ALMA MATER STUDIORUM UNIVERSITA' DI BOLOGNA
COMPUTER SCIENCE DEPARTMENT

PH.D. THESIS

Multidimensional analysis of complex networks

Possamai Lino

Ph.D. Supervisors:

Prof. Massimo Marchiori

Prof. Alessandro Sperduti

Ph.D. Coordinator:
Prof. Maurizio Gabbrielli

DOTTORATO DI RICERCA IN INFORMATICA - Ciclo XXV
Settore Concorsuale di afferenza: 01/B1
Settore Scientifico disciplinare: INF/01

Esame Finale Anno 2013

Ai miei genitori.

Abstract

The analysis of Complex Networks turn out to be a very promising field of research, testified by many research projects and works that span different fields. Until recently, those analysis have been usually focused on deeply characterize a single aspect of the system, therefore a study that considers many informative axes along with a network evolve is lacking. In this Thesis, we propose a new multidimensional analysis that is able to inspect networks in the two most important dimensions of a system, namely space and time. In order to achieve this goal, we studied them singularly and investigated how the variation of the constituting parameters drives changes to the network behavior as a whole.

By focusing on space dimension, we were able to characterize spatial alteration in terms of abstraction levels. We propose a novel algorithm that, by applying a fuzziness function, can reconstruct networks under different level of details. We call this analysis telescopic as it recalls the magnification and reduction process of the lens. Through this line of research we have successfully verified that statistical indicators, that are frequently used in many complex networks researches, depends strongly on the granularity (i.e., the detail level) with which a system is described and on the class of networks considered.

We keep fixed the space axes (that is, nodes' coordinates) and we isolated the dynamics behind networks evolution process. In particular, our goal is to detect new instincts that trigger online social networks utilization and spread the adoption of novel communities. We formalized this enhanced social network evolution by adopting new special nodes, called "sirens" that, thanks to their ability to attract new links, were also able to construct efficient connection patterns. We both simulate the dynamics of individuals and sirens by considering three most known growth models, namely random, preferential attachment and social.

Applying this new framework to real and synthetic social networks, we have shown that the sirens, even when used for a limited period of time, effectively shrink the time needed to get a network in mature state. In order to provide a concrete context of our findings, we formalized the cost of setting up such enhancement and we provided the best combinations of system's parameters, such as number of sirens, time span of utilization and attractiveness, which minimize this cost.

Contents

Abstract **5**

1 Introduction **15**

1.1 Outline of the thesis . . . 17

2 Background **21**

2.1 Graph Theory . . . 21

2.1.1 Statistical properties . . . 24

2.2 Network Modeling . . . 27

2.2.1 Erdős-Rényi Networks . . . 28

2.2.2 Small World Networks . . . 29

2.2.3 Scale-Free Networks . . . 31

2.3 Longitudinal analysis . . . 34

2.4 Critical analysis of complex networks . . . 36

2.4.1 Static robustness . . . 36

2.4.2 Dynamical robustness . . . 38

2.5 Spreading processes . . . 39

2.5.1 Epidemic spreading . . . 39

2.5.2 Rumors spreading . . . 40

3 Multidimensional analysis of networks **43**

4 The space dimension **47**

4.1 Related work . . . 49

4.2 Telescopic Algorithm . . . 51

4.3 Experimental Results . . . 54

4.3.1 Datasets . . . 54

4.3.2 Results . . . 63

4.4 Network perturbations . . . 74

4.4.1 Null models as a function of networks' size . . . 83

4.4.2 Null model and real networks comparison . . . 91

5 The time dimension **95**

5.1 Introduction 95

5.2 Network growth models 96

5.2.1 Evolutionary models: serial and parallel 98

5.3 Sirens 100

5.3.1 Network Cost 103

5.4 Experimental Results 103

5.4.1 Datasets 104

5.4.2 Results 108

6 Conclusions and Future Directions **127**

A Supplementary material for space dimension **133**

List of Figures

2.1 Graph abstraction of the seven bridges of Königsberg problem. 22
2.2 Graph types. 23
2.3 Regular to random networks via small-world 30
2.4 Random and scale-free degree distributions 32
2.5 Impact of cumulative degree distribution on noise reduction in the data 33
2.6 Random failure and attacks on networks 37

3.1 High level overview of multidimensional analysis on complex networks 44

4.1 Example of telescopic analysis 48
4.2 Example of box-covering on networks 49
4.3 Spatial coarse graining for brain voxels 50
4.4 One-step abstraction process 52
4.5 Sequence of box covering iteration for telescopic analysis 53
4.6 Box covering issue 53
4.7 Example of box covering on real networks 55
4.8 Subway network maps 56
4.9 Examples of city-based networks 57
4.10 Example of city-based networks 58
4.11 P_{cum} distribution of subways, transportation and social networks. . . . 62
4.12 Number of collapsed nodes and edges as a function of f in log-log axes 64
4.13 Effect of the telescopic abstraction on the diameter as a function of f 65
4.14 Effect of the telescopic analysis on the degree 66
4.15 Impact of the telescopic analysis on the degree correlations ρ 67
4.16 Effect of the telescopic process on E_{glob} 68
4.17 Effect of the telescopic analysis on E_{loc} 70
4.18 Effect of the telescopic analysis on c_t and c_m 71
4.19 Effect of the telescopic analysis on C/E_{glob} 72
4.20 Effect of the abstraction process on the degree distribution $P(k)$ . . . 73
4.21 Number of collapsed nodes and edges with different perturbations . . 75
4.22 Telescopic analysis on various degree based indexes and perturbations 77
4.23 Effect of the telescopic analysis on ρ with different perturbations . . . 78
4.24 Effect of the telescopic process on E_{glob} as a function of perturbations 79

4.25 Effect of the telescopic process on E_{loc} as a function of perturbations . 80
4.26 How different perturbations influence costs in the telescopic analysis . 81
4.27 How different perturbations influence C/E_{glob} in the telescopic analysis 82
4.28 Effect of the telescopic analysis on random networks on nodes and edges 84
4.29 Effect of the telescopic abstraction on diameter on null models 85
4.30 Effect of the telescopic abstraction on degree in null models 86
4.31 Effect of the telescopic abstraction on E_{glob} on random networks . . . 87
4.32 Effect of the telescopic abstraction on E_{loc} 88
4.33 Effect of the telescopic analysis on c_t and c_m in random networks. . . 89
4.34 How the telescopic analysis influence C/E_{glob} 90
4.35 Comparison between original and randomized subway networks 92
4.36 Comparison between original and randomized city-based networks . . 93

5.1 Example of social rule in evolving social networks 98
5.2 Impact of C scale parameter on overall simulated time 99
5.3 Example of on line social network with sirens 100
5.4 Examples of sirens' social rule 102
5.5 Cumulative degree distributions of CM and VT datasets 106
5.6 Node degree correlations in CM and VT datasets 106
5.7 Serial network growth model 110
5.8 Parallel network growth model 111
5.9 Parallel network growth model of random networks 111
5.10 How attractiveness influence the overall network behavior 113
5.11 E_{glob} varying the cost, accelerated growth, word of mouth model (CM) 114
5.12 E_{glob} varying the cost, accelerated growth, preferential model (CM) . . 115
5.13 Effect of multiple runs of simulations on VT 115
5.14 How cost affects E_{glob} in accelerated growth, broadcast model (VT) . 116
5.15 How cost affects E_{glob} in accelerated growth, word of mouth model(VT) 117
5.16 How cost affects E_{glob} in accelerated growth, preferential model (VT) 118
5.17 E_{glob} as a function of cost keeping fixed the number of sirens (CM) . . 120
5.18 Effect of attractiveness and time in CM, word of mouth model 120
5.19 Effect of attractiveness and time in CM, preferential model 121
5.20 Effect of attractiveness and time in VT, broadcast model 121
5.21 Effect of attractiveness and time in VT, word of mouth model 122
5.22 Effect of attractiveness and time in VT, preferential model 122
5.23 Correlation between cost and T_{min} in CM 124
5.24 Correlation between cost and T_{min} in VT 125

A.1 Effect of the telescopic analysis on nodes of perturbed subways 134
A.2 Effect of the telescopic analysis on edges of perturbed subways 135
A.3 Effect of the telescopic analysis on k_{max} of perturbed subways 136
A.4 Effect of the telescopic analysis on k_{mean} of perturbed subways 137
A.5 Effect of the telescopic analysis on σ_k of perturbed subways 138
A.6 Effect of the telescopic analysis on E_{glob} of perturbed subways 139
A.7 Effect of the telescopic analysis on E_{glob} of perturbed subways 140

A.8 Effect of the telescopic analysis on E_{loc} of perturbed subways 141
A.9 Effect of the telescopic analysis on E_{loc} of perturbed subways 142
A.10 Effect of the telescopic analysis on c_t of perturbed subways 143
A.11 Effect of the telescopic analysis on c_m of perturbed subways 144
A.12 Effect of the telescopic analysis on C_t/E_{glob} of perturbed subways . . 145
A.13 Effect of the telescopic analysis on C_m/E_{glob} of perturbed subways . . 146

List of Tables

4.1 Statistical features of transportation and city-based social networks . . 59
4.2 Statistical features of randomized networks 60

5.1 Dataset statistics of CM and VT on line social networks 105
5.2 Summarize of attractiveness values 108
5.3 Summarize of configurations 108
5.4 Summarize of T_{min} for accelerated simulations 110
5.5 Summarize of T_{min} for Communities 124
5.6 Summarize of T_{min} for VirtualTourist 125
5.7 Summarize of T'_{min} and β. 126
5.8 Total cost as a function of C_s 126

Chapter 1

Introduction

For many years, researchers studied the components of the nature singularly omitting to observe the whole system behavior. For instance, considering water molecules, the state transformation from liquid to solid happens through a topological modification of the interaction network of molecules. So, a study that takes into account only the singular molecular behavior will necessarily be limited.

Complex systems have attracted so much attention in the last decade because of the direct implications of results to medicine, business, computing, physics, mass transportation, power distribution, law, and many other fields. For example, vaccination campaigns against hard to eradicate viruses, such as smallpox, is more effective by first treating hub nodes of the social network, i.e., nodes that have many social ties compared to the average. Unfortunately, identifying such individuals can be a challenging task.

Mapping out the interactions among genes, proteins and other molecules regulating cell's activity in living organisms were fundamental to aid researchers uncovering and controlling the side effects of drugs, before release to the market. Furthermore, the critical analysis of these networks lead to the identification of new regulatory relations, functional to the development of diseases [20]. On the business field, understanding how companies, industries and people are connected each other could help researchers to monitor and avoid cascading financial failures[91]. Moreover, understanding how global market works, and in particular studying the spread of a contagion on a topological structured network could offer new effective ways to promote and sell products. On the World Wide Web, identifying which the most critical nodes helps to eradicate computer virus and defend from attacks and sabotages.

