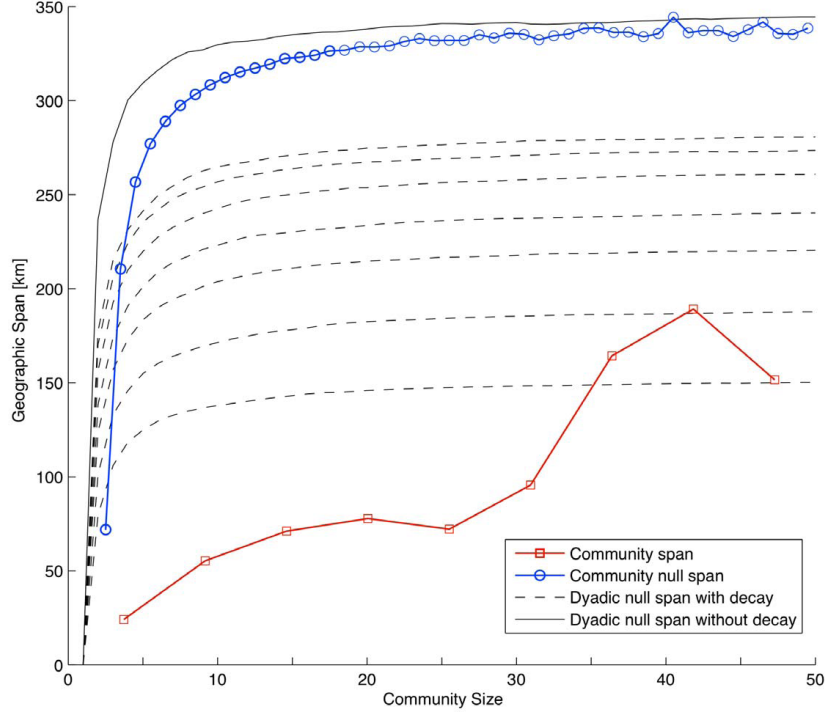


Figure 3.7: Average geographic span (red) for communities and average geographic span for the null model (blue). A bump is observed for communities of size 30 and more, which could not be reproduced by the different null models. Figure reproduced from [144].

he lives [64]. Using both CDRs of fixed landline (99% coverage) and mobile phones (90% coverage), they showed that the social and geographical diversity of nodes' contacts, measured using the entropy of contact frequencies, correlates positively with a socio-economic factor of the neighbourhood. Given a node i , calling each of his d_i neighbours j at frequency p_{ij} , and calling each of the A locations a at frequency p_{ia} , his social and spatial diversity is given by

$$D_{\text{social}}(i) = \frac{-\sum_j p_{ij} \log p_{ij}}{\log d} \quad D_{\text{spatial}}(i) = \frac{-\sum_a p_{ia} \log p_{ia}}{\log A}, \quad (3.4)$$

which is 1 if the node has diversified contacts. On Figure 3.3, the authors

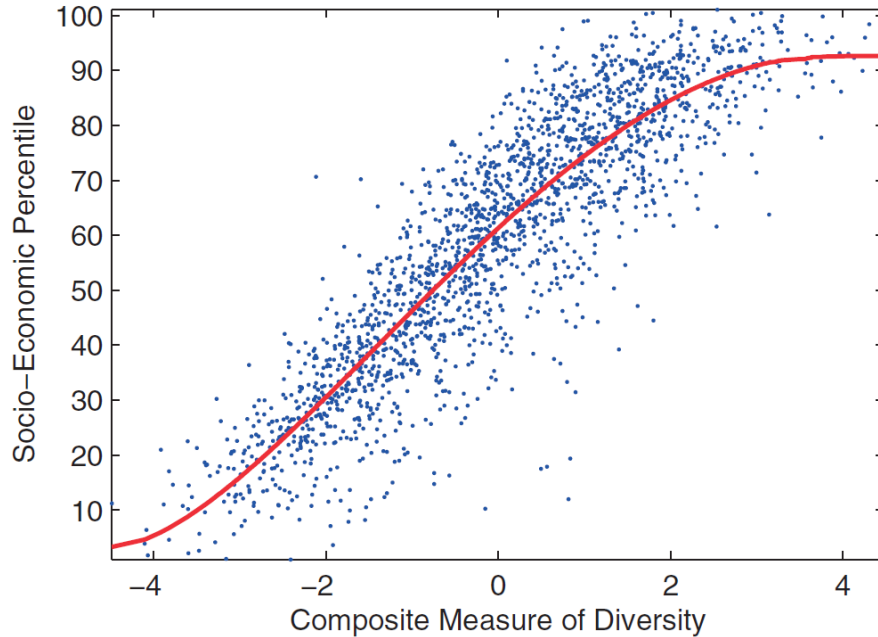


Figure 3.8: Average social wealth as a function of social and geographic diversity. Figure reproduced from [64].

compare a composite measure of both diversities with the socio-economic factor of the neighbourhood.

Geographic partitioning The availability to place customers in higher level entities, such as communes or counties, gave researchers the idea of drawing the “social borders” inside a country based on the interactions between those entities [79]. Individual call patterns of users are aggregated at a higher level to a network of entities, which can in turn be partitioned into a set of communities based on the intensities of calls between the nodes of this *macroscopic* network. It is important to notice that, in contrast with the *microscopic* network (the network of users), the macroscopic network is not a sparse network at all. Since the nodes represent the aggregated behaviour of many users, there is a high chance of having a link between most pairs of communes or counties. Hence,

the weights on the links of the macroscopic network are of crucial importance, since they define the complete structure of the network. To our knowledge, such a partition exercise using CDR datasets has been applied on three countries: in Belgium, France and Great Britain [30]. An initial study of the communities in Belgium [32] used the Louvain method optimizing modularity for weighted directed networks to partition the Belgian communes based on two link weights: the frequency of calls between two communes and the average duration of a call. The obtained partitions were geographically connected, with the influence of distance, of influential cities, and the cultural barrier of language being observable in the optimal partitions.

Given that the intensity of communication between two cities can be well-modelled by a gravity law, Expert *et al.* [69] proposed to replace Newman’s modularity by a more appropriate null model, given that geographic information was available. The spatial modularity (SPA) compares the intensity between communes to a null model influenced both by the sizes c_a and c_b of the communes and the distance that separates them

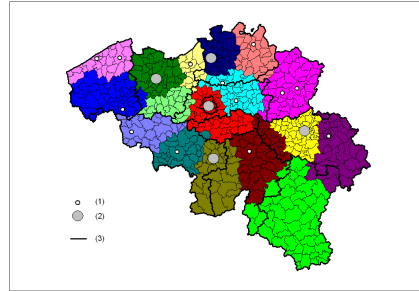
$$p_{ab}^{Spa} = c_a c_b f(r_{ab}). \quad (3.5)$$

The influence of distance is estimated from the data by a function f , which is calculated for distance bins $[r - \epsilon, r + \epsilon]$ as

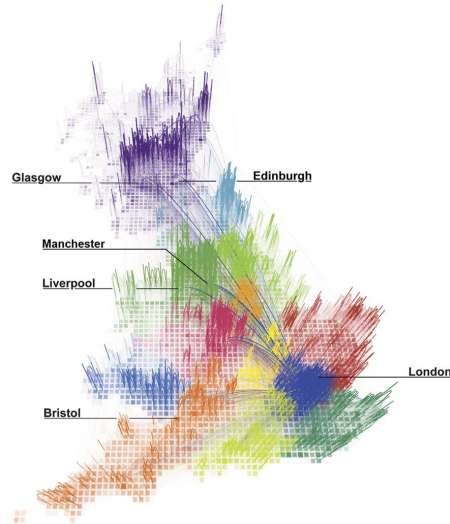
$$f(r) = \frac{\sum_{a,b|r_{ab} \in [r-\epsilon, r+\epsilon]} A_{ab}}{\sum_{a,b|r_{ab} \in [r-\epsilon, r+\epsilon]} c_a c_b}. \quad (3.6)$$

Using their null model, the authors obtained an almost perfect bipartition of the Belgian communes which renders the Belgian linguistic border. Moreover, they showed with a simple example that such a null model allows to remove the influence of geography and obtain communities showing geography-independent features.

On an identical topic, Ratti *et al.* used an algorithm of spectral modularity optimization, to partition the map of Great Britain [156] based on phone calls between geographic locations. Similarly to results obtained on Belgium, they obtained spatially connected communities after a fine-grain tuning of their algorithm, which correspond to meaningful areas, such as Scotland or Greater London. A stability analysis of



(a)



(b)

Figure 3.9: (a) Communities in Belgium, with modularity optimization. Communities are geographically well-balanced and are centred around important cities (gray dots). Figure reproduced from [32]. (b) Communication network in Great Britain (80% of strongest links). The colors correspond to the communities found by spectral modularity optimization. Figure reproduced from [156].

the obtained partition showed that while some variation appears on the boundary of communities, the obtained communities are geographically

centred at the same place. The intersections between several results of the same algorithm showed 11 spatially well-defined “cores” corresponding to densely populated areas of Great Britain. Interestingly, the map of the cores loosely corresponds to the historical British regions.

3.4 Adding time – Dynamical networks

A particularity of a mobile call graph is that the links are very precisely located in time. Although each call has a precise time stamp and duration, the previously presented studies consider mobile call graphs as static networks, where edges are aggregated over time. This aggregation leads on the one hand to a loss of information on the dynamics of the links (some may appear or disappear during the collection period) but on the other hand about the dynamics *on* the links. Recently, some authors have attempted to avoid this issue by taking the dynamical component of links into account in the definition of such networks. The topic of dynamical – or temporal – networks is quite new, but given their inherent dynamical nature, mobile call graphs are excellent sources of information for such studies.

Dynamics of structural properties One such question regards the persistence of links in a mobile phone network. How long does a link last in a network? By analysing slices of 2 weeks of a mobile phone network, Hidalgo and Rodriguez-Sickert observed that the frequency of presence of links in the different slices, the *persistence*, followed a bimodal distribution [82].

$$p_{ij} = \frac{\sum_T A_{ij}(T)}{M}, \quad (3.7)$$

where $A_{ij}(T)$ is 1 if the link (i, j) is in slice T and 0 otherwise, M being the number of slices. Most links in the network are only present in one window, and the probability of a link to be observed in several windows decreases with the number of windows, but there is an unexpectedly large number of links that are present in all windows. These highly recurrent links represent thus strong temporally consistent relationships, in contrast with the large number of volatile connections appearing only once in the slices. A deeper analysis of correlations between the persistence and static measures show moreover that clustering, reciprocity and

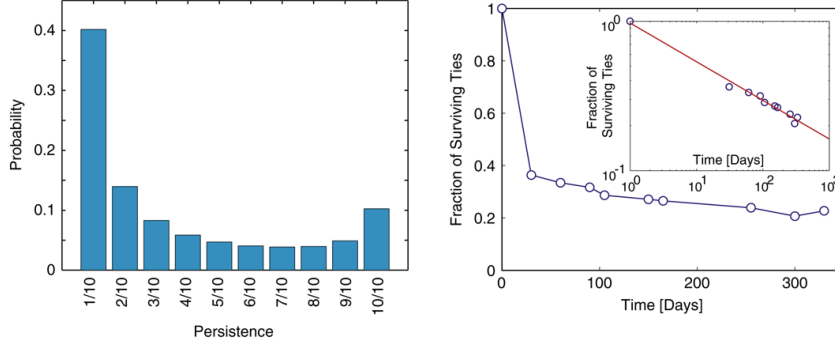


Figure 3.10: Left: the distribution of the persistence of links. Right: the fraction of surviving links as a function of time follow a power-law like decrease. Figures reproduced from [82]

high topological overlap are usually associated with a strong persistence.

Reader *et al.* [158] dig a bit deeper into that last topic, by attempting to predict which link will decay and which will persist, based on several local indicators. They quantify the information provided by each indicator with the decrease of entropy on the probability of an edge to persist, and obtain that the most informative indicators are the number of calls passed between both nodes as well as its scaled version. By trying both a decision-tree classifier and a logistic regression classifier, they manage to predict correctly about 70% of the persistent edges and decays.

From an evolving network perspective, the question of stability and survival of communities is closely linked to the previous questions. Palla *et al.* studied the temporal stability of a mobile phone network [148], analyzing communities detected on slices of two weeks. They observed that communities have different conditions to survive, depending on their size; small communities require to be stable, while large groups require to be highly dynamic and often change their composition.

The availability of timestamps in datasets allow to segment the calls between office hours and home hours. By supposing that calls made

during office hours are for a purpose of business, while private calls are made early morning, in the evening or over the weekend, Cebrian *et al.* managed to build two separate networks based on a mobile and landline dataset from the UK [45]. The degree and clustering coefficient distributions of both networks are mostly similar, but a deeper analysis of the network structure shows that some important differences exist between them. By decomposing the network into k-cores and monitoring the speed of information diffusion, they observe that the work network is much more connected than the leisure network, and that information diffuses almost twice as fast.

Burstiness The dynamics of many random systems are modelled by a Poisson process, where the average interval between two events is distributed following an exponential, well-characterized by its average. However, it has appeared that human interactions show a different temporal pattern, with many interactions happening in very short times, separated by less frequent long waiting times [12].

The same holds for mobile phone calls. Karsai *et al.* studied the implications the bursty patterns on the links of a mobile call graph [96]. They observed that indeed, the inter-event time ranges over a multiple orders of magnitude, and in particular, coupled with the network topology, the burstiness of human communication slows down the spreading of information over the network. Most bursts of activity are located inside communities, which leads them to stay trapped inside them, taking a long time to move to other parts of the network.

It is interesting to note that in the previous paper, the authors observed the inter-event time on links, by sorting links by weight. In [43], Candia *et al.* perform a similar task but for nodes, and measure the inter-event time for nodes, by grouping them based on the number of calls they made. Similarly to Karsai *et al.*'s observations, the inter-event times range over several orders of magnitude, and the distribution is shifted to higher inter-event times for nodes of lower activity. By rescaling with the average of each distribution, the inter-event time distributions collapse into a single curve fitted by a power law with exponent 0.9 followed by an exponential cutoff at 48 days.

$$p(\Delta T) = (\Delta T)^{-\alpha} \exp(\Delta T/\tau_c). \quad (3.8)$$

Mobile phone networks are composed of complex patterns and interac-

tions, but still only few work has been done yet in order to characterize these interactions. The temporal arrival and disappearance of edges or more complex structures and the timescales of human communication are only two examples of the wide possible research that still need to be done in this matter.

3.5 Adding space and time – Mobility

Given their portability, mobile phones are trusty devices to record mobility traces of users. The availability of spatio-temporal information of mobile phone users has led into a large number of potential applications, which would be too large to review here. In particular, since the mid-90's, attempts have been made to monitor car traffic using mobile phone traces, with evident potential applications in accident management and traffic jam prevention. Over the last (almost) 20 years, a large number of attempts have been made to enhance prediction using mobile phone data. This topic is not addressed here, we will refer the interested reader to a recent review of the contributions in this field [178].

Individual mobility is far from random A mobility trace is represented as a sequence of cell phone towers at which a specific user has been recorded while making a phone call. By studying the traces of 100,000 mobile phone users over 6 months, González *et al.* found that human trajectories show a high degree of temporal and spatial regularity [78]. This result contrasts with usual approximations of human motions by random walks or Lévy flights. Their main results showed that all users show very similar patterns of motion, up to a parameter defining their radius of gyration. The regularity is mainly due to the fact that users spend most of their time in a small number of locations. If rescaled and oriented following its principal axis, the mobility of all users can then be described by a single function. These findings are supported by an additional work produced by Song *et al.* [175], who identify significant differences between observational data and two typical models of human displacement: the continuous time random walk and the Lévy flight. Instead, the authors show that a model mixing the propensity of users to return to previously visited locations and a drift for exploration manages to reproduce characteristics present in their data but absent from

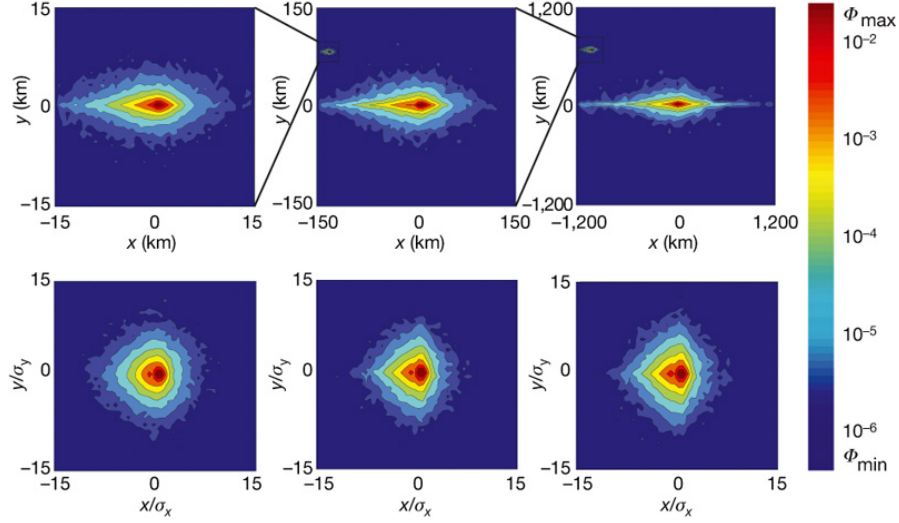


Figure 3.11: Probability density function $\Phi(x, y)$ of finding a mobile phone user in location (x, y) . The plots, from left to right, were generated for users having a different radius of gyration. After rescaling based on the variance of each distribution, the resulting distribution show approximately the same shape. Figure reproduced from [78].

traditional models. In their model, each time a user decides to change location, they can either choose a new location with a probability that decreases with the number of already visited locations ($p_{new} \propto S^{-\gamma}$, where S is the number of visited locations, and γ a constant), or he can return to a previously visited location. Despite the simplicity of this model they manage to explain the temporal growth of the number of distinct locations, the shape of the probability distribution of presence in each location, and the slowness of diffusion.

If mobility traces are not random, and if users often return to their previous visited locations, could one state that human mobility could be predicted? In a recent study, Song *et al.* [176] investigate to what extent one could predict the subsequent location of a user based on the sequence of his previous visited locations. This predictability is given by the entropy rate of the sequence of locations at which the user is

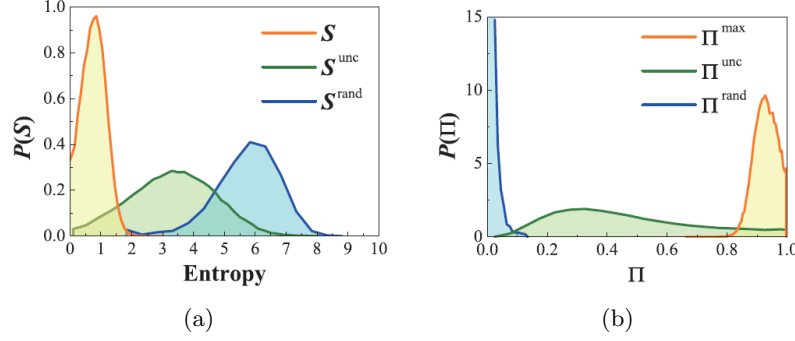


Figure 3.12: (a) Entropy rate of the location of users, for the real, uncorrelated and random data. (b) Maximal predictability of the location of users, for the real, uncorrelated and random data. Figures reproduced from [176].

observed. Importantly, one has to point out that not only the frequency of visit at each location is taken into account, but also the temporal correlations between those visits. Their results show that the temporal correlations of the users' displacements reduces drastically the uncertainty on the presence of a mobile phone user. Using Fano's inequality, they deduce that an appropriate algorithm could predict up to 93% of a user's location on average. The most surprising finding is that not only users are highly predictable on average, but this predictability remains constant across the whole population, whatever distance users use to travel. While one would expect that people travelling often and far would be less predictable than those who stay in their neighbourhood, Song's results seem to point out that there is no variation in predictability in the population.

While the aim of the previous work was to show how predictable human motion could be, the authors did not provide any prediction algorithm, keeping their contribution on the theoretical side. Calabrese *et al.* went a step further and proposed in [39] a predictive model for the location of people. Their algorithm is both based on the past trajectory of the targeted user and on a general drift of the collectivity, imposed by geographical features and points of interest. The prediction is then a weighted average between an individual behaviour and a col-

lective behaviour. The individual behaviour is modelled as a first-order approximation of the concept proposed by Song [176], building a Markov chain where states are locations visited by the user and the probability of moving from state i to state j is proportional to the number of times it has been observed in the data. The collective behaviour is then modelled as a weighted average between the influence of distance, points of interest and land use. The predictions of their model on a sample of a dataset containing the records of 1 million people on 4 months shows that in 60% of their predictions, they manage to predict correctly the next location of a user.

The Markov chain approach used by Calabrese *et al.* for modelling the individual behaviour is also at the base of the study proposed by Park *et al.* [150]. They showed how the temporal evolution of the radius of gyration of a user can be explained by the eigenmode analysis of the transition matrix of the Markov chain. More precisely, the eigenvectors of the transition matrix provide fine-grain information on the traces of individuals.

As it appears, the massive amount of mobility data, which would on first view be considered as random motion, respects a strict routine. Mathematical models, prediction algorithms and visualisation tools (see for example Martino’s work [131]) have recently shed light on this routine, allowing to construct better human displacement models which can be used to predict epidemics outbreaks. At individual level, this routine appears to be strictly ruling our daily behaviour, as Eagle and Pentland [65] show that six eigenvectors of the mobility patterns of users are sufficient to reconstruct 90% of the variance observed. They also observed that individuals tend to have synchronized behaviours, which will be described in the next paragraph.

Aggregate mobility reveal synchronized behaviour of populations At a higher level, those datasets allow to consider whole populations from a God-eye point of view. More practically, the availability of such massive data allows us first to observe and quantify the interaction of people with their environment, and second to quantify the synchronicity of those interactions.

Initial projects, such as the Mobile Landscapes [157] project and Real

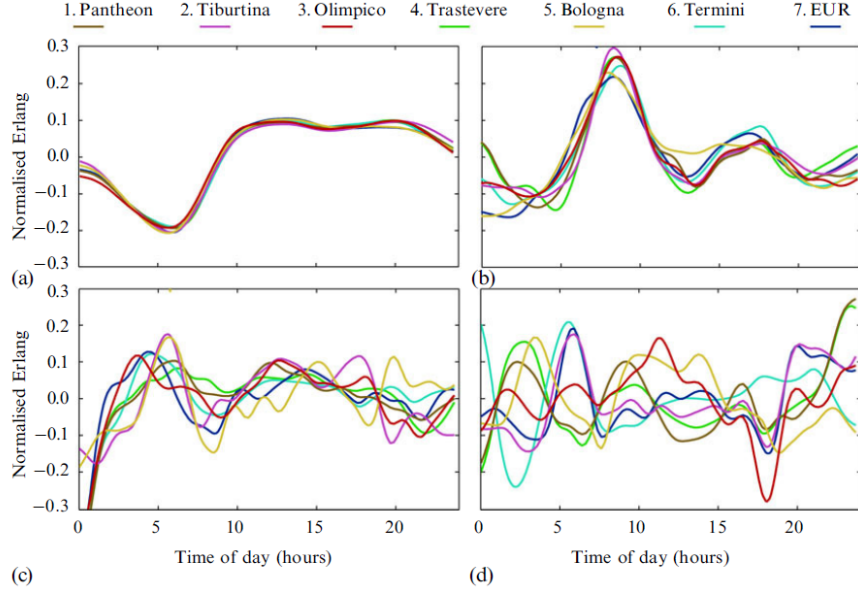


Figure 3.13: Four principal eigenvectors of the Erlang signature for a weekday of 7 places in Rome. While most of the variance is dominated by the principal eigenvector, representing the normal daily activity, the differences between other eigenvectors indicate differences in space usage. Figure reproduced from [159].

Time Rome [38] have shed light on the potential of such an approach, contributions being essentially visual. However, the next step has been made by Reades *et al.* [160], who used tower signals as a digital signature of the neighbourhood. They showed how similar locations presented similar signatures, which implies that a clustering of the urban space is possible, based on the phone usage recorded by its antennas. In particular, the obtained clusters reveal known segmentations of the town, such as residential areas, commercial areas, bars or parks. In short, such a technique may be used as a cheap census method on area usage, which could be of great interest to local authorities. Going a bit further, the same team showed how using an eigendecomposition [159] of the signatures of different locations in town it is possible to extract significant information on differences and similarities in space usage.

Beyond the analysis of a single city, Isaacman *et al.* explored behavioural differences between inhabitants of different cities [91]. By analysing the mobility of hundreds of thousands of inhabitants of Los Angeles and New York City, they showed that Angelenos travel on average twice as far as New Yorkers. Finding an explanation for such a significant difference seems possible, if the inhomogeneities of population density and city surfaces are taken into account. See, for example the work of Noulas *et al.* [143], who show using Foursquare location data that using a rank-based distance, the differences between cities are levelled. A rank-based distance measures the distance between two places i and j as the number of potential opportunities (people, places of interest) being closer to i than j . Given the geographic distance r_{ij} and the density of opportunities expressed in radial coordinates and centred in i , $p_i(r, \theta)$, such a distance reads

$$\text{rank}(i, j) = \int_0^{2\pi} \int_0^{r_{ij}} p_i(r, \theta) r dr d\theta \quad (3.9)$$

In a city of large population density, there will be more opportunities at short geographical distance than in a city with low population density. Hence, users are likely to travel over shorter distances in city of large population density. These distortions of the use of geographical distance are here levelled by the rank-based distance.

Extreme situation monitoring If the availability of data containing the time-stamped activity of a large population allows to perform monitoring of routine in population activities, it also enables to observe the population's collective response to emergencies. Two papers addressed this interesting question. Candia *et al.*, for first, focused on the temporal activity of users at antennas [43]. They propose a method that is based on the study of the statistical fluctuations of individual users behaviours with respect to their average behaviour. As shown on Figure 3.14, in an anomalous case, users show many high fluctuations from their average, while the overall average is close to that of a normal activity.

$$\sigma(a, t, T) = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (n_i(a, t, T) - \langle n(a, t, T) \rangle)^2} \quad (3.10)$$

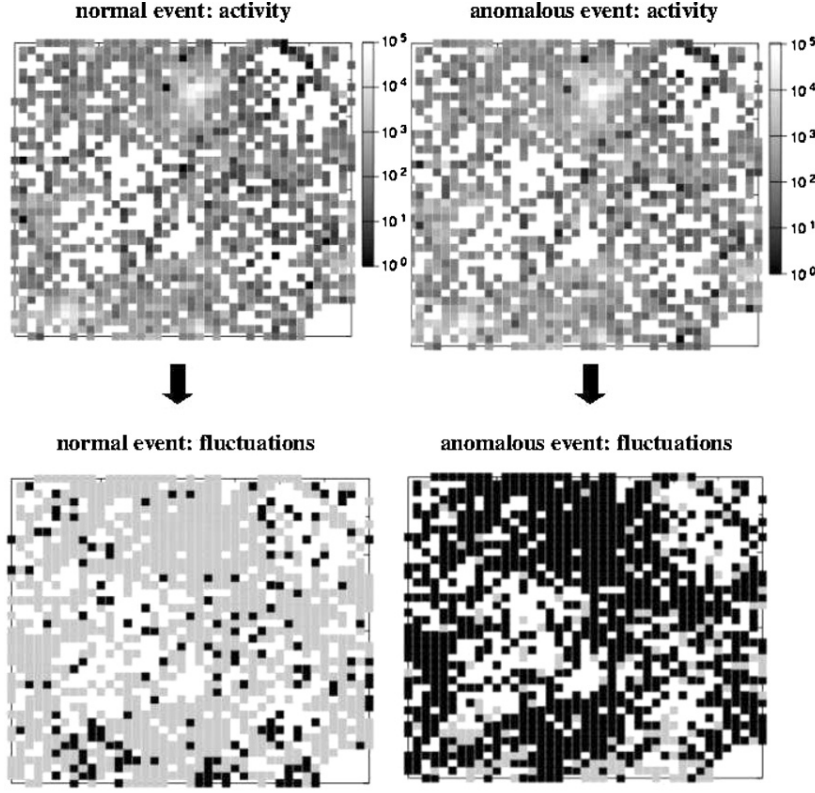


Figure 3.14: activity (top) and fluctuations (bottom) for a normal day (left) and an anomalous event (right). Note that even if no difference is observed on activity, fluctuations are significantly different. Figure reproduced from [43].

The variance is computed for each place a , for the time interval $[t, t + T]$ between the different individual behaviours $n_i(a, t, T)$ and the average expected behaviour. Comparing this variance with the normally expected variance allows to identify locations where users are acting abnormally, and that such locations are, in case of emergencies, spatially clustered. In cases of extreme emergencies, the response of populations can even be monitored as geographically and temporally located spikes of activity. In a recent paper, Bagrow *et al.* [10] analyzed the reaction

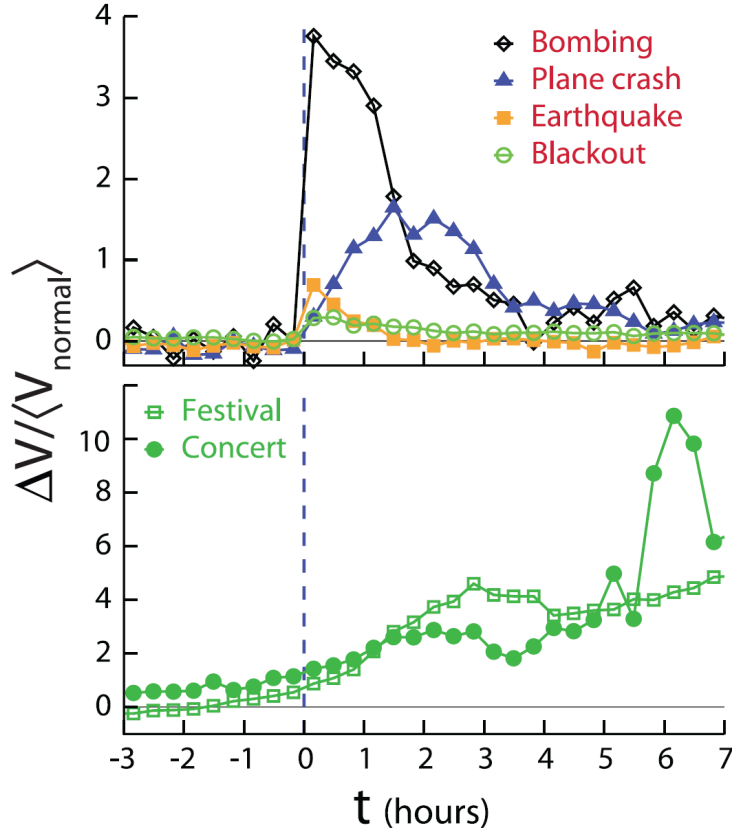


Figure 3.15: Spikes of activity during emergency situations. The activity has been recorded for users close to the center of activity of several emergency situations, relatively to the normal activity. Figure reproduced from [10].

of populations on different emergency situations, such as a bombing, a plane crash or an earthquake (Figure 3.15). They observed such spikes of information when eye witnesses and their neighbours reacted almost directly after the event. The reaction was mostly driven by calls made by nodes who don't call usually at that time, rather than an increase of call rate of active nodes. A detailed study of the paths followed by the information during its propagation shows the efficiency of the collective

response, with 3 to 4 degrees from eye witnesses being contacted within minutes from the situation.

Mobility and social ties The common availability of mobility traces and social interactions in the same dataset allows to address causality questions on the creation of social links. From the work of Calabrese *et al.*, it appears that users who call each other have almost always physically met at least once in one year[41]. Users call each other mostly right before or after physical co-location, and interestingly, the frequency of meetings between users is highly correlated with their frequency of calls as well as the distance separating them.

Going a step further, one may wonder if social ties could be predicted using mobility data. Wang *et al.* showed [187] that indeed, nodes that are not connected in the network, but topologically close, and who show similar mobility patterns are likely to create a link. By combining the mobility similarity and the topological distances in a decision-tree classifier, they manage to improve significantly classical link prediction algorithms, yielding in an average precision of 75% and a recall of 66%. Closely related, Eagle *et al.* showed on 4 years of data how the social network of people changes drastically when moving from one environment to another [63].

Side-questions: what should one do with this data? As showed in each previous paragraphs, mobility data allows to observe and quantify human behaviour as never before. Besides purely sociological questions, this data also opens a number of potential applications, which provides to this data an intrinsic economical value, thinking of geo-localized advertising applications [8].

We have previously addressed the possibility of using mobile phone signatures as a cheap census technique, while Isaacman *et al.* show how one can derive the home and workplace of people based on their mobile phone activity, and from there derive their carbon footprint emissions [90].

By monitoring the movements of people towards special planned events, Calabrese *et al.* [40] show that the type of events highly correlates to the neighbourhood of origin of the users. Such a cartography of taste can be used by authorities when planning the congestion effects of large

events, or for targeted advertising of events (see Quercia *et al.* [155]).

Finally, mobility traces can also be used to monitor temporal populations [128], such as tourists. Kuusik *et al.* [112] studied the mobility of roaming numbers in Estonia for 5 consecutive years, showing the potential for authorities to understand and efficiently target visiting tourists.

3.6 Dynamics on mobile phone networks

Many networks represent a transport between nodes via their links. In mobile phone networks, the links transport either information (exchanged during phone calls or contained in messages) or non-voice exchanges (SMS, MMS). Information diffusion has opened questions on the speed of the diffusion or on the presence of super-spreaders, with applications in viral marketing or crowd management. The transmission of data has been at the centre of attention only recently, with the raise of new types of computer viruses running on smartphones.

Information diffusion A phone call is associated to the transfer of information between caller and callee. However, as paradoxical as it may sound, mobile phone datasets are not appropriate to *observe* real propagations of information. The content of phone calls or text messages is, for evident privacy reasons, unknown. Yet, without having access to the content, it is impossible to decide if an observed pattern of calls reflects the transmission of information or if it happens by chance. One can imagine a network with a number of indistinguishable balls circulating between the nodes. Each time a node receives a ball from one of his neighbours, he decides to keep it for a random time interval and after that to transmit it to one of his neighbours. Suppose now that one decides to track the movement of one specific ball. If the number of balls is small compared to the number of nodes, this can still be doable, as long as each node has maximum one ball in his possession. However, if the number of balls increases to become equivalent to the number of nodes, there is a high chance to confuse the paths of several balls. Add to this that balls might be added, removed or duplicated during the process, and one gets a similar situation as trying to track a piece of information in a mobile phone network.

This artificial example reflects well the issue of tracking information. Peruani and Tabourier addressed this issue and showed that cascades of information, such as observed in mobile call graphs are statistically irrelevant, and correspond thus probably not to real propagations (paper still under submission).

In a small number of cases, however, the actual observation of diffusion of information might be possible. Studying the case of emergencies, such as a plane crash or a bombing, Bagrow *et al.* [10] observed an unusual activity in the geographical neighbourhood of the catastrophe. In this case, the knowledge of both the temporal and spatial localization of an unexpected event that is likely to generate a cascade of information allows to assume that the observed sequences of calls are correlated.

If in most cases the observation of real propagations seems an unreachable objective, a more complete research has been driven in the simulation of propagation of information on complex networks, which results have been extended to questions related to mobile phone networks. There are several ways of modelling information diffusion on networks. A simple way is used in [146] with an SI or SIR model where at each time step, infectious nodes try to infect their neighbours with a probability proportional to the link weight, which correspond to a sequence of percolation processes on the network. However, mobile phone networks are known to have very particular dynamics (recall Section 3.4), which are not taken into account here. Miritello et al. [133] used a formalism similar to the one presented by Newman [137] for epidemics, to characterize the *dynamical strength* of a link, which can be used as link weight to map the dynamical process onto a static percolation problem. The dynamical strength, given an SIR model of recovery time T and probability of transmission λ , is given by

$$\mathcal{T}_{ij}[\lambda, T] = \sum_{n=0}^{\infty} P(w_{ij} = n; T) [1 - (1 - \lambda)^n], \quad (3.11)$$

which is the expected probability of having n calls between i and j in a time range of T multiplied by the probability of propagation given these n calls, summed over all possible values for n . Using an approximation of this expression, they manage to link the observed outbreaks to classical percolation theory tools.

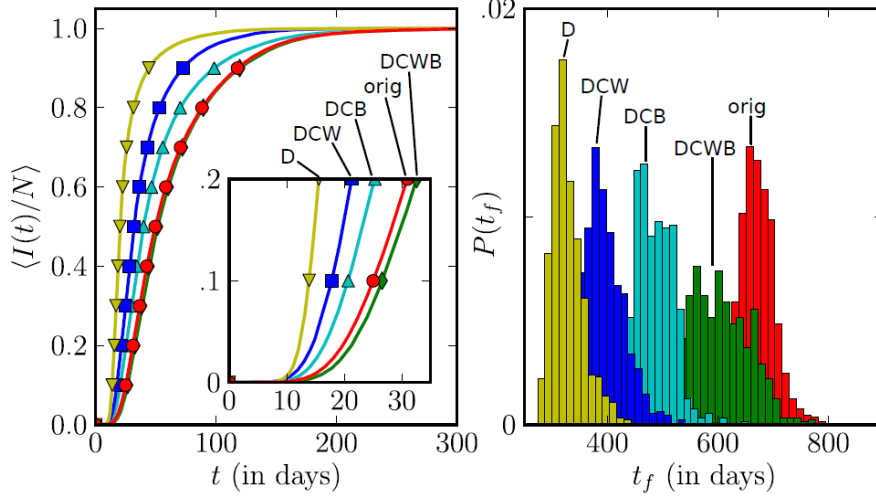


Figure 3.16: (left) Fraction of infected nodes as a function of time for the real (red) data and different randomization schemes. (right) Average prevalence time distribution for nodes. Figure reproduced from [96].

However, such a formalism still neglects the impact of temporal correlations between calls, which, together with network topology, slow significantly down the transmission of information over a network. This has been observed by Karsai *et al.* [96] who used different randomization schemes to show that social networks might be small-world networks, but also much slower as the simulated diffusions.

From a more theoretical point of view, diffusion processes can be seen as particular cases of dynamical systems. Liu *et al.* [125] questioned in this framework the controllability of complex networks. The problem was stated as follows; given a linear dynamical system with time-invariant dynamics

$$\frac{d\mathbf{x}(t)}{dt} = A\mathbf{x}(t) + B\mathbf{u}(t), \quad (3.12)$$

where $\mathbf{x}(t) = (x_1(t), \dots, x_N(t))^T$ defines the state of the nodes of the network at time t , A is the (possibly weighted) adjacency matrix of the network, and B an input matrix, what is the minimal number of nodes needed for the input such that the state of each node is controllable,

i.e. the system is entirely controllable? From control theory, one knows that a sufficient and necessary condition is that the *reachability matrix* $C = (B, AB, A^2B, \dots, A^{N-1}B)$ is of full rank. From previous work, it is known that the minimal number of nodes required is related to the maximal matching in the network, which can be computed with a reasonable complexity. For example, the authors show that in a mobile phone network, one needs to control about 20% of the nodes in order to achieve full controllability of the system. Surprisingly, most nodes needed for controlling the network are low-degree networks, while hubs, which are commonly used as efficient spreaders, are under-represented in the set of input nodes. While the practical interest of this research still needs to be defined, this first result on controllability of networks might open new ideas in the domain of information propagation.