A complex network is a system containing many interdependent units that interact in a non-linear way. With this definition at hand, researchers have constructed complex networked systems based on data gathered from many different fields: the brain is a network of nerve cells connected by axons; cells themselves are networks of molecules connected by biochemical reactions. Societies too are networks of people connected by friendship, familial and professional relationships. On large scale, food webs and ecosystems can be represented as networks of predator-prey relations. Indeed, networks are presents also in technology: the Internet as web pages connected by hyper links, the routers network (AS), power grids, and transportation systems are but a few examples. Even the language we are using everyday is a network, made up of words connected by semantic relationships. The most famous motto utilized is: *Networks are everywhere.*

The first step toward understanding the principles that govern the creation, maintenance, dissolution and reconstitution of links and nodes of the graphs was to create models that hypothetically generate such structures. The first scientists that developed models of network growth were mathematicians and physicists. Initially, they pictured real world networks as they were random and supposed that nodes were pair wisely connected with fixed probability p. In this way, the parameter p determines the level of connectedness of the networks (i.e. the number of edges). When $p \to 0$ the network becomes a cloud of disconnected nodes whereas in the opposite situation, when $p \to 1$, the outcome graph will be complete, i.e. every node is connected to every others. For intermediate values of p, the network will have approximately $pn\cdot(n-1)/2$ links (for undirected networks) on average and nodes are connected to randomly selected neighbors without any geographical or social constraints.

The availability of network datasets, allowed scientist to empirically verify that random models do not have much in common with real networks nor with regular one whose connection pattern is fixed (every node is connected to k close neighbors). Interestingly, Watts and Strogatz [110] found that real complex systems are structurally similar to graphs that sit in between these two extremes: on one hand complete disorder and on the other regularity (see Background chapter 2 and in particular figure 2.3). They indeed proposed a model that, starting from a ring lattice, rewires each edge at random with probability p. This parameter allows shifting between the two extreme patterns. When $0 < p < 1$, the properties of the output graph strongly match with that of real networks. This finding was the beginning of a series of unexpected and very important results.

Parallelly to the first discoveries on complex networks, researchers' efforts were focused also in defining and quantitatively characterize new measures that describe the topology of real networks. The main result has been the identification of a series of unifying principles and statistical properties common to most of the real networks considered. Two fundamental elements that characterize networks are nodes' degree, i.e. the number of its adjacent connections, and degree distribution $P(k)$. The latter is defined as the probability that a node chosen uniformly at random has degree k or, equivalently, as the fraction of nodes in the graph having degree k. Another important result was to find that the degree distribution $P(k)$ of real networks significantly deviates from Poisson distribution (expected for random graphs) exhibiting, instead, a power law $P(k) \sim \gamma$ (scale-free) behavior (especially in the tail of the distribution) with an exponent γ that empirically belongs to the range $[2, 3]$. Moreover, real networks are characterized by many other common features, such as correlations in the node degrees, by average short paths between any two nodes (small-world property), and by the presence of a large number of short cycles or specific motifs.

Furthermore, it was shown that the coupling architecture (for example degree-degree correlation) has important consequences on the network functional robustness and resilience to external perturbations, as random failures, or targeted attacks [105][50][87]. At the same time, correlations have a strong impact on network dynamical properties such as epidemic diffusions. Real world networks are extremely sensitive to assortativity by degree correlations.

This led to a series of evidences pointing to the crucial role played by the network topology in determining the emergence of collective dynamical behavior, such as synchronization, or in governing the main features of relevant processes that take place in complex networks, such as the spreading of epidemics, information and rumors.

1.1 Outline of the thesis

Researches on complex systems of the last decade was permeated by static analysis of graphs and generation models that aim at producing the structure that matches features specific to real world dataset. However, few of them consider that real complex systems are embedded and constrained by space and time. Because of that, it is of fundamental importance that new analysis will be able to deal with these informative axes in order to get a better understanding of the entire system.

This Thesis aims at filling this gap and in particular, we propose a novel multidimensional analysis of complex systems in which we separately study networks along the two most important informational axes, time and space, and subsequently we lay the foundations for a complete, and first time seen, analysis that considers both of them.

By fixing the time dimension and by focusing on spatial x, y axes, we were able to propose a new analysis that quantitatively describes how the metrical and topological structure of complex networks varies as a function of the detail levels. To the best of our knowledge, this is the first attempt to use such approach in the study of complex networks. To achieve this goal, we developed a new algorithm (based on the resolution power of human eyes) that abstracts from the local properties of nodes and reconstructs the structure of the original graph as it was placed at specific distance (fuzziness) from a point of view. This analysis, called *telescopic*, was fundamental to uncover that not all statistical quantities are spatially universal but instead, they strongly depends on the specific detail level with which a network is described.

On the other hand, by keeping fixed the space dimension we can concentrate on studying how networks evolve as a function of time. Several works in the literature has been devoted to understand how structure evolves and to depict the underlying rules behind links and nodes creation (these analysis are referred as *longitudinal*). An important, but by far underestimated aspect that has never been studied before is the detection of the instincts that trigger and consequently spread the users' commitment in a new on line community. We already know that the adoption of new on line social sites is a slow process. Because of that, we identify a method that successfully fosters individuals' network utilization.

The thesis is organized as follow. Chapter 2 reviews some basic definitions, statistical measures and models of complex networks. In particular, we introduce the notations that have been adopted throughout this Thesis. For the sake of clarity, we focus only on the concepts that will be used extensively in the rest of the Thesis. A series of review articles and books referenced can be found in this chapter.

In Chapter 3 we present an introduction to the framework for multidimensional analysis and the sub analysis that is formed by.

In Chapter 4 we introduce and characterize the telescopic framework that is able to describe networks as a function of the detail levels. We tested our algorithm on rapid transportation networks such as European and American subways and two on line social networks. We also studied the abstraction process on synthetic networks by systematically testing different types of perturbations on nodes' positions and edges connectivity and finally compared the results with those obtained in real world networks.

Chapter 5 is devoted to present the analysis of the second informative axis though which networks evolve: time. In particular, throughout all the simulations and experiments carried out in this Thesis we review and apply the most straightforward techniques known that are at the root of the network growth namely *random* (every edge has the same probability to exists), *aristocratic* (the most connected nodes are those with higher probability to acquire new links) and *social* (two friend of a person are likely to know each other). These mechanisms are extensively applied in conjunction with a set of special nodes, called *sirens*, that effectively amplify network's utilization. The results obtained for two on line social networks, as well as artificial networks, confirm that this is an effective method that governs the system's evolution.

Finally, conclusions and future directions based on results obtained in this Thesis are presented in chapter 6.

Chapter 2

Background

In this section, we introduce some background concepts that will be useful to understand the remainder of the Thesis. In particular, we will provide some definitions originally proposed in the context of graph theory and now widely used in network analysis.

2.1 Graph Theory

In mathematics, graph theory[40] has been around for almost three hundred years. Leonhard Euler in 1736 was the first that applied graph theory concepts to solve real world problems. He showed that citizens of Königsberg can not traverse their seven bridges without back-tracking. This result was accomplished by reformulating the main problem into the settings of the graph theory and by considering nodes as lands (letter A, B, C and D in figure 2.1) and bridges as edges. Euler's demonstration was based on considering the degree of nodes, i.e. the number of its neighbors. In particular, he observed that, for a walk of the desired form, the graph must be connected and its nodes must have odd degree.

The mathematical object that naturally fits in the previous example as well as in all systems of interacting elements is the graph. It is utilized as an abstraction of a system where each element of the network is a node and relations between them are edges/arcs. This is the most trivial definition of graph though. The graph theory community developed more expressive versions of graphs that account for edges' orientation (directed and undirected), nodes' strength, Euclidean nodes' coordinates, and so on. The following paragraphs are devoted to delve deeply into these concepts in a more formal way.

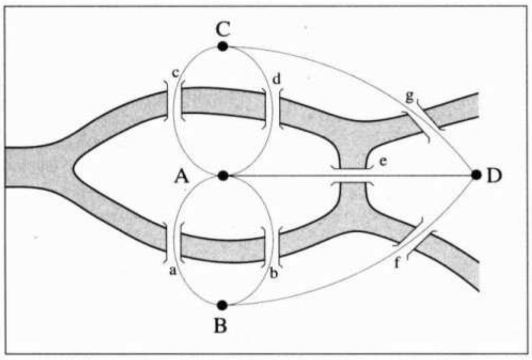


Figure 2.1: Graph abstraction of the seven bridges of Königsberg problem.

A (*simple*) graph **G** is a pair (V, E) where $V = \{u_1, u_2, ..., u_n\}$, $|V| = n$ is a finite set of vertices and $E \subseteq V \times V$, $E = \{(u_i, u_j), i \neq j\}$, $|E| = m$ is the set of edges that links couples of nodes. These graphs are called *topological*. A graph G could be represented by a $n \times n$ adjacency matrix A with entries $a_{ij} = 1$ when $(u_i, u_j) \in E$, $a_{ij} = 0$ otherwise. $a_{ii} = 1$ denotes self loops. A weighted graph is defined as $G = (V, E, w)$ where $w : E \to \mathcal{R}$ is a function that assigns real values to edges. In undirected graphs, $(u, v) \in E \Leftrightarrow (v, u) \in E$ and adjacency matrix A will be symmetric (with respect to its diagonal, that consists of all zeros if self loops are not allowed). Conversely, in directed graphs, or *digraphs*, each edge (sometimes referred as arc or link) has a orientation, so $(u_i, u_j) \neq (u_j, u_i)$.

Metrical graphs (also known as spatial) extend weighted graphs as they are spatially embedded, that is, every node exists in a Euclidean coordinates' space. Specifically, $G = (V, E, C, w)$ where $C = \{(x_1, y_1), (x_2, y_2) \cdots, (x_n, y_n)\}$ is the set of nodes' coordinates and the function w might assigns, for instance, Euclidean distances between nodes. *Multigraphs* are generalized graphs in which the same couple of nodes might be connected by more than one edge. Even though many real world complex systems could be represented by multigraphs, in many occasions, these networks are transformed into weighted graphs in such a way that the number of edges connecting two nodes is reflected in the edge weight of the new graph (see figure 2.2 for a summary of of graph classes).