Finally, one may wonder if the patterns of phone usage are efficient in a collaborative scheme. Cebrian *et al.* [44] studied this with a small model, where each node of a mobile phone graph is represented as an agent assorted with a state represented by a binary string. The agents are all given the same function f , that takes their binary string as input and which is hard to optimize, and which computes their personal score. At each communication, the two communicating agents can modify their state in order to increase their personal score. This modification is done with a simple genetic algorithm, which simulates a cross-over of the states of both agents.

Practically, suppose that two agents i and j are respectively in state $\mathbf{x}_i^{(t)}$ and $\mathbf{x}_j^{(t)}$ at time t . These states are both binary strings of length T . The agents choose a random integer c in the interval $[1, T]$ and both update their state as

$$\mathbf{x}_i^{(t+1)} = \arg \max_{x \in \{\mathbf{x}_i^{(t)}, \mathbf{y}_1, \mathbf{y}_2\}} f(x) \quad (3.13)$$

$$\mathbf{x}_j^{(t+1)} = \arg \max_{x \in \{\mathbf{x}_j^{(t)}, \mathbf{y}_1, \mathbf{y}_2\}} f(x) \quad (3.14)$$

where $\mathbf{y}_1$ is the vector with the c first entries of $\mathbf{x}_i^{(t)}$ and the $T - c$ last entries of $\mathbf{x}_j^{(t)}$ and $\mathbf{y}_2$ is the vector with the c first entries of $\mathbf{x}_j^{(t)}$ and the $T - c$ last entries of $\mathbf{x}_i^{(t)}$.

The authors observe with this model that the average score on all agents

obtained in the real dataset is smaller than for a random topology, which is in line with similar known results from population genetics. Also, perturbation of the time sequence of calls produces a small enhancing of the global fitness.

Viral marketing In 1955, Katz and Lazarsfeld introduced the breakthrough idea that, more than mass media, the neighbourhood of an individual is influencing their decisions [97]. This idea has induced the concept of opinion leaders –persons who have a high influence on their neighbourhood–, although some debate exists on the exact role played by opinion leaders [191], and introduced the concept of viral marketing. In opposition to direct marketing, the principle of viral marketing is that consumers respond better to an information accessed from a friend than to information provided through direct means of communication. Viral marketing searches thus for means of making people communicate about a brand, in order to push friends of an early adopter to adopt the product in turn. In particular, mobile viral marketing has proved to be an effective mean of propagation of such marketing campaigns. The influence of one’s neighbours can be observed using CDR data coupled to data on product adoption. In a study of the adoption of 4 mobile services, Szabó and Barabási [180] showed that the adoption of a product by a user was highly correlated to the adoption of their neighbours for some services only, while other services were not showing any viral attribute. A similar study by Hill *et al.* [83] on the adoption of an undisclosed technological service showed again that neighbours of nodes who had adopted the service were 3 to 5 times more likely to adopt the service than the best-practice selection of the company’s marketing service. A related result was also obtained in the FunF project by Aharony *et al.* [2], who showed that the number of common installed applications was significantly larger for pairs of users having often physical encounters.

The identification of “good” spreaders for a viral marketing campaign is a tough work, especially given large datasets, where it is hard to extract informational data in a small time frame. Having this in mind, the authors of [29] proposed a local definition of *social leaders*, nodes which are expected to play an influential role on their neighbourhood. They defined the social degree of a node as the number of triangles in which the node participates, and social leaders as nodes that have a

higher social degree than their neighbours. This definition has its use in marketing campaigns, to identify the customers where to start the campaign, which proved to be efficient [55]. Moreover, social leaders can also be used to reduce the complexity of a network, by only considering the network of social leaders, with possible uses in visualisation and community detection.

Mobile viruses The study of virus propagations has a long history, may it be biological viruses or more recently computer viruses. Wang *et al.* [188] studied recently a new kind of virus, which spreads over mobile phone networks. Their work is motivated by the increase of smartphones, which have high-level operating systems like computers, which leads to a higher risk of an outbreak. For now on, despite the large number of known *mobile viruses*, no real outbreak has been noticed yet. The reason for this is that mobile viruses function only on the operating system for which they are designed for. An infected phone can hence only transfer the virus to his contacts running on the same operating system. As exposed by Wang *et al.*, this situation corresponds to a site percolation procedure on the network of possible contacts. Given the actual market shares of the main operating systems, the authors showed that those were below the percolation transition of the contact network. The study concerns two types of spread available for viruses: the diffusion via Bluetooth and via Multimedia Messaging System (MMS). Both diffusions show major differences of spreading patterns; Bluetooth viruses spread relatively slow and depend on user mobility. In contrast, MMS epidemics spread extremely fast and can potentially reach the whole network in short times, but currently they are contained in small parts of the network, due to the different operating systems. As a conclusion, the authors deduce thus that if no outbreak has taken place so far, it is not due to the lack of efficient viruses, but it is rooted in the fragmentation of the call graph. However, the current evolution of the market leads to a situation where some operating systems are gaining a large market share, which could lead to a risky situation.

In a subsequent study, Wang *et al.* [189] show how the scanning technique, where MMS malware generate random phone numbers to which they try to propagate instead of using the address book of their host, increase the probability of a major outbreak, even when the market share of operating systems are too low for having a giant component.

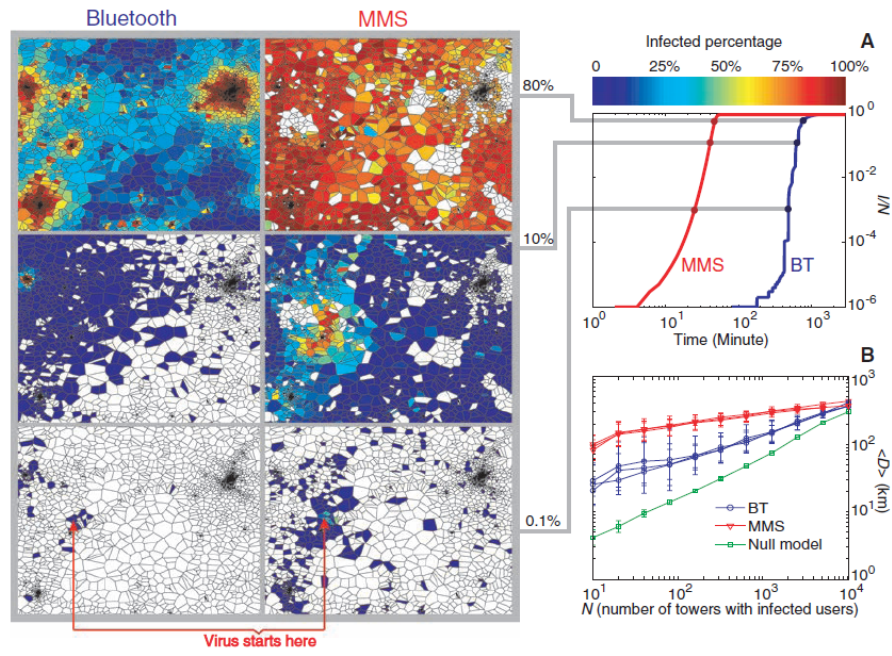


Figure 3.17: Propagation of a mobile virus, either via MMS or Bluetooth service, over the observed area. Figure reproduced from [188].

Operators can detect such outbreaks by monitoring the MMS traffic of their network and observe suspicious increases of volume. However, given enough time, viruses can infect a large fraction of the network without being detected by operators. Smart anomaly detection schemes may prevent such outbreaks, as well as a reduction of market shares of operating systems.

3.7 Privacy issues

The collection and availability of personal behavioural data such as phone calls or mobility patterns raises evident questions on the security of users' privacy. The content of phone calls or text messages is not recorded, but even the simple knowledge of communication patterns between individuals or their mobility traces contains highly personal information that one typically does not want to be disclosed. The recent raise of amount of personal data made available to researchers, via among other CDR datasets has carried on a number of research on the threads of such availability and their possible countermeasures.

The procedure often used when a company shares private data with a third party such as a research group is the following: the company keeps on secured machines the exact private information such as names, addresses or phone numbers on their customers, as well as the CDRs, which contain the phone number of the caller, the callee, the time stamp of the call, the tower at which the caller was connected, idem for the callee, and additional information such as special service usage and so on. The anonymisation procedure consists then to identify each user by a randomly generated number, such that each user has a unique random ID, from which it is impossible to retrieve the original phone number by reverse engineering procedures. The CDRs are then modified such that phone numbers are replaced by the corresponding ID. After this procedure, the CDRs are anonymised, and can be transferred to a third party. Standard procedure then imply that the third party signs a non-disclosure agreement, stipulating that it cannot make the CDR data available, and usually also restricts the range of potential research questions to be dealt with the data. The safety of users privacy is then guaranteed both by the removal of information allowing to identify users and by the assumption that the third party doesn't make use of

the data for any malicious intent.

De-anonymisation attacks Some research has been produced on mobile phone datasets to challenge this apparent feeling of security, however, recent results are opening new ways of considering the privacy problem. Using CDR data containing mobility traces, Zang and Bolot [194] show how it is possible to uniquely identify a large fraction of users with a small number of preferred locations. Their methodology goes as follows: for each user, it is possible to list the top N locations at which calls have been recorded. The authors show then that depending on the granularity of the locations, a non-negligible fraction of users may be uniquely identified by only 2 locations. For example, if locations are taken at cell level, up to 35% of the users of a 25 million communication network can be uniquely identified with 2 locations, which will be likely to correspond to home and work. Thus, while the anonymisation procedure is intended to impeach any linkage between the dataset and individuals, using this procedure allows to potentially retrieve the mobility and calling pattern of targeted users given the access to as little information as home and work addresses. If additional data, such as year of birth or gender of users would be available – which is common in most datasets – it would be possible to identify large fractions of the network. In the previous attack schemes, the exact timing of calls has not been taken into account. Again, using the timing of each call to reconstruct the path followed by a user would produce even a higher risk of de-anonymisation, since it would be easy to recognize significant locations, such as home location (mostly spent at early morning, in the evening and at night) or work.

Other possible attacks have also been considered on anonymised online social networks. Although those attacks are not likely to be applied in the case of mobile phone data, we quickly mention some of them, as it is likely that breaches found in different applications might be similar to potential breaches in mobile phone datasets.

For example, Backstrom *et al.* [9] describe a family of local attacks, which enable to retrieve the position of some targets in the network, and hence to uncover the connections between those patterns. The authors showed that on a network of 4.4 million nodes, by controlling the links of

7 dummy nodes they manage to uncover the presence or absence of 2,400 links between 70 target nodes, without being detected by the database manager. On a wider scale, Narayannan and Shmatikov [136] show that it is possible to retrieve the identity of a large part of a social network by combining it with an auxiliary network. Such a situation happens when users are present in two separate datasets. The authors show then that even if this overlap is available for only a fraction of the users, it is still possible to retrieve the information for a large part of the network.

Against these possible threats of privacy breach, one may wonder if solutions are proposed to counter such attacks. If research on mobile datasets only consider average behaviours, rather than exact patterns, a simple countermeasure is to perform small modifications of the dataset, that would not alter the general aspect of it but that would have dramatic consequences on the algorithms used by attackers, who search for exact matchings between statistics on the network and a priori known properties of the targets.

Another protection against such attacks, and particularly when mobility data is involved, is to produce new random identifiers for each user at regular time intervals. By regenerating random identifiers, it makes it impossible to use longitudinal information in order to assess the preferred locations of a user. As shown by Zang and Bolot [194], by changing every day the ID of each user, only 3% of the nodes can still be identified using their top 2 locations. While this method seems efficient to protect the privacy of users, it reduces substantially the possible information to retrieve from such a dataset for research purpose.

Personal data: ownership, usage, privacy Phone companies collect data about their users, about their habits, their mobility, their acquaintances. Still, the actual legislation on the ownership and usage possibilities is fuzzy [127], chilling companies to share such data for research and making customers feeling as if George Orwell's predictions come true.

Such data represents an enormous added value, both to companies for marketing purposes and client screening, as to authorities for traffic management or epidemic outbreak prevention. It is often forgotten, but the use of mobile phone datasets also has a huge positive potential in the

developing world, may it be for supervising the health status of populations, generating census data or tracking the dynamical evolution of slums¹.

Such opportunities, both for corporates and authorities need to develop standardized procedures for the acquisition, conservation and usage of personal data, which is not the case here. The communication about these procedures to customers hasn't been clear, as are the possibilities for a user to "opt-out" if they don't want to have their personal data released.

With this intent, several voices have recently been raised in order to urge authorities to develop a "New Deal" [152] on data ownership, in which users would own their personal data as well as the decisions to provide it –in exchange of payment– to companies interested in their usage. A transparent system armed with the necessary protocols and regulation for a transparent use of personal data would also facilitate the access to data for researchers [62], and could so benefit to the entire society.

3.8 Conclusion and research questions

The first analyses of mobile phone datasets appeared in the late 90's, and the result of this decade of research contains a large number of surprises and several promising directions for the future. In this chapter, we have reviewed the most prominent results obtained so far, in particular in the analysis of the structure of our social networks, and human mobility. We decided not to cover some closely related questions, such as churn prediction (see [88, 51, 164, 58]) or dynamic pricing [70, 101], which are rather business-related topics, and for which a vast literature is available.

The recent availability of mobile phone datasets have led to many discoveries on human behaviour. We are not all similar in our ways of communicating, and differences between users can range to several orders of magnitudes. Our networks are clustered in well-structured groups, which are spatially well-located. With the raise of communication technology, some have predicted that the barrier of distance would fall, shrinking the world into a small village. However, mobile phone data suggests instead that distance still plays a role, but that its impact

¹See examples of projects on <http://AI-D.org>

is nuanced by the varying population density. Regarding our mobility behaviour, individuals appear to have highly predictable movements [14], while as populations we act and react in a remarkable synchronized way. In this context, the availability of mobile phone data has for first allowed to observe populations from a God-eye point of view, monitoring the pace of daily life or the response to catastrophes.

The ubiquity of mobile phones – there are nowadays more mobile phones than personal computers in use – which allows us to obtain such precise results raises also the thread of viral outbreaks, from which mobile phones have been safe until now. Mobile viruses could be a potential risk for users' privacy, as it is also the case that the anonymised datasets provided by operators to third parties for research could potentially be deanonymised too.

The availability of such enormous datasets creates a huge potential that could benefit to society. The research that has been conducted so far only represents the tip of the iceberg of what could potentially be done, when adequately exploited. However, it is the necessity of authorities to ensure that such datasets could not be misused.

Further research The number of possible research questions on mobile phone datasets is gigantic. In this last part, we will present one research direction that we believe to be highly important and still not addressed in its most general form.

A large number of research has been conducted on the analysis of social networks, based on CDRs. As it appears from the different publications on this topic, there exist some common features but also many differences in the structure of the constructed network. Recall as simplest example the degree distributions, which show different functional forms for most datasets.

These differences may, of course, be linked to cultural differences between the different countries of interest, but there are probably other, quantifiable, reasons. The datasets differ greatly in the market shares of the operators, in the time span of the data collection period, in the size of the network and in the geographical span of the considered country. The method of network construction is also always different and has a tangible impact on the network structure. The use of directed or

undirected links, weights and thresholds for removing low-intensity or non-mutual links all greatly impact the structure and hence the statistical features of the obtained network.

Hence, we believe that a serious analysis, both on theoretical and on empirical side of the influence of these factors on the general structure of mobile phone networks may lead to a general framework, allowing to interpret differences between results obtained on several datasets with the knowledge of potential side-effects.

This question is closely related to the even more general question of the significance of information provided by CDR data. Recalling what was said in Section 3.2, CDR datasets are noisy data, some links appear there by chance, while other have not been captured in the dataset. It would thus be interesting to question the stability of the obtained results, provided that the real network is different from what has been observed in the data. This links with the recent work of Gourab [75], who analyzed the stability of PageRank under random noise on the network structure. Again, in this framework, no real theoretical result has been achieved, allowing to characterize which result is significant, and which is not.

Chapter 4

Geographic distribution of networks

In this chapter, we study the geographical distribution of a **large mobile phone network**. This research results from a collaboration with a Belgian mobile operator, and treats two aspects of the relationship between space and mobile phone networks. First, the communication between cities is modelled. Second, the way internal communications, i.e. communication between people of the same city, scales with city size is in turn analysed. Finally, the influence of external factors, that create small deviations from these general laws will be illustrated.

4.1 Introduction

Motivation There are two positioning data that can be used with mobile phone datasets: *real-time location* and *centralized location*.

Real-time location data consists either of GPS traces of mobile users, or of the sequence of antennas at which the user has made or received its calls. Such datasets allow to infer laws on the displacement of users [78], to estimate the predictability of their displacement [176] or to detect special events happening in a city and the origin of their participants [40].

Centralized location data assigns one fixed location to every customer, typically its home or office location. This data is extracted from billing information, where the ZIP code of the billing address is sometimes

available. The idea behind the use of such data is that the activity of a customer is centralized in the given location, and allows then to perform different studies, such as the geographical dispersal of social ties [115]. Datasets – other than mobile phones – using centralized locations have been used in various applications, such as for modelling epidemics propagations [11, 47], studying the dispersal of internet or airline traffic [73, 18] or inferring laws on commuting traffic [94, 11]. However, in the purpose of mobile phone analysis, these datasets have been a bit underused, essentially due to the lack of available data.

In this chapter, we analyse a large dataset with centralized location data and explore subsequently the influence of geography on the communication *between* Belgian communes, then the communication *inside* those communes. Finally, we show how the influence of geography and other external parameters may be visualized with an analysis of the *communities* of Belgian communes.

Notations We consider two directed weighted networks G and G' . The microscopic network $G = (N, L)$ is the network of customers. The macroscopic network $G' = (N', L')$ is the network of communes. In both networks, links may be weighted in different ways (in the following, x and y denote either customers or communes):

l_{xy} is the number of links going from node x to node y . In the microscopic network, this variable is binary. In the macroscopic network, it is the number of links in the microscopic network for which the originating node is associated to commune x and the destination node associated to commune y .

n_{xy} is the number of calls passed from node x to node y . In the macroscopic network, it is the total number of calls passed from customers associated to commune x to customers associated to commune y .

t_{xy} is the total duration of calls from node x to node y . In the macroscopic network, it is the sum of the length of all the calls from a customer associated to commune x to a customer associated to commune y .

w_{xy} is the average length of a call from node x to node y . In the macroscopic network, it is the average length of all calls from a customer associated to commune i to a customer associated to commune j . The average is taken over the total number of calls.

In addition, we have the following data available:

r_{ab} is the distance (in kilometers) between the center of commune a and the center of commune b .

c_a is the number of customers associated to the commune a , which we will refer to as the population of the commune.

4.2 Presentation of the data

The data that we consider consists of the communications made by more than 2.5 million customers of a Belgian mobile phone operator over a period of 6 months in 2006 [115]. Every customer is identified by a surrogate key and to every customer we associate its corresponding billing address zip code. In order to construct the communication network, we have filtered out calls involving other operators (there are three main operators in Belgium), incoming or outgoing, and we have kept only those transactions in which both the caller and the callee are customers of the mobile phone company. In order to eliminate “accidental calls”, we have kept links between two customers i and j only if they have made at least six calls during the six months time interval and that there has been at least one call in each direction. The resulting microscopic network is composed of 2.5 million nodes and 38 million links.

In order to analyze the relationship between this social network and geographical positioning, we associate customers to communes based on their billing address zip code. Belgium is a country of approximately 10.5 million inhabitants, with a high population density of 344 inhab./km². The Belgian National Institute of Statistics (NIS) [21] divides this population into 589 communes, whose sizes show an overall lognormal population distribution with approximate parameters $\mu = 4.05$ and $\sigma = 0.37$ ¹.

The analysed communication network provides information for the operator’s customers rather than for the entire population. However, the

¹The lognormal distribution of city size is consistent with similar data on US cities [67].

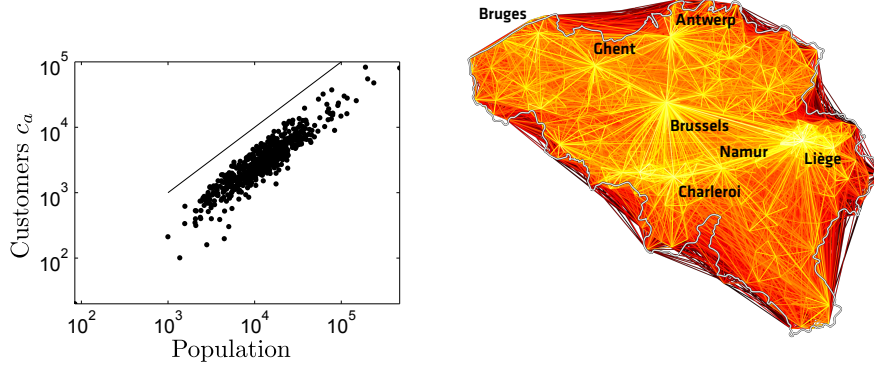


Figure 4.1: (left) The fraction of customers (black dots) is stable in all communes. The black line shows a linear increase, and is intended as a guide to the eye. (right) The Belgian network, colors represent the total time of communication between communes. Only the 20% strongest edges are shown.

average market share of the operator in the Belgian communes is stable (see Fig. 4.1, left) and so this suggests that our dataset is not structurally biased by particular user-groups and market shares. In the following, we always compare the intensities of calls made between two communes to the number of customers that are associated to them rather than their population. In this way, the possible bias induced by the market shares of the operator have no impact on the results observed in this study. To avoid confusions in the rest of this chapter, when we use the term *population* of a commune, we are referring to the number of *customers* associated to it, even if they are not connected to any other customer in the microscopic network of human-to-human communications defined above. These customers are still taken into account for the population size, since their presence is of interest for the normalization of the communication data.

The distance between each municipality was inferred from the geographical coordinates of their centres. This results in a network of 589 nodes and 167,646 weighted directed

links. Each node corresponds to a commune of Belgium, identified by its code (as given by the NIS), the weights of each link can be weighted following the weights presented in the previous section.

4.3 Gravitational law for inter-city communication

Many studies have been made on human-to-human communications but few analyses are available on how these communications, once aggregated at the city level, are related to the properties of that city. In the following, we model the communication intensity between cities as a function of the population sizes and of the distance between them.

First, we analyse how communication flowing into and out of cities, scale with population size. For doing this, for each of the 589 cities we compare the total incoming (t_{*a}) and outgoing (t_{a*}) communication intensities, as defined below, to the population sizes of these cities.

$$t_{*a} = \sum_{i \notin a, j \in a} t_{ij}, \quad t_{a*} = \sum_{i \in a, j \notin a} t_{ij}. \quad (4.1)$$

As shown on Figure 4.2(a), both incoming and outgoing inter-city communication intensities scale linearly with city size ($t_{a*}, n_{*a} = kc_a^\beta$, $\beta = 0.96$, confidence interval: $[0.93 \ 0.99]$, $R^2 = 0.87$). Also, incoming and outgoing communications are strongly symmetric ($t_{a*} \approx t_{*a}$, $\forall a$), that is, calls in one direction usually find a match in the opposite direction.

Another parameter that influences communication intensity between cities is distance. It seems reasonable to expect that the intensity of communication between two cities will decay when distance increases and other parameters are kept unchanged. This idea is supported by several studies that suggest gravity-like models for car traffic [94], trade [24] or economic activity [182]. In all these cases, the intensity at distance r is proportional to r^{-2} . For the case of communication, a similar model has been presented in [196], but with an intensity evolving like r^{-1} , though the authors acknowledge that their model does not seem to fit well with the data. Explanations for such a power-law decrease of the communication intensity with respect to distance have been proposed, in particular, Hu *et al.* [87] propose a network model where each node has

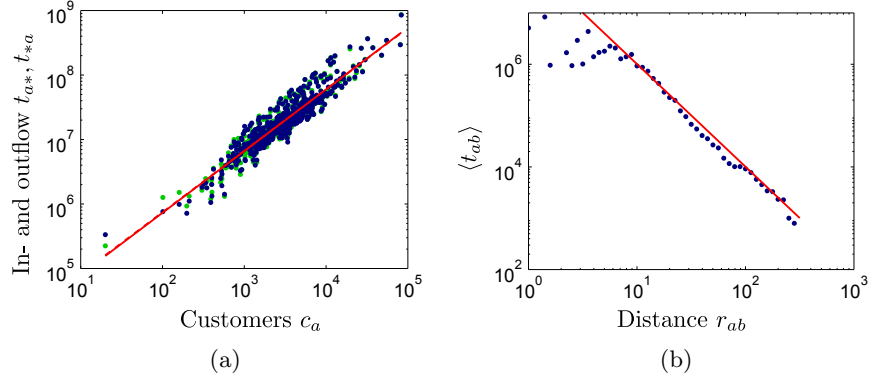


Figure 4.2: (a) Relation between the outgoing intensity t_{a*} (blue), the incoming intensity t_{*a} (green) and the number of customers in the commune c_a . The red line shows the best least squares fit. (b) Dependence of the average communication intensity between pairs of communes and the average distance separating them. The red line is intended as a guide to the eye, and shows a r^{-2} decrease.

a limited amount of resources, and where each link costs an amount of resources proportional to the geographic distance between both nodes. If nodes connect to their neighbourhood so as to maximize the information they can collect from their neighbours, the optimal network shows the same power-law decrease as observed in real networks.

To test the influence of distance on the Belgian network, we measure it as the distance between the centroids of each communes area. The communication intensity between two communes is then compared with the distance between them, where a power law decrease, with an exponent close to -2 is shown (see Figure 4.2(b)). The power law fits very well for inter-city distances larger than 10 km.

These results suggest that the communication between communes is ruled by a gravity model, which is symmetric, scales linearly with the

population sizes and decreases with r , with an exponent close to -2 :

$$t_{ab} = K \frac{c_a^{\beta_1} c_b^{\beta_1}}{r_{ab}^{\beta_2}}, \quad (4.2)$$

there, the scaling constant K is the gravity constant for a timespan of 6 months of calling activity, β_1 and β_2 are the exponents of respectively the size of the commune and the distance that separates them. To retrieve the parameters K , β_1 and β_2 , we take the logarithm of all considered quantities. By writing $\beta_0 = \log K$, it reads

$$\log t_{ab} = \beta_0 + \beta_1(\log c_a + \log c_b) - \beta_2 \log r_{ab} \quad (4.3)$$

The parameters are then easily calculated with a least squares regression. It provides the following results: $\beta_0 = 0.01$, $\beta_1 = 0.96$, $\beta_2 = 1.67$. The value of β_2 diverges from 2, due to the deviation of the gravity law at low distance. This issue will be further developed in the next section. If the extreme values of distances are neglected ($10 \text{ km} \leq r \leq 100 \text{ km}$), the fit to the gravity law is better ($\beta_0 = -0.01$, $\beta_1 = 1.00$, $\beta_2 = 1.87$).

To ensure the validity of our results, we plotted the estimated intensity given by the gravity model versus the observed intensity. As shown on Figure 4.3, the model fits well to the dataset above a certain threshold. It is possible to provide explanations for prediction errors for low intensities. First, those correspond to a very small sample of the dataset (the first 4 points correspond to respectively 2, 4, 163 and 1,745 pairs out of the 167,045 pairs that are used to build the model). Secondly, this overestimation corresponds to pairs of communes for which the predicted amount of communication is lower than 1000s in 6 months. For those low values (where the model predicts the communication to be from 10s to 1000s) one may expect the errors to correspond to a “discretization error”, in the way that there is a lower bound on the intensity of communication between people – *if* a communication occurs between two customers, it will last *at least* a certain amount of time. For the general case (which accounts for about 99% of the dataset) the gravity law fits well to the observed average trend.

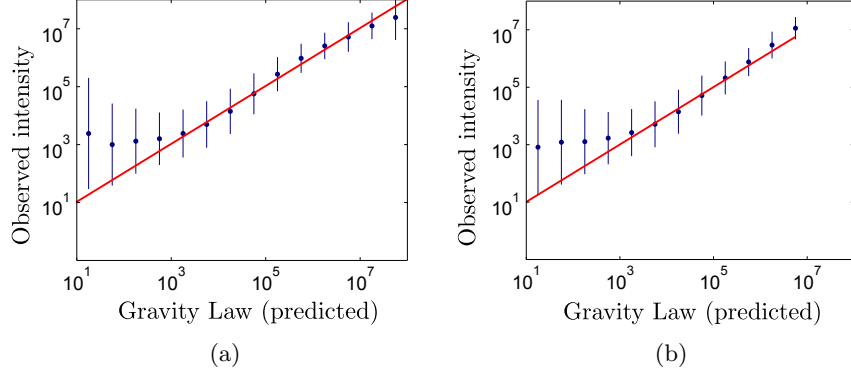


Figure 4.3: Predicted and observed communication intensity (in seconds). (a) Fit on the complete dataset. (b) Fit on data for intermediate distances ($10 \text{ km} \leq r \leq 100 \text{ km}$). The red line shows the fitted prediction, the vertical lines show the standard error.

4.4 Scaling properties in communication networks

In the previous section, we have focused on the communication *between* communes. Now, we will consider the communication *inside* those communes. In a recent study, researchers have shown that several indicators of wealth, energy consumption, or productivity scale with the size of cities [23]. Scaling properties have been widely studied in the context of living entities, the first discovery being that the metabolic rates of mammals scale with masses with an exponent of $3/4$ [102]. In the case of cities, the scaling exponent depends on the nature of the indicator; energy consumption indicators scale sublinearly ($\alpha < 1$) which reflect economies of scale, while indicators of productivity scale superlinearly ($\alpha > 1$). In a larger city, there will be a larger productivity *per capita*. Similar results are obtained with cell phone data using centralized location data (Figure 4.4). The total number of edges, calls or the total time of communication inside a commune a (so, between two customers belonging to the same commune) can be expressed as a power of its number of customers c_a . The explanation of these exponents is not immediately trivial. In the following, we use a simple model of a geographical network

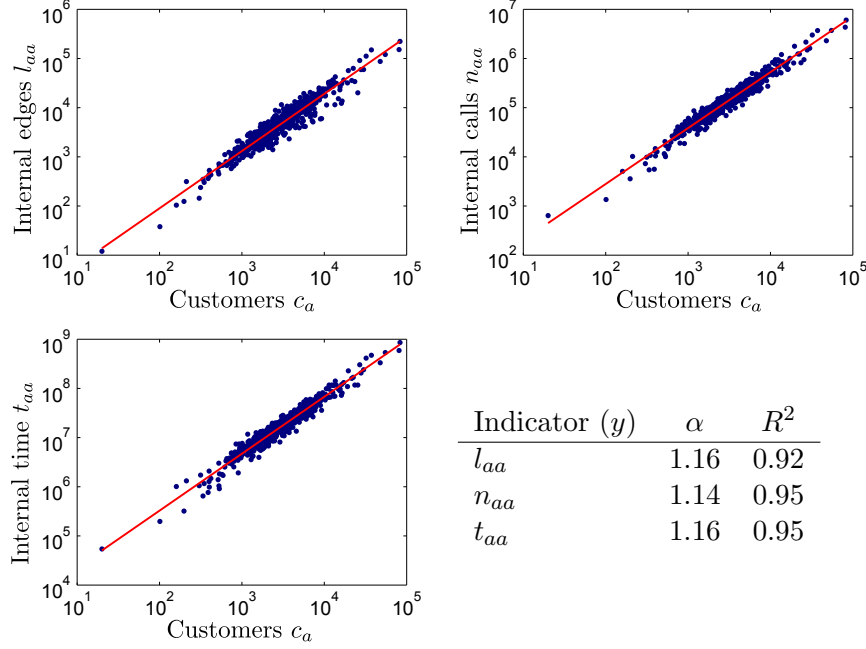


Figure 4.4: Three different weights scale with the number of customers inside the commune with the relationship $y = bc_a^\alpha$

to explore under which conditions such a scaling can occur.

Geographical network model Let us consider a network where nodes (customers) are distributed over space at random according to a given population density function $h(x, y)$. The plane is partitioned into a set of connected tiles (by connected we mean that one tile cannot be made of two disconnected parts), representing the communes. In the following, we will abuse notations by denoting by $h_a(r, \theta)$ the function $h(x, y)$, written in radial coordinates and centred around the center of commune a . Finally, the network is constructed by placing links randomly between nodes following a given probability function $p^{conn}(r)$, which should depend on the distance separating the customers, and their respective communes.

We intend to calculate the expected number of edges inside a commune, given by the sum of the expected degrees of the customers in the com-

munne

$$l_{aa} = \sum_{i \in a} d_i = c_a \langle d \rangle_a$$

The average degree of a node inside a commune $\langle d \rangle_a$ is obtained using the probability $p^{conn}(r)$ of making a contact at distance r and the density of population $h(r, \theta)$. Here, we suppose that the probability of making a contact solely depends on distance, and is not influenced by the orientation θ . By approximating the position of the average user at the centre of the commune, the average degree reads

$$\langle d \rangle_a = \int_0^{2\pi} \int_0^\infty h_a(r, \theta) f_a(r, \theta) p^{conn}(r) r dr d\theta \quad (4.4)$$

The function f_a is, like h_a , centred at the centre of commune a and defines the shape of the commune:

$$f_a(r, \theta) = \begin{cases} 1 & \text{if the point } (r, \theta) \text{ is in commune } a \\ 0 & \text{otherwise} \end{cases}$$

In this framework, the population of a commune is completely defined by its shape $f_a(r, \theta)$ and its population density $h_a(r, \theta)$.

Influence of shape and population density The variety of available functions for f_a and h_a makes it impossible to test out all cases. In the following, we will focus on two realistic test cases. First, we will simplify the equation by supposing that the communes have a circular shape of radius R_a , that may depend on the commune. Second, we will assume that population density follows the same radial profile for each commune, up to a multiplicative constant. This is, $h_a(r, \theta) = h_a g(r, \theta)$, where g can be any function and h_a is a constant that may depend on the commune. Let us now see how those two parameters influence the scaling of the average degree with respect to population.

Supposing that **the radius of each commune is constant** ($R_a = R \ \forall a$), the population of a commune is proportional to its density ($c_a \propto h_a$). In such a case, equation (4.4) simplifies to

$$\langle d \rangle_a = h_a 2\pi \int_0^R g(r) p^{conn}(r) r dr d\theta. \quad (4.5)$$

The integral is constant for all communes, thus the average degree scales linearly with population size, and the number of internal edges scales quadratically with c_a .

The opposite case happens when **the population density is identical in each commune** ($h_a(r, \theta) = hg(r, \theta) \quad \forall a$), while the radius of communes varies. The exact relationship between R_a and c_a depends on the nature of g , but given the assumption that the connection probability $p^{conn}(r)$ is a non-increasing function of r , one can derive a general idea of the scaling of the average degree with c_a .

The average degree reads in this case

$$\langle d \rangle_a = h \int_0^{2\pi} \int_0^{R_a} g(r) p^{conn}(r) r dr d\theta. \quad (4.6)$$

Its dependence on R_a is then given by

$$\frac{\partial \langle d \rangle_a}{\partial R_a} = 2\pi h g(R_a) p^{conn}(R_a) R_a, \quad (4.7)$$

$$= \frac{\partial c_a}{\partial R_a} p^{conn}(R_a). \quad (4.8)$$

Thus, the ratio between $\frac{\partial \langle d \rangle_a}{\partial R_a}$ and $\frac{\partial c_a}{\partial R_a}$ decreases with R_a , as long as the probability of connection decreases with distance. In this case, the average degree of a commune scales sublinearly with population, and the scaling exponent of l_{aa} ranges between 1 and 2, depending on the shape of $p(r)$ (the exponent is equal to 2 if $p(r)$ is a constant function and equal to 1 if all nodes make a fixed number of contacts in their commune, independently of population).

These two cases suggest that one should observe a situation close to the second case for the Belgian communes. Communes with a large population should cover a wider area than smaller ones. However, one can see on Figure 4.5(a) that the area of communes is almost constant ($\alpha \approx 0.2$) with population size (data provided by the NIS), hence the population density increases almost linearly with population size. The model proposed here is thus not sufficient to explain the scaling exponents observed in Figure 4.4. In the following, we show how a small modification of the model is sufficient to bridge the gap.

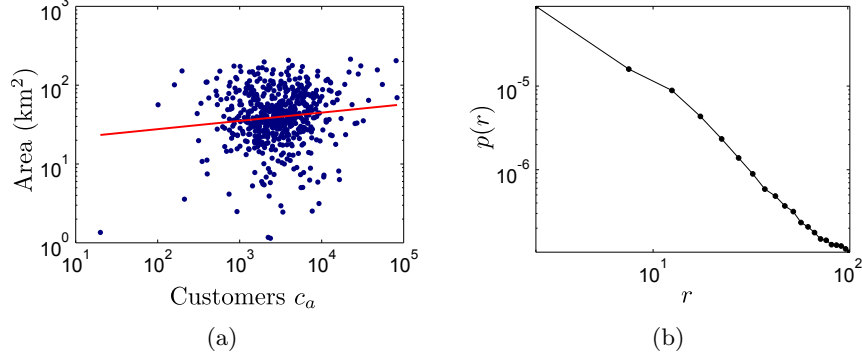


Figure 4.5: (a) Area of Belgian communes versus their population. There is almost no correlation between both variables ($\alpha = 0.1$, $R^2 = 0.1$). (b) Connection probability for two customers versus of the distance separating them (reproduced from Lambiotte et al. [115]). The distribution is close to a power law with exponent 2.