A *path* is a non empty graph $P = (V, E)$ in the form of

$$V = \{u_0, u_1, ..., u_k\} \qquad E = \{(u_0, u_1), (u_1, u_2), ..., (u_{k-1}, u_k)\};$$

simple paths are those in which all vertices u_i are distinct. The number of edges in a path determines its length and a path of length k is defined as P^k. A path from a to

b of length k is a path P^k in which $u_0 = a$ and $u_k = b$. A graph $\mathbf{G}$ is *connected* if for each $u_i, u_j \in V, i \neq j$, there exists a simple path from u_i to u_j (denoted as $u_i \rightsquigarrow u_j$). In order to simplify notations, in this Thesis we equivalently specify nodes as i or u_i.

An important graph property is the *shortest path* between two vertices, d_{ij} (also known as *geodesic*). The definition of shortest path depends on the class of graphs we are dealing with. In simple graphs, the shortest path between nodes i and j represents the minimum number of traversed nodes (hops) to reach j from i. If the graph is connected, it is natural to observe $d_{ij} \geq 1 \ \forall i, j$ and $d_{ij} = 1$ if node i is directly connected to node j. If there are no paths between i and j then $d_{ij} = \infty$. Indeed, in weighted graphs, the shortest path is calculated taking into account the weights on edges such that $d_{ij} = \min\{w_p | p \text{ is a path between } i \text{ and } j\}$ where

$$w_p = \sum_{e \in E(p)} w(e)$$

is the sum of edge weights along path p.

The *diameter* D of graphs is usually defined as the maximum d_{ij} between every couple of nodes. However, since d_{ij} depends on the graph type, D could also have the following meanings: the number of hops that separates two vertices, the maximum shortest weighted path or the maximum Euclidean distance between the farthest nodes, without considering the underlying topological structure (in this case we refer to *physical diameter*).

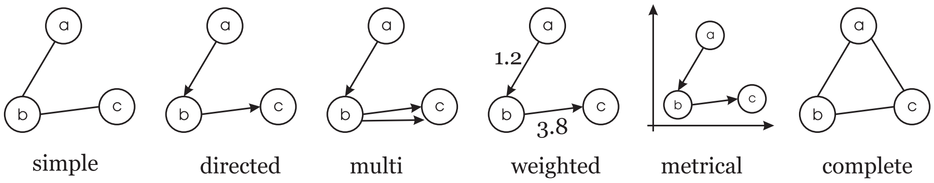


Figure 2.2: Graph types.

The *degree* of a node u in a graph corresponds to the cardinality of the set $N(u) = \{v \in V \mid (u, v) \in E\} = deg(u) = k_u$ and $\sum_{u \in V} deg(u) = 2|E|$. When $deg(u) = 0$, then u is said to be *isolated.* In directed graphs, it is customary to split node degree into inbound k^{in} and outbound k^{out} degree. Indeed, the degree distribution $P(k)$ that corresponds to the probability of having a node with degree k, has to be split into two parts, inbound $P^{in}(k)$ and outbound $P^{out}(k)$ degree distribution. The *average degree* of a graph $\langle k \rangle$ (or k_{mean}) is $1/n \sum_{i=1}^{n} k_i$ and the strength[16] s_i of node i is the sum of the weights of the edges incident on i, $s_i = \sum_j w_{ij}$. In directed graphs, the strength can be split relative to the edges directions, reflecting the total inbound and outbound weight, as for the nodes' degree and the degree distribution.

A graph $\mathbf{G}$ is *complete* if for each $i, j \in V(i \neq j) \Rightarrow (i, j) \in E$. In the literature, complete graphs are usually denoted as $\mathbf{K}_n$, with n representing the total number of nodes and $|E| = \frac{n(n-1)}{2}$ if the graph is undirected, $n(n-1)$ otherwise. Recent experiments showed that those graphs are rare to find in nature mainly because of the inherent high cost of creation and maintaining such a redundant structure. Think, for instance, of having a telephone network in which there exist direct connections between every user. This class of networks are usually used in ideal contexts or as normalizing factor in formulas (see section § 2.1.1).

A graph T is a *subgraph* of a graph G, denoted by $T \subseteq G$, when $V_T \subseteq V_G$ and $E_T \subseteq E_G$ holds. V_T and V_G are the set of nodes of G and V respectively. A graph $T \subseteq G$ is said to be *induced* when $E_T = \{(u_i, u_j) \in E_G | u_i \in V_T, u_j \in V_T\}$.

The previous definitions are only a subset of all concepts and ideas that have been developed in the graph theory literature. For interested readers, Diestel's [40] book is a one of the best reference in this field.

2.1.1 Statistical properties

Before delving into the details of network modeling, we present an overview of metrics that catch the most important network properties.

Watts and Strogatz [110] proposed two effective and intuitive metrics, namely the characteristic path length L and the clustering coefficient C. The first measures the typical separation between two vertices in a graph (a global quantitative measure of graphs), whereas the second measures the cliquishness of a typical neighborhood (a local property) [102]. More formally, the former is calculated as

$$L(\mathbf{G}) = \frac{1}{n(n-1)} \sum_{i \neq j \in V} d_{ij}.$$

Since real world networks might have disconnected subgraphs (for example Escherichia coli [10] or some protein to protein networks[20]), network scientists usually restrict their study to the largest connected component (LCC), in which $d_{ij} < \infty$ for each $(u_i, u_j) \in E$. The results, in order to be significant have to be calculated on big LCC, i.e. the fraction of nodes that belongs to it must be very high.

Vice versa, clustering coefficient C is formally described as the mean of all C_i's, namely:

$$C(\mathbf{G}) = \frac{1}{n} \sum_{i \in \mathbf{V}} C_i \qquad C_i = \frac{E[\mathbf{G}_i]}{k_i(k_i - 1)/2}$$

where C_i is the fraction between the numbers of edges of the subgraph $\mathbf{G}_i$ over the total number of edges of $\mathbf{K}_i$. Subgraph $\mathbf{G}_i$ is the graph of the neighbors of node i (i

excluded).

Latora and Marchiori [67] developed a set of metrics, based on the concept of efficiency ϵ, that allow considering both connected and disconnected graphs. They define global efficiency of a graph G as:

$$E_{glob}(\mathbf{G}) = \frac{\sum_{i \neq j \in V} \epsilon_{ij}}{n(n-1)} = \frac{1}{n(n-1)} \sum_{i \neq j \in \mathbf{G}} \frac{1}{d_{ij}}$$

as the average of efficiency ϵ_{ij} of the graph. Here, they assumed that efficiency ϵ_{ij} and distance d_{ij} are inversely proportional. However, other relationships might be used (instead of d_{ij}), especially when justified by a more specific knowledge about the system. Nevertheless, d_{ij} will have different meanings in weighted and unweighted networks. In the first case, it corresponds to the number of hops between two nodes in the shortest path (*topological efficiency*) whereas in the second one is the sum of all edge weights in the shortest path (*metrical efficiency*). Global efficiency, as defined above, ranges from 0 to $+\infty$. In practical applications, it is convenient to normalize it by the ideal network K_n, namely $E_{glob}(\mathbf{G})/E_{glob}(K_n)$ such as $0 \leq E_{glob}(\mathbf{G}) \leq 1$, therefore it can be used to compare efficiency of different graphs.

On the other side of the same measure, the efficiency can be used to evaluate any subgraph of $\mathbf{G}$, and therefore to characterize the local properties of a network as the following:

$$E_{loc}(\mathbf{G}) = \frac{1}{N} \sum_{i \in \mathbf{V}} \frac{E_{glob}(\mathbf{G}_i)}{E_{glob}(\mathbf{G}_i^{ideal})}$$

that is merely the average of the global efficiency applied to each subgraph $\mathbf{G}_i$, normalized by the referring ideal graph $\mathbf{G}_i^{ideal}$.

Moreover, the same authors proposed a statistical property that accounts for the *cost* of a network, defined as:

$$Cost(\mathbf{G}) = \frac{2m}{n \cdot (n-1)}, \qquad Cost(\mathbf{G}) = \frac{\sum_{i \neq j \in \mathbf{G}} a_{ij} \gamma(d_{ij})}{\sum_{i \neq j \in \mathbf{G}} \gamma(d_{ij})}$$

The leftmost formula is used in unweighted networks and is usually known as *density* whereas the rightmost accounts for weighted networks where a_{ij} is an element of the graph adjacency matrix A and γ is the cost evaluator function which calculates the cost needed to build up a connection with a given distance (length) d_{ij}.

In order to depict the most important nodes of the network, the following indexes are widely used in the literature. The *closeness centrality* defined as [109]:

$$C_i^C = \frac{n-1}{\sum_{j \in G, i \neq j} d_{ij}}$$

represents how far nodes' neighborhood are from i. In undirected graphs, when a node i is connected to every other nodes (i.e. obtaining a star network), the closeness centrality is equal to 1.

Betweenness [109][46][47], C^B, express the centrality of a node by calculating how many shortest paths pass through a node, namely:

$$C_i^B = \frac{1}{(N-1)(N-2)} \sum_{j,k \in G, j \neq k \neq i} n_{jk}(i)/n_{jk}$$

where n_{jk} is the number of shortest paths between j and k, and $n_{jk}(i)$ is the number of shortest paths between j and k that contains node i.

The *straightness centrality*, C^S, represent how far are neighborhood's nodes from an origin considering ideal straight connections. In other words, this index captures to which extent the connecting route between nodes i and j deviates from the virtual straight line [108], namely:

$$C_i^S = \frac{1}{N-1} \sum_{j \in G, j \neq i} d_{ij}^{Eucl}/d_{ij}$$

where d_{ij}^{Eucl} is the Euclidean distance between nodes i and j along a straight lines.

The *information centrality* [68], C^I, allow to identify a central nodes by calculating the relative drop in the global efficiency of the network caused by the removal of the edges incident in i from G [103]. It is defined as the following:

$$C_i^I = \frac{\Delta E}{E} = \frac{E_{glob}(\mathbf{G}) - E_{glob}(\mathbf{G}')}{E_{glob}(\mathbf{G})}$$

and $\mathbf{G}'$ is the graph with n nodes and $m - k_i$ edges.

Another fundamental property of networks is the *degree-degree correlation* (also known as network *assortativity*). This feature is extremely important in the resilience of networks[105] [98] but it has also strong impact on the network dynamical properties, such as spreading processes. In *assortative* networks, most edges connect nodes that exhibit similar degrees (nodes aristocracy). On the other hand, *disassortative* networks are such that high-degree nodes are connected to low-degree nodes.