In the model, we have supposed that the probability of two customers to connect solely depends on the distance that separates them. This assumption is supported by the previous analysis by Lambiotte et al [115] who have shown on the same dataset that the probability of two users to connect is well fitted by a gravitational law ($p^{conn}(r) \propto r^{-2}$, see Figure 4.5(b)). However, we obtain a different picture by segmenting customers based on the population size of the commune they are associated to. We divide customers in different classes based on the population size of their commune, and for each class, we compute the probability that customers from that class connect to any other customer located at distance r . If customers of different classes behave on the same way, these probabilities should be identical for each class, and correspond to the curve shown in Figure 4.5(b). Instead, as is shown on Figure 4.6, we see a clear shift in the connection probability; customers associated to communes of small sizes have a larger connection probability than customers associated to communes with a high population. This is particularly noticeable at short distance, which we highlight in the the inset of Figure 4.6, where we show that the probability to connect with a customer located at less than 5 km (y-axis) decreases with the population size of the commune

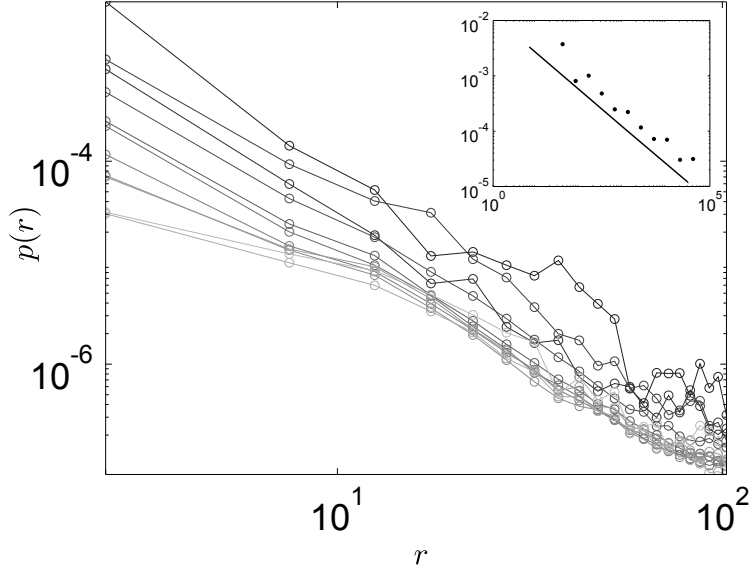


Figure 4.6: The main figure shows the probability for a customer to connect with a customer located at distance r . The customers have been segmented based on the size of the commune to which they are associated. Hence, the darkest line shows the probability for customers associated to communes of small size to connect to any other customer located at distance r . The brightest line shows the same probability for customers associated to communes of large size. Inset: Probability for a customer to make a link with a customer within a range of 5km, as a function of the size of the commune of origin. The black line shows a slope of $-2/3$, and is intended as a guide to the eye.

of the caller (x-axis).

This deviation of the connection probability at short distance can be explained in terms of number of intervening opportunities. Let us compare two customers, associated to two different communes, such that one commune has a ten times larger population than the other. Then, the customer associated to the large commune has ten times more possible connections at short distance than the other customer. If both customers make their connections following the same gravity law, the cus-

tomer from the large commune would have ten times more connections in his commune than the customer associated to the smaller commune. It is clear that this situation cannot hold in a real situation. Hence, the customer associated to the smaller commune has a larger probability to connect at short distance than the customer associated to the larger commune. We write this dependency of connection probability on size as $p^{conn}(r < 5 \text{ km}) \propto c_a^{-\delta}$. The value of δ determines then how different both customers from the example behave. If $\delta = 0$, the gravity law is the same for both customers, and the customer associated to the larger commune has ten times more connections in his commune than the customer in the smaller commune. If $\delta = 1$, both customers make the same number of connections in their respective commune. In the inset of Figure 4.6, we see that from the data δ is close to $2/3$.

In the following, we show how, by taking the deviation of the connection probability into account in the model, we manage to explain the scaling exponents observed in Figure 4.4.

Influence of a modified connection probability If we change the connection probability so as to make it depend on the population size of the commune $p_a^{conn}(r) = K_a p^{conn}(r)$, where K_a decreases with c_a as $c_a^{-\delta}$, and retain again the hypothesis of constant radius, equation 4.4 becomes

$$\langle d \rangle_a = d_a K_a \int_0^{2\pi} \int_0^R g(r) p^{conn}(r) r dr d\theta, \quad (4.9)$$

and thus l_{aa} scales with c_a as $c_a^{2-\delta}$. Here, δ is the exponent that is empirically observed in the inset of Figure 4.6 (in this case, $2/3$).

The deviation of the probability of connection explains well the scaling exponent. In particular, let us point out that the connection probability only deviates for small distances, at which the local population density plays an important role. At longer distances, the population density becomes independent of the commune of origin, which is why the connection probabilities become identical for different commune sizes.

To conclude, let us notice that the deviation on of the classical gravity law due to population density has already been observed in different

datasets, and is a long-standing subject of debate in geographical sciences. In studies on human mobility, two diverging viewpoints on how distance should be considered have emerged. One camp argues that mobility is driven by costs, hence by geographical distance, while the other camp considers instead that the distance does not play any role in mobility, but instead the number of *intervening opportunities* [179] that separates them. This has led to the rank-based distance (see [124, 143] and references therein) that measures the distance between two nodes i and j as the number of people that are closer from i than j . The usage of a rank-based distance instead of a gravity law would make sense in the case under study, however, due to a lack of fine-grain data, it is hard to be applied on the dataset under study.

4.5 External factors influencing geographical networks

The general laws described in the two preceding sections fit well to the *average* behaviour of the network. However, the underlying process, namely people making calls, is driven by many external factors other than distance and population density that may influence locally the structure of the network. For example, the presence of a river or a mountain range might create a physical –and thus sociological– barrier between two towns, that is not taken into account by the geographical distance. For this particular issue, researchers have attempted to tackle it by using the commute distance (the time to e.g. drive from one city to the other) rather than geographical distance.

Another factor of crucial importance is the cultural difference between populations. This factor has not been studied so far using mobile phone datasets, since those datasets are usually only available for one single country, and that cultural difference is not easily measurable. For such a case, Belgium is an interesting country, given its particular socio-demographical configuration.

Belgium is geographically divided in three linguistic communities: a Flemish-speaking part in the north, a French-speaking part in the south and a small German-speaking part in the east. The only exception

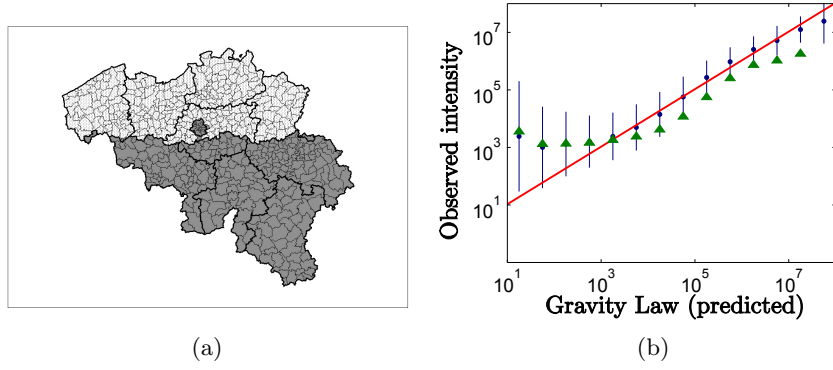


Figure 4.7: Influence of language on communication. (a) Partition of the Belgian communes in two communities, based on the average communication length. (b) Deviation from the gravity law for communes of different linguistic groups. The blue circles and the red line show the same information as in Figure 4.3(a), while the green triangles show the average communication intensity for links that go from the dark community to the bright community and vice-versa.

is Brussels, which is bilingual and located inside the Flemish-speaking part, close to the *linguistic border* with the French-speaking part. We can thus easily quantify the cultural difference by associating to each commune the language of its inhabitants, and observe how the communication varies inside or between cultural groups.

In Figure 4.7(a), we visualize the influence of language by showing a partition in communities of the network of communes when the links are weighted by the average duration of a call. The communities have been obtained with the Louvain method, which is a heuristic method for approximating the partition of largest modularity [32]. The partition shown here is a very strong one: it is independent of the node ordering, and has also been obtained with different community detection methods.

Strikingly, the optimal partition separates the country in two communities, and the separation follows almost perfectly the linguistic border. In this partition, Brussels is part of the southern community as well as

8 Flemish-speaking communes, which are known to have large French-speaking populations.

This result is not a surprise, since language difference is known to be a strong barrier to communication. With this partition, 98% of the links in the network are located within the communities, and only 2% go from one community to the other. A similar separation has already been observed at the level of the customers by Blondel et al [31].

The impact of language on communication is important: on Figure 4.7(b) the gravity law, as presented in Section 4.3 is shown, where the green triangles represent the average communication intensity of the edges going from one community to the other. On average, the intensity of communication is significantly lower than what predicts the gravity law based on distance and population.

4.6 Conclusions

There is a complex interplay between the way we communicate and the surrounding geography. Some general trends can be well described by simple statistical laws, but those laws are valid only for large average behaviours of populations.

Results In this chapter, we have shown with a wide-scale dataset how interpersonal communication is influenced by the surrounding geography, and how, once aggregated to commune level, the intensity of communication can be well-described by a **gravity-like law**. This law is valid for the average behaviour of communes that are expected to communicate at least 15 minutes in 6 months, which makes the model valid on average for about 99% of the observed pairs of communes.

By zooming inside each commune, we observed that intra-communal communication scales with population with a non-trivial exponent that can be explained by a **deviation of the connection probability** at small distance. Customers from small communes appear to have a higher probability to connect at small distance than customers from large communes, which indicates that a rank-based distance would fit better than

a geography-based distance.

Finally, the particularity that Belgium is a bilingual country, with two communities of different cultural background, allows to show the **influence of the cultural differences** on communications. To our knowledge, there are no other datasets that cover populations having such a cultural diversity, and where the cultural difference may be easily quantified. In the last section we have shown the influence of language, both by showing that a partition in communities separates the country in two communities based on language usage, and by showing that the intensities of communication between communes of different language are significantly lower than predicted by the gravity law.

Further work The variation of the connection probability with respect to the gravity law, observed in Section 4.4, indicates that the geographical distance is not an appropriate distance measure as soon as the country has a heterogeneous population density. When such a case is encountered, a rank-based distance is likely to be more appropriate. A rank-based distance measures the distance between two nodes i and j as the number of nodes that are geographically closer from i than j . In regions of high population density, the rank-based distance grows faster with geographical distance than in low-density regions. Such a distance has already been proposed by Liben-Nowell et al. [124], but cannot be easily implemented with the data available, which is too coarse-grained. Given some simple assumptions on the geographical shape of the communes and on their population density function, one should be able to obtain an analytical model of connection probability, and its link with the gravity model seems to be an interesting research topic.

In the last section, the influence of external factors, such as cultural differences has been shown. However, no proper analysis has yet been performed in this scope. Since the influence of external factors might be non-trivial, a simple way to visualize this is with the geographical dispersal of communities. Some early work has already been done, Brockmann et al. [37] noticed that mountain ranges appeared as the border of communities in a money circulation network. Sociological separations, as well as the influence of cities playing as economical attractors have

also been shown in Belgium [32], while Expert et al. have proposed models for community detection decorrelated from the influence of population and distance, which allows to highlight such external influences [69]. Finally, issues related to the sampling of the dataset should also be addressed. Given that any dataset represents only a fraction of the population behaviour, one should ensure that the results are not biased by a population structure that differs from the total population of the country under consideration.

Chapter 5

Time scales of dynamical networks

In this chapter, we address the issue of **time scales** in evolving networks. We first show evidence for the role played by time scales in the process of network growth, and illustrate the effect of time scales with a dataset of mobile phone communications. Secondly, we characterize the growth of the network with time, i.e. we propose an **analytical form** as well as a **stochastic model** that fit the observed number of links in the network versus the number of events used to build it. Finally, we present some exploratory work on how it is possible to quantify the **information** provided by additional events using the **entropy** of a growing network

5.1 Introduction

In this chapter, we study networks where links have a dynamical component. In practice, a large quantity of the available data on real-world networks intrinsically contains a dynamical component, let us cite the internet, the world wide web, social networks, phone call networks, citation and scientific collaboration networks, or airline networks. Most studies have addressed these networks as static entities, while in the recent years some research has been pursued to study the dynamics of such datasets, such as the evolution of communities [148], the persistence of ties [82] or the microdynamics of apparently stationary networks [74]. These studies have in common that they split their dataset into a series

of time windows, which are used to construct a sequence of networks. The dynamical component is then extracted from similarities and differences between consecutive networks.

Another common feature of these studies is that the length of the time windows used in the studies have been fixed in advance, often under technical constraints such as the temporal granularity of the collected data. One may question if the results obtained in these studies would have been similar if different time window lengths had been used. For many real-world complex networks, there are a number of mechanisms that are expected to contribute to the emergence of the network structure as a function of aggregation time, which don't all function following the same time scales. First, the distribution of link weights is often broad [146, 145]. Thus there are high-weight links that should on average be observed earlier on in the aggregation process, and many links of low weight that take a long time to be observed. Second, link weights are correlated with network topology, such that high-weight links are associated with denser network neighbourhoods [146]. Third, for links of any weight, it is known that for some datasets the distributions of inter-event times are also broad, *i.e.* event sequences are bursty [43, 96], giving rise to longer-than-Poissonian waiting times between events. Fourth, there are circadian patterns [93], where the overall level of human activity varies by hour, as well as weekly patterns where human behaviour depends on the day of the week. Fifth, there are naturally changes in the network itself too – connections grow and wane in strength, new links appear, and old ones are terminated. The aggregated network structure then reflects the joint effect of these properties that are associated with time scales of varying length.

In this chapter, we show this complex interweaving of time scales with a detailed study of a dynamical network of 6 months of mobile phone calls. We show that at different time scales, different structural features of the network emerge, leading to a strong variation of the network structure depending on the aggregation interval. Next, we show that the size of the network grows as a power of the number of events used to build the network, such growth being easily reproducible with a simple network growth model. Finally, we end this chapter with some exploratory work on how information theoretical concepts can be used

to quantify the information contained in dynamical networks.

Data Our data consist of the mobile telephone call records of the customers of a Belgian mobile operator from October 1, 2006 to March 31, 2007. Each customer is uniquely identified, and every call is associated with a time stamp and a duration. This data set has already been studied from a static perspective in several papers [115, 108, 32]. Since our focus is on link dynamics, we filter out all customers who have modified their subscription plan during the data collection period. This removes new customers, and customers who have cancelled their subscription during the period and are no longer customers. This leaves us with a network of 2.1 millions customers, making over 170 millions calls during the collection period.

For reference, we also construct two randomized ensembles, based on two randomization techniques of the time stamps. Thus for both cases the resulting randomized reference sequences contain the same number of calls between the same individuals as the original data. In the first ensemble, the time stamps have been generated uniformly over the complete time range, in order to remove the system-level call frequency pattern. The second ensemble has the time stamps of the original calls but randomly reshuffled, hence the system-level call frequency pattern (daily and weekly patterns) is retained. When aggregating over the entire observation period, such reference call sequences produce networks that are equal to the network from aggregating the original data. In the remaining, we will refer to these references as respectively the “uniform” and the “shuffled” references.

Notations and definitions Contrary to static networks, there is no clear consensus for standard notations for dynamical networks. This chapter focuses on growing networks, i.e. networks that are defined by the aggregation, or integration, of a growing set of events.

Definition 3. An event is a triplet (i, j, t) , such that there is a connection between nodes i and j at time t , t being here a discrete time stamp, representing the start of the call.

In the *aggregated network* $G(t)$, a link is established between i and j if at least one event between them is observed during the time interval $(0, t)$, i.e. given the collection of events $\mathcal{E}(t) = \{(i, j, s) : 0 \leq s \leq t\}$,

$G(t) = (N(t), L(t))$ is the network composed of the set of links $L(t) = \{(i, j) : (i, j, \cdot) \in \mathcal{E}(t)\}$ and $N(t)$ is the set of nodes supporting those links. The dependence of $N(t)$ on t will be used in the first section. We define in the same way $G(x) = (N(x), L(x))$, where x is the size of $\mathcal{E}(x)$, to define the network constructed with x events. This notation will be used in the second section, where time is removed from the analysis. Similarly to static network notations, $n(t)$ and $m(t)$ define respectively the number of nodes $|N(t)|$ and links $|L(t)|$, and $n(x)$ and $m(x)$ define the same quantities depending on x . The weight w_{ij} of a link is defined as the total number of events between i and j during the aggregation period. The strength s_i of node i is then defined [17] as the total number of events where i participates, and the degree d_i is the number of links from node i .

5.2 Influence of time scales on network structure

Consider a network constructed with all calls between users during one single day. Is its structure expected to be similar to a network constructed with all calls of one week? The answer is probably no. Interactions of different kinds happen at different time scales, and some structural properties emerging after a week might not be noticeable in a single day. This section covers a study of the influence of time scales on the structure of dynamical networks, the text being for a large part a reproduction of the text of the original article, which is submitted to EPJ Data Science [109].

Structural and temporal inhomogeneities We begin our investigations by addressing the inhomogeneities that are expected to play an important role in the evolution of the structural properties of networks aggregated over growing time intervals. The fundamental structural inhomogeneities are reflected in the standard statistical distributions for the call network, aggregated over the entire 6-month period of observation. In the aggregated network $G(t)$, a link is established between nodes i and j if a call is observed between them at any point during the interval $(0, t)$. As expected on the basis of earlier results [146, 145, 115], the probability density distributions of degree, strength, and link weights

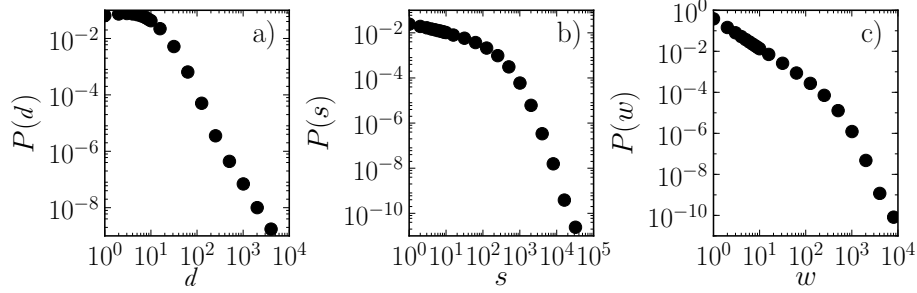


Figure 5.1: The degree (a), strength (b) and weight (c) distributions of the aggregated network when the aggregation period covers the whole 6 months of data.

are all broad (see Figure 5.1). Thus there is a large number of nodes that make only infrequent calls, and a large number of links that carry only a few calls. When aggregating the network over shorter time intervals, one thus expects to first discover the high-strength nodes and high-weight links that are associated with the tails of these distributions.

In the time domain, the two main inhomogeneities are related to burstiness of calls forming the links, and the overall circadian pattern of the system-wide call frequency. Burstiness of the calls is reflected in the probability distribution $p(\tau)$ of the times τ between consecutive calls on individual links. In Fig. 5.2a), we see that in line with earlier observations [43, 96, 133], the distribution $p(\tau)$ in our empirical data has a broader-than-Poissonian tail, a signature of burstiness. Such an inter-call time distribution gives rise to longer waiting times than expected if the calls were placed uniformly in time. Because of this, we expect to see slower network growth than for the uniform case. Further, as seen in Fig. 5.2, the network-level call frequency clearly displays the usual daily and weekly pattern [96, 93], where the frequency shows two daily peaks followed by a decrease during nights. In addition, weekend activity is lower, especially for Sundays.

Evolution of network structure All of the above features are expected to have an effect on the properties of networks aggregated over growing time intervals. Let us first monitor the growth of the aggregated network in terms of the numbers of nodes and links and the average de-

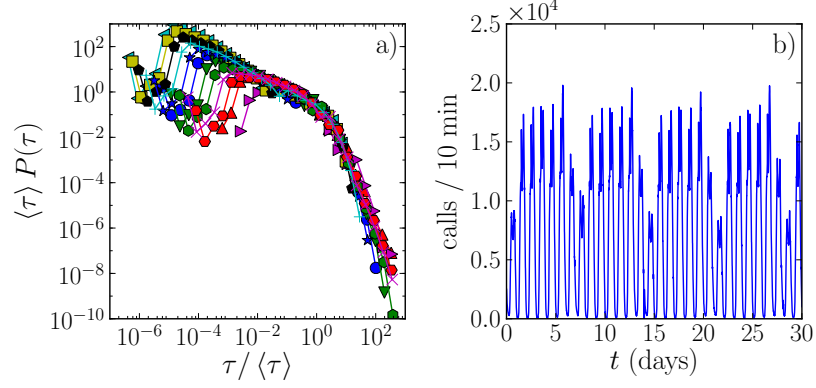


Figure 5.2: Temporal inhomogeneities affecting network growth. a) The scaled inter-event time distribution of links $p(\tau)$, divided into bins of different numbers of calls and scaled by the average inter-event time $\langle\tau\rangle$ of each bin, following [43]. b) Total call rate in the network as a function of time, for the first 30 days, displaying the circadian and weekly patterns.

gree, when the network $G(t)$ is aggregated up to a time t . As seen in Fig. 5.3 a), the number of observed nodes $N(t)$ displays a rapid increase in the beginning of the aggregation process, such that the aggregated network contains 90% of the nodes after $t \sim 30$ days. This rapid increase is followed by slower growth as nodes with low call activity are gradually observed to make calls, joining them to the aggregated network. When compared to the uniform reference, where the time stamps of all calls are drawn uniformly at random from the entire 6-month interval, it is seen that the growth of $N(t)$ is slightly slower; however, for longer aggregation times, the difference can be considered negligible and thus the time-domain heterogeneities have a visible effect only for short time windows. For short time windows, in addition to the slowing-down effect of burstiness, the daily pattern is seen to give rise to a stepped shape of the $N(t)$ curve (see the inset of Fig. 5.3 a).

In contrast, the growth in the number of links $L(t)$ is much more gradual, as seen in Fig. 5.3 b). Here, an aggregation time of $t \sim 149$ days is required for catching 90% of the links of the final 6-month aggregated

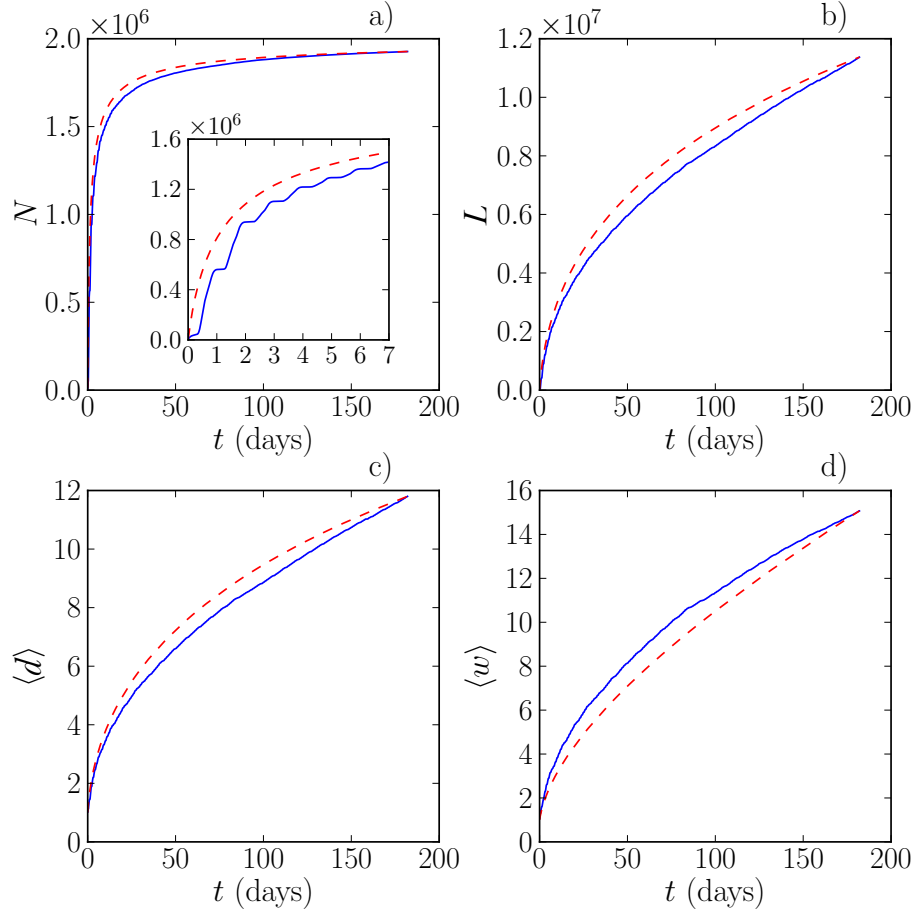


Figure 5.3: Total numbers of (a) nodes and (b) links (c) average degree and (d) average link weight in the aggregated network as a function of aggregation time. The solid (blue) line denotes original empirical data, while the dashed (red) line denotes the uniform reference. The inset in panel (a) displays the number of nodes for the first 7 days.

network. In addition, unlike for the number of nodes, the number of links keeps on growing steadily and no saturation in growth is observed. This is also reflected in the growth of the average degree (Fig. 5.3 c). Hence, even though the number of nodes becomes fairly stable in an aggregation

period of 6 months, one cannot claim to have captured all the links in the “underlying” network, and for longer windows, the average degree would still increase. This reflects the joint effect of several factors: first, as the link weight distribution is broad, there are large numbers of links with very low call frequencies, and observing those evidently takes a long time; there may be many links where calls take place less frequently than once in six months. In addition, the ubiquitous burstiness that results in longer waiting times between calls slows down the growth in the number of links especially for the low-weight links – this effect is visible in Fig. 5.3 b), although it is not very strong. Second, for such long observation periods, one can argue that the changes in the network structure should already have a visible effect: new social ties are formed while older ties wane in strength and may even cease to exist. Third, as the data contains all the calls made by the subscribers, many of the calls may be “random” in the sense that they do not represent real social interactions; as there is no background information on the nature of the calls, a random call to one’s dentist or a call in response to an advertisement on used car sales are counted as links, just as calls to one’s friends or relatives. This third mechanism would naturally result in an ever-growing number of links.

As a result of the interplay of the above mechanisms, the network itself keeps changing while it is being aggregated, and while some of its links are stable in the sense that they remain active for prolonged periods of time, others come and go. One way of estimating the related time scales is to compare the similarity of networks aggregated for different periods of time, when the observation period is divided into multiple consecutive aggregation windows. We calculate the similarity σ of two networks $G_1 = (N_1, E_1)$ and $G_2 = (N_2, E_2)$ as

$$\sigma(G_1, G_2) = \frac{|E_1 \cap E_2|}{|E_1 \cup E_2|}, \quad (5.1)$$

such that $\sigma = 1$ if the networks are the same, and $\sigma = 0$ if they share no links. Fig. 5.4 displays the average similarity σ of networks in consecutive windows of different durations W . When the windows are very short, the networks are very sparse and the number of common links is low. Then, the similarity increases with increasing window duration, reaching a maximum at ~ 30 days; then, the similarity begins slowly decreasing as the aggregation process captures more and more of the “random”,

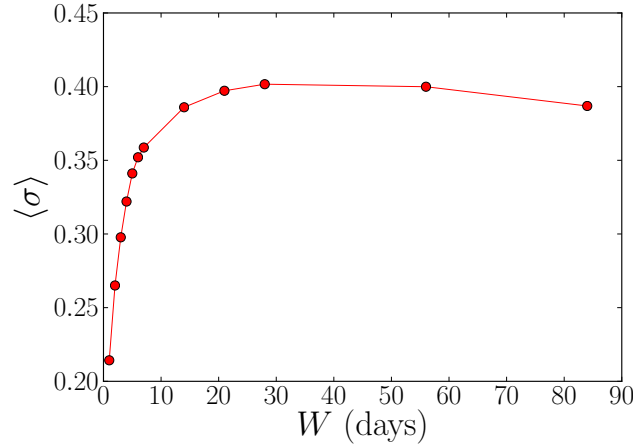


Figure 5.4: The average fraction of links f common to consecutive aggregation windows of duration W .

one-off links.

As the growth of the number of links does not saturate, it is of importance to understand the characteristics of links that emerge early on in the process. It is known from previous investigations [146] that link weights correlate with the network topology such that high-weight links are associated with dense network neighbourhoods, whereas low-weight links connect such neighbourhoods, in line with the Granovetter hypothesis [80]. This is directly related to the presence of *community structure* [71] in social networks; links within communities are stronger and have higher-than-average weights [181]. For the network aggregation, this means that clusters and communities containing high-weight links are likely to appear early on in the process. In order to investigate this effect, we measure the evolution of the network-level clustering coefficient $C(t)$, given by $3 \times$ the number of triangles divided by the number of connected triplets in the network. As seen in Fig. 5.5 a), the clustering coefficient does indeed show a rapid increase as a function of the aggregation interval length, and then decreases after a peak at around $t = 9$ days. This decrease can be attributed to the weak links observed

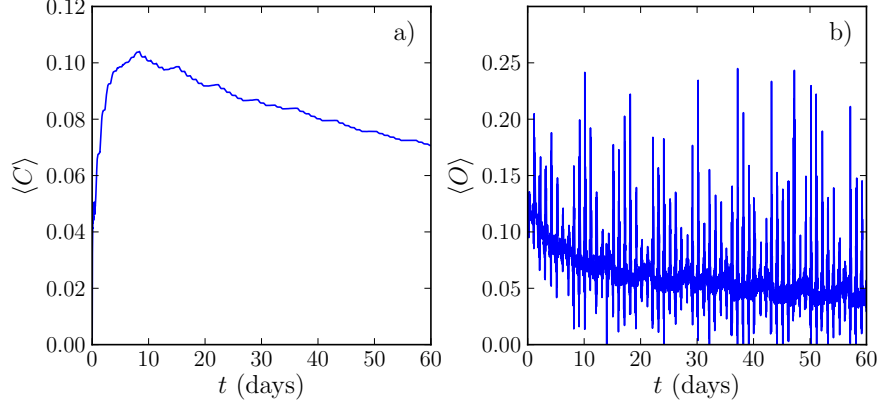


Figure 5.5: a) Global clustering coefficient of the network and b) average final overlap of added links as a function of aggregation time, for the first 2 months. The global clustering coefficient is computed as the number of triangles divided by the number of connected triplets. For the overlap, we calculate the final overlap of links in the 6-month aggregated network, and average over the final overlap values of newly added links at each time point.

later in the process; newly observed links do not contribute to triangles. Hence, for detecting nodes and links associated with dense clusters of strong links, even fairly short aggregation periods appear sufficient.

The fact that the links observed early on in the aggregation process are related to the community structure is also visible when monitoring the *overlap* [146] of the added links. The overlap of a link connecting nodes i and j is defined as $O_{ij} = n_{ij} / [(d_i - 1) + (d_j - 1) - n_{ij}]$, where n_{ij} is the number of common neighbours of i and j , and d_i and d_j are their degrees. Thus the overlap measures the fraction of common neighbours out of all neighbours of the two connected nodes. Fig. 5.5 b) displays the average overlap of the added links as a function of aggregation time. Here we have calculated the overlap of each link in the final 6-month aggregated network, and averaged over these values for links that are added to the network at time t . It is seen that the links that are added early on in the aggregation process have on average a higher

overlap than those added later. Hence, even when the aggregation times are short, the networks capture much of the community structure of the final aggregated networks. Interestingly, the overlap also shows a strong circadian and weekly pattern – its highest peaks correspond to the early morning when the overall call rate is very low. Thus, if calls are made during these hours, they are likely to be targeted towards people in the strongest clusters of friends and family.

In order to illustrate the network growth, we have visualized small subnetworks corresponding to different aggregation times t (Fig. 5.6). Here, the subnetwork has been obtained by selecting all individuals whose subscriptions are associated with a certain postal code – this method of sampling yields better results than e.g. snowball sampling. Panels a) to d) of Fig. 5.6 show the growth of the network, such that links that participate in triangles in the final 6-month aggregated network are coloured red. For the shortest aggregation periods (panels a and b), most of the added links are in this set, reflecting the above observations on the early appearance of links connected to communities and clusters. It should be noted that not all community-internal links are discovered early on; rather, those links that appear early are associated with communities with a high probability.

Behaviour of statistical distributions

Above, it was seen that even for fairly long aggregation intervals, the average degree of the aggregated networks still keeps increasing as a function of the interval length. Likewise, because of how the aggregation process is defined, the average weight necessarily keeps on growing as well. Next, we turn to the statistical distributions of these quantities, and ask when they become descriptive of the underlying network. Evidently, as the averages keep increasing with the aggregation interval length, the probability density distributions of degrees and weights also change and shift towards higher degrees/weights (see Figure 5.7). However, for such distributions to be meaningful descriptors of network structure, their underlying forms should for long enough intervals become stationary and depend only on their average values. In order to study the stationarity of the underlying dynamics one can compare such distributions by rescaling them as $p(x, \langle x \rangle) \sim \langle x \rangle p(x / \langle x \rangle)$, where $\langle x \rangle = \sum x p(x)$. We show the rescaled versions of the degree and weight

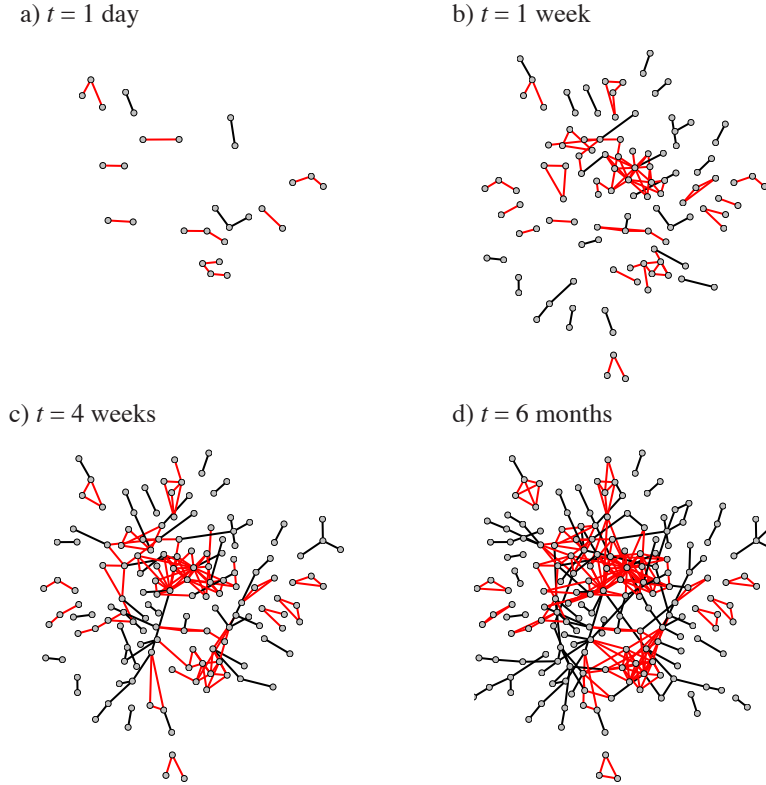


Figure 5.6: Series of aggregated networks with a growing aggregation interval. The network aggregated here represents a small subnetwork, obtained by picking individuals from a single postal code. Links that participate in triangle(s) in the final 6-month aggregated network are colored red, while the rest are black.

distributions on Figure 5.8, where the distributions have been rescaled with respect to their averages.

For degrees, one sees that the rescaled distributions collapse onto a single curve, surprisingly already for aggregation intervals over a few days, while for the weights, the distributions converge slightly slower but eventually also collapse for window sizes of the order of weeks.

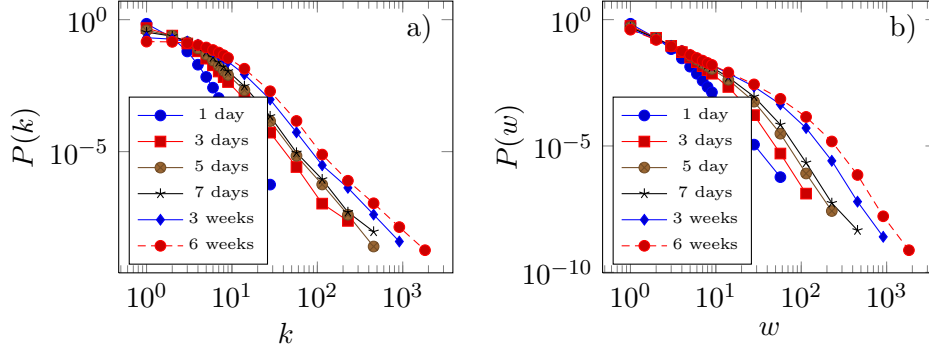


Figure 5.7: (a) Degree and (b) weight distributions for windows of different lengths. As the averages increase (see Figure 5.1), the statistical distributions shift to the right.

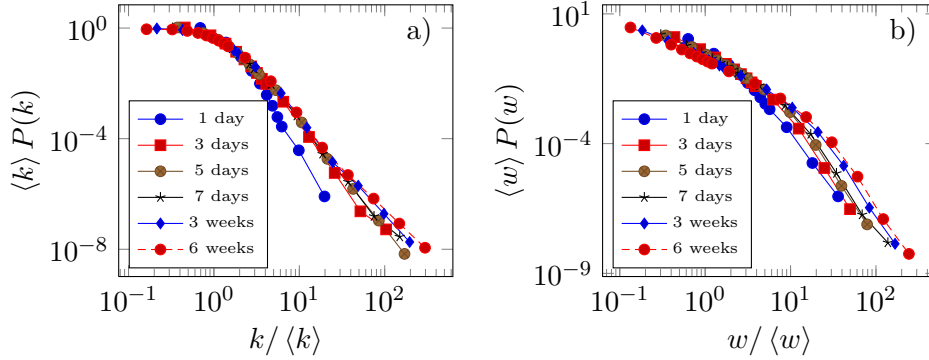


Figure 5.8: After rescaling the distributions with respect to their averages, the statistical distributions all collapse on the same functional form, after a sufficient long time. For (a) degrees, a week is sufficient for distributions to rescale, while for (b) weights, the subsequent distributions converge more slowly to one functional form.

The convergence of the rescaled degree distributions onto a single curve suggests that the underlying dynamics that is being aggregated has some stationary properties. A similar observation was made in [74] for different aggregation windows of the same length, with the conclusion

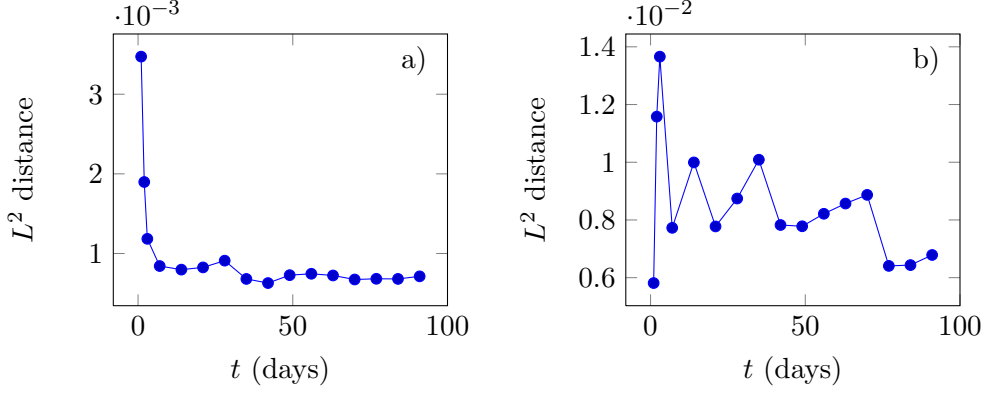


Figure 5.9: The L^2 distance between the degree (a) and weight (b) distributions of the networks aggregated over a time interval of length t and $2t$. That is, the first point is the distance between the rescaled distributions of the network built with the first day and the network built with the two first days. The second point is the distance between the rescaled distributions of the two first days and the four first days, etc. The last point is the distance between the distributions of the network built with the 91 first days and the network built with the 182 first days.

that the underlying process is stationary.

To quantify the convergence of the rescaled distributions to a single curve, we measure this convergence with the help of the L^2 distance between two distributions. The L^2 distance between two scalar functions f and g defined over the same domain $D \subseteq \mathbb{R}$ is computed as

$$d(f, g) = \left(\int_D [f(x) - g(x)]^2 dx \right)^{\frac{1}{2}}, \quad (5.2)$$

To measure the convergence, we have successively calculated the distance between the rescaled degree (weight) distributions of networks aggregated over an interval of length t and networks aggregated over twice longer intervals of length $2t$.

In Figure 5.9, it is seen that the L^2 distance between the degree distributions decreases and converges to a roughly constant value already for aggregation intervals longer than a few days. Hence, already after a short period, a sufficient amount of data has been collected to correctly

estimate the stationary degree distribution of the network, and the shape of the degree remains similar for longer intervals. The weight distribution displays a slightly slower convergence, and is still slowly changing even for aggregation intervals of around 3 months; this is also related to the curvature seen in the evolution of the average link weight (Fig. 3 d). There are several possible explanations for this slower convergence: first, the number of links is increasing and each newly added link brings in a unit of weight. Second, the weights of existing links may display very different time dynamics, and third, the growth of link weights is affected by the burstiness of the call sequences and the resulting long inter-call times. Nevertheless, even the weight distributions do not change much, and care is needed only when interpreting the weight distributions of call networks aggregated over very short periods of time.

On the effects of aggregation window placement

In all analysis so far, we have assumed that the exact placement of the aggregation window, *i.e.* the time point of its beginning, plays no role in the results. However, as the characteristic daily and weekly patterns of Fig. 5.2 b) indicate, the overall level of call activity in the network displays large variations by hour and day, and this is expected to have an effect on the aggregated networks, at least on shorter time scales. In addition, there may be less trivial effects if the actual behavioural patterns of individuals – affecting *who* they call – are also time-dependent.

Let us first focus on short time scales, and illustrate the growth of the network with a 2D heat map plot, displaying the number of nodes in aggregated networks as a function of the aggregation time and the beginning point of the aggregation window 5.10. The daily and weekly pattern is clearly visible in the plot – the network grows fastest during weekdays, especially Fridays. The growth is slowest when the aggregation begins on Saturdays and especially Sundays; the difference between Saturdays and Sundays is fairly large. However, this observation might be explained by the network-wide variation in the call frequency alone.

In order to have a closer look at the actual call patterns of individuals, we monitor the growth of the giant connected component of the network.

As seen above, the links that emerge early on in the aggregation process typically have high overlap, *i.e.* are associated with dense network neighbourhoods; however, the overlap also displays a time-dependent

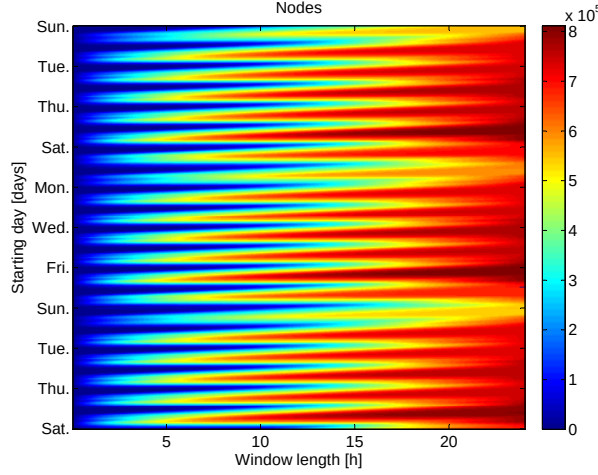


Figure 5.10: The number of nodes in the aggregated networks as a function of the aggregation time (horizontal axis) and the beginning point of aggregation (vertical axis). Day 0 on the vertical axis is a Sunday, and the first time point is midnight.

pattern. Roughly speaking, if the majority of calls in a given window is directed towards high-overlap individuals that are part of the same neighbourhood or cluster, the giant component should grow more slowly than if the calls are directed towards far-away nodes. We measure this effect by monitoring the average size of the emerging largest components; for reference, we also calculate their size relative to those in the time-shuffled reference ensemble that only contains the daily/weekly pattern at the system level, but where the timings of calls do not correlate with topology. Figure 5.11 displays the absolute and relative size of the giant component similarly to the 2D heat map plot for nodes. The absolute size displays a clear daily and weekly pattern much akin to that seen for the number of nodes in Fig. 5.10. However, the behaviour of the average giant component size relative to the time-shuffled reference ensemble reveals an interesting feature: when the aggregation begins in the weekends, especially Sundays, the giant component grows more slowly than it does in the reference. Likewise, for weekdays, it grows far faster

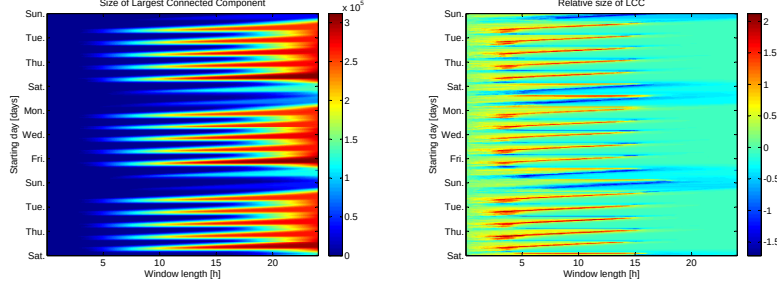


Figure 5.11: The absolute (left) and relative (right) size of the largest connected component as a function of the aggregation time (horizontal axis) and the beginning point of aggregation (vertical axis). The relative values of the bottom panel are calculated with respect to the shuffled reference model, where call times do not correlate with network structure.

than it does in the reference. This points towards different behavioural modes – weekday calls are frequently related to links joining nodes that would otherwise remain disconnected within the aggregation window, whereas weekend calls appear to relate to high-overlap links and dense clusters and thus contribute less to the growth of the largest connected components. In other words, friends and relatives with shared social circles are called more frequently in the weekends.

Changes in behaviour very similar to the daily and weekly patterns also appear on longer time scales, as human communication patterns are affected by *e.g.* holiday periods. In order to observe such changes, we divide the data into non-overlapping time windows of 1 day, 1 week, and 2 weeks. We then aggregate networks corresponding to each time window, and monitor the similarity of the networks in two consecutive windows as a function of time. For this, we employ a weighted generalization of the Jaccard index of Eq. 5.1:

$$\sigma_w = \frac{\sum_{(i,j) \in E} \min(w_{ij}^1, w_{ij}^2)}{\sum_{(i,j) \in E} \max(w_{ij}^1, w_{ij}^2)}, \quad (5.3)$$

where w_{ij}^1 and w_{ij}^2 are the weights of links between nodes i and j in the consecutive windows ($w_{ij} = 0$ if there is no link). Fig. 5.12 displays

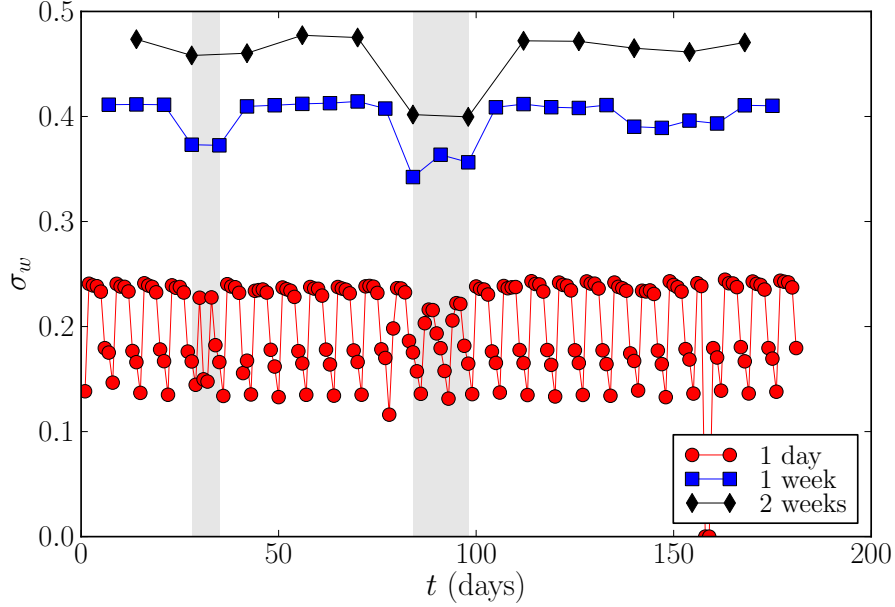


Figure 5.12: Similarity between networks corresponding to two consecutive aggregation windows of 1 day, 1 week, and 2 weeks. The shaded areas correspond to national holiday periods: the autumn holiday (left), and holidays around Christmas and New Year (right).

the weighted network similarities for the three different window sizes as a function of time. On average, the similarities remain level and no general trends can be observed. The 1-day similarity displays a clear daily and weekly pattern, in line with our observations on the connected component sizes. Here, for each week, the consecutive windows corresponding to Monday to Thursday have highest similarities, while similarities are far smaller between the 1-day networks for Friday, Saturday and Sunday. Some larger changes can be seen already for the 1-day measure during the holiday periods (autumn holiday, Christmas season) indicated by the grey shaded areas. The differences corresponding to holiday periods become much more apparent when aggregation window sizes of 1 week and especially 2 weeks are used.

5.3 A model of growing networks

In the previous section, we focused on the impact of external factors such as daily/weekly activity on the growing network structure.

In the following, we will approach growing networks from a different angle in order to quantify their growth. Keeping in mind that we consider data-driven evolving networks, they are generated by a list of events, further denoted $\mathcal{E}$, of size $x = |\mathcal{E}|$, where each event is an observation of a link at a precise time stamp.

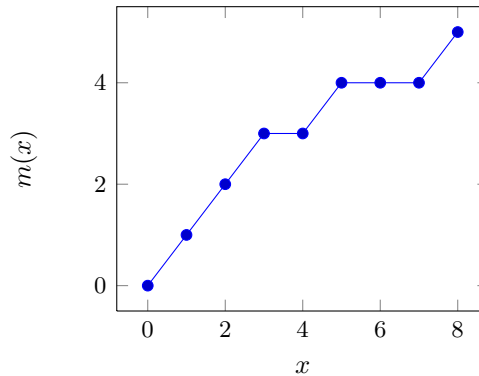
Quantifying the growth process of a network consists then in characterizing the relationship between the size of $\mathcal{E}$, when additional events are added to the list and the size m of the corresponding set of links L . Formally, L and $\mathcal{E}$ are related as

$$L = \{(i, j) : (i, j, \cdot) \in \mathcal{E}\}, \quad |L| = m.$$

In the following, the time stamp of each event is dropped, which reduces $\mathcal{E}$ to a list of links, with possibility of having multiple occurrences of the same link in $\mathcal{E}$. The growth of the network is then characterized by the function $m(x)$, which is the size of L growing with the size of $\mathcal{E}$. For example, given the following list of events

$$\{(1, 2), (2, 3), (1, 4), (1, 2), (3, 1), (2, 3), (1, 4), (4, 5)\}.$$

is the size of L generated by the x first events is given by the function $m(x)$:



The exact nature of events depends on the nature of the observed network; they may be phone calls, emails, flights, ... When being added to the network, an event has two possible effects: it can add a new link to the network, or it can increase the weight of the link, if this was already existing. In the general case, events could have varying weights, but in this framework we will consider equally unit-weighted events.

The function $m(x)$ characterizes the growth of the network, its shape can vary depending on the type of process generating the network. A citation network, for example, where each event is the citation of a paper by another, shows a linear growth of the network ($m(x) = x$), since no paper cites twice the same source. Conversely, a network of train transportation, where each event is a train connection between two stations, grows until reaching a constant size after typically one week, since each connection has a frequency of at most once per week. Notice that we don't consider the deletion of links in this framework, so m can be any positive non-decreasing function of x , having as constraint that $0 \leq \frac{m(x_2) - m(x_1)}{x_2 - x_1} \leq 1, \forall x_1, x_2$.

Authors studying the growth of networks have already considered the speed of growth of networks by comparing the speed of growth of the number of nodes and links, leading to the notion of *accelerating* networks [60, 172], i.e. networks which are densifying over time [123], for which each event is associated to the creation of a new link. The present analysis comes in addition to these results, for networks where events may not contribute to the growth of the network because they represent a link that already exists in the network. For these, the study of the growth of the network in function of events (phone calls, emails, tweets, ...) provide useful insights into the dynamics of social contacts, which we will illustrate with the mobile phone dataset.

The mobile phone network shows a continuous growth (Figure 5.13), links being uninterruptedly added to the network. For this particular dataset, the curve shows two distinct phases in the growth: a first phase that we will call a *sharp start* followed by a second phase that we will call the *steady tail*, which can be well-approximated by a power-law ($m(x) = Kx^\alpha$) with exponent close to 0.5. As an illustration, we show that it is possible to predict the shape of the curve based on a small portion after the sharp increase (see Figure 5.13).

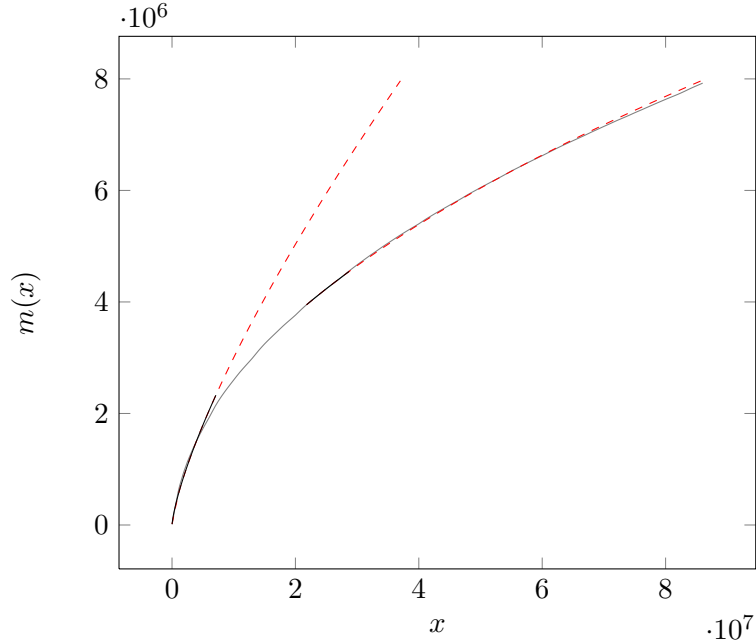


Figure 5.13: Prediction of the size of the network. The exponent has been fitted with the black portions, which corresponds to the data from day 22 to 28. The prediction fails when the prediction is made on an early portion of the dataset.

The previous notion of “sharp start” and “steady tail” are rather informal but one can characterize the transition between those phases as the moment where the derivative of m (approximated with finite differences, Figure 5.14) reaches a shape close to the derivative of a power-law. We introduce for this the *growth rate* $r(x)$, defined as follows:

$$r(x) = \frac{x}{m(x)} \frac{dm(x)}{dx}, \quad (5.4)$$

If $m(x)$ grows as a power-law, this quantity is expected to be constant and equal to the exponent. The growth rate $r(x)$ does indeed stabilize on average (Figure 5.14), the variations around the average are explained in detail at the end of this section.

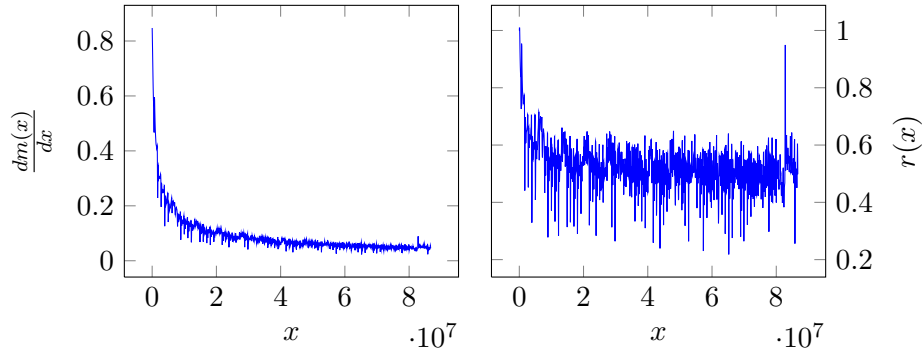


Figure 5.14: (a) Derivative of the network size function. The derivative has been approximated using first-order finite differences, with a step of 10,000 events. (b) The growth rate of the network decreases and reaches an average stable value around 0.5.

The power-law shape of the network growth is not a trivial behaviour. A similar growth curve has been observed in the growth of the number of words in text books and have been described by the Heaps' law [81], an empirical law that has to date been explained in a limited number of theoretical situations [184, 126]. The observed shape of $m(x)$ raises two questions: first, which sort of process generates a power-law growth of a network, and second, how can the transition between the sharp start and steady growth been explained?

In the following, we first propose a simple model of network growth that reproduces a power-law growth, and we show how a modified version of the model, that mimics what is expected to happen in real data collection procedures, explains as well the initial sharp increase followed by a steady growth.

A simple model of network growth: PU (Preferential Usage)

There exists a wide literature about models of network growth, such as preferential attachment models, node copying models or triangle closure (see Chapter 14 of [141] for a review), but these models aim at reproducing structural features of existing networks such as degree distribution or clustering and don't focus on growth speed.

We bridge this gap by suggesting a model that solely focuses on the speed of growth of the networks, without suggesting a network structure. In order to keep our model tractable, we will state some simplifying assumptions on how the network exactly grows, and how the weights of links are dynamically updated by the model. In this model, we assume that each event is independent of the preceding events, hence bursts and temporal correlations of events are not handled in this model. Since no network structure is given alongside this model, the link weights are also independent of each other. At each step, the model decides, based on link weights of the network and a single parameter, either to create a new link or to increase the weight of an already existing link, without determining to which nodes each link is connected. This leaves some freedom to the user to couple this network growth model with any other model, such as preferential attachment or triangle closure.

The dynamical model takes as input a non-empty set of links L and a corresponding non-empty set of events $\mathcal{E}$. Each link l of L has a non-zero weight w_l such that $\sum_l w_l = |\mathcal{E}|$. Initially, the network is hence composed of $m_0 = |L|$ link and $x_0 = |\mathcal{E}|$ events. We set $x = x_0$. The model takes then as parameter $K \in [0, 1]$ and repeats the following steps (each iteration of those 3 steps represents one event):

1. choose a link $l \in L$ proportionally to its weight w_l : the probability of a link l to be chosen is $\frac{w_l}{x}$;
2. with probability $\frac{K}{w_l}$, create a new link l^{new} of unit weight, and add it to $\mathcal{E}$ and L , otherwise, add l to $\mathcal{E}$ and increase w_l by one unit;
3. increase x by one unit.

The model is explained as follows: given a previous sequence of events, the next event is chosen in the already known set of links such that links that are more frequent are more likely to be used again. In reference to preferential attachment, which uses a similar concept, we call this preferential usage. However, instead of using the chosen link it is possible to create a new link, with a probability that decreases with its weight. This can be seen as a “price of treason”: if two nodes have a strong connection, they are likely to be eager to connect again, while

if two nodes have a weak connection, one of them might be tempted to contact another node instead. Hence, the higher the link weight, the lower the probability to create a new link.

Each iteration representing the addition of an event to $\mathcal{E}$, what is the expected growth of L ? The probability of a link to be chosen is given by $\frac{w_l}{x}$. The probability of creating a new link, knowing that link l has been chosen is $\frac{K}{w_l}$. Then, the expectation of the probability of creating a new link is given by

$$\text{Exp}[p_{\text{new}}] = \sum_{l \in L} \frac{w_l}{x} \frac{K}{w_l} = \frac{m(x)K}{x}. \quad (5.5)$$

As this probability is also the average derivative of $m(x)$, we have

$$\frac{dm(x)}{dx} = \frac{m(x)K}{x}, \quad (5.6)$$

and we deduce

$$m(x) = m_1 x^K. \quad (5.7)$$

From the initial condition $m(x_0) = m_0$, we have $m_1 = m_0 x_0^{-K}$, and the expression becomes:

$$m(x) = m_0 \left(\frac{x}{x_0} \right)^K. \quad (5.8)$$

The tuning of parameter K allows to simulate different speeds of growth, from a constant network size ($K = 0$) to a linear growth ($K = 1$). In particular, if $m_0 = x_0$ and $K = 1$, we obtain the growth of a citation-like network ($m(x) = x$).

The average result of 10^2 realizations of PU with parameters $m_0 = x_0 = 10^3$ and $K = 0.2$ is illustrated at Figure 5.15. The results show a clear power-law growth, with the same parameters as expected by the model.

DPU (Discovery and Preferential Usage) Model The previous model manages to simulate the growth of a network, but needs a non-empty set of events and links to initiate the growth. However, this situation differs from the exact situation in which a network is aggregated “from scratch”, which is what happens when analysing data. This can be simulated by modifying slightly the previous algorithm: instead of

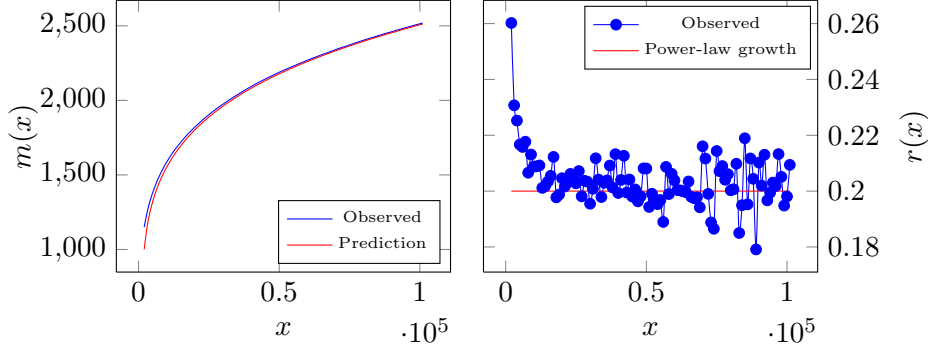


Figure 5.15: Average result of 10^2 realizations of PU with $m_0 = x_0 = 10^3$ and $K = 0.2$. (left) Size of the network versus x (blue). The growth fits well to the predicted power-law (red) with exponent 0.2. (right) $r(x)$ versus x . The blue curve shows the observed value (as given by eq. (5.4)), the red curve is the predicted value when using $\frac{dm(x)}{dx} = K \frac{m(x)}{x}$ instead of finite differences approximations.

initiating the algorithm with non-empty sets $\mathcal{E}$ and L , the dynamics are simulated by non-empty “hidden” sets of events and links $\mathcal{E}^h$ and L^h . Such a “hidden” network represents then the set of connections that are already existing at the start of the data collection, while the new links, that are created later during the dynamical process represent new connections. The events generated by PU using $\mathcal{E}^h$ and L^h are then added to an empty list of events $\mathcal{E}$ and links L . The difference lies here between the observed network size $m = |L|$ and the unknown hidden network size $m^h = |L^h|$. This separates the growth process of the network in two parts: first, a *discovery phase*, during which the growth mostly discovers L^h , followed by a *steady phase*, where L^h has been (almost totally) unveiled and new links are only created by the dynamical model. Depending on the parameter K and on the initial weights w_l^h of the links in the hidden network, the discovery phase can be unexistent, short or long.

The model is described hereafter:

We start from a non-empty hidden set of events $\mathcal{E}^h$ and links L^h (these sets will grow at each iteration, in the following, we will refer the exact initial sets as $\mathcal{E}_0^h$ and L_0^h). Each link l of L^h has a non-zero weight

w_l . Initially, this hidden network is then composed of $m_0 = |L_0^h|$ links and $x_0 = |\mathcal{E}_0^h|$ events. The observed network is initiated as an empty set $L = \mathcal{E} = \emptyset$, and $x^h = x_0$ and $x = 0$.

Then, we repeat the following steps:

1. choose a link l in L^h proportionally to its weight w_l , the probability of a link l to be chosen is then $\frac{w_l}{x^h}$;
2. with probability $\frac{K}{w_l}$, create a new link of unit weight, and add it to L^h , $\mathcal{E}^h$, L and to $\mathcal{E}$, otherwise, increase w_l by one unit, add l to $\mathcal{E}^h$ and $\mathcal{E}$ and if $l \notin L$ add l to L ;
3. increase x and x^h by one unit.

At the initiation of the algorithm, the first observed links are either from the hidden network, or newly created links. This mixed situation persists until no more links of the hidden network are discovered after which the network grows as a power-law. This bimodal situation explains the phenomenon of *sharp start* (discovery) and *steady tail* (preferential usage), described at the start of this section.

The average result of 10^2 realizations of DPU with parameters $m_0^h = x_0^h = 10^3$ and $K = 0.2$ is illustrated at Figure 5.16. The initial fast growth of L is the conjunction of the discovery of L^h and the creation of new links. Once L^h has been uncovered, the growth follows a power-law.

Transition from discovery to steady phase The goal of this model is to reproduce the dynamics observed in real data, and in particular the transition between the sharp start and the steady tail where the growth can be modelled as a power law. Intuitively, one can expect that the model will eventually transit from a sharp start to a steady growth, however the time needed before the transition occurs varies in function of the parameters. We show in the following how the different parameters influence the speed at which the transition from discovery phase to steady state occurs.

In two particular cases, the transition never happens. First if $x_0^h = m_0^h$ (hence $w_l = 1 \forall l$) and $K = 1$, in which case each new event adds a new link to the network. This simulates the dynamics of a citation

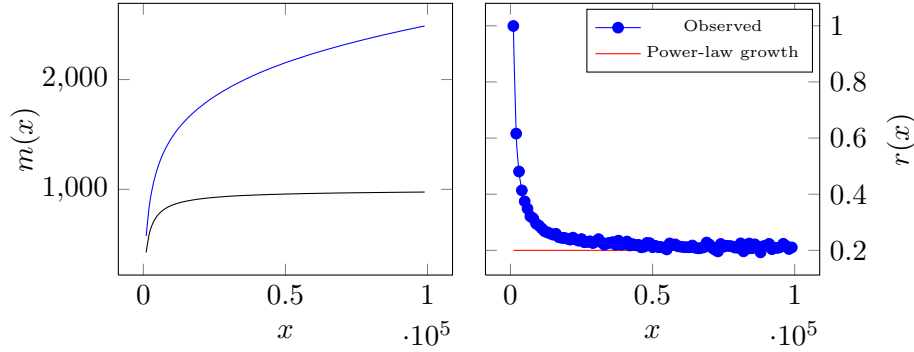


Figure 5.16: Average result of 10^2 realizations of DPU with parameters $m_0^h = x_0^h = 10^3$ and $K = 0.2$. (a) size of the network as a function of x (blue). The first sharp increase results from the discovery of both new links and links from the hidden network. The black curve indicates the number of links from the hidden network that have been observed. Once that value has stabilized, the remaining growth is a power-law of exponent 0.2. (b) $r(x)$ as a function of the number of events. A first sharp decrease is followed by a stabilized value (blue), close to the one predicted by the power-law growth model (as given by eq. 5.4) (red).

network, for example, where each new citation creates a new link in the network. Second, less practical, if $x_0^h = \infty$, either $l_0^h < \infty$, and some links have infinite weight, hence new links never appear, or $l_0^h = \infty$ and links from the hidden network are continuously discovered.

Excepted for those two cases, a transition always occurs, with each parameter playing a different role in this transition. To convince ourselves that this will effectively happen, let us remark that the derivative of m is composed of two parts

$$\frac{dm(x)}{dx} = d^{growth}(x) + d^{disc}(x), \quad (5.9)$$

where $d^{growth}(x)$ is the result from the dynamical creation of links and $d^{disc}(x)$ gives the number of links from L_0^h that are detected at x . The transition occurs then when $d^{disc}(x) \rightarrow 0$. The average expected number of links of L_0^h that are discovered after x steps is modelled by a non-homogeneous Poisson process of rate $\frac{w_l - K}{x + x_0}$, which can be approximated

in continuous time by

$$\left\langle L_0^{disc}(x) \right\rangle = \sum_{l \in L_0^h} 1 - e^{(w_l - K) \int_0^x \frac{dy}{y + x_0}}, \quad (5.10)$$

$$= \sum_{l \in L_0^h} 1 - \left(\frac{x_0}{x + x_0} \right)^{w_l - K}. \quad (5.11)$$

Its contribution to $\frac{dm(x)}{dx}$ is hence given by

$$d^{disc}(x) = \sum_{l \in L_0^h} (w_l - K) \left(\frac{x_0}{x + x_0} \right)^{1 + w_l - K}, \quad (5.12)$$

which indeed tends to 0 when x increases. The exact dynamics depend on the weight distribution w_l , but if one considers the average case where each link has the same weight w , the role of each parameter is clearly shown in (5.12). If we fix the number of links, the exponent K slows down the discovery period in the exponent of (5.12), while increasing x_0 accelerates the discovery period (since an increase of x_0 also implies an increase of w).

To sum up, the DPU model allows to simulate different dynamics of network growth, which can be modified by tuning the growth rate K , which slows down the dynamics. The hidden network size m_0^h and the generating events x_0^h also influence the speed at which the transition occurs, in particular, if the network is composed of links of high weights, the discovery period is short. Finally, let us also notice that the weight distributions produced by the model are observed to be broad, which is also in accordance with what is usually observed in real data. However, this observation is only empirical, we do not provide a functional form for the weight distribution.

Anomalies in real data The DPU model succeeds in reproducing a similar average growth as observed in mobile phone networks. The knowledge of the growth factor K allows then also to predict the future growth of a network, under normal circumstances. However, in

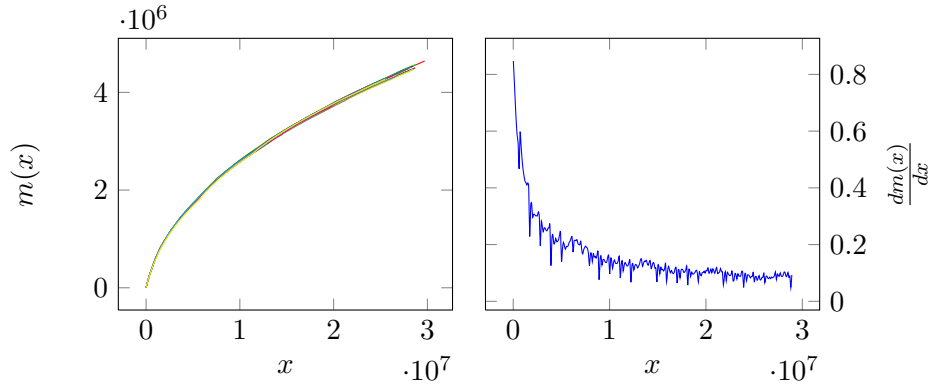


Figure 5.17: (left) The network growth for different starting days. The similar shapes suggest that network growth is a stationary process. (right) $\frac{dm(x)}{dx}$ for the first four weeks of data. Drops are observed, which correspond to calls made late at night and early morning.

some cases, divergences are observed between the average predicted behaviour and the real dataset, which turn out to be a significant source of side-information on collective human behaviour, with interesting (and even amusing) conclusions. We will show here three examples of these side-questions.

The derivative of $m(x)$, for which a zoom on the first 4 weeks of data is shown in Figure 5.17 (right), shows deep drops at a regular frequency. Those drops correspond to events happening late at night or early in the morning. Recall that here, $m(x)$ is decorrelated from circadian patterns, the drop of its derivative shows that all calls passed early morning and late at night produce between 40% to 60% fewer new links than on average. The explanation is simple, and has already been stated in the first section; early morning and late at night, people only dare to call close relationships, hence it is likely that the events produced early morning and late at night contain links that have already been seen before in the data. Besides these deep drops, smaller drops are also observed at two other significant periods of the day: noon (lunch time) and 4PM (school closures).

In Figure 5.13, we showed that it is possible to predict the growth

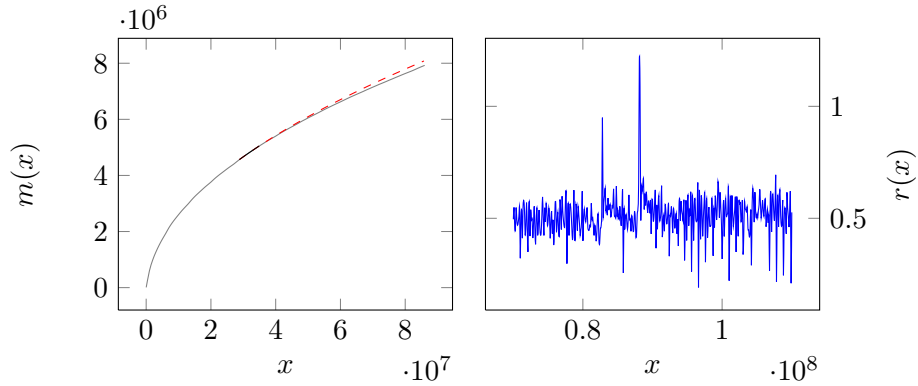


Figure 5.18: (left) The prediction using a power-law fails when the fitting is made on a holiday week (right) Peaks appear in $r(x)$ during Christmas night and New Year’s Eve. During these evenings, phone calls made by people are anomalously more directed to people who are not in their contacts.

of the network by using a power-law fit. However, for some particular periods in the dataset, this prediction is not possible, in particular for the fifth week of the dataset, in which the size of the network is overestimated (see Figure 5.18, left). The reason for this is that the fifth week in the dataset corresponds to a holiday week, more particularly the autumn holiday. During that week, the rate of creation of new links is larger than an average week. We deduce from this that people tend to call different acquaintances during holidays than during normal weeks.

Finally, and more surprising, the growth rate $r(x)$ allows to detect anomalous events, when a collective inhabitual behaviour is observed in the population. For example, $r(x)$ shows two large peaks in the dataset (see Figure 5.18, right), indicating two moments where the creation of new links is much larger than expected. Those two peaks correspond to December 24, at 10:30PM and December 31 at 11:50PM. This is the moment at which most people call each other to wish a Merry Christmas or a Happy New Year. Hence, it appears that most “Merry Christmas” and “Happy New Year” wished by people over the phone are sent to acquaintances they didn’t call to in the three previous months.

In conclusion, modelling the growth of a mobile allows to grasp an un-

derstanding on basic principles that could influence the creation of new links, but also helps at the detection of global anomalous behaviours. Research has already been conducted in detecting global collective responses from populations to anomalous events [10] or to detect unnatural behaviours from users [43], both results working on local scales both in temporal and geographical space. In this work, however, we show that it is also possible to detect behaviours diverging from the usual routine, even if the diverging behaviour happens on the long run (such as a whole week).

5.4 Information contained in growing networks

In the previous section, we showed that it is possible to predict how many links will be contained in a network after a precise time. Pushing the reasoning a bit further, we are tempted to ask: could one know *which* link will appear at the next event? In most cases, the answer to that question is certainly no. While transport networks are highly regulated with precise schedules to follow, social networks, instead, behave much more in a random-like way. However, we show in this section that some structure can be extracted from this apparently random growth.

In this section, we consider a growing network as a growing sequence of symbols, each symbol representing a different link. The prediction of the next link in the network is then rephrased as; given the ordered sequence of symbols produced by the x first events, how much uncertainty do we have on the next event? We define the ordered sequence of events of the whole network as $S_G = X_1, X_2 \dots$ $X_i \in L \forall i$, and the ordered sequence of events generated by a single node (in a directed network, this contains only the outlinks of the node) S_i . We use notations similar to those used in the previous section, the links of G observed after x events is $L(x)$, and the weight $w_l(x)$ is the number of times that l appears in S_G after x events. We then have the following assumption on S_G :

Assumption A. The events of S_G are independent.

In a social network, this assumption holds if the number of nodes is large. In each individual sequence S_i , the events are mostly correlated, but S_G is composed of a large number of such sequences, all mixed into one large sequence. In S_G , two consecutive events are very likely to be

uncorrelated, from which the independence assumption results.

A naive way of quantifying the uncertainty on the next event would be to approximate it by the uncertainty on a random choice of an element of S_G , which is given by the Shannon entropy of the frequency distribution of elements of S_G . We write this frequency distribution as $\mathcal{P}(x) = \{p_l(x)\}_{l \in L(x)}$, $p_l(x) = \frac{w_l(x)}{x}$, where $x = |S_G| = |\mathcal{E}|$ is defined as in the previous section. The Shannon entropy of $\mathcal{P}$ is given by

$$H(\mathcal{P}(x)) = - \sum_{l \in L} p_l(x) \log p_l(x). \quad (5.13)$$

Let us note that in this document, we always consider logarithms in base 2, hence the results are expressed in bits. We consider only finite networks, hence $\mathcal{P}(x)$ is a finite set. Moreover $p_l(x) \geq 0 \forall l$, and $\sum_l p_l(x) = 1$. $\mathcal{P}(x)$ is thus a finite discrete probability distribution, over which the Shannon entropy is well defined.

Approximating in this way the uncertainty of the next event to be added to S_G works well in non-growing networks, i.e. networks where additional events do not produce new links. In the case of growing networks, approximating the uncertainty on the next event by (5.13) does not work. The Shannon entropy is defined on probability distributions on the outcome of a complete random variable, i.e. the probability on the total set of possible outcomes (here $L(x)$) is 1. However, in growing networks, the probability that the next outcome is in $L(x)$ is rather given by $1 - \frac{dm(x)}{dx}$, where $\frac{dm(x)}{dx}$ is the average growth of $L(x)$. The remaining probability is $\frac{dm(x)}{dx}$, and represents the case where the next event is a new link. We take this into account by using a modified version of $\mathcal{P}(x)$, the generalized probability distribution $\mathcal{Q}(x) = \{q_l(x)\}_{l \in L(x)}$ $q_l(x) = p_l(x)(1 - \frac{dm(x)}{dx})$. The entropy on a distribution of an *incomplete random variable* has been defined by Rényi [163] as

$$H(\mathcal{Q}) = \frac{\sum_l -q_l \log q_l}{\sum_l q_l}, \quad (5.14)$$

which satisfies the same properties as the Shannon entropy, such as symmetry with respect to the elements of $\mathcal{Q}$. This entropy can be interpreted in the topic of this work as follows: an incomplete random variable is a

variable where the probability on its entire space of possible outcomes can be less than one. In such a case, the remaining probability represents an outcome that is not part of the space of possible outcomes (i.e. an event creates a new link rather than increasing the weight of an already known link), on which there is an infinite uncertainty (since it can be anything). Developing (5.14) with the expression of $\mathcal{Q}(x)$ leads to

$$H(\mathcal{Q}(x)) = H(\mathcal{P}(x)) - \log\left(1 - \frac{dm(x)}{dx}\right). \quad (5.15)$$

The entropy for this generalized distribution is composed of two parts, one representing the uncertainty on the set of known events, and one on the uncertainty due to the growth of the network. Note that if the network is of constant size ($\frac{dm(x)}{dx} = 0$), (5.14) reduces to $H(\mathcal{P}(x))$, and if instead $\frac{dm(x)}{dx} = 1$, $H(\mathcal{Q}(x)) = \infty$, since the next event could be anything with probability 1.