More analytically, the network correlation k_{nn} between vertices is calculated as $k_{nn}(k) = \sum_{k'} k' P(k'|k)$ where $P(k'|k)$ is the conditional probability that a node with degree k is connected to a node with degree k'. If there is no degree correlation, the formula simplifies to $k_{nn}(k) = \langle k^2 \rangle / \langle k \rangle$, i.e. is independent of k. Positively correlated graphs are classified as assortative if k_{nn} is an increasing function of k, whereas they are referred to disassortative when $k_{nn}(k)$ is a decreasing function of k [84]. Degree correlations are usually quantified by reporting the numerical value of the slope of $k_{nn}(k)$ as a function of k or by calculating the Pearson correlation coefficient of the degrees at either ends of a link [85]. ER graphs are, by definition, uncorrelated graphs, since the edges are connected to nodes regardless of their degree. Consequently, the assortative-mixing value is neutral (zero). This holds also for the preferential attachment model proposed by Barabási-Albert [15].

Along with assortativity by degree, Barrat et al. [16] defined the weighted nearest-neighbors degree:

$$k^w_{nn,i} = \frac{1}{s_i}\sum_{j=1}^{n} a_{ij} w_{ij} k_j.$$

They perform a local weighted average of the nearest-neighbor degree according to the normalized weight of the connecting edges, w_{ij}/s_i. The $k^w_{nn,i}$ thus measures the effective affinity to connect with high- or low-degree neighbors according to the magnitude of the actual interactions. As well, the behavior of the function $k^w_{nn}(k)$ marks the weighted assortative or disassortative properties considering the actual interactions among the system's elements.

Interestingly, Newman found that physics co-authorship [82], biology [82], mathematics [53], film actors [110] and company directors [37] networks all share the same assortative behavior. On the contrary, Internet[90], the World Wide Web[15], protein to protein interactions[60], neural[67] and food [73] network are disassortative. For a survey of this and the previous statistical properties we refer to the following references [21][26].

2.2 Network Modeling

Systems' modeling is a fundamental element that is extensively used in many fields of science. The main reasons for using modeling in network science are twofold. First, it formalizes a system in order to being able to use mathematical and analytical tools to describe properties in a precise way. Furthermore, is widely used as a prediction tool for the systems behavior.

Generation models were the first ones to be introduced in network science. These models produce a network from an initial core (seed) graph and generate output graphs whose structure matches real data properties such as degree distribution $P(k)$, local clustering, assortativity, etc at a specific time in the future (for instance when a system is in steady state). Even though these models were very precise in characterizing the statistical properties of mature graphs, they discard completely the history and all the events that govern and constraints the systems' evolution. However, thanks to the digitalization and the availability of big datasets and the possibility of knowing the exact timestamp of every element's interaction, it is now possible to propose new researches that aim at analyzing networks as time evolves (the so-called *longitudinal analysis*, see section § 2.3) taking into account the history paths of those systems with the aim of developing models that takes into account also the history paths of those systems.

In the following sections, we review the most important and most used network models proposed during the last decade.

2.2.1 Erdős-Rényi Networks

The systematic study of random graphs was initiated by Erdős and Rényi in 1959 [42] with the original purpose of studying, by means of probabilistic methods, the properties of graphs as a function of the increasing number of random connections. The term random graph refers to the disordered nature of the connections between nodes. In their first article, Erdős and Rényi proposed a model to generate random graphs (the so-called ER graphs) with n nodes and m links. Starting with n disconnected nodes, the model randomly selects couples of nodes, connects them (prohibiting multiple connections), until the number of edges is equals to m [42]. An output graph is only one of the possible outcomes of various realizations, an element of the statistical ensemble of all possible combinations of connections.

An alternative model for ER random graphs consists in connecting each couple of nodes with a probability $0 < p < 1$. This procedure defines a different ensemble and contains graphs with different numbers of links: graphs with m links will appear in the ensemble with a probability $p^m(1-p)^{M-m}$ where $M = n(n-1)/2$ is the total number of edges in a complete graph[43]. These two models have a strong analogy [88] and coincide in the limit of large n [113]. Notice that the limit $n \to \infty$ is taken at fixed $\langle k \rangle$, which corresponds to $2m/n$ in the first model and $p(n-1)$ in the second one. Although the first model seems to be more close to applications, analytical calculations are easier and are usually performed in the second model.

Many properties were discovered by Erdős and Rényi[22] in the context of graph connectivity. For instance, if $p > 1/n$ the corresponding average degree $\langle k \rangle = 1$ and when $p \geq \frac{\ln n}{n}$ almost any random graph constructed is totally connected. Furthermore, it was demonstrated that when n is large, $\langle k \rangle \simeq p \cdot n$ and the degree distribution $P(k)$ (i.e. the probability of finding nodes with degree k) is approximated by a Poisson distribution:

$$P(k) = \frac{\langle k \rangle^k}{k!} e^{-\langle k \rangle}.$$

For this reason, ER graphs are sometimes called Poisson random graphs.

The topological diameter D varies in a small range of values around $\ln n/\ln(p{\cdot}n) = \ln n/\ln\langle k\rangle$ [29] while the average geodesic L has the same behavior as a function of n (as the diameter), $L \sim \ln N/\ln\langle k\rangle$ [17]. The clustering coefficient is equal to $C = p = \langle k\rangle/n$ (see [110]):

$$C^{rand} = \frac{1}{n}\sum C_i^{rand} \qquad\qquad C_i^{rand} = \frac{p(k_i(k_i-1)/2)}{k_i(k_i-1)/2}$$

Hence, ER random graphs have a decreasing C in the limit of large system size.

The US highway network [9] and some other rapid transportation networks like subways[66] are examples of networks whose properties (for instance the degree distribution) are comparable to those of random models. However, empirical studies show that the topology of the majority of real world networks does not share many common properties with ER graphs. For instance, in social networks, the probability of knowing your neighbors is not the same as knowing a person 10 thousands kilometers far from you, as random models suggest (even though in on line social network this constraints could be slightly relaxed).

2.2.2 Small World Networks

The ability of collecting large-scale datasets in all fields that was almost impossible until recently, the birth and development of computational and communication capabilities and the breakdown of boundaries between disciplines helped researchers to propose models more representative of the real world.

Motivated by local clustering, a feature that lacked on previous random graphs models, Watts and Strogatz [110], introduced a new model (in the following, WS model), that will be known as small world. The starting point is a n nodes ring (see figure 2.3), in which each node is symmetrically connected to its $2g$ nearest neighbors for a total of $k = gn$ edges. Then, for every node, each link connected to a clockwise neighbor is rewired to a randomly chosen node with a probability p, and preserved with probability $1 - p$. Notice that for $p = 0$ we have a regular lattice, while for $p = 1$ the model produces a random graph with the constraint that each node has a minimum connectivity $k_{min} = g$.

For intermediate values of p the procedure generates graphs with low average distance between nodes and high local clustering coefficient. Networks with such features are called small-world. Alternative procedures to construct small-world networks exist and are based on adding edges instead of rewiring them[36][79].

The richness of the WS model has stimulated the scientific community to study network's properties as a function of the rewiring probability p and the network size

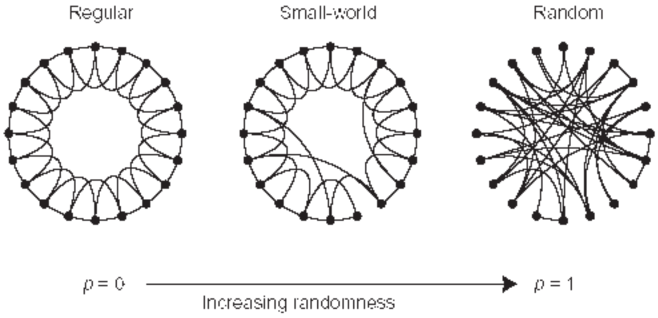


Figure 2.3: Random rewiring procedure for interpolating between a regular ring lattices and random networks, without altering the number of vertices or edges in the graph. The starting configuration is formed by a ring of n vertices, each connected to $2g$ nearest neighbors.

n [75][19]. As observed in [110], the small-world property results from the immediate drop in $L(p)$ as soon as p is slightly larger than zero. This is because the rewiring of links creates long-range edges (shortcuts) that connect otherwise distant nodes. The effect of the rewiring procedure is highly nonlinear on L, and not only affects the nearest neighbors' structure, but it also opens new shortest paths to the next-nearest neighbors and so on. Conversely, an edge redirected from a clustered neighborhood to another node has, at most, a linear effect on C. In other words, the transition from a linear to a logarithmic behavior in $L(p)$ is faster than the one associated with the clustering coefficient $C(p)$. This leads to the appearance of a region of small (but non-zero) values of p, where one has both small path lengths and high clustering.

Small world property have been widely founded in food web[80], the World Wide Web [4][96], power grid networks [110], words networks[101], dictionaries networks[70], transportation networks, biological networks [61], law networks[23], even though social networks [83] were the firsts to be analyzed. In fact, in the 1960s, an empirical study by Stanley Milgram[76] founds that every two people in the United States have an average acquaintanceship of six. Similar results are obtained through collaboration networks were actors are connected if they acted together, and scientists networks where exists a link whenever two authors publish a paper together[82].

Even though small-world model is considerably more adherent to real networks than random graphs, it has many limitations, for example:

- Small-world models do not follow the dynamics with which a real network evolves,
- The degree distribution of many real networks is not bell shaped, instead it is a

power law indicating the presence of hubs in the networks

The class of models that account for networks with degree distribution that deviates from Poisson is called scale-free and are presented in the following section.

2.2.3 Scale-Free Networks

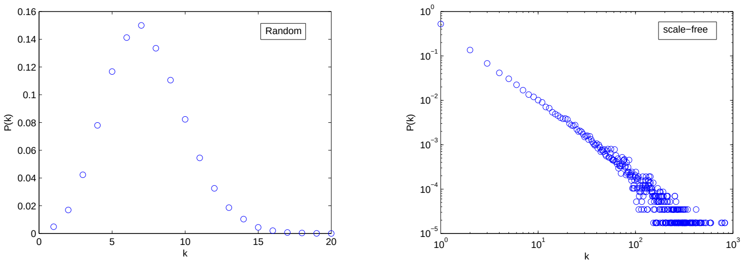
The models presented previously are devoted to construct networks that have small average shortest path L and high local clustering C, or equivalently, high E_{glob} and E_{loc} using Latora and Marchiori [67] framework metrics.

However, in many real world networks the degree distribution does not follow a bell curve (that for instance characterize the frequency of humans heights), but instead does follows a power law, i.e. $P(k) \sim c \cdot k^{-\gamma}$ where c is a constant and γ is a positive exponent that empirically varies between two and three. The reason why the exponent fits in that range is still unknown to network scientists and it remains an open question since the firsts discoveries on networks science. Having a $P(k)$ that has a decaying tail in the power law means that the vast majority of nodes have low degree and that there exist few nodes, the so-called *hubs*, that have an extremely high connectivity.

Even though one might expect a limited influence of hubs in the overall development and life of the networks because of the small cardinality, they play a fundamental role in the evolution, robustness and connectivity of the entire networks. For instance, in biology, hubs can represent genes that identify functional modules and whose removal can be the cause of specific diseases. Hubs are also very important in brain networks and represent functional areas that are anatomically highly connected with neighborhood. Their importance stems from the possibility of use the degree rank order in normal brain as an indicator of the first symptoms of diseases [3]. These special nodes are not captured by both random and small-world models [86][95] (see figure 2.4).

Power laws are not new in the literature. Pareto, back in 1900s, found that people's income is well approximated by a function that has long decaying tails. In other words, power laws guarantee that rare events, such as people with very high income, people that have many friends, popular web pages compared to the less popular ones, the most cited papers, or the earthquakes with high magnitude[74][13], have positive probability to happen.

Such networks have been named *scale-free* [7], because power-laws have the property of having the same functional form at all scales. In fact, power-laws are the only functional form $f(x)$ that remains unchanged, apart from a multiplicative factor, under a rescaling of the independent variable x, being the only solution to the equation



(a) (b)

Figure 2.4: Example of random (left) and scale-free (right) degree distributions on linear and log-log scale respectively. One of the networks whose node distribution linkage follows a bell-shape curve is the U.S. highway system[9].

$f(\alpha x) = \beta f(x)$. Power-laws have a particular role not only in complex system field but also in statistical physics because of their connections to phase transitions [99] and fractals [45].

When working with real networks, it may happen that the data have a rather strong intrinsic noise due to the finiteness of the sampling. Therefore, when the system size is small and the degree distribution $P(k)$ is heavy-tailed, it is sometimes advisable [86] to measure the cumulative degree distribution $P_{cum}(k) = \sum_{k'=k}^{\infty} P(k')$. Indeed, when summing up the original distribution $P(k)$, the statistical fluctuations generally present in the tails of the distribution will be smoothed. Consequently the exponent γ of $P(k) \sim k^{-\gamma}$ can be obtained from $P_{cum}(k)$ as one plus the slope of $P_{cum}(k)$ in a log-log plot, i.e., $\gamma = 1 + \gamma_{cum}$.

There are two types of scale-free models available in the literature: the first one that creates static scale-free networks and the second that creates evolving scale-free networks. The former is simply generated as a special case of random graphs with a given degree distribution. A model that belongs to this category is for instance the so-called *fitness model*[27]; it starts from n isolated nodes, and associates at every node i a fitness η_i, which is a real number taken from a fitness distribution $\rho(\eta)$. For each couple of nodes, i and j, a link is drawn with a probability $f(\eta_i, \eta_j)$, with f being a symmetric function of its arguments. The model generates power-law $P(k)$ for various fitness distributions and attaching rules, while it gives ER random graph if $f(\eta_i, \eta_j) = p$ for each i, j.

Conversely, in the evolving scale-free category, the growth process that determines the structural properties of the network is taken into account. The Barabási-Albert (BA)[15] network growth model was inspired from the formation of the World Wide

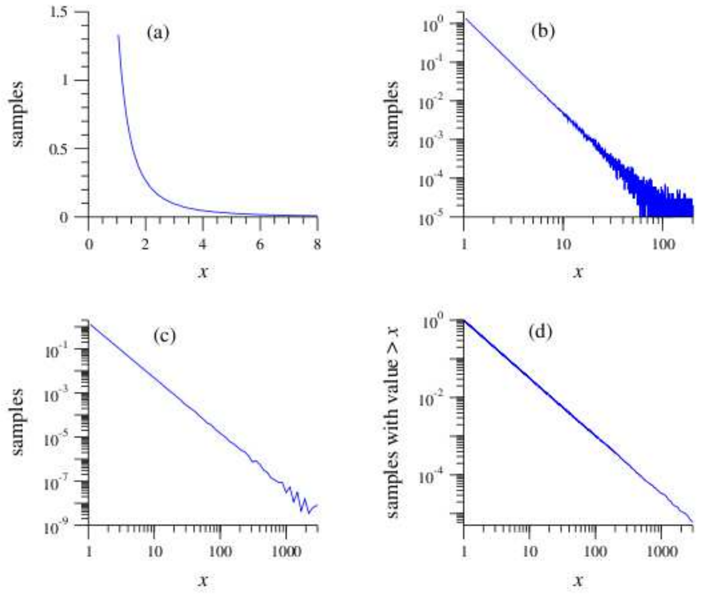


Figure 2.5: Example of noise reduction in the data calculating cumulative degree distribution $P_{cum}(x)$ [86]. A set of one million random numbers, power-law distributed, with scaling exponent $\gamma = 2.5$ is considered (a). Plot of the original data (b). Same histogram on logarithmic scale. Note the noise in the tail of the curve (c). A histogram with logarithmic binning (d). A cumulative histogram of the same data. This curve follows a power law, but the scaling exponent of the original curve can be obtained as the exponent of $P_{cum}(x)$ minus one, i.e., $\gamma = 2.5 - 1$.

Web and it is based on two basic ingredients: growth and preferential attachment. The basic idea is that in the WWW, web sites with high popularity (high degree) acquire new hyper links at higher rate than unpopular web sites (low-degree). More precisely, an undirected graph is constructed as follows: starting with m_0 isolated nodes, at each time step $t = 1, 2, 3, ..., n - m_0$ a new node j with $h \leq m_0$ links is added to the network. The probability that a link will connect j to an existing node i is linearly proportional to the actual degree of i:

$$\prod_{j \to i} = \frac{k_i}{\sum_l k_l}.$$

As every new node has h links, the network at time t will have $n = m_0 + t$ nodes and $m = h \cdot t$ links, corresponding to an average degree $\langle k \rangle = 2m$ for large times.

The Barabási-Albert model has been solved in the mean-field approximation and, exactly, by means of rate equation [64] and master equation approaches. In the limit $t \sim \infty$, the model produces a degree distribution $P(k) \sim k^{-\gamma}$, with an exponent $\gamma = 3$. On the contrary, the case of a growing network with a constant attachment probability $\prod_{j \to i} = 1/(m_0 + t - 1)$ produces a degree distribution $P(k) = e/h \cdot \exp(-k/h)$. This implies that the preferential attachment is an essential ingredient of the model.

The BA model shares many similarities with the model developed by Price [38] in 1976. The author's theory explains the power laws in citation networks he found out a decade before. In Price's model, the probability that a new published paper cites a previous one is taken to be proportional to $k_{in} + 1$, where k_{in} is the number of times that the paper has already been cited. Price's model is a reformulation, in terms of network growth, of a model developed by Simon in 1955 to explain the power laws appearing in a wide range of empirical data, as well as in the distribution of words in prose samples by their frequency of occurrence, or in the distributions of cities by population. Here, we simply mention that the model differs from the BA model in two main aspects: on one hand, it builds a directed graph and, on the other hand, the number of edges added with each new node is not a fixed quantity.

2.3 Longitudinal analysis

The previous models are classified into two classes of network generator [55]: editing and generative. The models that belong to the first class (for example the WS small world model), start with a given network and apply some modifications to the system in order to obtain specific features. Conversely, in the second class, the models produce a network from a small initial seed graph (like in evolving scale-free model). The goal of such models is to reproduce the structure that matches real data in prespecified properties, such as degree distribution [71], local clustering[14], assortativity by degree, average degree, etc.

In principle, scale-free networks presented above can be considered good candidates as network evolution models. However, this approach although simple and intuitive, can not resolve the full spectrum of interactions that thrives the whole complex system development. Because of that, scientists are now focused on tracking all the most important statistical quantities as the time runs using frequent static snapshots of the system [59] (even though studies [49] reveals that not all the tendencies can correctly be estimated using static snapshots only) or, in a more precise way, by using a complete fine level description of the system's evolution.

The longitudinal analysis is the consequence of the advent of the digitalization of everyday's actions that allowed to track in greater detail different types of mechanisms underlying networks (this trend started with on line social networks [49] but nowadays it is not limited to them only). This new framework of analysis was a fundamental key in order not only to precisely map system's evolution at individual level, but also to develop more accurate prediction models. In particular, they also describe the probabilistic tendencies of creation, maintenance, dissolution, and reconstitution of interpersonal ties during the evolution of social networks[93]. For instance, computational social scientists, thanks to longitudinal datasets, were able to classify social ties into five types of probabilistic tendencies, namely:

- *social exchange*, that corresponds to reciprocate a link that has been previously established,
- *balance*, that represents the social rule by which two person are likely to know each other if they have at least one common friend (also known as triadic closure),
- *distant connection*, as the name suggests, is a type of connection with a not close node,
- *collective action*, represent a connection to a node whose connectivity is beyond the average $\langle k \rangle$,
- *structural hole*, that corresponds to a link connecting two cluster not otherwise connected.

The previous works aim at investigating on the rules that govern interactions in a social setting. However, these represent only a limited list of all the features (such as like densification of edges and shrinking average distance) that complex systems such as social networks exhibit. Work of Lescovec et al. [69], Holme et al. [58] and Kumar et al. [65] are good starting points on that topic.

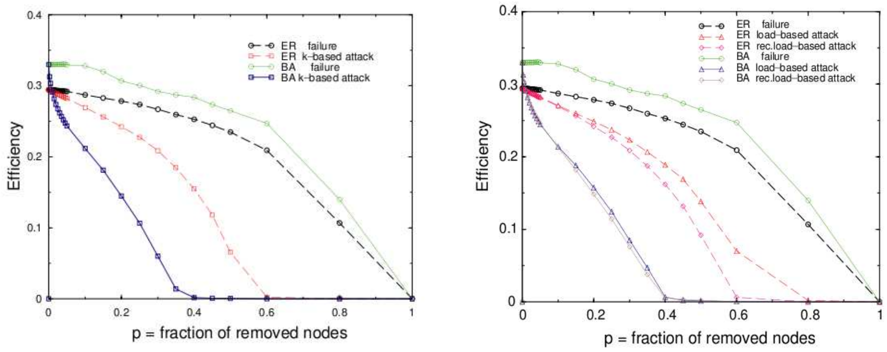
2.4 Critical analysis of complex networks

Robustness refers to the ability of a network to avoid malfunctioning when a fraction of its constituents is damaged. This is a very interesting topic, since it affects directly the efficiency of any process running on top of the network. In the literature, two are the main results concerning the resilience of networks: *random failures* and *intentional attacks*. The problem can be encountered in two different variants. The first one, referred to as *static robustness*, is meant as the act of removing nodes without the need of redistributing any quantity transported by the network. The second one, *dynamical robustness* refers to the case in which the dynamics of redistribution of flows is to be taken into account. An example can be found in the Internet at Autonomous System (AS) level where nodes can fail to routes packets they receive and its load has to be forwarded through alternative paths. Both types of robustness are similar in spirit, but while the first can be analytically treated (i.e. by using the tools of statistical physics such as percolation theory), the analytical treatment of the second case is harder and in almost all cases one has to rely on numerical simulations.