To understand the expected behaviour of this measure on a dynamical network, let us consider a finite network composed of a set of links L , where every link appears in the sequence of events S_G with frequency λ_l following a homogeneous Poisson process. To facilitate calculations, the process is approximated in continuous time t . We search for the expressions of $p_l(t)$ and $dm(t)/dt$, needed in (5.14) and (5.15). We suppose that $\sum_{l \in L} \lambda_l = 1$, which can always be achieved with an appropriate scaling of time. We define $m^{obs}(t)$ as the size of the set $L(t)$ of distinct links that have appeared in the time interval $[0, t]$. This value is given by

$$m^{obs}(t) = \sum_{l \in L} (1 - e^{-\lambda_l t}). \quad (5.16)$$

The expected number of events generated by link l in the interval $[0, t]$ corresponds to $\lambda_l t$. The expected frequency at which l is then observed in the generated set of events is given by

$$p_l^{exp}(t) = \frac{\lambda_l t}{(1 - e^{-\lambda_l t}) \sum_{k \in L} \lambda_k t}. \quad (5.17)$$

The entropy of $\mathcal{P}(t)$ then reads

$$\begin{aligned}
H(\mathcal{P}(t)) &= - \sum_{l \in L} \text{Prob}[l \in L(t)] p_l^{\text{exp}}(t) \log p_l^{\text{exp}}(t), \\
&= \sum_{l \in L} (1 - e^{-\lambda_l t}) \frac{\lambda_l t}{(1 - e^{-\lambda_l t}) \sum_{k \in L} \lambda_k t} \log \frac{\lambda_l t}{(1 - e^{-\lambda_l t}) \sum_{k \in L} \lambda_k t}, \\
&= - \underbrace{\sum_{l \in L} \lambda_l \log \lambda_l}_{H_{\text{stat}}} + \underbrace{\log \sum_{l \in L} \lambda_l}_0 + \underbrace{\sum_{l \in L} \lambda_l \log(1 - e^{-\lambda_l t})}_{-H_{\text{var}}(t)}. \quad (5.18)
\end{aligned}$$

$H(\mathcal{P}(t))$ decomposes thus into the difference between a stationary part H_{stat} , which corresponds to the entropy of the stationary frequency distribution and a variable part $H_{\text{var}}(t)$ which reflects the loss of entropy due to the fact that all links are not observed in the early stages of the process.

To complete the characterization of $H(\mathcal{Q}(t))$, one finally needs to quantify $dm^{\text{obs}}(t)/dt$, which is straightforward given its expression, to obtain the second term of (5.14).

$$H_{\text{growth}}(t) = -\log(1 - \sum_{l \in L} \lambda_l e^{-\lambda_l t}), \quad (5.19)$$

and we obtain

$$H(\mathcal{Q}(t)) = H_{\text{stab}} - H_{\text{var}}(t) + H_{\text{growth}}(t). \quad (5.20)$$

The forms of $H_{\text{var}}(t)$ and $H_{\text{growth}}(t)$ are very similar, since they are respectively equal to $-\text{Exp}[\log(1 - e^{-\lambda t})]$ and $-\log(1 - \text{Exp}[e^{-\lambda t}])$, where the expectations are taken over the probability distribution $\mathcal{P}$. They can be interpreted as follows: the total uncertainty on the outcome of the next event is the stationary uncertainty H_{stab} of the final distribution $\mathcal{P}$, from which a part is removed given that all links from L haven't been uncovered yet and a part is added due to the growth of the set $L(t)$.

We illustrate this with the example shown at Figure 5.19, which is simulated with a network of 10,000 links whose weights are distributed following a power-law of exponent 2 ($P(w_l = w) \propto w^{-2}$). The entropy grows quickly as more and more links are discovered. However, the entropy stabilizes after 250,000 events, even though a large number of links

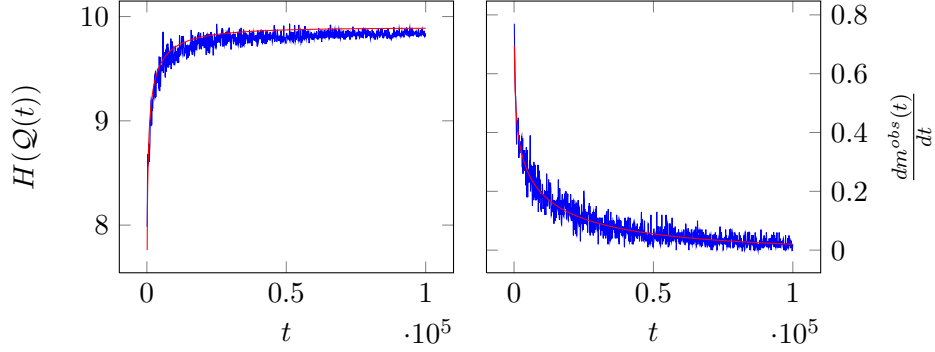


Figure 5.19: Simulation (blue) and prediction (red, with (5.18) and (5.16)) of the entropy of a growing network and the derivative of its size. The weight distribution follows a power law with exponent -2.

haven't been uncovered yet. Even though many links are added to the system, the entropy does not grow in consequence, as if the uncertainty brought by these new links does not weigh much in comparison with the contribution of more frequent links.

A similar behaviour is observed in the real dataset (Figure 5.20); although the number of links in the network is constantly increasing, the quantity of information contained in the sequence S_G is almost constant after one week of data (~ 7 million events, 98.5% of the final entropy). For comparison, the Shannon entropy is shown, as well as the entropy if each link has the same weight. After two days of data ($15 \cdot 10^6$ events), 3.5% of the links made in one month are uncovered, taking into account for more than 90% of the final entropy.

The stabilisation of the entropy after a small time confirms the observations from the first section. Keeping in mind that the contribution of every link to the total entropy is given by $-p_l \log p_l$, which is an increasing function of p_l in the interval $[0, \frac{1}{e} \approx 0.37]$, the contribution of high weight links (having thus a high frequency) is much larger than the contribution of low weight links (low frequency). Recalling the results from Section 5.2, links discovered at the start of the aggregation period are located in dense parts of the network, and have large weights. Later on, instead, links that emerge are rather bridges between those dense

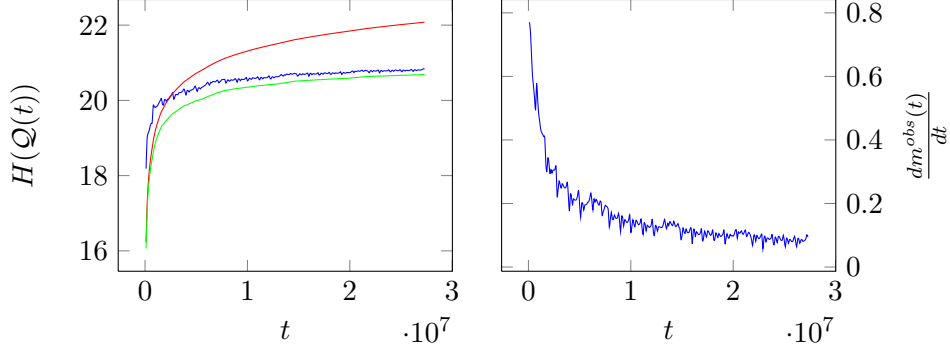


Figure 5.20: (left) The generalized entropy (blue), the Shannon entropy (green) and the entropy of a uniform distribution over all the links of the network (red). (right) $\frac{dm(x)}{dx}$, for the same data. Although the network continuously increases, most of the information is gathered during the initial phase of the growth. After about two weeks of data (half of the shown period), the generalized entropy has stabilized; additional links don't bring significant information to the system.

parts, and are lesser frequently used. Their contribution in the total network entropy is therefore smaller than the strong links, which is observed in the flat part of the entropy.

The apparent randomness in the growth of social networks shows hence some well-defined structure: the contribution of each link to the complexity of the growing network is not equal among the links, some contribute more than others, and most of these strong contributors appear early in the aggregation process. This idea is not new, we have already shown evidence of this fact in the two previous sections of this chapter, this analysis complements it as a consolidation of the findings.

To conclude this section, we go a step further in the analysis of sequences of events, and study instead the individual sequences S_i , the sequences of calls initiated by one single node. This is similar to the study performed in [176], but instead of observing sequences of localizations of nodes, we observe sequences of contacts.

When we focus on the sequence of calls of one single node, the assumption of independence of the calls doesn't remain valid, since two consecutive calls of one single user have a high probability to be correlated. This induces temporal correlations in the sequence of events, which need to be taken into account in the entropy measure. The Shannon entropy, as well as the extension to incomplete random variables, don't take these temporal correlations into account, since they solely rely on the average frequency of each link to compute the expected uncertainty. To take these temporal correlations into account, we use the *entropy rate*¹

$$H_r(X) = \lim_{n \rightarrow \infty} \frac{1}{n} H(X_1, X_2, X_3, \dots, X_n), \quad (5.21)$$

when the limit exists, where $H(X_1, X_2, X_3, \dots, X_n)$ is the joint entropy of the sequence $X_1, \dots, X_n$. If events are correlated over time, the entropy rate will be lower than the Shannon entropy of the same sequence, since the temporal correlations remove a part of uncertainty on the outcome of the next event.

Calculating the exact value of the entropy rate is not possible, we will estimate it using an estimator based on the Lempel-Ziv compression algorithm [104], which shows a rapid convergence to the entropy rate of the sequence. For n timesteps, the entropy rate is estimated by

$$H_r(X) \approx \left(\frac{1}{n} \sum_{i=2}^n \Lambda_i^i \right)^{-1} \log n, \quad (5.22)$$

with Λ_i^i being the length of the shortest substring starting at position i that has not appeared yet in the i previous steps.

We show the entropy rate over the 6 months of data averaged over the call sequences of 719 users in the network. The users were selected such that they have initiated at least one call in the first week of data (to ensure activity from the start of the dataset) and such that they made at least one call every two hours on average over the 6 months of data (which ensures that they made on average 3000 calls in 6 months).

¹In most textbooks, the entropy rate is notated H , but to avoid confusion with the Shannon entropy, we added the index r .

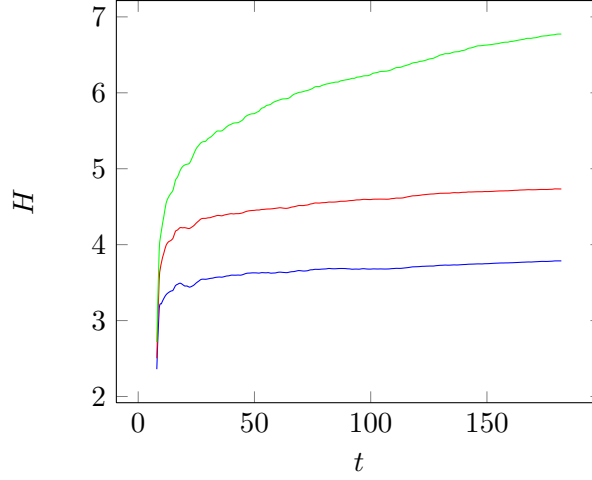


Figure 5.21: The average entropy rate H_r (blue), the average Shannon entropy (red) and the average degree-based entropy (green) of 719 nodes in the network.

The first week of the data is not taken into consideration, since it is used to detect active users. For comparison, we also show the Shannon entropy of the sequences (the entropy of the frequency distribution of the calls in each sequence, which would be the entropy rate if no temporal correlations exist in the data) as well as the average of the logarithm of the degree of the nodes, which would be the entropy if the nodes call each of their neighbours at the same frequency (that we will further call the degree-based entropy).

The average results (Figure 5.21) show that all entropies increase strongly during the first week of data, and after about 7 days only the degree-based entropy continues its increase. The sharp increase at the start reflects that the number of distinct neighbours increases, leading to a more complex landscape of calling partners. After one week, however, the entropy rate and the Shannon entropy stabilize, while the degree-based entropy continues its increase. This is only possible if the additional links are weak contributors to the entropy, understand that those links are not frequently used by nodes. As an illustration, after one week of data, on average 25% of the neighbours of a node are dis-

covered, as well as 90% of its entropy rate. The 75% remaining contacts are responsible for at most 10% of the information contained in the total time series.

5.5 Further work

The proposed models, PU and DPU generate a growing sequence of links, but also of their corresponding weights. Empirically, one can verify that after a sufficiently long time, the weight distribution produced by the model corresponds to a power law, independently of the initial weight distribution. Still, a formal proof that in all cases the weight distribution will converge to a power-law distribution is not straightforward and still missing. In this work, we did not manage to provide such a proof, however we give here the first steps of what we believe to be a promising start of the proof.

The weight distribution produced by PU can be modelled by a dynamical system, let us call y_k the number of links having weight k ($k > 1$). At each timestep, a link is chosen to either increase its weight, or to create a new link with unit weight. The probability to choose a link of weight k is then given by $\frac{ky_k}{\sum_k ky_k}$. The probability that a link increases its weight is given by $1 - \frac{K}{k}$. Finally, supposing that one starts the model with an initial graph of weight m_0 , the total weight of the graph at time t ($t = 0$ at initialisation) is given by $m_0 + t$.

The expected dynamics of these variables can then be described with the following dynamical system

$$\dot{y}_1 = \sum_i \left[\frac{iy_i}{\sum_j jy_j} \frac{K}{i} \right] - y_1 \frac{y_1}{\sum_j jy_j} (1 - K) \quad (5.23)$$

$$\dot{y}_k = y_{k-1} \frac{y_{k-1}(k-1)}{\sum_j jy_j} \left(1 - \frac{K}{k-1}\right) - y_k \frac{y_k k}{\sum_j jy_j} \left(1 - \frac{K}{k}\right) \quad \forall k > 1 \quad (5.24)$$

which directly simplifies to

$$\dot{y}_1 = \frac{1}{t + m_0} \left(K \sum_i y_i - (1 - K)y_1^2 \right) \quad (5.25)$$

$$\dot{y}_k = \frac{1}{t + m_0} ((k - 1 - K)y_{k-1}^2 - (k - K)y_k^2) \quad \forall k > 1 \quad (5.26)$$

we can get rid of the time-dependent component by scaling the time on a logarithmic scale: $s = \ln(t + m_0)$, leading to

$$\partial_s y_1 = K \sum_i y_i - (1 - K)y_1^2 \quad (5.27)$$

$$\partial_s y_k = (k - 1 - K)y_{k-1}^2 - (k - K)y_k^2 \quad \forall k > 1 \quad (5.28)$$

Let us notice that this system increases continuously; the derivatives annihilate each other, such that in the end, only the first term of the derivative of y_1 remains. Let us call $Y = \sum y_k$, the number of links in the graph, its size grows as $\partial_s Y = KY$, so $Y = e^{Ks}$. We recover the expected growth given earlier.

The weight distribution produced by the PU model after s steps is given by $\frac{y_k}{e^{Ks}}$. However, it is unclear whether this distribution reaches a stationary state, or if the average weight constantly increases. A tentative approach of this problem would be to concentrate on the dynamics of y_1 in order to search for a closed form solution. If y_1 reaches a stationary state, it is likely that the consecutive variables will do the same.

5.6 Conclusions

Network theory has been at the centre of the attention for the last decade, and has provided tools for understanding and modelling complex interactions in a well-defined environment. However, while most studied processes are dynamical, research has until recently been mostly conducted by modelling these systems by static networks. Recent advances in dynamical network studies have shown the usefulness of such an approach, may it be to observe the evolution of communities [148, 134], to understand why some links persist over time and others not [82, 74],

or to understand how the dynamics of social interactions influence the transmission of information. In this chapter, we complemented these previous research by showing how timescales play a crucial role in the structure of networks.

Most research conducted on data-driven dynamical networks base their construction method on time windows of a predefined length, from which they extract a sequence of networks. The differences and similarities between these networks are then expected to illustrate dynamical properties of the network under study. In this chapter, we show that such results are highly dependent on the length of these window sizes, and that at different time scales, different structural properties of the network emerge. Illustrating our findings with a mobile phone dataset of 6 months of data, we show that the network grows continuously, and that the average degree and the average strength of links never stabilize. After nine days, *the global clustering coefficient* of the network reaches a maximum, moment at which the dense parts (likely to be close to communities) of the network are completely unveiled. For studies aiming at studying communities, such a time scale is likely to be sufficient to unveil meaningful structures. In the meanwhile, it appears that one week of data is sufficient to *estimate correctly the degree distribution*, which rescales perfectly with distributions of longer time windows. Hence after one week some statistical properties seem to reach a stationary state, but not all, as for the weight distributions, which appear to converge more slowly to a stationary distribution. Another time scale is of high importance in this dataset: the networks built with windows of one month of data show *the highest network similarity between two consecutive windows*. Such an information would be of great interest for someone who wants to extract the “stable backbone” of the network. For smaller windows, the similarity is lower because some stable links have not been unveiled yet, while for longer windows, random links are added to the network, which leads to a lower similarity.

Besides the importance of time scales, we also show that the exact placement of time windows is of high importance. We show that networks constructed on different days of the week are likely to show very different properties, and that even holidays have a high impact on the structure of the obtained networks. Given that some studies have been

conducted on datasets as small as one day of data [4], this information needs to be taken into account when designing the study of a dataset.

In the second part of this chapter, we show that the size of such a dynamical network can be well-approximated by a power law, and provide a model that reproduces such a growth. The aim of the proposed model is not to provide a realistic explanation of the power-law growth of the network, but rather to show that with some very simple assumptions, such a growth can be obtained. By modifying slightly the proposed model, we also manage to reproduce the growth pattern that is often obtained when studying data-driven dynamic networks. In a first phase, one observes a sharp start, followed by a steady growth that can be represented by a power-law growth. We suggest that the initial strong increase is due to a large number of links that are present in the network before the start of the data collection period. These connections are all likely to be uncovered quickly after the start of the data collection period, at the same time as newly created connections also appear. After this discovery period, these “hidden” links are all uncovered, and only newly created links contribute then to the growth of the network, which is then characterized by a power-law growth of the network. Using this model, it is then possible to predict the future growth of the network, but also to detect exceptional events and uncover temporal regularities in the creation of new connections.

Even though a dynamical network is constantly growing, most information is gathered in the early steps of the aggregation. The last part of this chapter is strongly related to the finding that the most frequent links appear early in the aggregation process. We quantify the amount of information gathered in this aggregation process with a variant of the classical Shannon entropy of the link weight distribution that takes the dynamical component into account. Applied on the mobile phone dataset, this measure shows that after a week, 98.5% of the information contained in the network constructed in one month. Going a step further, we show that if one focuses on the series of calls made by single nodes over 6 months of data, 25% of their contacts are gathered in one week, which account for 90% of the information contained in their sequence of calls. Hence after one week, the most significant neighbours of each node are already uncovered, and the remaining 75% can be con-

sidered as of less importance.

The initial aim to study dynamical networks was to find the right time scale to study real-world datasets. Clearly, a dynamical network is not constructed around one time scale, and there is not one correct time window to study such networks. Instead, dynamical networks are constructed around a complex interweaving of different processes, which lead to a set of different time scales which each uncover different properties of the underlying network. For one who wishes to uncover communities or to study the statistical distributions, a first study of the time scales of the dataset is necessary to ensure that the correct time windows are used for the purpose of the analysis. As a final point, let us note that the present study has been limited to the study of one dataset, and evidently calls for further work. Although one may expect to observe similarly different time scales in other types of data, it is less evident to predict which results achieved on this dataset will be reproduced in other datasets. In particular, order of appearance of the different time scales (rescaling distributions, communities, similar time windows) might change.

Chapter 6

Multiscale community detection

In this chapter, we address the issue of **multi-level community detection**, more precisely the detection of communities by using a scale parameter. It presents two important issues, one concerning the **quality** of the results, and one on the **computational effort** needed to obtain informative results. We address the first point by showing that in addition to the well-known resolution limit a modularity-based method suffers of an **additional bound** on the size of communities. We also propose a way to reduce the computational effort by proposing a **fast multi-level algorithm**, that also provides information on the significance of the different levels of communities.

6.1 Introduction

Motivation Community detection in graphs has for long time been identified as a **multi-level problem**, meaning that different levels of organization can be represented by different levels of communities. The lowest levels correspond thus to small and tightly bound communities, while the highest levels correspond to large groups of nodes that are weakly connected - but still more than expected in a random network. Methods have been developed since the beginnings of community detection to uncover this multi-level structure, such as *agglomerative* or *divisive* algorithms [76] which produce a dendrogram of partitions. Other

methods use successive optimizations to produce a hierarchical tree, where the results of a level are used as input for the subsequent level [31]. Those methods implicitly constrain the results to be **hierarchically** organized, meaning that there is a strict relationship of inclusion between different levels.

More recently, methods that were developed to detect communities at one single level have been revisited by adding a scale parameter into the quality function to optimize. It is for example the case of *modularity* [140]. In this case, the scale parameter has been added in order to tune the resolution at which the communities are detected, and to handle in this way the issue of the so-called resolution limit [72]. There are many similar quality optimization methods, most are particular cases of the very general Potts model [162].

So, while fast algorithms have already been developed to detect communities given a precise scale parameter, there doesn't exist a way to know in advance what **scale parameter** to use. This has pushed researchers to test a large number of scale parameters, and to decide *a posteriori* which of them are significant, based on a quality criterion. Given the large number of possible parameters, the computational effort is extremely large only because a large number of detections are needed to test the whole range of parameters. In this chapter, we address this issue and propose a multi-level algorithm that uses the results produced at intermediate levels to accelerate the search. The resulting algorithm is particularly efficient if the network has a multi-level structure.

Since the introduction of the scale parameter was made in order to handle the issue of resolution limit, we will address this issue shortly in the next section. In particular, we will show how the introduction of the scaling parameter leads to an additional bound on the size of the optimal communities. However, the intent of this chapter is not to give an advice on which method is appropriate for research and which is not. In some applications, the resolution limit may not be a problem, thus methods suffering from a resolution limit may be totally appropriate in some cases. The fact that a method suffers from a resolution limit should not be seen as a red flag for using the method, but indicates that the results have to be considered with caution. In the practical application that

we will present at the end of the chapter, we use a resolution limit-free method, for reasons linked to the application itself.

Notations In this chapter, we consider a community detection method that treats communities as partitions of the nodes N of the network. A partition P of N is a set of non-overlapping groups of nodes, such that $\bigcup_{\pi \in P} \pi = N$, hence each node of N belongs to one and only one group. Each group π is a community.

The multi-level methods addressed in this chapter uses a scaling parameter, denoted γ hereafter. Its interpretation varies from author to author, here we will use the most common interpretation: if $\gamma = 0$, large communities with weak density conditions are favoured (usually each connected component of the network is one community), if $\gamma \gg 0$, high density conditions are favoured, usually each community is a clique or a single node. The scale parameter can be used in conjunction with various quality functions. In this chapter, we will mainly address the scaled version of modularity and the Constant Potts Model (CPM), introduced in [183]. The scaled version of modularity will be written Q_γ (abusively, Q is used for traditional modularity, when $\gamma = 1$), while CPM will be defined by H_γ .

Structure of the chapter The chapter is structured as follows. In the first section we explain the origins of multi-level community detection, especially in the case of scaled modularity, and we show how the resolution limit problem is not completely solved by the scale parameter. In the second section, we propose a fast algorithm for multi-level community detection. We provide proofs of the correctness of the algorithm, and show how the results of the algorithm indicate the presence or absence of communities. Finally, in the third section we show a practical application of the algorithm on a network of criminal offenders, made in collaboration with the Belgian Federal Police. Since community detection has still many open questions, we conclude and suggest further work in the last section.

6.2 Multi-level community detection

Many popular methods for detecting communities in networks are based on the optimization of the modularity function, which is a measure of the quality of a network partition into communities. The modularity of a partition compares the density of links inside communities to the corresponding density expected in a null model [140]. It has been shown by Fortunato and Barthélemy [72] that modularity suffers from a so-called resolution limit: modularity optimization methods often fail to identify small communities.

Several authors have proposed objective functions for community detection that incorporate a tunable resolution parameter so as to allow community detection at different scales. One such function introduced by Reichardt and Bornholdt in 2006 [161], can be written in the following form:

$$Q_\gamma = \sum_{\pi} \frac{l_{\pi}}{m} - \gamma \left(\frac{d_{\pi}}{2m} \right)^2. \quad (6.1)$$

In this expression, the sum is over the communities π , l_{π} is the number of links inside community π , d_{π} is the sum of the degrees of the nodes in π and m is the total number of links in the network. The case $\gamma = 1$ corresponds to modularity for the configuration model as defined by Newman [140]. Higher resolutions are obtained by choosing higher values for the resolution parameter γ in Equation 6.1. Objective functions that are mathematically equivalent to Q_γ have been proposed in a number of other contexts. In particular, Delvenne *et al.* have shown [56] that the function Q_γ corresponds to the first-order approximation of a dynamical process driven by the Laplacian of the graph where the resolution parameter plays the role of a timescale (see also [116]). The function Q_γ is also a special case (for $\omega = 0$) of the function used by Mucha *et al.* [134] to study so-called multislice networks. It has been shown by Kumpula *et al.* [111] that methods based on the optimization of Q_γ suffer from a resolution limit similar to the one reported [72] for $\gamma = 1$.

An upper bound on scalable community detection While it is known that an increase of the resolution parameter γ lowers the resolution limit of the method, we can show that any resolution parameter

value $\gamma > 1$ imposes a non-trivial upper bound on the size of communities. To establish this bound, consider two communities whose node degrees sum to, respectively, d_1 and d_2 and contain, respectively, l_1 and l_2 internal links. Let also e be the number of links connecting the two communities. Compare now the situation where the communities are separate with the one where the two communities are merged into one. In the latter case, the total degree of the community is given by $d = d_1 + d_2$ and the total number of links is equal to $l_1 + l_2 + e$. An elementary calculation shows that the difference in the objective function between these two situations is given by

$$\Delta Q_\gamma = \frac{1}{m} \left(e - \gamma \frac{d_1 d_2}{2m} \right) \quad (6.2)$$

with separate communities leading to a larger value of the objective function when $\Delta Q_\gamma < 0$.

Since $e \leq d_1$, we have

$$\Delta Q_\gamma \leq \frac{1}{m} \left(d_1 - \gamma \frac{d_1 d_2}{2m} \right) = \frac{d_1}{m} \left(1 - \gamma \frac{d_2}{2m} \right) \quad (6.3)$$

and so $\Delta Q_\gamma < 0$ when $d_2/(2m) > 1/\gamma$. Thus, if one can find a set of nodes in a community whose total node degrees exceed $1/\gamma$ of the total node degrees in the network, then the value of the objective function increases when making this set of nodes a separate community. This imposes a non-trivial upper bound on community sizes as soon as $\gamma > 1$. In particular, a community of n nodes may not contain a fraction of the total degree (or of the total number of links) larger than $n/((n-1)\gamma)$.

Let us now show with an example that the use of a resolution parameter may disaggregate large and densely connected communities. Consider the network consisting of a clique of 16 nodes and of four cliques of 4 nodes each. There is one edge between the clique of 16 nodes and each of the cliques of 4 nodes. All pairs of cliques of 4 nodes are connected to each others with 2 links (Figure 6.1). The partition of optimal modularity ($\gamma = 1$) consists of two communities of 16 nodes each, as shown on the left of Figure 6.1. This is a typical illustration of the resolution limit where modularity optimization fails to detect the four small cliques of 4 nodes. When γ is increased to enable the detection of the smaller cliques, the larger clique of 16 nodes splits into 16 distinct communities of one node each (middle of Figure 6.1, $\gamma = 1.5$). As γ is further increased,

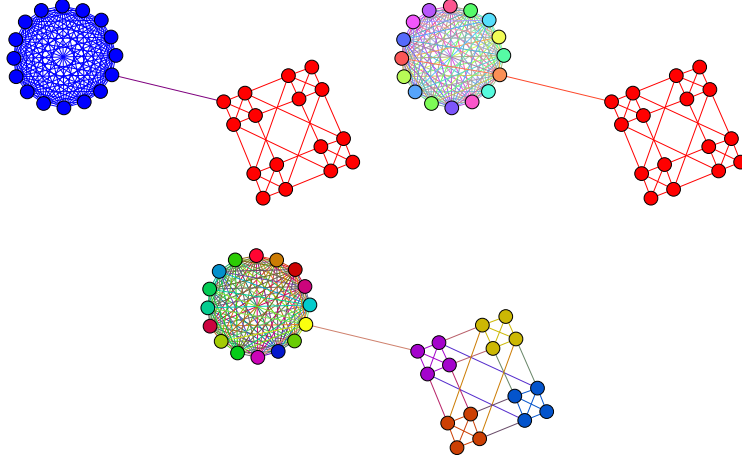


Figure 6.1: Community partitioning of the same network for different values of the tunable resolution parameter γ . On the left, the partition obtained for $\gamma = 1$ consists of two communities of 16 nodes each; this is a typical example of the resolution limit of modularity where modularity optimization fails to detect the four small cliques of 4 nodes. As the resolution parameter is increased to $\gamma = 1.5$, the method still fails to detect the four small cliques but the nodes in the large clique now form sixteen distinct one-node communities. When $\gamma = 2$ (right) the four small cliques are finally separated.

the second community finally splits into 4 cliques (right of Figure 6.1, $\gamma = 2$).

As this simple example shows, when optimizing the objective function Q_γ for $\gamma > 1$ one should be aware of the tendency of the resulting optimum to disaggregate large and dense communities and be cautious when interpreting the partitions obtained.

Issues on the resolution limit The preceding example could be reproduced for other quality functions, many global quality functions suffer from a resolution limit, even when a scale parameter is introduced. In fact, even the definition of a resolution limit-free method is not trivial. While it is easy to show, with some appropriate examples, that a method suffers from a resolution limit, there has not been, until recently, a proper

definition of the conditions for being free of any resolution limit. In a recent paper, Traag et al. [183] propose a mathematical formulation of conditions for resolution limit-free methods, and show that only a few methods fulfil those conditions. They also propose a quality function, the Constant Potts Model (CPM), that is resolution limit-free following their definition.

This model is striking of simplicity, but shows excellent performances on traditional benchmark graphs. It is built in the same philosophy as scaled modularity, but instead of comparing the density of a community to the configuration model, it compares it to a constant value

$$\mathcal{H}_\gamma = \sum_{\pi} l_{\pi} - \gamma \sum_{\pi} n_{\pi}^2 \quad (6.4)$$

Note that in expression, the normalization $\frac{1}{m}$ has disappeared. The quantity n_{π}^2 is the square of the number of nodes of community π , corresponding to the maximal possible number of links in the community. Note also that this expression is very similar to a measure proposed by Reichardt and Bornholdt [161], in which γ is multiplied by a parameter p , which represents the average probability of two nodes to be connected in the whole network.

The algorithm that we will present in Section 6.3 may be applied to a large variety of quality functions, but we will apply it with CPM in the case of the examples.

Computational challenges in multi-level community detection

Besides the qualitative issue of the influence of a resolution limit on the solutions, the introduction of the scale parameter provides an additional computational difficulty. The scaling parameter γ represents one degree of freedom of the detection method, and optimal partitions are detected independently for different values of γ by using the user's favourite optimization method. This procedure leads to a sequence of "optimal" partitions associated to a sequence of corresponding values of γ . However, it is clear that not all values of γ represent a significant level of organization of the network. Thus, the question of the choice of the "right" scale parameter has become a central question in multi-level community detection. Let us illustrate this with an example.

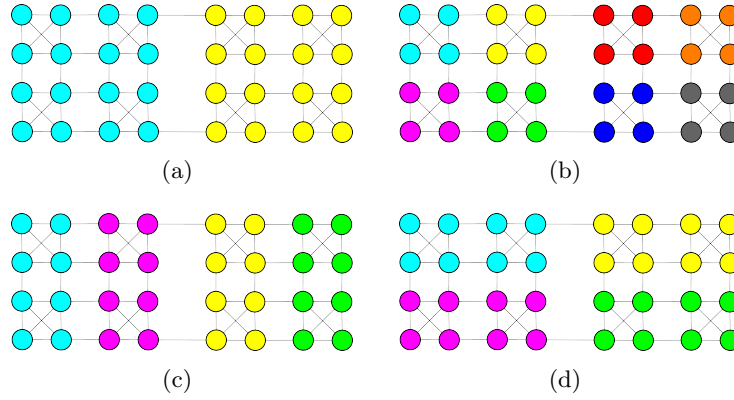


Figure 6.2: Communities at several levels. (a) partition for a low value of γ . (b) partition for a high value of γ . (c) and (d) two equivalent partitions for an intermediate value of γ .

The network shown on Figure 6.2 is composed of 8 cliques of 4 nodes, which are disposed in two groups of 4 cliques (left and right). Inside the two groups, each clique has two links to its two adjacent cliques, and there are only two links between the two groups. For a low value of γ , the only optimal partition is the left-right partition showed on Figure 6.2(a). For a high value of γ instead, the optimal partitions will be the cliques (Figure 6.2(b)). In those two cases, there is no alternative for the optimal partition, essentially because those partitions correspond to a real structural organization of the network. For an intermediate value of γ , however, there is no such structural organization, and the partitions shown on Figure 6.2(c) and 6.2(d) are completely equivalent from the perspective of the quality function (this holds for any quality function). This intermediate level is then possibly less representative of a strong structural organization than the two other levels. It is important to note that there is no actual consensus on whether degeneracy can be used as an actual proxy for badness of the partition. In some cases, there might exist two perpendicular segmentations of the nodes of a network, that both represent a significant partitioning, but that are equivalent. Several attempts have been made to provide a robust method that determines of the significance of γ . The main idea behind most proposed solutions is that for being significant, the partition at a given γ needs to be ro-

bust to perturbations on the problem. If the optimal solution of the perturbed problem is close to the optimal solution of the original problem, the original solution is said to be significant. There exist several ways to introduce perturbations in the community detection problem: by making small random modifications of the structure of the network [95], by modifying slightly the scale parameter [6], or by modifying the optimization algorithm used for detection communities. A more extensive list of proposed modifications is available in [114]. This method will be used in the remainder of this work, let us note, however, that this is not the only available way. Seen for example, Lancichinetti *et al.* [118] who propose instead to determine the significance of communities with the probability that the same group of nodes would be detected in a random network and references herein.

While the nature of the perturbation differs between the methods, they all follow the same procedure: given a range of possible values for γ , they detect the optimal communities for a sequence of values of γ , each γ being tested one or several times, depending on the type of perturbation. Then, based on a robustness criterion, some values of γ are said to be robust and others not. In such a situation, the researcher needs to make a choice between testing a large number of samples for γ and working fast. Knowing that for large networks, the possible values for γ may range over 10 or more orders of magnitude, the desired precision on γ has a high impact on the total computation time. This might sound like a never-ending story; a few years ago, it was hard to find good communities in a reasonable time on a large network. Nowadays, fast algorithms producing results of good quality are available (see for example [31]), but a large number of detections are needed, which leads again to an excessive computational time.

We propose a solution to this problem in the next section with an algorithm that allows to significantly reduce the number of scale parameters to test when the network has a real community structure at multiple levels.

Quality function	L_P	D_P
CPM	$\sum_{\pi} l_{\pi}$	$\sum_{\pi} n_{\pi}^2$
Scaled modularity	$\sum_{\pi} \frac{l_{\pi}}{m}$	$\sum_{\pi} \left(\frac{d_{\pi}}{2m}\right)^2$
Reichardt and Bornholdt (RB)	$\sum_{\pi} \frac{l_{\pi}}{m}$	$\sum_{\pi} p_{\pi}$
Ronhovde and Nussinov (RN)	$\sum_{\pi} l_{\pi}$	$\sum_{\pi} n_{\pi}^2 - l_{\pi}$
Label propagation	$\sum_{\pi} l_{\pi}$	0

Table 6.1: Expression taken for L_P and D_P for different quality functions that apply to the form (6.5). The RB model has been introduced in [162] where p_{π} stands for the expected number of links inside community π given a null model. The RN model has been introduced in [165]. For the sake of simplicity, only the unweighted versions of the quality functions are given here, the extension to weighted networks is straightforward.

6.3 Fast multilevel community detection

The present algorithm has been initially designed using CPM as quality function, but it can be applied to a range of quality functions, that all take a similar functional form, namely

$$\mathcal{H}_{\gamma} = L_P - \gamma D_P \quad (6.5)$$

where L_P and D_P are constants determined by the partition in communities P . The significance of L_P and D_P varies depending on the quality function, the most notable of them are given in Table 6.1. The quality of a partition thus depends only of those two values, and on γ . The following proposition is then straightforward:

Proposition 1. *Given P, P' two partitions on the same network, and γ, γ' two scale parameters such that*

$$\gamma < \gamma'; \quad (6.6)$$

$$\mathcal{H}_{\gamma}(P) \geq \mathcal{H}_{\gamma}(P'); \quad (6.7)$$

$$\mathcal{H}_{\gamma'}(P) \leq \mathcal{H}_{\gamma'}(P'). \quad (6.8)$$

Then, $D_P \geq D_{P'}$.

The three inequalities imply that P is a better partition than P' for γ and P' is a better partition than P for γ' .

The proof is immediate, considering two linear functions, it is a simple condition on their derivatives.

Proof: Let us write $\gamma' = \gamma + \epsilon$, $\epsilon > 0$.

We have then

$$\mathcal{H}_{\gamma'}(P) = \mathcal{H}_{\gamma}(P) - \epsilon D_P \quad (6.9)$$

$$\mathcal{H}_{\gamma'}(P') = \mathcal{H}_{\gamma}(P') - \epsilon D_{P'} \quad (6.10)$$

Supposing that $D_P < D_{P'}$, we have immediately a contradiction with 6.7 and 6.8. ■

Since this is true for any pair of partitions respecting the 3 conditions, it is also true for two optimal partitions for two different levels of γ . We can thus deduce the profile of D_P for a sequence of optimal partitions:

Lemma 1. *Given a set $\mathcal{P}$ of partitions on the same network and an ordered sequence of scale parameters $\Gamma = \{\gamma_0 < \gamma_1 < \dots\}$. Let the sequence $\{P_0, P_1, \dots\}$ be a sequence of the best partitions of $\mathcal{P}$ for the scale parameters of Γ . Then, the corresponding sequence $\{D_{P_0}, D_{P_1}, \dots\}$ is monotonically decreasing.*

Also, in the case of two partitions having the same value of D_C , it is always possible to rule out one partition.

Lemma 2. *If two partitions P and P' are such that $D_P = D_{P'}$, they are either equivalent ($L_P = L_{P'}$), or one is always dominating the other ($H_{\gamma}(P) > H_{\gamma}(P') \quad \forall \gamma$ or $H_{\gamma}(P') > H_{\gamma}(P) \quad \forall \gamma$).*

Let us finally remark that if a partition P is optimal for γ and γ' ($\gamma < \gamma'$), it is also optimal for the interval $[\gamma, \gamma']$. If this wasn't the case, if there existed γ'' and P' such that $H_{\gamma''}(P') > H_{\gamma''}(P)$, by applying twice Proposition 1 we would obtain $D_P > D_{P'} > D_P$, which is impossible.