2.4.1 Static robustness

Static tolerance to errors (or random failures) [8] is understood as the ability of the system to maintain its connectivity properties after the random deletion of a fraction r of its nodes or edges. On the other hand, we refer to an attack when such a deletion process is targeted to a particular class of nodes or edges, for instance to the highly connected nodes, the most central nodes, the nodes whose deletion cause the bigger drop in the global efficiency, the bridge nodes or edges, etc.

Crucitti et al. [33] have conducted some numerical simulations to test how random networks and Barabási-Albert (BA) scale-free model are robust to random failure and intentional attacks. The importance of a node, that is at the basis of attacks simulations, were determined by the following three criteria: the node degree, the *betweenness centrality* (that is the number of shortest paths over all pairs of nodes of the network that pass through a node) evaluated before any removal is performed and the recalculated betweenness centrality, the same quantity as the previous one except that shortest paths were recalculated after a node removal. As showed in figure 2.6 (a), the BA model shows highly different behavior with respect to attacks and random errors: by removing approximately 15% of the nodes in a targeted way, the network



(a)

(b)

Figure 2.6: Random failures and attacks considering random networks (ER) and scale-free network (Barabási-Albert). The plots show how global efficiency E_{glob} changes as a function of removed nodes. Attacks are based on the degree (left panel) or on betweenness and betweenness recalculated (right panel).

efficiency is reduced to about half of the initial value ($E_{glob} = 0.33$) and it is sufficient to remove 35% of the nodes in order to destroy completely the system; on the contrary, when nodes are removed in a random way, the network efficiency decreases very slowly and the system maintains a considerable efficiency ($E_{glob} = 0.15$) even though the vast majority of the nodes were removed ($p = 80\%$).

This behavior is rooted in the heterogeneity of the scale-free model where few nodes are responsible for the interconnectedness of the network: their (targeted) removal causes a rapid drop in the capability of communicating and exchanging information of the system. As far as the ER graph is concerned, differences of tolerance to attacks and to errors are much less pronounced. In this case, in fact, there is not a substantial variability in the degree: the removal of a node in a targeted or in a random way produces similar, though not equal, behaviors.

Figure 2.6(b) shows the global efficiency for the Barabási-Albert and random model as a function of the number of node removed. This time the attacks were based on the betweenness centrality (recalculated or not). The curves concerning the recalculated-load-based attacks for the BA model do not differ substantially from those relating to degree-based attacks of figure 2.6(a). Larger differences are visible in the ER graph, for which the recalculated-load-based attack causes a greater amount of damage than the degree-based attack. This is because in the BA model many nodes with the highest load are also those with the highest degree, while in the ER model there is not a perfect parallel among load and degree. A different tolerance to recalculated and non-recalculated-load-based attacks is also evident for the ER model: it means that load redistributes over the networks after removal, i.e., when an attack is performed, shortest paths that passed through the removed nodes are not redistributed in a uniform way and therefore nodes with low betweenness may become those that carry the highest load.

Besides these numerical simulations, a series of analytical approaches to study tolerance to errors and attacks in complex networks have been proposed [30].

As stated at the beginning of this section, the main idea is that random failure can be analytically treated by using *percolation theory* [100][18]. A standard percolation process can be, in general, of two types: site or bond. Site percolation on a given graph means that the vertices are empty with a given probability f (or occupied with a probability $1 - f$), while bond percolation refers to the existence or not of an edge between two arbitrarily chosen nodes. Once the random deletion (or placement) of nodes or edges is done, several quantities allow the characterization of the network properties. One usually looks at the existence and size of the giant component (LCC) as a function of f, and at the average size and fluctuations in the size of finite components. In such a way is possible to define a critical probability f_c below which the network percolates, and a set of critical exponents characterizing the phase transition.

2.4.2 Dynamical robustness

In the previous section, we have focused on static properties of a network as a function of the fraction nodes' removal. In particular, we presented two types of deletions: random and targeted.

In real world systems, however, there is another important dimension to add to the problem: it refers to modeling the dynamics of flows of the physical quantities of interest over the network. A full characterization of a network cannot be fully accounted for without considering the interplay between structural and dynamical aspects. When it comes to modeling the dynamics, the situation is very complicate, since the components of a network may have different transmission capacities, and the load is often highly variable both in space and in time.

Fluctuations due to external factors can have much more devastating consequences when the dynamics of flows of physical quantities in the network is properly accounted for.

Avalanche of breakdowns over the network is a serious threat when node and links are sensitive to overloading. In a power transmission grid, for instance, each node (power station) deals with a load of power. The removal of nodes, caused either by random breakdown or intentional attacks, changes the balance of flows and leads to a global redistribution of loads over the entire network that could not be, in some cases, tolerated and might trigger a cascade of overload failures, disabling the majority of facilities. The largest blackouts as the time of writing, were in India in 2012, Indonesia in 2005 and Brazil were 670, 100 and 97 millions of people respectively were affected. These, and other similar examples, constituted the motivation for the study of how the extent and the dynamics of the avalanches are dependent on the network structure.

In Ref. [81] the fiber bundle model on scale-free networks is introduced as a conceptual framework useful to describe cascading failures. In this model, the system is subjected to an external pressure (load). The behavior of several network quantities indicates that the system exhibits a sort of critical point which depends on both the architecture of the underlying network and the heterogeneity of nodes' capacity. More important, the results point out that, in order to prevent the breakdown of scale-free networks, it is necessary to find an optimal criterion that takes into account two factors: the robustness of the system itself under repeated failures, and the possibility of knowing in advance that the collapse of the system is approaching (prediction).

2.5 Spreading processes

2.5.1 Epidemic spreading

Epidemic spreading deals with the modeling of the spread of a particular infectious disease in a population, with the aim of reproducing the actual dynamics of the disease and designing the strategies to control and possibly eradicate the infection. Conversely, the aim of rumor spreading is to spread the "rumors" as fast and efficiently as possible, not to prevent it from spreading. Practical examples are the design of protocols for data dissemination on the Internet, or strategies of marketing campaigns. The inclusion of complex topologies in standard epidemic and rumors models radically changes the results previously established for random graphs and regular lattices.

The study of epidemiological models is a subject of great interest that has attracted the attention of epidemiologists since long time ago. Mathematical epidemiology has grown exponentially since the middle of the 20th century, so that a tremendous variety of models has now been formulated, mathematically analyzed, and applied to field research [56]. Epidemiology modeling has been used in planning, implementing and evaluating various prevention, therapy and control programs.

At the same time, physicists became interested in these kinds of models when it was pointed out that epidemiological processes could be regarded as percolation like processes [52]. More recently, starting with the works by Pastor-Satorras and Vespignani [89], there has been a burst of activity on understanding the effects of the network topology on the rate and patterns of disease spread. The two most important models that describe disease spreading through a population by contacts between infected and healthy individuals are: the susceptible-infected-removed (SIR) and the susceptible-infected-susceptible (SIS) model.

The *SIR model* describes diseases resulting in the immunization or death of infected individuals, and assumes that each individual can be in one of three possible states, susceptible (denoted by S), infected (I), or removed (R). Susceptible individuals are healthy persons that can catch the disease, if exposed to infected individuals. Once an individual catches the infection, it moves into the infected (and infective) class, and then, after some time, into the removed class (becomes immune). The model is based on two parameters, the transmission rate λ, and the recovery rate μ.

The *SIS model*, a more realistic disease model, considers only two states: susceptible (S) and infected (I). The former are the individuals that catch the infection, move into the infected/infectious state and become again susceptible after a period of time in which they recover. At the end, they are again exposed to the epidemic. Scientists want to find a critical value σ_c (epidemic transition) of the ratio $\sigma = \lambda/\mu$ such that, for $\sigma < \sigma_c$ no infinite epidemic is possible, while for $\sigma > \sigma_c$ an infinite epidemic occurs with a finite probability.

Authors have observed that in random networks, the infection spreads to all the nodes if λ exceeds a critic epidemic threshold, otherwise it remains confined in limited zones of the network. In scale-free networks, when $n \to \infty$, there is no epidemic threshold, the epidemic becomes pandemic and spreads to all over the network, for all the propagation ratio. At the same time, in finite size networks epidemic threshold must exist otherwise all biological scale-free networks should be died. It is well known that the topology of the scale-free networks helps the spreading process, but at the same time, helps us to cure the system in a more effective way. In fact, it was shown that we can obtain a better control over a disease by start treating the most central nodes in the network instead of treating them randomly. This finding is directly correlated to the studies of network robustness (section 2.4) where intentional attacks are used to eradicate the infection/disease.

2.5.2 Rumors spreading

As seen before, the inclusion of complex topologies in standard epidemic models radically changes the results previously established for random and regular graphs and helped scientists, for example, to improve patients' treatments or control outbreaks. Nevertheless, in a number of important technological and commercial applications, it is desirable to achieve the opposite outcome. Instead of preventing an outbreak, we want to spread the epidemic as efficiently as possible.

Important examples of such applications are epidemic (or rumor-based) protocols for data dissemination and resource discovery on the Internet [107], as well as marketing campaigns using rumor-like strategies (viral marketing). The above applications and their dynamics have passed almost unnoticed to the physics community working on complex networks even though computer scientists and sociologists [39] have extensively studied them. The problem here consists of designing an epidemic algorithm in such a way that the dissemination of data or information from any node of a network reaches the largest as possible number of remaining nodes. The main difference with respect to epidemic models is that scientist are free to design the rules of the dynamics in order to reach the desired result, instead of having to model an existing process. Furthermore, in a number of applications, such as peer-to-peer file sharing systems and grid computing built on top of the Internet, the connectivity distribution of the nodes can also be changed dynamically in order to maximize the performance of such protocols. The standard rumor model is the so-called *DK model*, proposed several decades ago by Daley and Kendal [35]. The basic DK rumor model is defined as follows. Each of the n elements of the network can be in one of three possible states. Following the original terminology, these three classes correspond to ignorant (denoted by I), spreader (S) and stifler (R) nodes. Ignorants are those individuals who have not heard the rumor and hence they are susceptible to be informed.