Given that the number of partitions on a finite set of nodes is finite, the sequence of optimal partitions for γ in the continuous interval $[\gamma_{min}, \gamma_{max}]$ is also finite. We can thus completely characterize the profile of D_P .

Theorem 1. *Given the finite set $\mathcal{P}_{opt}$ of partitions such that $\forall \gamma \in [\gamma_{min}, \gamma_{max}]$, the partition P which maximizes $\mathcal{H}_\gamma(P)$ is in $\mathcal{P}_{opt}$. Then, we can order the partitions of $\mathcal{P}_{opt}$ as $\{P_0, P_1, \dots, P_k\}$ such that $\{D_{P_0}, D_{P_1}, \dots, D_{P_k}\}$ is a decreasing sequence and $\exists$ a sequence $\{\gamma_{min} = \gamma_0, \gamma_1, \dots, \gamma_{k+1} = \gamma_{max}\}$ such that P_i is the optimal partition for the closed interval $[\gamma_i, \gamma_{i+1}]$.*

Proof: Since there are an infinite number of possible values for γ , and only a finite number of partitions in $\mathcal{P}_{opt}$, there exists at least one partition P in $\mathcal{P}_{opt}$ that is optimal for more than one value of γ . Let us take the minimum and maximum values, $\gamma_{P,min}$ and $\gamma_{P,max}$ for which P is optimal.

P is then optimal for the interval $[\gamma_{P,min}, \gamma_{P,max}]$. We can then remove P from $\mathcal{P}_{opt}$ and the open interval $] \gamma_{P,min}, \gamma_{P,max} [$ from $[0, \gamma_{max}]$, and start again the same reasoning. Since $\mathcal{P}_{opt}$ is finite and $\forall \gamma \in [0, \gamma_{max}]$, the partition C which maximises $\mathcal{H}_\gamma(P)$ is in $\mathcal{P}_{opt}$, in a finite number of steps, we will have built the desired sequence. From Proposition 1, we know that the sequence $\{D_{P_0}, D_{P_1}, \dots, D_{P_k}\}$ is a decreasing sequence. Finally, we need to prove what happens at the border of the intervals. Let us suppose that P is optimal for γ and that P' is optimal for all γ' such that $\gamma' < \gamma$. We also know that $D_{P'} > D_P$ and $\mathcal{H}'_\gamma(P') > \mathcal{H}'_\gamma(P)$. Let us write then $\gamma = \gamma' + \epsilon$.

We can rewrite $\mathcal{H}_\gamma(P) = \mathcal{H}_{\gamma'}(P) + \epsilon D_P$ and $\mathcal{H}_\gamma(P') = \mathcal{H}_{\gamma'}(P') + \epsilon D_{P'}$. Since ϵ can be as small as we want, and that we always need to have $\mathcal{H}'_\gamma(P') > \mathcal{H}'_\gamma(P)$, the only possibility is that $\mathcal{H}_\gamma(P) = \mathcal{H}_\gamma(P')$, and thus both partitions are optimal for γ . As a consequence, both partitions are optimal at the intersections of the intervals. ■

This result provides the profile of the optimal partitions: for a given interval $[\gamma_{min}, \gamma_{max}]$ of possible values for γ , the optimal partitions are a finite sequence of partitions for which the sequence of $D_P(\gamma)$ is a decreasing function in the shape of stairs. Each plateau $D_P(\gamma)$ corresponds to an optimal partition. The profile of $\mathcal{H}_\gamma$ is then a convex piecewise linear monotonically decreasing function of γ .

We can exploit this result to reduce the number of detections needed. Suppose there exists an algorithm that finds the partition that maximizes $\mathcal{H}_\gamma$ in polynomial time. If, for two different values γ and γ' , this algorithm returns optimal partitions with identical D_P , one knows that these partitions are also optimal for the intermediate values between γ and γ' . It is thus not necessary to detect communities for any value inside the interval, since an optimal partition is already known.

Such an algorithm capable of returning the exact optimal partition in polynomial time is not known to date. The best available algorithms are heuristics that produce in a short time partitions of good quality. The algorithm given hereunder takes this into account in order to detect as well as possible the plateaus of the function $D_P(\gamma)$.

Algorithm 1 Multi-Level Community Detection (MLCD)

Require: $G, \gamma_0, \gamma_{max}, \epsilon_\gamma, \epsilon_D$
Partitions $\leftarrow []$
ToDoList $\leftarrow$ Pile()
 $P_0 \leftarrow \text{BestPartition}(G, \gamma_0)$
 $P_{max} \leftarrow \text{BestPartition}(G, \gamma_{max})$
ToDoList.push($[\gamma_0, \gamma_{max}]$)
Partitions.insert($((P_0, \gamma_0))$)
Partitions.insert($((P_{max}, \gamma_{max}))$)
while length(ToDoList) > 0 **do**
 next $\leftarrow$ ToDoList.pop()
 ToDoList.push(OLCD(next))
end while
return Partitions

MLCD is a recursive algorithm, which calls a sub-routine. This sub-routine (OLCD, described below) takes as parameter an interval and the best known solution on its borders.

OLCD may encounter three different cases. The first case is that one solution on one side of the interval dominates (with respect to H_γ) the solution on the other side. This may happen since the community detection algorithm *BestPartition* is a heuristic and does not return the optimal solution. In that case, the dominated solution is replaced by the dominating solution, and the interval adjacent to the dominated solution is returned for check. This step is called *propagation*.

The second case, if no propagation occurred, is that either both partitions are similar (their values of D_P do not differ by more than ϵ_D) or that the width of the interval is smaller than ϵ_γ (the width is calculated on a logarithmic scale). In this case, the algorithm stops for this interval.

Finally, if none of the two previous cases happen, OLCD uses the community detection method to find the best partition for γ at the geometric centre of the interval, and returns as next intervals to check the first and second half of the original interval. This case is called *exploration*.

Algorithm 2 One Level Community Detection (OLCD)

Require: $G, [\gamma_i, \gamma_{i+1}], P_i, P_{i+1}, \epsilon_\gamma, \epsilon_D, \text{Partitions}$

Case 1: Propagation

if $\mathcal{H}_{\gamma_{i+1}}(P_i) > \mathcal{H}_{\gamma_{i+1}}(P_{i+1})$ **then**

$P_{i+1} \leftarrow P_i$

return $[\gamma_{i+1}, \gamma_{i+2}]$

end if

if $\mathcal{H}_{\gamma_i}(P_{i+1}) > \mathcal{H}_{\gamma_i}(P_i)$ **then**

$P_i \leftarrow P_{i+1}$

return $[\gamma_{i-1}, \gamma_i]$

end if

Case 2: End of algorithm

if $D_{P_i} - D_{P_{i+1}} < \epsilon_D$ **then**

return $\emptyset$

end if

if $\log \gamma_{i+1} - \log \gamma_i < \epsilon_\gamma$ **then**

return $\emptyset$

end if

Case 3: Exploration

$\gamma_{new} \leftarrow \sqrt{\gamma_i \gamma_{i+1}}$

$P_{new} \leftarrow \text{BestPartition}(G, \gamma_{new})$

$\text{Partitions.insert}(P_{new}, \gamma_{new})$

return $[[\gamma_i, \gamma_{new}], [\gamma_{new}, \gamma_{i+1}]]$

To sum up, the MLCD algorithm functions as a dichotomical search for plateaus in the interval $[\gamma_0, \gamma_{max}]$. It starts by detecting the communities on its borders γ_0 and γ_{max} , and if both solutions differ (if the difference between their values of D_P is larger than ϵ_D), it searches for the best partition at the centre of the interval, and recursively runs again on the two halves. The dichotomical search is made on a logarithmic axis.

The advantage of the algorithm is that it avoids searching for communities in the middle of plateaus. If the solutions left and right of an interval are similar, there is no need to search for an optimal solution inside it, which saves a large number of unnecessary trials. Moreover, the obtained plateaus provide in turn important information on the stability of their partitions. If a plateau is large, the corresponding partition is robust to a small perturbation on γ . Finally, one needs to determine which of the plateaus are significant and which are not. In practice, one expects that the whole range of possible values of γ is composed of a small number of intervals over which the partitions are stable, separated by “transitive” values of γ for which the obtained partitions are less stable. The stable intervals are then expected to appear as intervals where the degeneracy of the solutions is – relatively speaking – lower than the degeneracy of their surrounding transitive partition levels. We explain in the following how the degeneracy of solutions can be quantified using a measure coming from information theory.

Normalized variation of information to compare partitions To detect partitions that are likely to be degenerated, we apply several times the MLCD algorithm, and compare for each γ the optimal partitions given by the different runs. To compare the different partitions, we use a metric called the normalized variation of information (NVI), an information-theoretic based measure introduced by Meila [132]. Each partition can be associated to a random variable that gives the probability that a randomly chosen node will be in each community. We associate partitions to P and P' the probability distributions of two random discrete variables X and Y . The probability $p(X = x)$ is the probability that a randomly chosen node belongs to community x of P , and $p(Y = y)$ represents the same for communities y of P' . If n_x is the size of community x , $p(X = x) = \frac{n_x}{n}$.

If both partitions are identical, the outcome of X provides us a complete information on the outcome of Y , *i.e.* if one knows the outcome of X for one particular node, one also knows the outcome of Y for that node and vice-versa, while instead if both partitions are uncorrelated, the outcome of one variable does not give any information about the other variable. This is captured in the NVI, which measures the difference between the uncertainty on the outcome of each variable and how much information

they share.

The variation of information (VI) measures the uncertainty on the outcome of one variable knowing the outcome of the other.

$$VI(X, Y) = H(X) + H(Y) - 2I(X, Y) \quad (6.11)$$

where $H(\cdot)$ is the Shannon entropy of a probability distribution and the information I is given by

$$I(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(X = x, Y = y) \log \frac{p(X = x, Y = y)}{p(X = x)P(Y = y)} \quad (6.12)$$

where $p(X = x, Y = y)$ is the joint probability of having a node being in community x of P and at the same time in community y of P' . If both partitions are uncorrelated, this probability is $p(X = x)p(Y = y)$, and hence the information is 0. The NVI is then the normalized version of (6.11), which allows the results to be robust to a change of the entropy of both distributions

$$NVI(X, Y) = 1 - 2 \frac{I(X, Y)}{H(X) + H(Y)} \quad (6.13)$$

With the use of this tool, we compare partitions in the same way as what is done in [114] by taking the average NVI over all pairs of partitions returned by the different runs of MLCD for the same parameter γ .

Zachary's karate club network In order to help the reader to understand how the results of the algorithm need to be decoded, we show here a first small example, after which we will show the performance of the algorithm on graphs that are either known to show community structure, or on the contrary that do not have any community structure. In what follows, we have used as community detection method the algorithm proposed by Traag et al. [183], which optimizes CPM with a heuristic routine very similar to the Louvain method [31]. The original code of the method has been provided by Vincent Traag, who implemented the Louvain method to detect partitions using the CPM on unipartite networks. We modified this method to allow detection on bipartite networks (see the application in the next section) and implemented the multi-level algorithm around this one-level method.

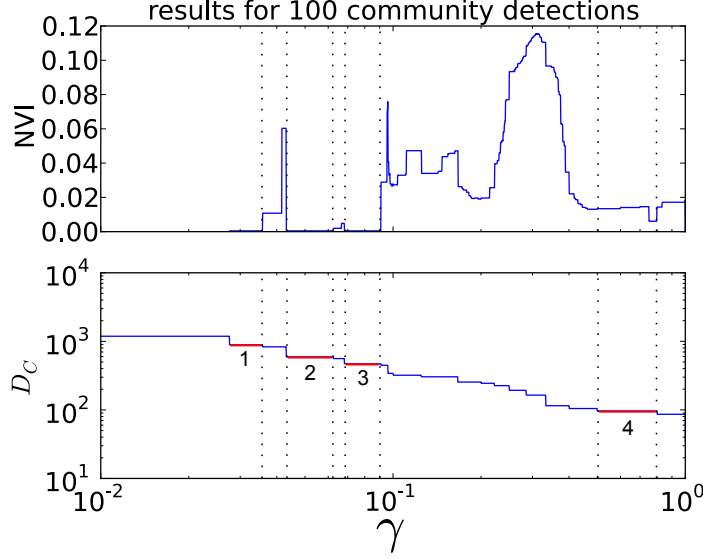


Figure 6.3: NVI and D_P profile for Zachary's karate club network, for 100 runs of MLCD, with parameters $\epsilon_D = 1$ and $\epsilon_\gamma = 10^{-3}$. The most stable plateaus are shown in red. The dashed lines are intended as a guide to the eye.

We show the results on the well known Zachary's Karate club network [193] (see Figure 6.3). The MLCD algorithm has been applied 100 times on the network, and for each possible value of γ , the NVI between the 100 optimal partitions has been computed.

In order to consider a partition as stable, one needs to have a large plateau for D_P , and a low NVI. On Figure 6.3, the most stable partitions – except for the trivial partition where all nodes are in the same community – are shown in red in the graph of D_P . The partition designated by number 2 shows both a large plateau and a NVI of 0 value, which means that the 100 runs have returned the same partition for the corresponding values of γ . This partition, shown on Figure 6.4(b), corresponds to the historical separation of the club. Partitions 1 and 3 show intermediate situations; both partitions are also returned by each run

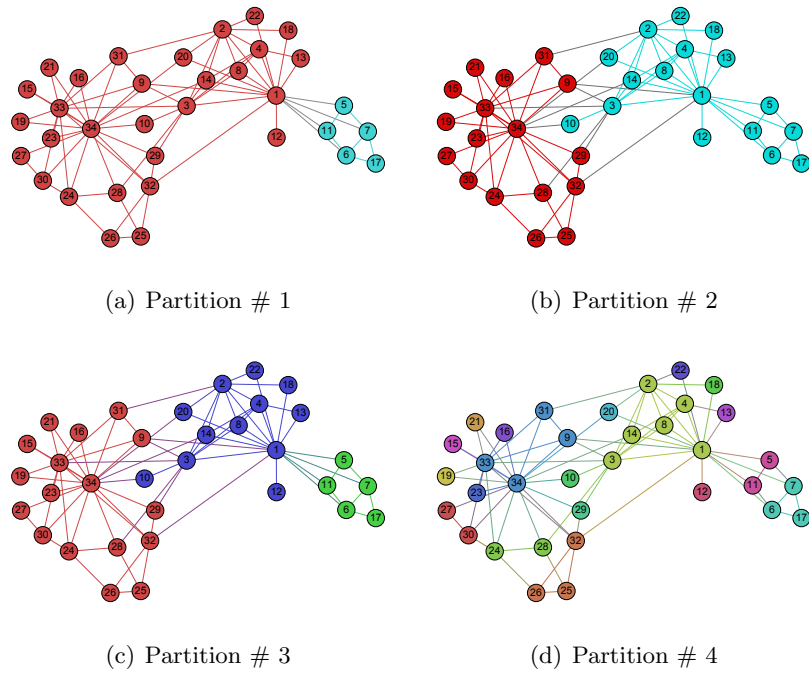


Figure 6.4: Partitions on the Zachary's karate club network, corresponding to the plateaus shown in Figure 6.3. Partition (b) corresponds to the historical separation of the club.

(NVI is 0), but their plateaus are narrower than partition 2, meaning less stability with respect to a perturbation on γ . Finally, partition 4 has a large plateau, but corresponds to a situation where several optimal partitions exist (degeneracy of solutions), as is shown by the non-zero NVI.

The results on the karate club network illustrate the ambiguity of the obtained solutions: two different indicators are taken into account to quantify the quality of a partition, hence it is not always possible to give a clear ranking between the candidate plateaus. Is it better to have a larger plateau or a lower NVI? In this example, is partition 2 better than partition 4? So far, we cannot propose a clear solution to this issue.

Results on Lancichinetti-Fortunato benchmark networks The algorithm has been tested on several instances of the benchmark networks introduced by Lancichinetti and Fortunato [117]. These benchmarks are constructed based on an *a priori* known community structure. The size n_π of the communities is decided such as to fit a power law ($p(n_\pi = n) \propto n^{-\tau_2}$). The degrees of the nodes are also decided such that they fit a power law ($p(d_i = d) \propto d^{-\tau_1}$). Finally, the stubs are connected randomly, with probability $1 - \mu$ inside communities and with probability μ between communities. The tuning of the topological mixing μ allows then to produce networks of different configurations. The interest of these benchmarks is that they have a community structure at only one level; since links are drawn randomly, so as to match a certain density inside the communities, there is no additional hierarchical structure in the communities. On Figure 6.5, we show the results produced by the algorithm for 100 runs. Strikingly, the only stable partition corresponds indeed to the partition *a priori* decided. Besides this partition, no other partition is found by the algorithm. The method also shows the range of scale parameter values for which the correct communities are detected.

In the absence of communities... A good community detection method finds communities when communities are present in the network, and does not find any when the network has no community structure. We show this by applying MLCD on Erdős-Rényi random networks, which are known not to have any community structure. Indeed, MLCD shows no stable partition on the Erdos-Renyi network, as expected. Some small

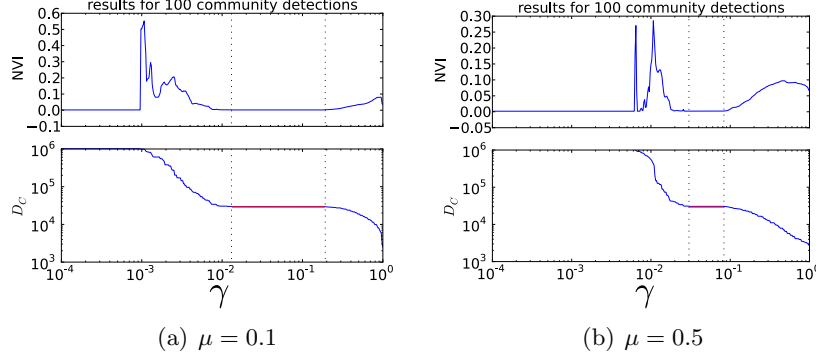


Figure 6.5: NVI and D_P profile for two instances of the Lancichinetti-Fortunato benchmark. Both networks have been generated with 1000 nodes, an average degree of 20 and a maximum degree of 50. Degrees fit a power-law distribution with exponent -2, and community sizes range from 10 nodes to 50 nodes, and fit also a power-law distribution with exponent -1. The network in (a) has a topological mixing of 0.1 (for each node, 90% of its links are within its community) and in (b) has a topological mixing of 0.5. In both cases, the NVI and D_P profile have been computed over 100 runs, with parameters $\epsilon_D = 1$ and $\epsilon_\gamma = 10^{-3}$.

locally stable partitions are observed, which comes from the fact that the network are not purely random, and some locally dense substructures appear in the generated networks.

Computational advantages of the algorithm The advantage of MLCD over classical multilevel methods, such as presented in [114] is that MLCD detects only communities at the boundaries of plateaus of D_P , which reduces the number of calls to the community detection method. However, it is not possible to quantify exactly what gain in computational time is achieved by this approach, since the number and width of plateaus depends on the community structure of the network. If the network is well-organized in multi-level community structures, D_P is composed of a few large plateaus, and hence MLCD needs only a few exploration steps to detect the transitions. If the network does not contain any community structure, such as an Erdos-Renyi network, D_P is

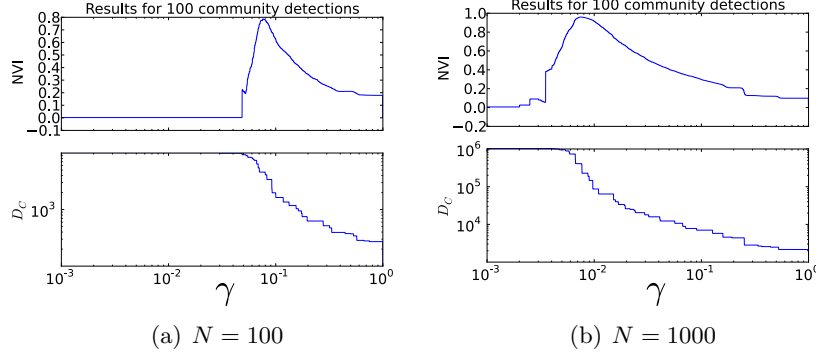


Figure 6.6: NVI and D_P profile for two instances of the Erdos-Renyi random network. In (a), the network is composed of 100 nodes, in (b), the network is composed of 1000 nodes. Both networks have an average degree of 10. In both cases, the NVI and D_P profile have been computed over 100 runs, with parameters $\epsilon_D = 1$ and $\epsilon_\gamma = 10^{-3}$.

composed of a large number of very short plateaus, and MLCD takes approximately the same time as methods that detect communities by sampling a large number of values of γ . To illustrate this, we applied MLCD on a sequence of networks showing an increasing level of randomness. The initial network is shown at Figure 6.7(a), it is composed of 256 nodes, organized in 3 levels of communities. The first level of communities are cliques of size 4. The second level of communities are groups of 4 communities of level 1, and the communities at level 3 are groups of 4 communities at level 2. To ensure connectedness between the communities at level $i - 1$ inside the level i , each pair of communities of level $i - 1$ belonging to the same community of level i are connected by two randomly placed links. In the following, we call this network $3CK_4$, in reference to the 3 levels of communities based on 4-cliques.

This hierarchically structured network is expected to perform very well using MLCD and to show three large plateaus. To increase the randomness of the network, we rewire randomly a fraction p of the links of the network, chosen at random. Increasing p leads then to a deterioration of the community structure, $p = 1$ producing a random network.

An example for $p = 0, 0.3$ and 1 is shown at Figure 6.7.

The shape of D_P and the NVI for increasing value of p changes from

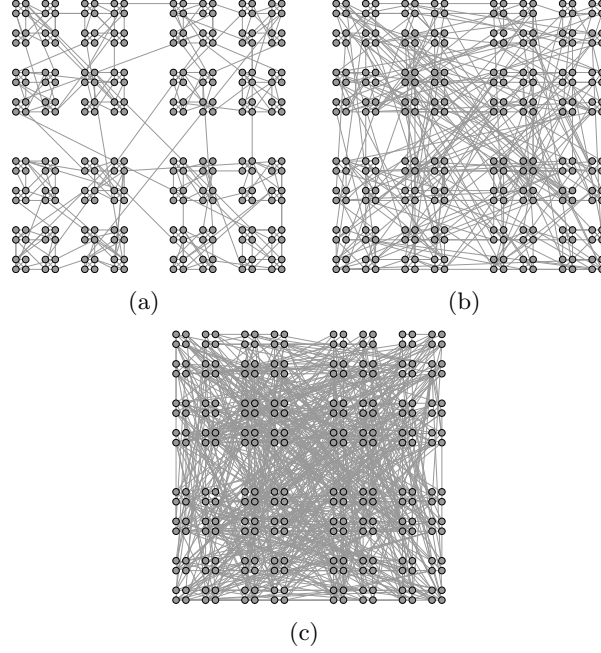


Figure 6.7: Example of a $3CK_4$ network with (a) $p = 0$, (b) $p = 0.3$, (c) $p = 1$. All networks are composed of 256 nodes and 636 links.

the three large plateaus, corresponding to the three levels of hierarchy ($p = 0$) to a shape similar to what has been obtained for Erdos-Renyi networks ($p = 1$). Figure 6.8 shows the increase of computational time for p varying from 0 to 1.

6.4 Results on a network of criminal offenders

MLCD was developed in a joint project with the Belgian Federal Police. The project's objective was to detect communities in a network of criminal offenders. Several criminal activities, such as burglary or drug traffic require the association of several people with different competences. Each time a group of people are caught for an illegal activity, their col-

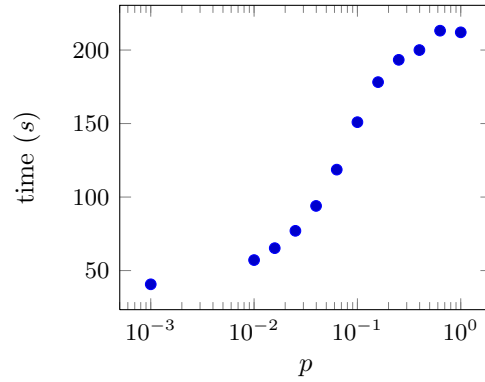


Figure 6.8: Average runtime for 100 runs of MLCD and computation of the NVI on $3CK_4$ networks with varying p . The point at $p = 10^{-3}$ represents the situation when no links are rewired from the original network.

laboration is recorded, which leads to a network of criminal offenders. A natural question then is if this network is organized in communities, which would reveal organized structures.

A dataset of criminal facts The studied dataset consists of the records of 5 years – from 2006 until 2010 included – of criminal facts known to the Belgian Federal Police. A *criminal fact* is an offence or attempt of offence against a law of the penal code or to a special law as defined by the ministry of interior, that is located in time and space and for which at least one person is known to be responsible for. Such persons are further called *actors*, criminal facts will be simply denoted by *facts*.

The dataset contains 1.4 million facts and 755,000 actors, every actor is involved in at least one fact. Every fact is identified by a unique identification number, and has a list of descriptors:

Time interval: The precise moment at which a fact has happened is often not known, most facts are therefore associated to a time interval in which it has certainly happened. The width of such a time interval may vary from a couple of minutes to several years.

Fact type: Every fact is designated by a numerical code coming from a dictionary of 287 different codes that describe the type of fact. These types are general descriptions.

Fact qualification: Every fact is designated by a numerical code from a dictionary of 1323 different codes that describe the qualification of the fact. These qualifications are more descriptive than types. For example, the fact type 52 is used for all facts that concern selling drugs, can relate to several different qualifications, such as unauthorised selling of drugs to a child under 12 years old, unauthorised selling of drugs to a child between 12 and 16 years old, unauthorised selling of drugs to a child above 16 years old, unauthorised selling of drugs to an adult. . .

NIS code: The commune code of the area where the fact has happened.

Fact environment description: Every fact has a numerical code that describes the place where the fact happened. The code comes from a dictionary of 717 different standard descriptions of places, with varying precisions. These descriptions can be anything: “a house”, “a castle”, “a kitchen”, “a bakery”, “a church”, . . .

It is important to note that every fact is completely described and that the descriptions are reliable, which will be useful for validation purposes.

Every actor is also identified with a unique identification number, and, in some cases, additional information about the actor is known. This information is in most cases incomplete, and its veracity is not always ensured, since the actors might provide false information about their identity. In some cases, it happens that the same actor has given several distinct dates of birth, several nationalities, several home addresses, etc. This is why most actor description fields contain more than one entry:

Gender: For most actors, their gender is specified. This information is encoded for 98.5% of the actors.

Nationalities: For some actors, one or several nationalities are known. Five fields are available, in which the five first known nationalities are recorded under a numerical code. The information for the first

to fifth nationality is encoded for respectively 96.5%, 0.6%, 0.04%, 0.005% and 0.001% of the actors.

NIS codes: Two fields are available for the official home address of the actor, and two other fields are available for the known address in use of the actor (if they do not live at their official address). The addresses of the two official and two known addresses are encoded for respectively 77.9%, 0.06%, 12.8% and 0.002% of the actors.

Years of birth: Five fields are available for the years of birth given by the actor, available for respectively 99.9%, 1.2%, 0.3%, 0.1% and 0.07% of the actors.

Specialities: Two fields are available for numerically-coded specialities coming from a dictionary of 43 different descriptions. This descriptor is used if the actor is known to be principally active in one particular activity, for instance strawpersons in financial criminal organizations, chemists or sellers in drug organizations... This information is available for respectively 6.0% and 0.003% of the actors.

Finally, each fact has a list of actors that are known to have participated to the fact, which will generate the network structure.

It is important to note that the available dataset represents only a small fraction of the reality. In this particular dataset, even more than any other, the nodes avoid to be seen. On average, the Police estimates that only 15% of the facts that are happening are finally assigned to at least one known offender in their database. This might have a high influence on the structure of the resulting network.

From the dataset, we build a bipartite network of actors and facts, where each actor is connected by a link with the facts in which he/she is involved. This results into a network of 591,707 connected components with sizes ranging from 2 to 270,615.

The high sparsity and disconnectedness of the network might be due to two particular reasons: either the available data is too poor, and the many missing links have disconnected the network, or some facts do not result from a networked organization. By grouping the connected

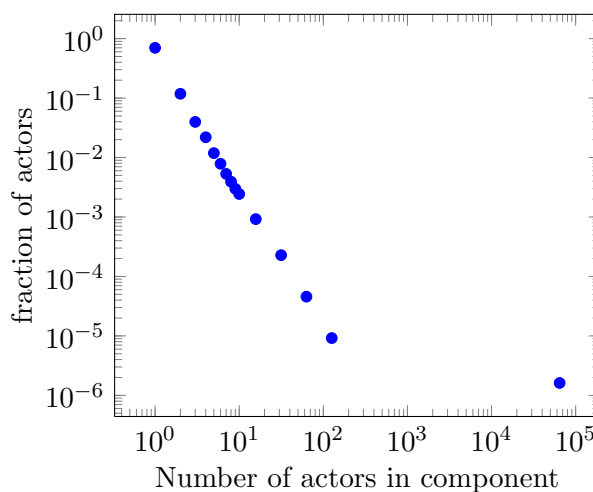


Figure 6.9: Distribution of the number of actors per connected component. 70% of the actors are the only actor of their component, only 7% are in the largest connected component.

components by their number of actors, and analysing the types of facts they contain, we see that the composition of the connected components varies following the number of actors they contain (see Figure 6.10). Connected components with one actor are dominated by facts such as driving license removal, or identification control. Those facts are indeed not related to any group activity, but rather to single persons. Larger connected components, instead, contain mostly facts related to theft, which is more related to a networked activity.

The sparsity of the network is probably due to both reasons, but the distribution of the fact types suggests that small connected components contain mostly facts that are not targeted by this study. In the remaining of the study, we have focused on the connected components that contain at least 5 actors. After this filtering, the network is composed of 4,856 connected components and contains 91,400 actors and 303,464 facts.

Structure of the network One of the reasons why complex networks have attracted the interest of researchers recently is that they all share a

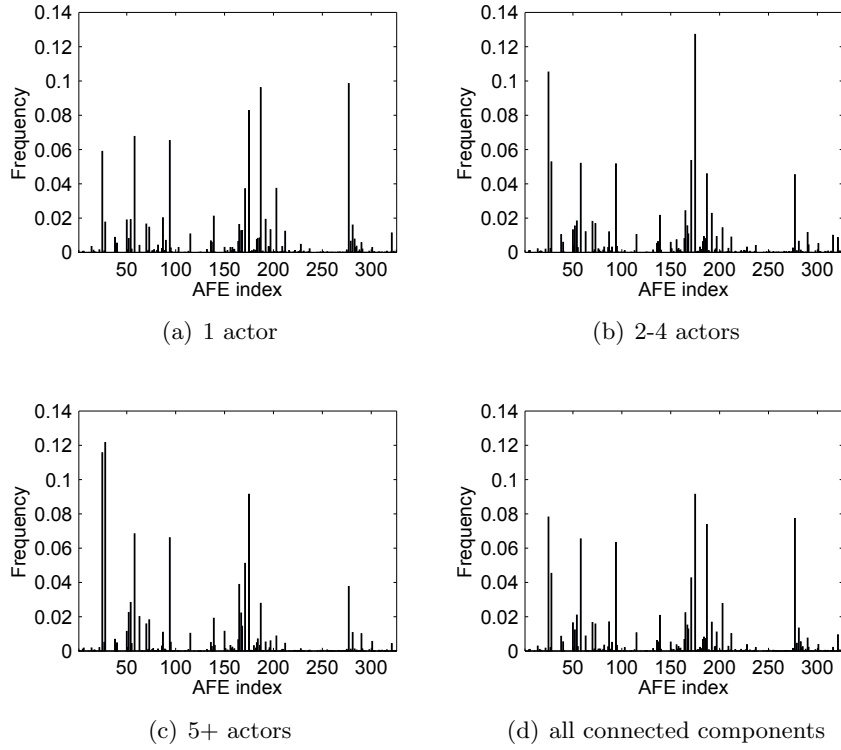


Figure 6.10: Frequency of each type of fact in the different kinds of connected components, sorted based on their number of actors. The connected components with one single actor are dominated by *losses of driving license* and the *special law on identification cards*, the components of 2 to 4 actors by *assault and battery* facts and the components of 5 and more actors by *simple and qualified theft*.

set of non-trivial characteristics, such as broad degree distributions and high clustering, that random networks are unable to reproduce. These features also characterize the network of criminal offenders. Since the network has a bipartite structure, the measures have been separated for both classes of nodes, such that any difference between classes can be detected. The degree distribution of actors and facts (see Figure 6.11) ranges over several orders of magnitude, there is a wide variety in the

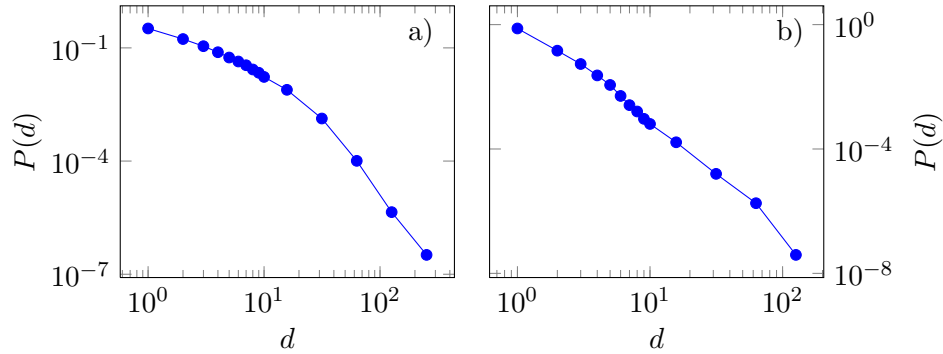


Figure 6.11: Degree distribution of (a) actors and (b) facts. Both distribution range over several orders of magnitude, some actors are involved in 300 different facts, and some facts involve up to 100 actors.

level of activity of actors and facts. However, it is important to note that contrarily to classical complex networks, such as social networks, the high-degree nodes might be of low importance. Since this network deals with criminal offenders, the most connected node is not necessary the most active node, but one that has often been caught. This can be viewed as a *godfather effect*: the most important nodes might be very weakly connected.

While the degree of an actor shows how many facts he/she is involved in, the **distance 2 degree** says how many collaborators he/she has. For facts, it represents the number of other facts in which its actors are involved in.

We compare the distance 2 degree distribution of actors and facts in the network to the same measure in a network in which the links have been randomly rewired so as to keep the same degree distribution than in the original network (i.e. the configuration model for bipartite networks).

As shown on Figure 6.12, the distance 2 degree is higher in the randomized network than in the original network. Knowing that the degree distributions of both networks are identical, this suggests that the network should be locally dense.

The **clustering coefficient** is a usual measure of the local density of

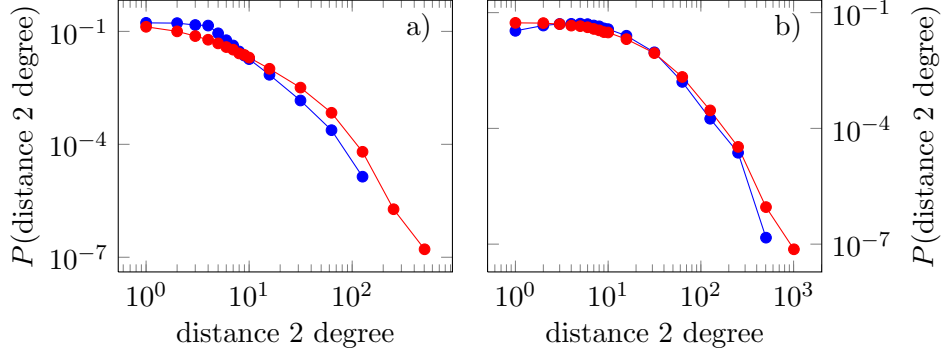


Figure 6.12: Distribution of distance 2 degree in the real network (blue) and randomized (red) for actors (a) and facts (b).

the network. In a classical unipartite network, it measures the transitivity of the network by counting the density of triangles around a node. In a bipartite network, however, triangles do not exist, and the clustering coefficient needs to be redefined. Borgatti and Everett have suggested an extension where the clustering coefficient is defined for pairs of nodes based on their number of common neighbours [35] and Latapy et al. have proposed to average the previous result on all distance 2 neighbours of a node to obtain a measure related to a single node [119]. However, none of these definitions have the same philosophy as the original clustering coefficient, which is why we introduce here a measure of clustering that captures properties similar to that of the usual clustering coefficient.

We define the clustering coefficient in a bipartite graph as the density of the network between the neighbours at distance 1 and 2 of the node. If a node has d_i neighbours at distance 1 and $d_{i,2}$ neighbours at distance 2, the minimal number of links between those two levels is $d_{i,2}$ (since each node at distance 2 needs to be connected to a node at distance 1). The maximal density is $d_i d_{i,2}$, in which case each node at distance 2 is connected to each node at distance 1. The clustering coefficient reads then

$$C(i) = \frac{z - d_{i,2}}{d_i d_{i,2} - d_{i,2}}, \quad (6.14)$$

where z is the number of observed links between the neighbours at dis-

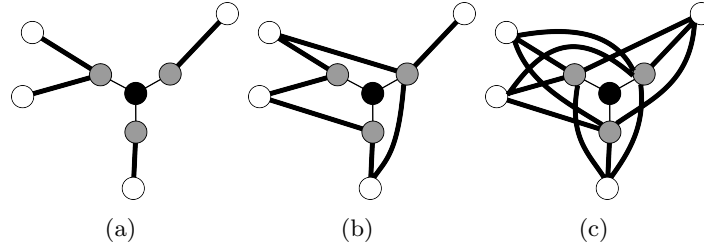


Figure 6.13: Example of clustering coefficient of the black node. It has 3 neighbours at distance 1 (grey) and 4 neighbours at distance 2 (white). The clustering coefficient is then computed by counting the number of thick links, between those two levels. In (a) the density is minimal $C = \frac{4-4}{12-4} = 0$. In (b), the clustering is $\frac{7-4}{12-4} = 0.375$. In (c), the network is as dense as possible $C = \frac{12-4}{12-4} = 1$.

tance 1 and distance 2 (see Figure 6.13 for an example)

We show in Figure 6.14 the distribution of the clustering coefficient in the network (blue) and in a network where the links have been randomly rewired so as to keep the degree distribution identical to the original network (red). We see that most nodes have a low clustering coefficient, but that in the real network, a significant number of nodes have a higher than expected clustering coefficient, which indicates that the network is locally dense. This is a good indicator of the presence of communities.

Bipartite communities Many unipartite networks are based on bipartite datasets, from which a *projection* is made in the space of one particular class of nodes. For example, a mobile phone network is originally a bipartite network of customers and phone calls, where each customer is connected to the phone calls in which he/she participates. While in that particular case the projection from a bipartite network to a unipartite network is not an issue because each call only involves exactly two customers, the projection process usually destroys some information on the network structure. This has led researchers to advocate the use of the bipartite structure of networks, instead of using one of their projections [122]. However, the difference between the community structure one would obtain in a bipartite network and in one of its projections is not

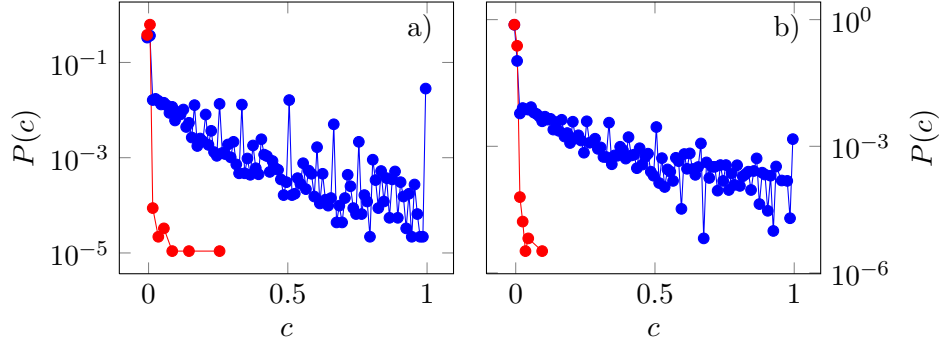


Figure 6.14: Distribution of the clustering coefficients in the real network (blue) and randomized (red) of the actors (a) and facts (b).

trivial. In this paragraph, we address this question and show how the result of CPM provide local information when detecting communities in a bipartite network and global information in one of its projections.

On a bipartite network, CPM has to be slightly modified. The second term, $\gamma \sum_{\pi} n_{\pi}^2$, represents the number of possible links inside the community π . In a bipartite network, where nodes are divided in two classes, the number of possible links between them is given by the product of their sizes, hence CPM reads

$$\mathcal{H}_{\gamma} = \sum_{\pi} l_{\pi} - \gamma \sum_{\pi} n_{\pi} n'_{\pi} \quad (6.15)$$

where n_{π} and n'_{π} are the number of nodes of both classes in the community π .

Consider a bipartite network with two classes of nodes Δ and Γ . This network is defined by its adjacency matrix $A \in R^{n_{\Delta} \times n_{\Gamma}}$ in which each row correspond to a node from Δ and each column to a node from Γ . The projection on Δ is given by $A^{\Delta} = AA^T$ and its projection on Γ by $A^{\Gamma} = A^T A$.

Let us remark that if a node in Δ is connected to k nodes in Γ , they will form a clique in the projection on Γ , hence a node of degree d creates

d^2 links in its projection.

A community π in the bipartite network is constituted of a group of nodes of Δ , called π_Δ of size n_π and a group of nodes of Γ , called π_Γ of size n'_π .

For a given γ , the detection of communities corresponds to the following optimization problem

$$\max_P \sum_{\pi \in P} \sum_{\substack{i \in \pi_\Delta \\ j \in \pi_\Gamma}} (A_{ij} - \gamma n_\pi n'_\pi) \quad (6.16)$$

There exists a wide literature on clustering methods on bipartite network, given the large number of possible applications *e.g.* in text mining or bioinformatics. Most methods developed in this area are known under the generic term of co-clustering, or bi-clustering. These methods have the same goal as community detection on a bipartite network, *i.e.* finding groups of nodes in a network that can be represented by a bipartite structure such that nodes of the same group share similar properties. However, the method proposed here differs from classical co-clustering methods in the same way as classical community detection methods differ from classical clustering methods: community detection methods use the network structure to detect dense parts of the network, while clustering methods often rely on distance measures between items to define the optimal groups.

One can rewrite the objective function (6.16), by writing $N_{\pi_\Delta}(j)$ for the number of neighbours of j in π_Δ , as

$$\max_P \sum_{\pi \in P} \sum_{j \in \pi_\Gamma} (N_{\pi_\Delta}(j) - \gamma n_\pi n'_\pi) \quad (6.17)$$

In the projection on Δ , this corresponds to

$$\max_P \sum_{\pi \in P} \sum_{j \in \pi_\Gamma} (N_{\pi_\Delta}(j)^2 - \gamma n_\pi n'^2_\pi) \quad (6.18)$$

Let us notice that here, we simply rewrite the significance of the function optimized in the bipartite network in the projection on Δ .

If one aims to detect communities directly in the projection on Δ without

taking into account the bipartite network, the maximization problem reads as

$$\max_{P_\Delta} \sum_{\pi \in P_\Delta} \sum_{\substack{i \in \pi \\ j \in \pi}} \left(A_{ij}^\Delta - \gamma n_\pi^2 \right) \quad (6.19)$$

where P_Δ is a partition on the nodes of Δ .

Knowing that $A_{ij}^\Delta = \sum_{k \in \Gamma} A_{ik} A_{jk}$, one obtains

$$\max_{P_\Delta} \sum_{\pi \in P_\Delta} \sum_{k \in \Gamma} \sum_{\substack{i \in \pi \\ j \in \pi}} \left(A_{ik} A_{jk} - \gamma n_\pi^2 \right) \quad (6.20)$$

which is equivalent to

$$\max_{P_\Delta} \sum_{\pi \in P_\Delta} \sum_{k \in \Gamma} \left(N_\pi(k)^2 - \gamma n_\pi^2 \right) \quad (6.21)$$

By writing $\gamma_\Delta = \frac{\gamma}{n_\Gamma}$, one finally obtains a form similar to (6.18):

$$\max_{P_\Delta} \sum_{\pi \in P_\Delta} \sum_{k \in \Gamma} \left(N_\pi(k)^2 - \gamma_\Delta n_\Gamma n_\pi^2 \right) \quad (6.22)$$

The maximization problems (6.18) and (6.22) correspond to the detection of communities, on a bipartite network (Eq. (6.18)) and on its projection (Eq. (6.22)). Both problems can be transposed *a posteriori* to a community detection problem in the projection network, but with different link weights. In problem (6.22), the link weight between two nodes is given by the number of common neighbours they have in the bipartite network. In problem (6.18), instead, the link weight between two nodes is given by the number of common neighbours *from their own community* they have in the bipartite network. In this sense, one could say that a community detection method on a bipartite network provides information on the local connectivity of nodes, while its equivalent in one projection provides information on their global connectivity.

Communities of criminal offenders We apply 5 times MLCD on the bipartite network of criminal offenders presented above, for which the algorithm finds about 10 large plateaus (see Figure 6.15) separated by large steps with tiny steps inside the plateaus. We recall that the

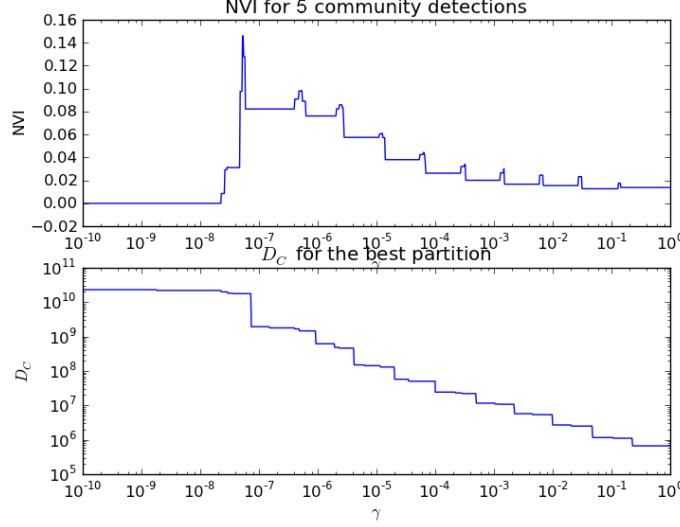


Figure 6.15: D_P and NVI for the network of criminal offenders. The plateau corresponding to the lowest NVI is centred around $\gamma = 0.1$.

profile of D_P is the profile of the optimal partition out of the 5 partitions returned by the 5 runs of MLCD. The jumps in the NVI are shifted with respect to the transitions between the plateaus. It is hard to give an exact ranking of which plateau corresponds to the most stable partition. Given that their width is approximately identical, we focus in the following on the plateau centred around $\gamma = 0.1$, which has the lowest NVI .

For that specific plateau, MLCD returns 46,077 different communities, containing between 1 and 53 actors and between 1 and 217 facts. A closer analysis of the communities shows that they correspond to significant communities. We illustrate this by computing for each community its most frequent fact type, *i.e.* the type of fact that is the most observed in the facts of the community. We show then the probability distribution p_f of the most frequent fact type frequency, *i.e.* the fraction of nodes that belong to a community that has as most frequent fact type frequency f . In Figure 6.16, we show the distribution for all communities that have at least two facts ($\bullet$), as well as the cumulative of this

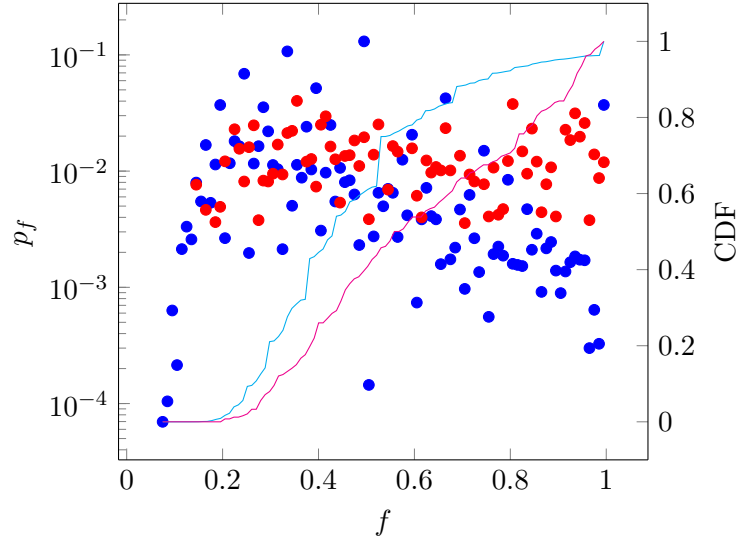


Figure 6.16: Probability density function (dots, left axis) and cumulative density function (straight lines, right axis) of the maximal fact type frequency for all communities in the network of at least two facts ($\bullet$), and for all communities of at least two facts and at least 50 nodes ($\bullet$).

distribution (---), and the same for communities that have at least two facts and 50 nodes in total (facts and actors undistinguished, ($\bullet$) and (---)).

For all community sizes, the probability peaks between 20% and 40%, and a particularly high peak is noticeable for 100%. We observe that if small communities are removed, the distribution tends to be close to a uniform probability. From the cumulative, we notice that half of the large communities have a maximal fact type frequency higher than 60%, for the small communities, this fraction is only 20%. Let us notice that these low values can partly be due to the variety of possible fact types. Remember that the fact types are numerical codes from a dictionary of 287 types, some of which concern similar activities. For instance there exist at least 7 different fact types related to drugs: import/export, manufacturing, sale, consumption, help to consumption, possession and other. Hence, if all these types were set under one single description, it is likely that for many communities, the maximal fact type frequency

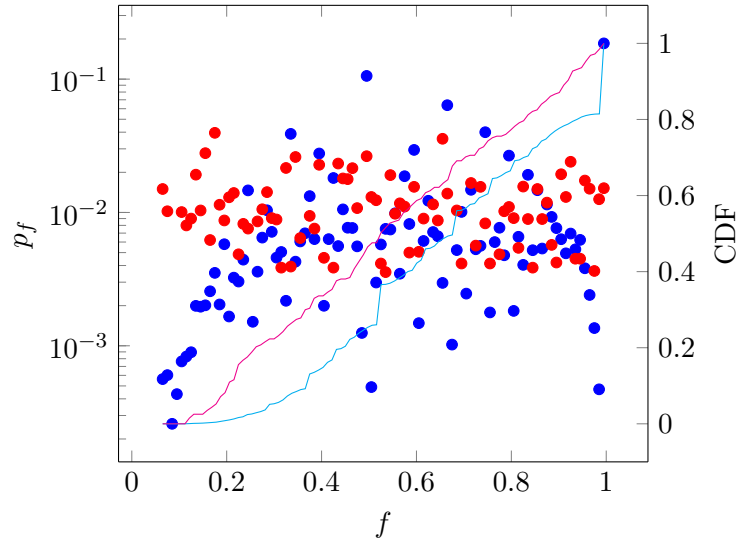


Figure 6.17: Probability density function (dots, left axis) and cumulative density function (straight lines, right axis) of the maximal commune frequency for all communities in the network of at least two facts (●), and for all communities of at least two facts and at least 50 nodes (●).

would be higher than what is actually observed. The same situation happens for other types of facts, which are subdivided into a number of sub-classes in the data.

Instead of using fact types, one can also measure the cohesion of communities based on the locality where the facts composing the community happen. We perform the same analysis as above, but using instead the commune code of the place where each fact happens. Hence, instead to have a maximal fact type frequency, we observe now the maximal commune frequency, which is the fraction of facts that happened in the most frequent commune. Again, we plot the probability density function of the maximal commune frequency p_f , where p_f is the fraction of nodes that belong to a community of maximal commune frequency f . This is shown on Figure 6.17, for all communities with at least two facts (●) and for all communities with at least two facts and 50 nodes (●), as well as their cumulatives. We observe that many communities are

geographically consistent: 60% of the communities have a maximal fact commune frequency higher than 60%. Again, when small communities are removed, the distribution evolves to a uniform distribution.

This network is too large for being displayed. We show instead a higher-level partition, corresponding to $\gamma = 10^{-3}$, which contains fewer communities and can be displayed (Figure 6.18). Only the communities belonging to the giant component, which is composed of 270,000 nodes (50,000 actors and 220,000 facts) are shown. The nodes in the network represent communities, their size is proportional to the number of nodes (facts and actors) that are part of the community, and their colour renders the fraction of those nodes that are facts (red for 0% and green for 100%). Since there are more facts than actors, it is expected to observe that communities are mostly composed of facts.

To show the potential use of these results, let us show how the structure of the obtained communities allow to uncover significant structures from this large database. A visual analysis of the obtained communities shows highlights the singularity of the brown-red structure that can be observed in the top-right corner of Figure 6.18. This group of communities seems to be well-structured but poorly connected to the remainder of the network (only one link connects it to the rest of the network). We show a zoom on these communities in Figure 6.19, notice that only links inside and between these communities have been kept, links to the rest of the network have been removed. A close analysis shows that this network corresponds to a chain of human trafficking, with all facts happening close to the Belgian coast – where trucks often cross the Channel to reach the British coasts – and all actors being foreigners coming from one Asian country. Although most of the actors in this subnetwork have nothing in common, the use of a community detection algorithm has allowed in this case to highlight their similarity.

While the structure of groups of communities allows to uncover some surprising similarities between actors, the study of single communities also show some significant structures, that would be hard to detect without these analysis tools. In Figure 6.20 we show a zoom on an other community, taken from Figure 6.18, again, notice that only the links inside the community have been kept for visualisation purposes. In this community, the actors are all connected by 6 facts in the centre of the

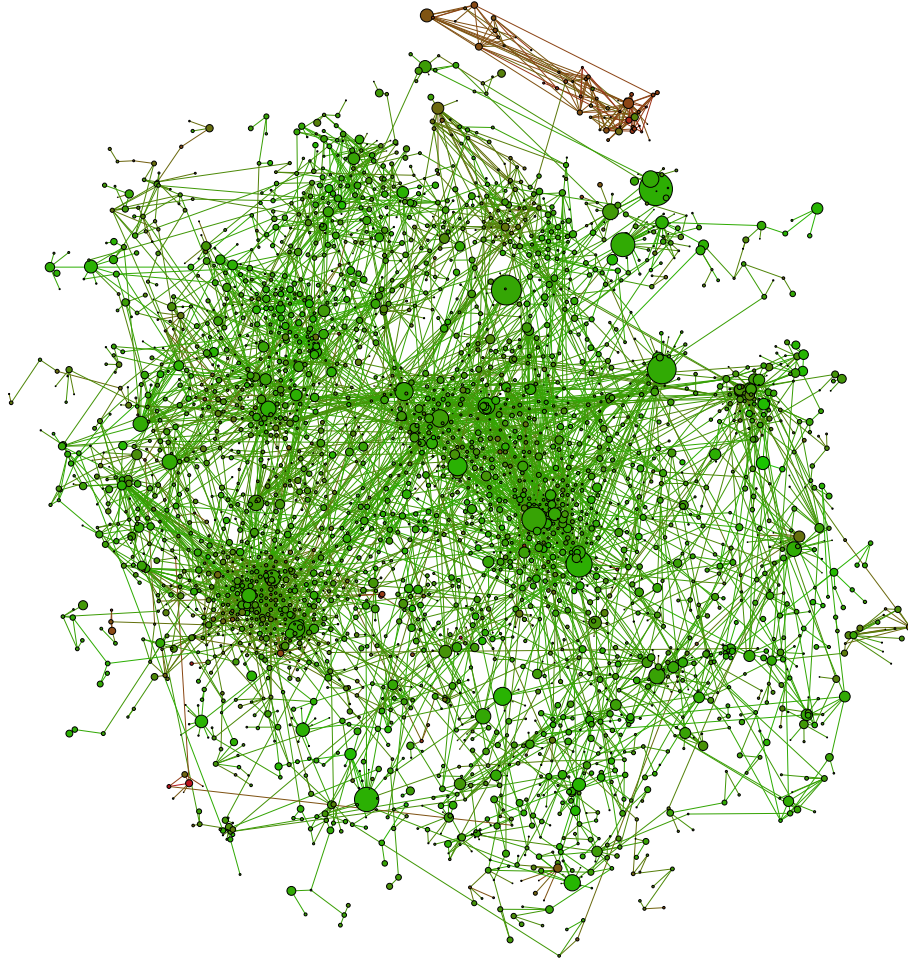


Figure 6.18: Network of communities of the giant connected component of the network for $\gamma = 10^{-3}$. The size of the nodes is proportional to the number of actors and facts that are part of the community. The colour of each node is given by the fraction of nodes in the community that are facts. Light green indicates that 100% of the nodes in the community are facts, light red indicates that 100% of the nodes in the community are actors. Of course, none of those two extreme cases happen.

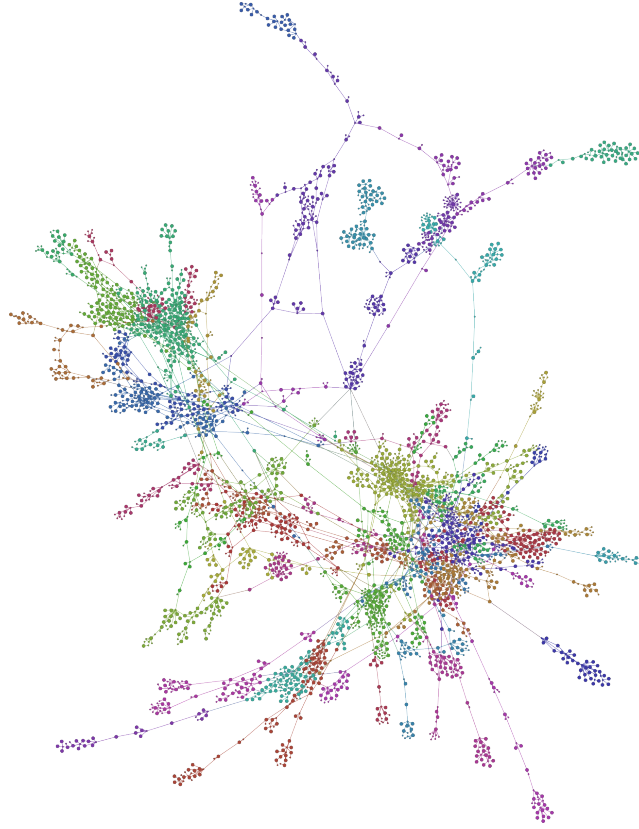


Figure 6.19: Zoom on the 51 communities composing the brown-red-coloured structure on the top right part of Figure 6.18. Large nodes correspond to actors, small nodes to facts, the colors represent the communities. This subnetwork is composed of 2109 actors and 1935 facts.

image, which are all related to a prisoners escape. Strikingly, a large majority of the other facts in this community take place in the same city, showing a remarkable geographical coherence of the community.

To give a grasp of the internal structure of those communities, we show on Figure 6.21 the partition of the second largest connected component for the plateaus corresponding to $\gamma = 10^{-3}$ and $\gamma = 10^{-1}$. Given that the communities have been detected using CPM, which is

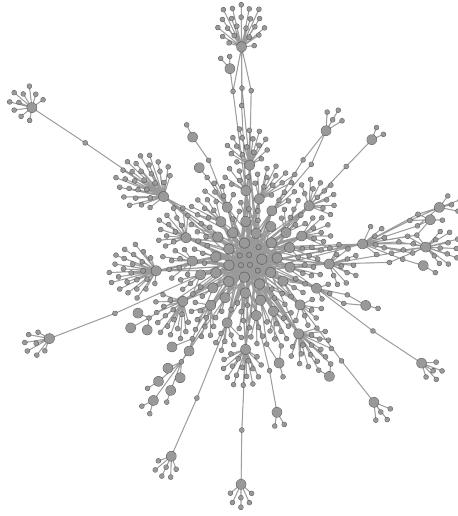


Figure 6.20: Zoom on one large community from Figure 6.18. Large nodes correspond to actors, small nodes to facts, the colors represent the communities. This community is composed of 80 actors and 459 facts. The 6 facts in the centre are all connected to the same 49 actors, and are related to prisoners escape.

a resolution-limit free method, the aspect of the communities in the giant component are expected to be similar to the ones observed in this component. In this figure, we see that communities can be constructed with diverse kinds of structures. In 6.21(a), the top blue community is composed of a sequence of actors connected in a linear-like structure, each pair of actors in this linear structure being connected by one single fact. The green community, instead, is composed of a small number of actors all sharing many common facts. By diving inside the descriptive data about these communities, one can see that the types of facts that compose both communities are also very different, reflecting different types of collaborations between actors depending on the type of fact they are involved in. The communities also show a strong homogeneity in the type of facts they contain, in the nationality of origin of its actors, and on the communes in which the facts happen.

In conclusion, it is worth to highlight that this work has allowed to

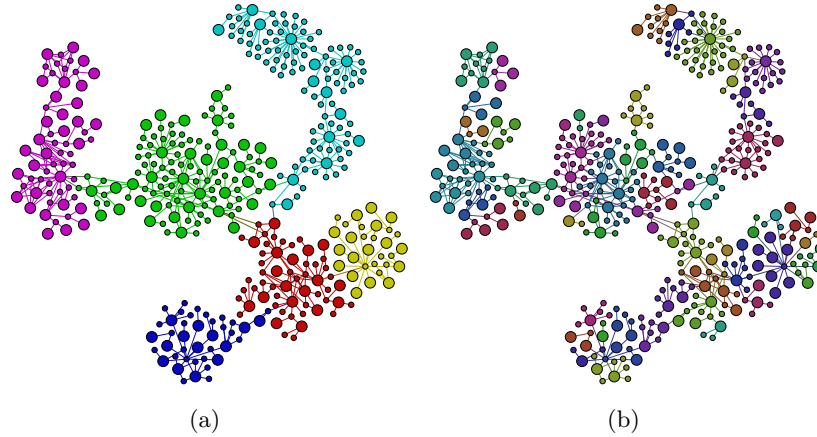


Figure 6.21: Partition in communities of the second largest connected component, for (a) $\gamma = 10^{-3}$ and (b) $\gamma = 10^{-1}$. Large nodes represent actors and small nodes represent facts. The network is composed of ~ 100 actors and ~ 500 facts.

show for the first time to draw a global picture of the criminal landscape in one complete country (which is materialized in Figure 6.18). The usage of this method by the Federal Police allows to handle large datasets in a very short time, and to integrate a large amount of descriptive data to understand the structure of organized crime, in a fashion that was not accessible so far.

6.5 Conclusions

The interest for community detection has boomed in the recent years, boosted by the availability of fast community detection algorithms that produce meaningful results on very large networks in a reasonable time. Although, no consensus exists on which algorithm performs best, in particular because there is no consensus on the clear definition of a community.

The recent finding that several quality function based methods suffer from a resolution limit, which introduces a non-trivial lower bound on the size of the optimal communities, has stimulated research in the de-

velopment of scaled versions of these quality functions.

The addition of a scaling parameter effectively allows to lower the resolution limit of the quality functions, but it introduces instead a non-trivial higher bound on the size of the optimal partitions. Thus, the addition of the scale parameter creates an interval of acceptable community sizes bounded below by the (lowered) resolution limit and above by this new bound. Each community falling out of that interval is then automatically split into smaller pieces or merged to another community.

In this chapter, we have focused our interest on the computational difficulty of detecting communities at multiple scales. We have proposed an algorithm that performs a binary search on the space of the possible scale parameters in order to detect intervals in which a given partition is optimal. The detection of such intervals allows to disregard the scale parameters inside these intervals, which represent an unnecessary computational load. We show that the method performs well at detecting stable partitions, in particular when the network is well-structured in communities at several scales. The results provided by the algorithm allow to quantify how meaningful the results are, but the interpretation is lacking a strict definition of which partition is the most meaningful. Instead, it provides a landscape of the possible partitions, which allows the user to decide on his own which partitions is most appropriate.

This method has been applied in a joint project with the Federal Belgian Police, in an analysis of a network of criminal offenders. We took this opportunity to present this particular dataset and its characteristics. The application of the multi-level detection method on this dataset has allowed to draw for the first time a large-scale map of organized crime in one whole country, integrating at once five years of data with no restrictions on the type of crime under study, as it is often the case in such studies. The obtained communities appear to be significant in the sense that they contain facts of similar kind and happening in the same neighbourhoods. A closer analysis of some communities returned by the method show indeed that the obtained structures show extremely relevant information about the internal structure and collaborations of the actors.

Further work The problem of community detection in complex networks is far from being solved. Out of the numerous possible directions of future research, we point out here one topic that has been, to our knowledge, not properly considered. The detection of communities in bipartite networks has never been addressed as a problem on its own. Most community detection methods designed for bipartite networks are adaptations of methods from unipartite networks (as is the method developed in this chapter). However, the meaning of communities may differ strongly in bipartite networks compared to their unipartite counterpart. The development of community detection methods for bipartite networks is an interesting future direction of research.

Chapter 7

General Conclusions

We conclude this thesis with a short summary of the contributions contained in each chapter of this thesis. We show current and potential applications of the results of this thesis, and conclude with some ideas for further research directions.

This thesis

The study of networks has started 350 years ago with Euler’s solution to the problem of Königsberg, which was the first known result in *graph theory*. However, it is only recently that the study of large networks has attracted the attention of a large scientific community, given the wide availability of large-scale databases of real networks. The insights provided by recent findings suggest that the structure of complex networks presents some non-trivial characteristics, and that a better understanding of these structural properties of networks has a large potential impact in many applications. In this thesis, we proposed several contributions in the analysis of geographic networks, dynamic networks and community detection.

In Chapter 3, we give a survey of the contributions made so far in the analysis of mobile phone datasets. Mobile phone datasets are excellent proxys for social networks, and have been studied in the recent years for their particular structure, for the geographical information that they contain and for the dynamics of human contacts. This survey is the first attempt to structure the contributions in this emerging field and to provide a global overview of the topics related to these datasets.

The addition of geographic information to a network provides valuable

information for the prediction and the analysis of networks. In Chapter 4, we have shown that the communication network of cities can well be predicted by a gravity model, where the size of cities play the role of masses. The knowledge of this average behaviour may reveal itself as a useful tool to detect particular features in the network. For instance, the gravity model proposed in Chapter 4 has been used by other authors as null model for community detection in geographic networks [69].

In Chapter 5, we contribute to the study of dynamical networks. The originality of our approach has been to challenge a simple principle in the network construction that has not been addressed in the previous contributions in this direction, namely that the length of time windows used in dynamic network analysis might significantly influence the results of such analyses. From the study of a mobile phone network, we observe that depending on the length of the time window, networks might have a different equilibrium between the local and global density, different forms of statistical distributions, and that the structure of networks might also be influenced by the position of the time window. Moreover, we obtain a way of quantifying the growth of the network size by removing the information about time.

Finally, Chapter 6 contains an improvement of actual community detection methods and a detailed analysis of a network of criminal offenders. The most-used methods that detect communities at multiple levels do not take into account the communities that have already been detected at different levels to accelerate the search. However, a large number of unnecessary detections can be avoided by performing the search efficiently, which is done in the method we propose. We provide an in-depth analysis of the structure of the network of criminal offenders, which has been studied in the context of a research contract with the Belgian Federal Police, and on which the community detection method is applied.

Applications

One tool developed in this thesis is a product of a research contract, and is used in as an analysis tool by the Federal Police. Besides this, other results from this thesis may also be applied in practical situations.

The gravity model presented in Chapter 4 is now in use as null model for geographic networks. A deviation from the model of link weights observed in real data indicate that external factors not taken into account by the gravity model may influence the strength of ties between

cities. In [69], the authors show that by using a modification of the modularity with the gravity law as null model, they manage to uncover the well-known linguistic border in Belgium, similarly to what we show in Section 4.5 with a different measure for link weight.

A detailed study of the dynamics of mobile phone calls may have a significant potential for marketing and dynamic pricing. In Chapter 5, we show that different types of calls are made at different times of the day. In particular early morning, late night, lunch time and school ends are typical times in Belgium when people call close relatives. This information can be leveraged by dynamic pricing strategies and marketing campaigns.

Finally, the community detection method described in Chapter 6 is an implemented tool used for longitudinal analysis of the evolution of criminal activities in Belgium, and is currently being applied in other practical applications too, such as the detection of well-connected coral reefs in the Great Barrier Reef in Australia, which could be served as information source for environment protection agencies.

Further directions

We end this thesis with some directions for future research that we estimate to be worth of interest.

Recalling results from five independent studies of mobile phone datasets [4, 135, 145, 169, 115], we observe that the results of these studies vary strongly. As an illustration, the degree distribution of the five studied networks have almost all different shapes. However, it is unclear if the observed differences emerge from cultural differences or from technical reasons, given that the datasets under study provide from different countries, that they are aggregated over different time periods, that they represent operators having different market shares, and finally that the networks are constructed using different construction schemes. A detailed study of these influences would provide a general framework for the analysis of mobile phone networks, which lacks nowadays.

In Chapter 5, we present initial results on the analysis of the influence of time scales on the structure of networks, however more complex analyses can be done in this direction. In a similar fashion as what was done in this thesis, one can detect communities over windows of growing length, which could provide useful information on how communities are created. This information could then in turn be used to develop efficient commu-

nity detection methods on growing networks.

Finally, the study of dynamical networks opens up the idea of studying the stability of nodes and links over time. This question was one of the initial questions that motivated this thesis, but that has not been answered since other questions popped out under way. In particular, a dynamic link is represented in a static network by its average frequency, although other measures may exist to take into account the burstiness of its activity. A first attempt can be found in [133], however there is still plenty of room for improvement in this direction. A clear definition of stable nodes or links in the network may then open new possibilities for definitions of opinion leaders in dynamical networks. Static definitions of leaders, with applications in mobile phone networks can be found in Cristobald de Kerchove's thesis [55], the extension to dynamical networks should take into account the stability of a node and its connections.

Bibliography

- [1] J. Abello, P.M. Pardalos, and M.G.C. Resende. On maximum clique problems in very large graphs. *External memory algorithms*, 50:119–130, 1999.
- [2] N. Aharony, W. Pan, C. Ip, and A. Pentland. Tracing mobile phone app installations in the” friends and family” study. In *Proceedings of the 2010 Workshop on Information in Networks (WIN’10)*, 2010.
- [3] Y.Y. Ahn, J.P. Bagrow, and S. Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466(7307):761–764, 2010.
- [4] W. Aiello, F. Chung, and L. Lu. A random graph model for massive graphs. In *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pages 171–180. Acm, 2000.
- [5] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97, Jan 2002.
- [6] A Arenas, A Fernández, and S Gómez. Analysis of the structure of complex networks at different resolution levels. *New Journal of Physics*, 10(5):053039, 2008.
- [7] T. Aynaud and J.L. Guillaume. Static community detection algorithms for evolving networks. In *Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt), 2010 Proceedings of the 8th International Symposium on*, pages 513–519. IEEE, 2010.

- [8] F. Baccelli and J. Bolot. Modeling the economic value of location and preference data of mobile users. *Proc. IEEE Infocom 2011*, 2011.
- [9] L. Backstrom, C. Dwork, and J. Kleinberg. Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography. In *Proceedings of the 16th international conference on World Wide Web*, pages 181–190. ACM, 2007.
- [10] J.-P. Bagrow, D. Wang, and A.-L. Barabási. Collective response of human populations to large-scale emergencies. *PloS one*, 6(3):e17680, 2011.
- [11] D. Balcan, V. Colizza, B. Goncalves, H. Hu, J.J. Ramasco, and A. Vespignani. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences*, 106(51):21484–21489, 2009.
- [12] A.-L. Barabási. The origin of bursts and heavy tails in human activity. *Nature*, 435:207, 2005.
- [13] A.-L. Barabási. Scale-free networks: a decade and beyond. *Science*, 325(5939):412, 2009.
- [14] A.-L. Barabási. You’re so predictable. *Physics World*, pages 22–26, 2010.
- [15] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439):509, 1999.
- [16] A.-L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311(3-4):590–614, 2002.
- [17] A. Barrat, M. Barthélemy, R. Pastor-Satorras, and A. Vespignani. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences*, 101(11):3747–3752, 2004.
- [18] A. Barrat, M. Barthélemy, and A. Vespignani. The effects of spatial constraints on the evolution of weighted complex networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(05):P05003, 2005.

- [19] M. Barthélemy. Spatial networks. *Physics Reports*, 499(1-3):1–101, 2011.
- [20] M. Bauer and D. Bernard. Maximal entropy random networks with given degree distribution. *Arxiv preprint cond-mat/0206150*, 2002.
- [21] Belgian national institute of statistics. http://www.statbel.fgov.be/home/_en.asp.
- [22] L. Benamara and C. Magnien. Removing bias due to finite measurement of dynamic systems: case study on p2p systems. *Arxiv preprint arXiv:1104.3694*, 2011.
- [23] L. Bettencourt, J. Lobo, D. Helbing, C. Kuhnert, and G.B. West. Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National Academy of Sciences*, 104(17):7301, 2007.
- [24] K. Bhattacharya, G. Mukherjee, J. Saramäki, K. Kaski, and SS Manna. The International Trade Network: weighted network analysis and modelling. *J. Stat. Mech*, page P02002, 2008.
- [25] G. Bianconi. The entropy of randomized network ensembles. *EPL (Europhysics Letters)*, 81:28005, 2008.
- [26] G. Bianconi. Entropy of network ensembles. *Physical Review E*, 79(3):036114, 2009.
- [27] G. Bianconi and A.-L. Barabási. Competition and multiscaling in evolving networks. *EPL (Europhysics Letters)*, 54:436, 2001.
- [28] G. Bianconi, A.C.C. Coolen, and C.J.P. Vicente. Entropies of complex networks with hierarchically constrained topologies. *Physical Review E*, 78(1):016114, 2008.
- [29] V.D. Blondel, C. de Kerchove, E. Huens, and P. Van Dooren. Social leaders in graphs. *Lecture notes in control and information sciences*, 341:231, 2006.
- [30] V.D. Blondel, P. Deville, F. Morlot, Z. Smoreda, P. Van Dooren, and C. Ziemlicki. Voice on the border: do cellphones redraw the maps? *Paris Tech Review*, 2011.

- [31] V.D. Blondel, J.L. Guillaume, R. Lambiotte, and E.L.J.S. Mech. Fast unfolding of communities in large networks. *J. Stat. Mech*, page P10008, 2008.
- [32] V.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.
- [33] J.E. Blumenstock, D. Gillick, and N. Eagle. Who’s calling? demographics of mobile phone use in rwanda. *Transportation*, 32:2–5, 2010.
- [34] L. Bogacz, Z. Burda, and B. Waław. Homogeneous complex networks. *Physica A: Statistical Mechanics and its Applications*, 366:587–607, 2006.
- [35] S.P. Borgatti and M.G. Everett. Network analysis of 2-mode data. *Social networks*, 19(3):243–269, 1997.
- [36] U. Brandes, D. Dellling, M. Gaertler, R. Görke, M. Hoefer, Z. Nikoloski, and D. Wagner. On modularity-np-completeness and beyond. *ITI Wagner, Faculty of Informatics, Universität Karlsruhe (TH), Tech. Rep*, 19:2006, 2006.
- [37] D. Brockmann and F. Theis. Money circulation, trackable items, and the emergence of universal human mobility patterns. *IEEE Pervasive Computing*, pages 28–35, 2008.
- [38] F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, and C. Ratti. Real-time urban monitoring using cell phones: A case study in rome. *Intelligent Transportation Systems, IEEE Transactions on*, 12(1):141–151, 2011.
- [39] F. Calabrese, G. Di Lorenzo, and C. Ratti. Human mobility prediction based on individual and collective geographical preferences. In *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*, pages 312–317. IEEE, 2010.
- [40] F. Calabrese, F. Pereira, G. Di Lorenzo, L. Liu, and C. Ratti. The geography of taste: analyzing cell-phone mobility and social events. *Pervasive Computing*, pages 22–37, 2010.

- [41] F. Calabrese, Z. Smoreda, V.D. Blondel, and C. Ratti. Interplay between telecommunications and face-to-face interactions: A study using mobile phone data. *PloS one*, 6(7):e20814, 2011.
- [42] R. Cancho and R. Solé. Optimization in complex networks. In Romualdo Pastor-Satorras, Miguel Rubi, and Albert Diaz-Guilera, editors, *Statistical Mechanics of Complex Networks*, volume 625 of *Lecture Notes in Physics*, pages 114–126. Springer Berlin / Heidelberg, 2003.
- [43] J. Candia, M.C. González, P. Wang, T. Schoenharl, G. Madey, and A.-L. Barabási. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, 41:224015, 2008.
- [44] M. Cebrian, M. Lahiri, N. Oliver, and A. Pentland. Measuring the collective potential of populations from dynamic social interaction data. *Selected Topics in Signal Processing, IEEE Journal of*, 4(4):677–686, 2010.
- [45] M. Cebrian, A. Pentland, and S. Kirkpatrick. Disentangling social networks inferred from call logs. *Arxiv preprint arXiv:1008.1357*, 2010.
- [46] A. Clauset, M.E.J. Newman, and C. Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004.
- [47] V. Colizza, A. Barrat, M. Barthélemy, and A. Vespignani. The role of the airline transportation network in the prediction and predictability of global epidemics. *Proceedings of the National Academy of Sciences of the United States of America*, 103(7):2015, 2006.
- [48] C. Cortes, D. Pregibon, and C. Volinsky. Communities of interest. *Advances in Intelligent Data Analysis*, pages 105–114, 2001.
- [49] L.F. Costa, F.A. Rodrigues, G. Travieso, and P.R.V. Boas. Characterization of complex networks: A survey of measurements. *Advances in Physics*, 56(1):167–242, 2007.