The second class comprises active individuals that are spreading the rumor. In the end, stiflers are those who know the rumor but that are no longer spreading it. The spreading take place in a social network (virtual or real) by pair-wise contacts between spreaders and other people. In case a spreader meets an ignorant, the latter become a spreader whereas he meets another spreader, this one might decide not to tell the rumor anymore, therefore becoming a stifler.

Chapter 3

Multidimensional analysis of networks

Real world dynamical complex networks are non linear systems. This means that the full set of elements that interact pairwise (even in a trivial way) will result in a behavior that is often unpredictable. Moreover, an aspect frequently neglected is that these systems evolve along many informational axes. The two most important and not completely explored are *spatial* and *temporal.* The first one is crucial because all networks such as protein-to-protein networks, brain network[41], transportation networks[54][34], social networks[24], power grids[110], the Internet at Autonomous System (AS), companies network [25], etc are all embedded in Euclidean space, and most interestingly, the space variable itself constraints the natural evolution. This work is also justified by the fact that the space dimension is frequently underestimated in the studies of complex systems. Conversely, when added as an evaluation parameter, could help to characterize how it triggers environment changes. Moreover, the second axis though which a network could change its structure (expand, shrink or die) is time. Different models of network growth can result in faster or slower time evolution of the whole system, so understanding how to accomplish the best connection pattern is crucial in the context of complex systems. The mix of these two dimensions, time and space, defines completely a network evolution (see figure 3.1).

In order to understand how these dimensions influence the complex systems as a whole, it is fundamental to focus on them singularly. The intent of our novel multidimensional analysis is to fix one of the two dimensions alternatively, and investigate how the variation of the other drives changes in the network behavior (as is usually done with experiments that involve multi parameters functions).

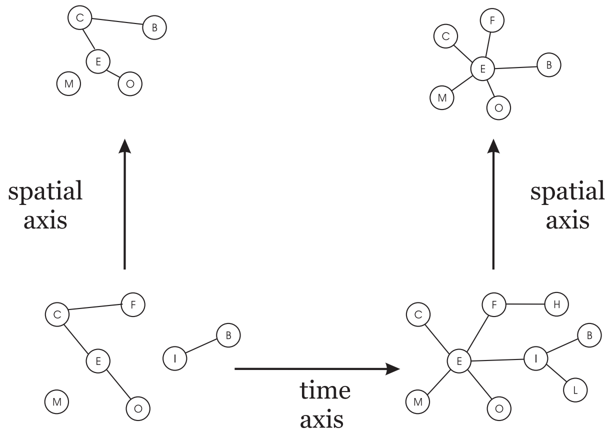


Figure 3.1: High level overview of multidimensional analysis on complex networks. The graph at the bottom of the panel evolves over time. The leftmost graph represent an hypothetical snapshot of a graph at time t_0 whereas on the rightmost the one at mature state. At every time step, the graph could also evolve along spatial axes (x and y, vertically in this sketch) by tuning the fuzziness f parameter that indicates the level of abstraction.

By keeping fixed the time dimension, and varying nodes' Euclidean coordinates, we focus our attention to spatial characterization of networks shedding light on how this can alter statistical measures of the graphs under study. In particular, we defined spatial characterization in terms of abstraction levels: we call this novel framework *telescopic* as it recalls the ability of magnifying and reducing lens. The parameter that accomplish for the capability of distinguish elements of the graph is the *fuzziness*. The more a graph is close to the point of view (low fuzziness), the more precise the connectivity and nodes will be. Conversely, the far it is the more imprecise and unclear the structure will become (high fuzziness). The family of graphs calculated at different resolution granularities forms the so-called telescopic spectrum. To the best of our knowledge, this is the first work that attempts to study spatial network in this way.

Through this line of research we were able to verify that statistical quantities depends strongly on the granularity (i.e., the detail level) with which a system is described and on the class of networks considered (for instance differences were detected by examining exponential and small world scale-free networks). We made simulations using two types of complex networks: rapid mass transportation systems (such as subways and airline) and on line social-based networks. Chapter § 4 is devoted to formalize this new analysis and reports on results obtained.

The second informative axis that our multidimensional analysis deals with is the temporal one. We fixed nodes' coordinates and study the network evolution as a function of time. In particular, our goal is to detect new instincts that trigger on line social systems utilization and spread the adoption of novel communities. To do so, we need to make assumptions on individuals' evolution on a social system. The order with which the nodes and edges are added into the network and the strength with which some nodes could acquire new connections (compared to the average) defines different temporal evolutions.

The most simple and straightforward growth model proposed was the one that randomly choose the edges that will join the network. Although many studies demonstrated that it is far from being realistic, this model still remains useful as a baseline on the experiments. Furthermore, we employed other classical growth models that are based on preferential attachment and on triadic closure respectively [6]. The first one (aristocratic rule) proposed as a natural model of scale-free networks, is characterized by older nodes that are more likely to acquire new links compared to the new ones. The so-called triadic closure (social rule) approach is based on an empirical observation, typically present in social networks, by which two friends of a person are more likely to know each other compared to two randomly chosen persons. In other than social networks, triadic closure is still present and forms the basis of local clustering.

The identification of new important instincts is accomplished by using special nodes, called "*sirens*", that help to spread users' adoption of social systems. Their importance stems from the ability of attract new links and consequently to construct an efficient connection pattern.

The use of additional nodes implies to consider also the dynamics of the sirens. We assumed that is independent of normal nodes because the true mission of these nodes' classes should be different. For simplicity of treatment, we decided to use similar rules to that of normal nodes, namely random, aristocratic and social. The simulation process starts with empty networks and iteratively adds links to both sirens' and individuals' subnetworks until a time threshold is exceeded (for the sirens') and until the original network is obtained (for normal nodes).

We applied the previous network growth principles to two important on line communities, VirtualTourist and Communities, and we verified the effectiveness of using these special nodes as a mean of increasing individuals' network engagement. To the best of our knowledge, this is the first attempt to quantitatively explore and formalize the spreading in the adoption in such a way. Details on this novel approach and reports on results are described in chapter § 5.

Chapter 4

The space dimension

As introduced in the previous chapter, multidimensional analysis consist of studying how networks evolve along the two most important informative axes, time and space. In order to attain this goal, we fixed one axis at time and see how the other one influence the overall behavior of the system.

This chapter is devoted to depict how space influences local and global properties of networks (so without considering growth over time) and in particular we study an important but far underestimated problem connected to the spatial property of many networks: the level of detail. Network modeling abstracts from a real system, and such abstraction can be modeled in several ways. Upside down, some real systems are, for various reasons, just too difficult to determine in a totally precise way, and so an apparently correct network modeling could just be a bad approximation.

Because of that, our intent was to propose a new spatial analysis that is able to arbitrarily model a network under different levels of abstraction by spatial modifications of nodes' coordinates and by reconstructing the network connectivity at a specific abstraction values. Doing so, we will be able to study, within the same framework, the effects of abstraction (space informative axis) in the study of complex networks.

In particular, we investigate the variations of the statistical properties not only when the network detail is high (the so-called "micro" view) or low ("macro" view), but also in between these two extremes. The final questions that we would able to answer are the following: Are the results of the static analysis that network scientists proposed in the last years dependent on specific detail levels? Which property is stable and which is not? In other words, what are the properties that are safe to consider when abstracting networks? Which topological structures better preserves system attributes?

By stability, in our framework, we mean that the variation of properties' values in the spectrum is small compared to the values reported at the two extremes (i.e. when the detail level is finer or coarser). Finding that many properties are unstable for a network during the abstraction process is clear evidence that a single detail level analysis could suffer from incompleteness and results will be consequently dependent on the granularity selected.

The algorithm (that we call *telescopic*) is inspired from the ability of human eyes to distinguish two points when placed at some distance from an observer. In this metaphor, the observation object is a graph and the points are nodes. The networks reconstruction is accomplished by tuning the distance parameter f (called equivalently fuzziness, detail or granularity level) in order to virtually place a graph far or close to a fixed point of view. Small f ($f \to 0$) yields clear networks with finer detail level, while big f ($f \to 1$) results in obfuscated networks, reassembling the abstraction process (see figure 4.1).

Our algorithm handles weighted and undirected graphs. It uses latitude and longitude nodes' coordinates and links information gathered from many databases that recently became freely accessible thanks to institutions, governments and companies that understood the benefit of open data.

We applied our framework to a number of networks, both real world networks (such as rapid transportation systems like subways, airline, and social-based networks) and synthetic. Indeed, we show how this novel analysis could provide great insights on what changes a network modeling bears with respect to the real system, and correspondingly on what part of network analysis, in our view, is unsafe under certain modelings.

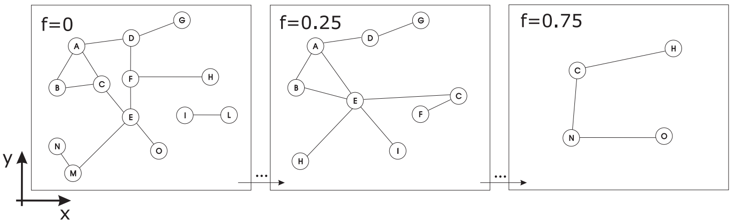


Figure 4.1: Example of multidimensional analysis by keeping fixed the time dimension and increasing (abstraction) fuzziness f. When $f = 0$, no abstraction is applied whereas at increasing values of f, the network will be more obfuscated and the structure will be simpler. In the extreme situation when f is maximum, $f = 1$ (not displayed in the figure), the original network will be collapsed into a one node graph.

4.1 Related work

Telescopic analysis is a novel technique of network analysis and because of that literature on this argument is lacking. The only topics that might have some analogies with our spatial abstraction are box covering and coarse graining of complex networks. *Coarse graining*, is the process that reduces the size of a network by preserving the most representative properties at the cost of throwing away some finer details of the system. Coarse graining is useful because of the computational impracticability of analyzing systems with a huge number of elements and, sometimes, from a statistical point of view, it is not even required to completely map an entire network with fine detail level [63].

One of the first method proposed by Song et al.[97] to coarse grain a network is the box-counting technique. The underlying ideas go back to the work about fractals and self-similarity under a renormalization procedure [72][106]. Since fractal's structure is similar no matter what length scale you choose, box-counting try to group together system units into boxes whose dimension determine the length scale at which the system is observed. The box covering procedure involves many steps of merging nodes until the network will collapse into one single vertex.

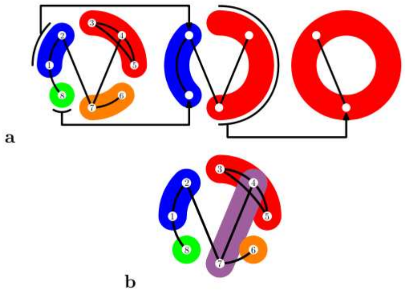


Figure 4.2: Two different covering examples of a graph with $l = 2$ [97].