- [50] T.M. Cover, J.A. Thomas, J. Wiley, et al. *Elements of information theory*, volume 6. Wiley Online Library, 1991.
- [51] K. Dasgupta, R. Singh, B. Viswanathan, D. Chakraborty, S. Mukherjee, A.A. Nanavati, and A. Joshi. Social ties and their relevance to churn in mobile telecom networks. In *Proceedings of the 11th international conference on Extending database technology: Advances in database technology*, pages 668–677. ACM, 2008.
- [52] J. Davidsen, H. Ebel, and S. Bornholdt. Emergence of a small world from local interactions: Modeling acquaintance networks. *Physical Review Letters*, 88(12):128701, 2002.
- [53] W.K. Davies. Urban connectivity in montana. *The Annals of Regional Science*, 13(2):29–46, 1979.
- [54] C. de Kerchove, G. Krings, R. Lambiotte, P. Van Dooren, and V.D. Blondel. Role of second trials in cascades of information over networks. *Physical Review E*, 79(1):016114, 2009.
- [55] C. de Kerchove d’Exaerde. Ranking large networks: Leadership, optimization and distrust (phd thesis). 2009.
- [56] J.-C. Delvenne, S.N. Yaliraki, and M. Barahona. Stability of graph communities across time scales. *Proceedings of the National Academy of Sciences*, 107(29):12755–12760, 2010.
- [57] L. Demetrius and T. Manke. Robustness and network evolution - an entropic principle. *Physica A: Statistical Mechanics and its Applications*, 346(3):682–696, 2005.
- [58] T. Dierkes, M. Bichler, and R. Krishnan. Estimating the effect of word of mouth on churn and cross-buying in the mobile phone market with markov logic networks. *Decision Support Systems*, 2011.
- [59] S.N. Dorogovtsev, A.V. Goltsev, and J.F.F. Mendes. Critical phenomena in complex networks. *Rev. Mod. Phys.*, 80:1275–1335, Oct 2008.
- [60] S.N. Dorogovtsev and J.F.F. Mendes. Effect of the accelerating growth of communications networks on their structure. *Physical review E*, 63(2):025101, 2001.

- [61] N. Du, C. Faloutsos, B. Wang, and L. Akoglu. Large human communication networks: patterns and a utility-driven generator. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 269–278. ACM, 2009.
- [62] N. Eagle. Engineering a common good: Fair use of aggregated, anonymized behavioral data. 2009.
- [63] N. Eagle, Y.-A. de Montjoye, and L.M.A. Bettencourt. Community computing: Comparisons between rural and urban societies using mobile phone data. In *2009 International Conference on Computational Science and Engineering*, pages 144–150. IEEE, 2009.
- [64] N. Eagle, M. Macy, and R. Claxton. Network diversity and economic development. *Science*, 328(5981):1029, 2010.
- [65] N. Eagle and A.S. Pentland. Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63(7):1057–1066, 2009.
- [66] N. Eagle, A.S. Pentland, and D. Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36):15274, 2009.
- [67] J. Eeckhout. Gibrat’s Law for (All) Cities. *American Economic Review*, 94(5):1429–1451, 2004.
- [68] P. Erdős and A. Rényi. *On the evolution of random graphs*. Akad. Kiadó, 1960.
- [69] P. Expert, T.S. Evans, V.D. Blondel, and R. Lambiotte. Uncovering space-independent communities in spatial networks. *Proceedings of the National Academy of Sciences*, 108(19):7663, 2011.
- [70] E.D. Fitkov-Norris and A. Khanifar. Dynamic pricing in cellular networks, a mobility model with a provider-oriented approach. In *3G Mobile Communication Technologies, 2001. Second International Conference on (Conf. Publ. No. 477)*, pages 63–67. IET, 2001.

- [71] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75 – 174, 2010.
- [72] S. Fortunato and M. Barthélemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36, 2007.
- [73] M.T. Gastner and M. E.J. Newman. The spatial structure of networks. *The European Physical Journal B - Condensed Matter and Complex Systems*, 49:247–252, 2006. 10.1140/epjb/e2006-00046-8.
- [74] A. Gautreau, A. Barrat, and M. Barthélemy. Microdynamics in stationary complex networks. *Proceedings of the National Academy of Sciences*, 106(22):8847, 2009.
- [75] G. Ghoshal and A.-L. Barabási. Ranking stability and super-stable nodes in complex networks. *Nature Communications*, 2:394, 2011.
- [76] M. Girvan and M.E.J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.
- [77] J. Gómez-Gardeñes and V. Latora. Entropy rate of diffusion processes on complex networks. *Physical Review E*, 78(6):065102, 2008.
- [78] M.C. González, C.A. Hidalgo, and A.-L. Barabási. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.
- [79] D. Grady, R. Brune, C. Thiemann, F. Theis, and D. Brockmann. Modularity maximization and tree clustering: Novel ways to determine effective geographic borders. *Arxiv preprint arXiv:1104.1200*, 2011.
- [80] M.S. Granovetter. The Strength of Weak Ties. *American Journal of Sociology*, 78:1360–1380, 1973.
- [81] H.S. Heaps. *Information retrieval: Computational and theoretical aspects*. Academic Press, Inc., 1978.

- [82] C.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.
- [83] S. Hill, F. Provost, and C. Volinsky. Network-based marketing: Identifying likely adopters via consumer networks. *Statistical Science*, 21(2):256–276, 2006.
- [84] P. Holme, C.R. Edling, and F. Liljeros. Structure and time evolution of an internet dating community. *Social Networks*, 26(2):155–174, 2004.
- [85] P. Holme and J. Saramäki. Temporal networks. *arXiv: 1108.1780*, 2011.
- [86] J. Hopcroft, O. Khan, B. Kulis, and B. Selman. Tracking evolving communities in large linked networks. *Proceedings of the national academy of sciences of the United States of America*, 101(Suppl 1):5249, 2004.
- [87] Y. Hu, Y. Wang, D. Li, S. Havlin, and Z. Di. Possible origin of efficient navigation in small worlds. *Phys. Rev. Lett.*, 106:108701, Mar 2011.
- [88] S.Y. Hung, D.C. Yen, and H.Y. Wang. Applying data mining to telecom churn management. *Expert Systems with Applications*, 31(3):515–524, 2006.
- [89] International telecommunication union. <http://www.itu.int/>.
- [90] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky. Identifying important places in people’s lives from cellular network data. *Pervasive Computing*, pages 133–151, 2011.
- [91] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, J. Rowland, and A. Varshavsky. A tale of two cities. In *Proceedings of the Eleventh Workshop on Mobile Computing Systems & Applications*, pages 19–24. ACM, 2010.
- [92] H. Jeong, Z. Neda, and A.-L. Barabási. Measuring preferential attachment in evolving networks. *EPL (Europhysics Letters)*, 61:567, 2003.

- [93] H.-H. Jo, M. Karsai, J. Kertész, and K. Kaski. Circadian pattern and burstiness in human communication activity. *Arxiv preprint arXiv:1101.0377*, 2011.
- [94] W.S. Jung, F. Wang, and H.E. Stanley. Gravity model in the Korean highway. *EPL-Europhysics Letters*, 81(4):48005–48005, 2008.
- [95] B. Karrer, E. Levina, and M.E.J. Newman. Robustness of community structure in networks. *Physical Review E*, 77(4):046119, 2008.
- [96] M. Karsai, M. Kivelä, R.K. Pan, K. Kaski, J. Kertész, A.-L. Barabási, and J. Saramäki. Small but slow world: How network topology and burstiness slow down spreading. *Physical Review E*, 83(2):025102, 2011.
- [97] E. Katz and P.F. Lazarsfeld. *Personal influence: the part played by people in the flow of mass communications*. Free Press, 1955.
- [98] D. Kempe, J. Kleinberg, and A. Kumar. Connectivity and inference problems for temporal networks. In *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pages 504–513. ACM, 2000.
- [99] B.W. Kernighan and S. Lin. An efficient heuristic procedure for partitioning graphs. *Bell System Technical Journal*, 49(2):291–307, 1970.
- [100] K. Kianmehr and R. Alhajj. Calling communities analysis and identification using machine learning techniques. *Expert Systems with Applications*, 36(3):6218–6226, 2009.
- [101] Y. Kim, R. Telang, W.B. Vogt, and R. Krishnan. An empirical analysis of mobile voice service and sms: A structural model. *Management Science*, 56(2):234–252, 2010.
- [102] M. Kleiber. Body size and metabolic rate. *Physiological Reviews*, 27(4):511, 1947.
- [103] J. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. The web as a graph: Measurements, models, and methods. *Computing and Combinatorics*, pages 1–17, 1999.

- [104] I. Kontoyiannis, P.H. Algoet, Y.M. Suhov, and A.J. Wyner. Non-parametric entropy estimation for stationary processes and random fields, with applications to english text. *Information Theory, IEEE Transactions on*, 44(3):1319–1327, 1998.
- [105] G. Kossinets and D.J. Watts. Empirical analysis of an evolving social network. *Science*, 311(5757):88, 2006.
- [106] L. Kovanen, J. Saramäki, and K. Kaski. Reciprocity of mobile phone calls. *JDySES*, 2(2):138–151, 2011.
- [107] G. Krings, F. Calabrese, C. Ratti, and V.D. Blondel. Scaling behaviors in the communication network between cities. In *2009 International Conference on Computational Science and Engineering*, pages 936–939. IEEE, 2009.
- [108] G. Krings, F. Calabrese, C. Ratti, and V.D. Blondel. Urban gravity: a model for inter-city telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment*, 2009:L07003, 2009.
- [109] G. Krings, M. Karsai, S. Bernhardsson, V.D. Blondel, and J. Saramäki. Effects of time window size and placement on the structure of aggregated networks. *Arxiv preprint arXiv:1202.1145*, 2012.
- [110] J.M. Kumpula, J.-P. Onnela, J. Saramäki, K. Kaski, and J. Kertész. Emergence of communities in weighted networks. *Physical review letters*, 99(22):228701, 2007.
- [111] J.M. Kumpula, J. Saramäki, K. Kaski, and J. Kertész. Limited resolution in complex network community detection with potts model approach. *The European Physical Journal B - Condensed Matter and Complex Systems*, 56:41–45, 2007.
- [112] A. Kuusik, R. Ahas, and M. Tiru. Analysing repeat visitation on country level with passive mobile positioning method: an estonian case study. In *XVII scientific conference on economic policy*, pages 1–3, 2009.
- [113] R. Kwok. Personal technology: Phoning in data. *Nature*, 458(7241):959, 2009.

- [114] R. Lambiotte. Multi-scale modularity in complex networks. In *Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt), 2010 Proceedings of the 8th International Symposium on*, pages 546–553. IEEE, 2010.
- [115] R. Lambiotte, V.D. Blondel, C. de Kerchove, E. Huens, C. Prieur, Z. Smoreda, and P. Van Dooren. Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications*, 387(21):5317–5325, 2008.
- [116] R. Lambiotte, J.-C. Delvenne, and M. Barahona. Laplacian dynamics and multiscale modular structure in networks. *arXiv*, 812, 2009.
- [117] A. Lancichinetti, S. Fortunato, and F. Radicchi. Benchmark graphs for testing community detection algorithms. *Phys. Rev. E*, 78(4):046110, Oct 2008.
- [118] A. Lancichinetti, F. Radicchi, and J.J. Ramasco. Statistical significance of communities in networks. *Physical Review E*, 81(4):046110, 2010.
- [119] M. Latapy, C. Magnien, and N.D. Vecchio. Basic notions for the analysis of large two-mode networks. *Social Networks*, 30(1):31–48, 2008.
- [120] D. Lazer, A. Pentland, L. Adamic, S.n Aral, A.-L. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. Van Alstyne. Computational social science. *Science*, 323(5915):721–723, 2009.
- [121] S. Le Blond, J.-L. Guillaume, and M. Latapy. Clustering in p2p exchanges and consequences on performances. *Peer-to-Peer Systems IV*, pages 193–204, 2005.
- [122] S. Lehmann, M. Schwartz, and L.K. Hansen. Biclique communities. *Phys. Rev. E*, 78(1):016108, Jul 2008.
- [123] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):2, 2007.

- [124] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic routing in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33):11623, 2005.
- [125] Y.Y. Liu, J.J. Slotine, and A.-L. Barabási. Controllability of complex networks. *Nature*, 473(7346):167–173, 2011.
- [126] L. Lü, Z.K. Zhang, and T. Zhou. Zipf’s law leads to heaps’ law: analyzing their relation in finite-size systems. *PloS one*, 5(12):e14139, 2010.
- [127] A. Madan, B.N. Waber, M. Ding, P. Kominers, and A.S. Pentland. Reality mining and personal privacy. 2009.
- [128] F. Manfredini, P. Tagliolato, and C. Di Rosa. Monitoring temporary populations through cellular core network data. *Computational Science and Its Applications-ICCSA 2011*, pages 151–161, 2011.
- [129] M. Marsili, F. Vega-Redondo, and F. Slanina. The rise and fall of a networked society: A formal model. *Proceedings of the National Academy of Sciences of the United States of America*, 101(6):1439, 2004.
- [130] J.L. Martin and K.T. Yeung. Persistence of close personal ties over a 12-year period. *Social networks*, 28(4):331–362, 2006.
- [131] M. Martino, F. Calabrese, G. Di Lorenzo, C. Andris, L. Liang, and C. Ratti. Ocean of information: fusing aggregate & individual dynamics for metropolitan analysis. In *Proceedings of the 15th international conference on Intelligent user interfaces*, pages 357–360. ACM, 2010.
- [132] M. Meila. Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, 98(5):873 – 895, 2007.
- [133] G. Miritello, E. Moro, and R. Lara. Dynamical strength of social ties in information spreading. *Physical Review E*, 83(4):045102, 2011.

- [134] P.J. Mucha, T. Richardson, K. Macon, M.A. Porter, and J.-P. Onnela. Community Structure in Time-Dependent, Multiscale, and Multiplex Networks. *Science*, 328(5980):876–878, 2010.
- [135] A.A. Nanavati, S. Gurumurthy, G. Das, D. Chakraborty, K. Dasgupta, S. Mukherjea, and A. Joshi. On the structural properties of massive telecom call graphs: findings and implications. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 435–444. ACM, 2006.
- [136] A. Narayanan and V. Shmatikov. De-anonymizing social networks. In *2009 30th IEEE Symposium on Security and Privacy*, pages 173–187. IEEE, 2009.
- [137] M.E.J. Newman. Spread of epidemic disease on networks. *Phys. Rev. E*, 66:016128, Jul 2002.
- [138] M.E.J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):pp. 167–256, 2003.
- [139] M.E.J. Newman. Fast algorithm for detecting community structure in networks. *Phys. Rev. E*, 69:066133, Jun 2004.
- [140] M.E.J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.
- [141] M.E.J. Newman. *Networks: An Introduction*. Oxford University Press, 2011.
- [142] M.E.J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.
- [143] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil, and C. Mascolo. A tale of many cities: universal patterns in human urban mobility. *Arxiv preprint arXiv:1108.5355*, 2011.
- [144] J.-P. Onnela, S. Arbesman, M.C. González, A.-L. Barabási, and N.A. Christakis. Geographic constraints on social network groups. *PloS one*, 6(4):e16939, 2011.

- [145] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, M.A. de Menezes, K. Kaski, A.-L. Barabási, and J. Kertész. Analysis of a large-scale weighted network of one-to-one human communication. *New Journal of Physics*, 9(6):179, 2007.
- [146] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18):7332, 2007.
- [147] T. Opsahl. Triadic closure in two-mode networks: Redefining the global and local clustering coefficients. *Social Networks*, 2011.
- [148] G. Palla, A.-L. Barabási, and T. Vicsek. Quantifying social group evolution. *Nature*, 446(7136):664, 2007.
- [149] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.
- [150] J. Park, D.S. Lee, and M.C. González. The eigenmode analysis of human motion. *Journal of Statistical Mechanics: Theory and Experiment*, 2010:P11021, 2010.
- [151] J. Park and M.E.J. Newman. Statistical mechanics of networks. *Physical Review E*, 70(6):066117, 2004.
- [152] A. Pentland. Reality mining of mobile communications: Toward a new deal on data. *The Global Information Technology Report 2008–2009*, page 1981, 2009.
- [153] M.A. Porter, J.-P. Onnela, and P.J. Mucha. Communities in networks. *Notices of the American Mathematical Society*, 56(9):1082–1097, 2009.
- [154] D.S. Price. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27(5):292–306, 1976.
- [155] D. Quercia, N. Lathia, F. Calabrese, G. Di Lorenzo, and J. Crowcroft. Recommending social events from mobile phone

- location data. In *2010 IEEE International Conference on Data Mining*, pages 971–976. IEEE, 2010.
- [156] 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.
 - [157] C. Ratti, S. Williams, D. Frenchman, and R.M. Pulselli. Mobile landscapes: using location data from cell phones for urban analysis. *Environment and planning B, Planning and design*, 33(5):727, 2006.
 - [158] T. Reader, O. Lizardo, D. Hachen, and N.V. Chawla. Predictors of short-term decay of cell phone contacts in a large scale communication network. *Arxiv preprint arXiv:1102.1753*, 2011.
 - [159] J. Reades, F. Calabrese, and C. Ratti. Eigenplaces: analysing cities using the space- time structure of the mobile phone network. *Environment and Planning B: Planning and Design*, 36(5):824–836, 2009.
 - [160] J. Reades, F. Calabrese, A. Sevtsuk, and C. Ratti. Cellular census: Explorations in urban data collection. *IEEE Pervasive Computing*, pages 30–38, 2007.
 - [161] J. Reichardt and S. Bornholdt. Statistical mechanics of community detection. *Phys. Rev. E*, 74(1):016110, Jul 2006.
 - [162] J. Reichardt and S. Bornholdt. Partitioning and modularity of graphs with arbitrary degree distribution. *Phys. Rev. E*, 76:015102, Jul 2007.
 - [163] A. Rényi. On measures of entropy and information. In *Fourth Berkeley Symposium on Mathematical Statistics and Probability*, pages 547–561, 1961.
 - [164] Y. Richter, E. Yom-Tov, and N. Slonim. Predicting customer churn in mobile networks through analysis of social groups. In *Proceedings of the 2010 SIAM International Conference on Data Mining (SDM 2010)*, 2010.

- [165] P. Ronhovde and Z. Nussinov. Local resolution-limit-free potts model for community detection. *Phys. Rev. E*, 81:046114, Apr 2010.
- [166] M. Rosvall and C.T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118, 2008.
- [167] M. Rosvall and K. Sneppen. Networks and our limited information horizon. *Arxiv preprint cond-mat/0604036*, 2006.
- [168] M. Rosvall, A. Trusina, P. Minnhagen, and K. Sneppen. Networks and cities: An information perspective. *Physical review letters*, 94(2):28701, 2005.
- [169] M. Seshadri, S. Machiraju, A. Sridharan, J. Bolot, C. Faloutsos, and J. Leskovec. Mobile call graphs: beyond power-law and log-normal distributions. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 596–604. ACM, 2008.
- [170] C.E. Shannon and W. Weaver. The mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 1948.
- [171] G. Simonyi. Graph entropy: A survey. *Combinatorial Optimization*, 20:399–441, 1995.
- [172] D. Smith, J.-P. Onnela, and N.F. Johnson. Accelerating networks. *New Journal of Physics*, 9:181, 2007.
- [173] K. Sneppen, A. Trusina, and M. Rosvall. Hide-and-seek on complex networks. *EPL (Europhysics Letters)*, 69:853, 2005.
- [174] R. Solé and S. Valverde. Information theory of complex networks: On evolution and architectural constraints. *Complex networks*, pages 189–207, 2004.
- [175] C. Song, T. Koren, P. Wang, and A.-L. Barabási. Modelling the scaling properties of human mobility. *Nature Physics*, 2010.
- [176] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018, 2010.

- [177] D.A. Spielmat and S.H. Teng. Spectral partitioning works: Planar graphs and finite element meshes. In *Foundations of Computer Science, 1996. Proceedings., 37th Annual Symposium on*, pages 96–105. IEEE, 1996.
- [178] J. Steenbruggen, M.T. Borzacchiello, P. Nijkamp, and H. Scholten. Mobile phone data from gsm networks for traffic parameter and urban spatial pattern assessment: a review of applications and opportunities. *GeoJournal*, pages 1–21, 2011.
- [179] S.A. Stouffer. Intervening opportunities: a theory relating mobility and distance. *American sociological review*, 5(6):845–867, 1940.
- [180] G. Szabó and A.-L. Barabási. Network effects in service usage. *Arxiv preprint physics/0611177*, 2006.
- [181] G. Tibély, L. Kovanen, M. Karsai, K. Kaski, J. Kertész, and J. Saramäki. Communities and beyond: mesoscopic analysis of a large social network with complementary methods. *Physical Review E*, 83(5):056125, 2011.
- [182] J. Tinbergen. An Analysis of World Trade Flows. *Shaping the World Economy, Tinbergen, ed. New York: Twentieth Century Fund*, 1962.
- [183] V.A. Traag, P. Van Dooren, and Y. Nesterov. Narrow scope for resolution-limit-free community detection. *Physical Review E*, 84(1):016114, 2011.
- [184] D.C. van Leijenhorst and T.P. van der Weide. A formal derivation of heaps’ law. *Information Sciences*, 170(263-272), 2005.
- [185] A. Vázquez. Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations. *Physical Review E*, 67(5):056104, 2003.
- [186] B. Wang, H. Tang, C. Guo, and Z. Xiu. Entropy optimization of scale-free networks’ robustness to random failures. *Physica A: Statistical Mechanics and its Applications*, 363(2):591–596, 2006.
- [187] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabási. Human mobility, social ties, and link prediction. In *17th ACM*

SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'11), 2011.

- [188] P. Wang, M.C. González, C.A. Hidalgo, and A.-L. Barabási. Understanding the spreading patterns of mobile phone viruses. *Science*, 324(5930):1071, 2009.
- [189] P. Wang, M.C. González, R. Menezes, and A.-L. Barabási. New generation of mobile phone viruses and corresponding countermeasures. *Arxiv preprint arXiv:1012.3156*, 2010.
- [190] Y. Wang, B. Wu, and N. Du. Community evolution of social network: feature, algorithm and model. *Arxiv preprint arXiv:0804.4356*, 2008.
- [191] D.J. Watts and P.S. Dodds. Influentials, networks, and public opinion formation. *Journal of Consumer Research*, 34(4):441–458, 2007.
- [192] D.J. Watts and S.H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393(6684):440–442, 1998.
- [193] W.W. Zachary. An information flow model for conflict and fission in small groups. *Journal of anthropological research*, pages 452–473, 1977.
- [194] H. Zang and J. Bolot. Anonymization of location data does not work: a large-scale measurement study. *submitted to ACM Mobicom*, 11, 2011.
- [195] H. Zhang and R. Dantu. Discovery of social groups using call detail records. In *On the Move to Meaningful Internet Systems: OTM 2008 Workshops*, pages 489–498. Springer, 2008.
- [196] G.K. Zipf. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley Press, 1949.

Appendix A

Identification of optimal parameters for DPU

The identification problem is the following: given a set of data points $\{(x_i, m_i)\}$ which are observations of x and $m(x)$ for a part of the growth curve, can we identify the parameters x_0, m_0 and K that generate this curve?

A.1 Identification based on $m(x)$

The expression of the size of the network, once the steady growth phase is reached, is given by

$$m(x) = m_0 \left(\frac{x + x_0}{x_0} \right)^K - m_{err}, \quad (\text{A.1})$$

where m_{err} is the number of links of the hidden network that have not been discovered. Those links are e.g. inactive links, and are estimated as constant. The identification problem consists thus to determine the value of the unknown parameters K, x_0, m_0 and m_{err} , in order to fit as best as possible the observed growth of the network.

The optimization problem reads then as follows:

$$\min_{\hat{w}, m_0, K, m_{err}} \sum_i R_i^2 \quad (\text{A.2})$$

s.t.

$$R_i = m_i - m_0 \left(\frac{x_i + \hat{w}m_0}{\hat{w}m_0} \right)^K - m_{err} \quad \forall i \quad (\text{A.3})$$

$$\hat{w} \in [1, \infty) \quad (\text{A.4})$$

$$m_0 \in [1, \infty) \quad (\text{A.5})$$

$$K \in [0, 1] \quad (\text{A.6})$$

$$m_{err} \in [0, \infty) \quad (\text{A.7})$$

Where the constraint $m_0 \leq x_0$ has been handled by writing $x_0 = \hat{w}m_0, \hat{w} \geq 1$. This problem is solved using the Nelder-Mead simplex method (NM). The initial iterate of the method is estimated as follows: since r has stabilized, it should be stabilized around $K \frac{m(x)}{x}$, hence the estimate $\tilde{K}$ of K is given by $\left\langle r_i \frac{x_i}{m_i} \right\rangle_i$, where r_i are the observations of $r(x)$ over the same data points as (x_i, m_i) . The estimate for m_0 is obtained by fitting a power-law with exponent $\tilde{K}$ to the data, m_0 being the intercept of this power-law and $\hat{w}$ and m_{err} are set to 1 and 0 respectively. The optimal parameters are then calculated by applying iteratively NM and by feeding the next NM run by the optimal solution of the previous run, until the difference between two consecutive optimal costs are smaller than a predefined tolerance.

To analyse the stability of the detection procedure, we used a sliding window and detected the optimal parameters for different positions of the window on the data generated by the DPU model using two different sets of parameters and on the real data.

The first sets of parameters of DPU is the same as presented in Figure 5.16, the parameters are $\hat{w} = 1$, $m_0 = 10^3$ and $K = 0.2$. The result has been averaged over 100 realizations to remove partly the noise, the final results are showed on Figure A.1.

We detect the optimal solution of the optimization problem for windows of varying lengths. In total, the data contains 100,000 events, the sliding windows used for the detection have respectively a width of

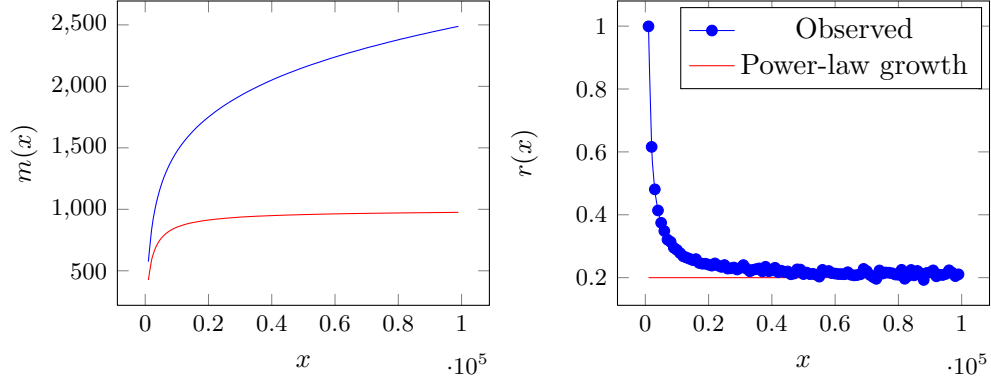


Figure A.1: Average result of 10^2 realizations of DPU with parameters $m_0^h = x_0^h = 10^3$ and $K = 0.2$. (a) size of the network as a function of x (blue). The first sharp increase results from the discovery of both new links and links from the hidden network. The red curve indicates the number of links from the hidden network that have been observed. Once that value has stabilized, the remaining growth is a power-law of exponent 0.2. (b) $r(x)$ as a function of the number of events. A first sharp decrease is followed by a stabilized value (blue), close to the one predicted by the power-law growth model (as given by eq. 5.4) (red).

20,000, 40,000 and 60,000 events. The results, shown for different positions of the sliding window, are shown at Figure A.2. We observe that windows at the start of the data systematically overestimate K , which is expected, since $r(x)$ has not stabilized yet, hence the growth is composed of a mix between links from the hidden network and new links. Towards the last windows, the exponent is still higher than the real exponent, which is due to the fact that m_0 is underestimated, since some links of the initial hidden network are never observed.

As expected, longer time windows lead to a more stable solution. In this first example, we observe that the prediction is of good quality: the optimal solutions are close to the real parameters and the solution is stable once $r(x)$ has stabilized. We try now the same analysis on a more complex example that induces a different dynamic in the growth process. Remember that in the previous example, $r(x)$ has stabilized

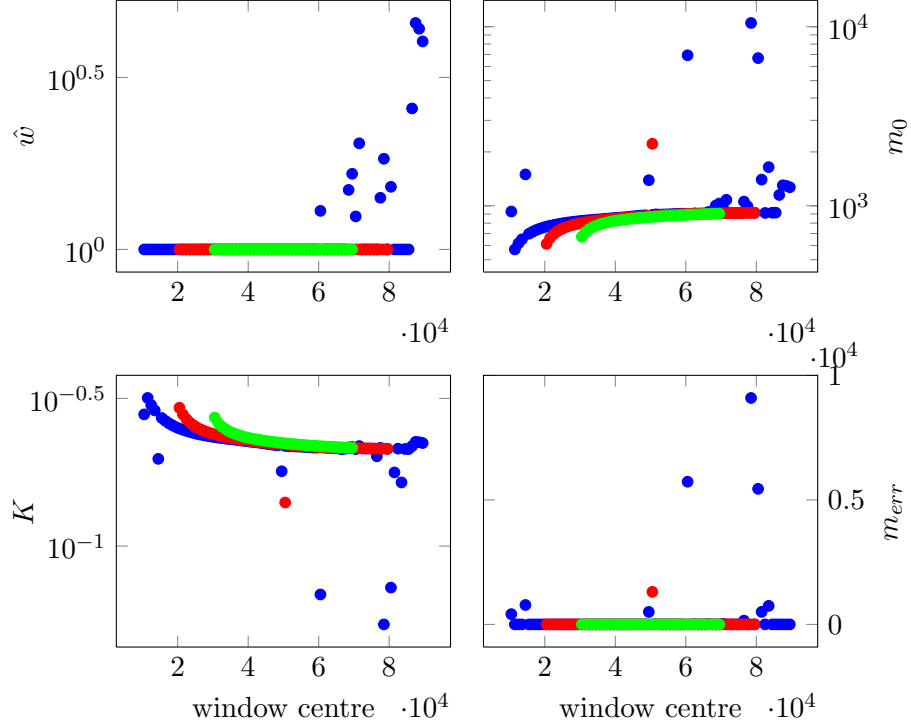


Figure A.2: Parameters obtained using the NM algorithm with windows of varying lengths and positions. The points show the value of the parameters for the window centred in x , with length 20,000 (•), 40,000 (•) and 60,000 (•) out of the 100,000 events of the simulation

once the number of discovered links from the hidden network became constant. In the following example, we show an instance of DPU where the influence of x_0^h is larger and impacts on the shape of $r(x)$ (Figure A.3).

The shape of $r(x)$ is here different to what was predicted by (??), since it decreases below the actual growth rate (0.2) and increases then and eventually stabilizes around that value. The reason why $r(x)$ decreases is related to the way it is estimated. The exact value for the growth rate is

$$r(x) = \frac{x^h}{m^h(x)} \frac{dm^h(x)}{dx}, \quad (\text{A.8})$$

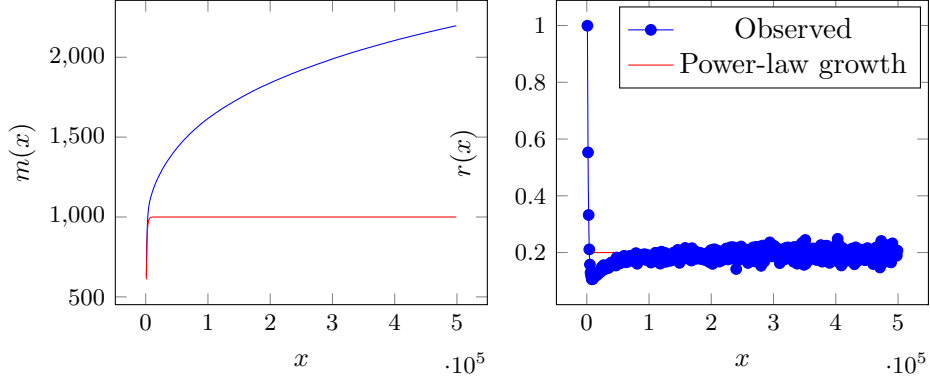


Figure A.3: Average result of 10^2 realizations of DPU with parameters $x_0^h = 10^4$, $m_0^h = 10^3$ and $K = 0.2$. (a) size of the network as a function of x (blue). The first sharp increase results from the discovery of both new links and links from the hidden network. The red curve indicates the number of links from the hidden network that have been observed. (b) $r(x)$ as a function of the number of events. The shape differs from the what is observed for in Figure 5.16, here $r(x)$ decreases further than 0.2, due to the influence of x_0^h .

which is expected to be equal to K . However, since x^h and $m^h(x)$ are unknown, they are estimated by x and $m(x)$. However, since in this example, each link in the hidden network has a high weight (since $x_0^h = 10m_0^h$), all m converges quickly to m^h while the difference between x and x^h stays constant. Hence, the estimation of $r(x)$

$$r(x) \approx \frac{x}{m(x)} \frac{dm(x)}{dx}, \quad (\text{A.9})$$

underestimates $r(x)$, and only stabilizes when $x \gg x_0^h$.

In this case, the stabilization of $r(x)$ is hence not a necessary condition for having a power-law growth.

Again, we observe that the small windows are much more unstable than the large windows, however here the overall instability of the results is much larger than in the first example. The largest window shows acceptable results for low x (i.e. the window is located at the start of the aggregation process), while for high values of x , the optimal parameters

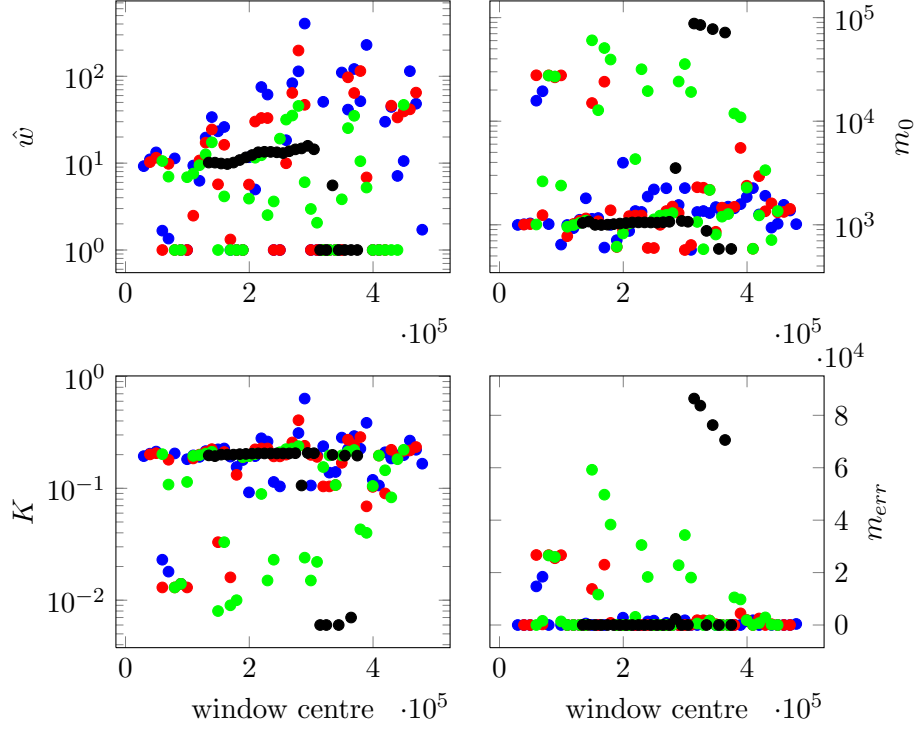


Figure A.4: Parameters obtained using the NM algorithm with windows of varying lengths and positions. The points show the value of the parameters for the window centred in x , with length 40,000 ($\bullet$), 60,000 ($\bullet$), 100,000 ($\bullet$) and 250,000 ($\bullet$) out of the 500,000 events of the simulation

found by the identification procedure are far from the exact parameters, because the fitted curve is very flat, and hence can be fitted by several different sets of parameters.

We finally apply the same procedure on the first three months of the real data, for windows of one and two weeks. Given the weekly patterns present in the data, we avoid to use windows having a length that is not a multiple of one week. It appears that for windows of two weeks, the results show a good stability over time, however there is a slight increase of the fitted hidden network size and a decreasing exponent. This non-constant behaviour seem to suggest that the parameters of the

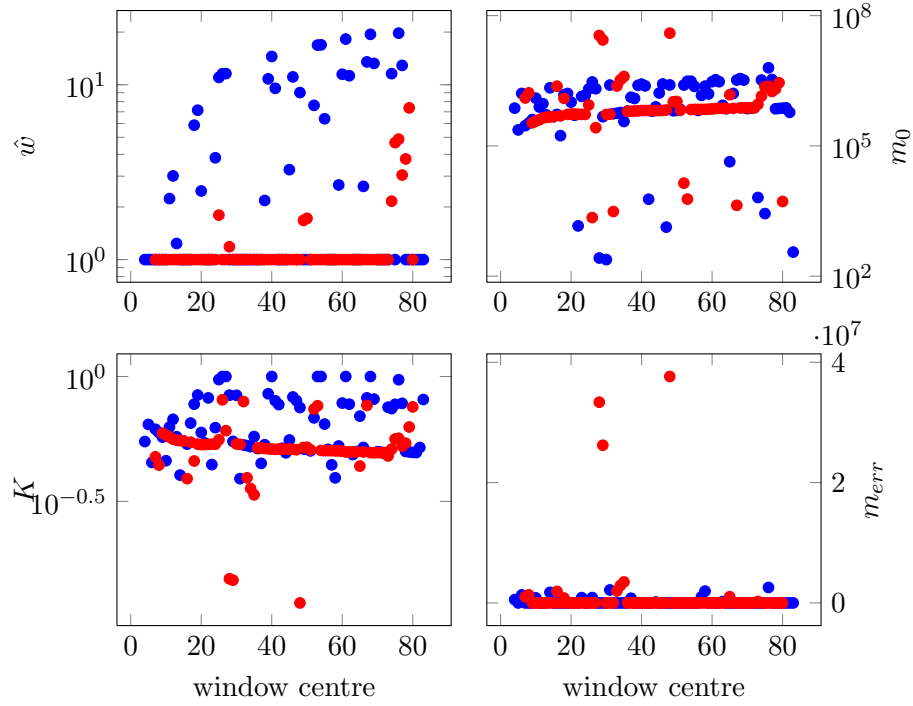


Figure A.5: Parameters identified on the 3 first months of data, for the Belgian mobile phone network, for windows of 1 week (•) and 3 weeks (•).

model are changing over the long term, hence the prediction from the DPU model can mostly be used for short-term prediction.

Interestingly, the parameters fitted on the two-week window are strongly divergent for the windows centred around the days 26 to 35 (the effect is particularly visible on the result of m_0 and K). That period corresponds exactly to the autumn holiday week, during which other measures also confirm that the behaviour of people is different to normal weeks.