For example, in figure 4.2 are represented two covering instances of a graph with eight nodes. Each box contains vertices whose distance (defined as shortest path length) is smaller than $l = 2$. The number of boxes N_B and the distance l are related as follow:

$$N_B \propto l^{-d}$$

with d the fractal dimension of the system. Box-counting and telescopic analyses share some common characteristics since they both group together nodes with different granularities. However, our analysis differs in the following points:

- The telescopic approach considers real nodes' Euclidean positions whereas box-counting technique uses only topological structure of the networks, throwing away useful information that come with spatial dimension of the vertices.

- In box counting, the number of boxes N_B varies according to length and fractal dimension. Vice versa, in the telescopic approach, the number of nodes belonging to boxes is not bounded: it depends on the spatial distribution of the nodes on the plane and on the fuzziness value. The maximum number of boxes is upper bounded and is inversely proportional to the fuzziness value (this concept will be extensively described in section § 4.2).

- Box covering and telescopic analysis differ in the way they consider input and output graphs. In the former, output and input graphs are the same, in the sense that the input graph corresponds to the output of the previous step. Conversely, in the latter, the same graph is provided as input but different abstraction parameters will be applied at every step.

Since Song's approach considered systems that are not embedded in Euclidean space, Kim [63] proposed a modification of the previous coarse graining process that deals with spatial networks. In particular, he studied the brain network formed by cubic cells (voxel) [3] linked by edges representing correlation in the activity of two partition of the brain. Indeed, he assessed how network properties like degree exponent, the clustering property, the assortativity and the hierarchical structure varies by repeatedly applying coarse graining.

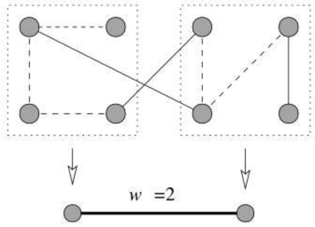


Figure 4.3: Spatial coarse graining method for a two voxels formed by 2×2 nodes each. Two voxels will be connected in the output graph if there exist at least one edge between the two boxes. The number of crossing edges will be expressed by the edge weight.

The coarse graining process is exemplified in the figure 4.3, in which 2×2 neighbor square boxes are merged together and the number of edge crossing the boxes is reflected in the edge weight of the resulting graph. Kim's work represents a spatial analysis of a complex network. However, his work is strongly different from ours in the following points:

- In the telescopic approach, the box dimension in the Euclidean space is not fixed and it is correlated to the spatial parameter f. This means that the number of nodes in a single box is variable and depends on the spatial distribution of the vertices across the plane.
- Our approach consider weights as real distances between vertexes whereas in Kim's, weights represent the number of crossing edges between voxels.
- We do not remove edges to avoid the creation of complete networks as Kim did.
- The overall functional behavior of the (brain) network that Kim consider in his paper is fundamentally different from ours. In fact, he do not consider actual path of voxels through which biochemical signal transfers.
- We repeatedly coarse grain networks by starting from the same graph but using different fuzziness values indicating the level of abstraction applied. Conversely, Kim's approach is similar to the original work of Song et al.[97] in which the output's graph at step t is used as input at the $t + 1$ iteration.

4.2 Telescopic Algorithm

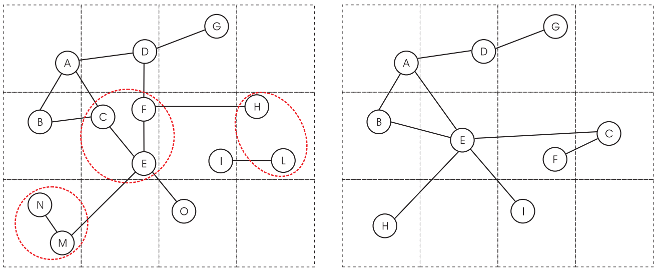
The aim of this section is (i) to describe our new analysis, which is called *telescopic*, that is capable of abstracting networks at various granularities and (ii) to assess whether statistical properties are affected by the multidimensional network analysis itself. We start by describing how the telescopic algorithm works, the phases that is composed by and subsequently present the dataset we used, describing statistical attributes of those networks.

Telescopic analysis is a technique that resemble from resolution power of human eyes, i.e., the ability to distinguish two points when placed at some *distance* from an observer. In this way, the distance corresponds to the level of fuzziness perceived by an observer, the more an object is far away from the viewer the more obfuscated will be. The observation objects in our context are networks and nodes are points in the human eyes resolution power metaphor. For instance, networks that are close to a virtual observer are clearly distinguishable and therefore are characterized by a finer level of details. Conversely, in networks far away from the point of view, the nodes will be obfuscated and the overall structure will be simpler than the original one (abstraction process). In the rest of this chapter, we indistinctly use fuzziness, distance, details or resolution level as synonyms of granularity with which a network has been described.

The networks we consider are defined as weighted and undirected graphs $G = (V, E, C)$. For simplicity, we assume that the latitude and longitude coordinates of the nodes $(x_i,y_i)_{i=1,\cdots n}$ (see section § 2.1), are normalized in $[0, 1]$. Edge weights are real normalized distances between nodes.

More precisely, we define the telescopic function as $t : (G \times f) \to G'$ that takes a graph G and a value of fuzziness f as parameters and return an abstracted graph G'. In this way, by applying repeatedly the function t with different value of f we obtain an abstraction spectrum that is formed by a family of networks $\mathcal{F}_G = \{G_1, G_2, \dots, G_k\}$ where $G_i = t(G, f_i)$, k sets the spectrum resolution and f_i is the fuzziness value of the i^{th} step. A small value of f_i leads to clear view of networks and thus the resulting graphs G_i will have the finer detail level. Conversely, when f_i is big $(f \to 1)$ the view will be obfuscated and in the limit when $f_i = 1$ only one node will belong to the outcome network.

Network abstraction is accomplished by two distinct phases. The first one deals with creating nodes in G_i while the second defines the topological structure. Intuitively, nodes in G_i are the result of collapsing nodes in G that are close each other, hence not clearly distinguishable from an observer. The number of nodes that has to be collapsed obviously depends on f and on their spatial distribution on the plane.



(a) (b)

Figure 4.4: One-step application of the abstraction process to a small graph. (a) Original graph G. Red (dashed) circles identify the group of nodes that will be merged together. (b) Output graph G_i in which nodes c, f, e, h, l and n, m are collapsed into new nodes $e, c, h \in V_i$ respectively. Coordinates are the barycenter of collapsed nodes. Three edges are then removed because they connect the collapsed nodes: (n, m), (c, e), and (f, e).

More technically, the process by which nodes in G_i are created is based on placing a virtual grid on top of G (see figure 4.4). This grid is formed by a set of square boxes whose spatial dimensions corresponds to the fuzziness f (see figure and 4.5). Since we assumed that $0 \leq f \leq 1$ and coordinates $0 \leq x_i, y_i \leq 1$, the total number of square boxes will be $N_B = f^{-2}$.

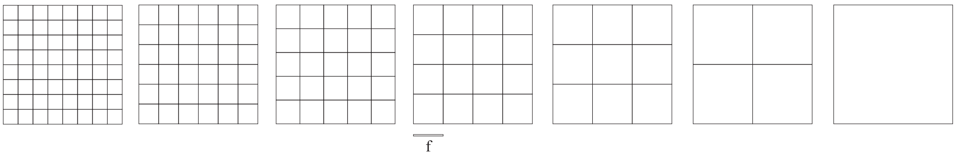


Figure 4.5: Example of grids applied on top of networks as a function of fuzziness. Leftmost grid has low fuzziness $f = 0.125$ whereas the rightmost has $f = 1$. The granularity of the spectrum in this example is equal to 7. In this Thesis, we only consider linear increase of f.

All nodes of G that belong to the same square cell are collapsed into a new node in G_i and new coordinates will be the barycenter of the collapsed nodes. The maximum number of nodes in G_i with fuzziness f is bounded to $|V_i|_f \leq f^{-2}$ (maximum one node per box). This procedure aims at grouping nodes that are far by almost f units (eventually $f\sqrt{2}$ if the two points are at the extremes of the diagonal). However, the limitation of this algorithm is that not all nodes that are close each other by almost f units will be collapsed. This circumstance occurs when the grid fall in between neighbors' nodes as figure 4.6(a) shows. For figuring out this issue, we applied a random grid shift that attenuates the bias introduced by grid displacement (see results in § 4.3) and take averaged results of the statistical properties considered.

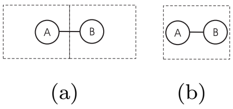


Figure 4.6: Grid displacement issue when the distance between two nodes is less than fuzziness value. Wrong (a) and correct (b) grid displacement.

In the second phase, once vertices of G_i are defined, we re-establish the network connectivity. Here we adopt the most straightforward rule that preserves network structure: if two clusters of collapsed nodes of G are connected by at least one path, then in G_i the two representative nodes will be connected. Let's define this concept in more detail using the notation presented above. Let Γ_i a set of nodes that belongs to box i and

$$g_{ij} = \{(k, m) \in E \mid k \in \Gamma_i,\ m \in \Gamma_j \text{ with } i \neq j\}$$

a set of edges whose source and target nodes belong to i and j box respectively. An edge $(u, v) \in E_i \Leftrightarrow |g_{uv}| > 0$.

Figure 4.7 shows an example of application of telescopic analysis to Boston and New York subway networks.

4.3 Experimental Results

In this section, we report our experimental analysis. The telescopic algorithm was implemented as a C module and used in a Python script. All the experiments were conducted on three Linux machines equipped with i5 Intel processors at 3.2 Ghz and 8Gb of RAM.

4.3.1 Datasets

We conducted experiments on several datasets: **rapid transportation networks**, **on line social networks** and **synthetic random networks**. We decided to consider subway networks because they are a fundamental element of mass transportation in urban areas and important means of cost reduction in transportation. Indeed, in the literature there exist some great network studies [94][32][66] that characterized the most important subway networks. The networks we used in our experiments are Paris Métro, Milan Metropolitana, New York and Boston Subways (see figure 4.8). Paris Métro, one of the densest and busiest networks in the world, has sixteen lines and the first line opened in 1900. It has 295 stations connected by 346 rail connections. Milan Metropolitana, the smallest subway network we will consider, but the biggest rapid transit system in Italy, opened in 1964, has 81 stations and 80 rail connections. New York subway is the most extensive rapid transportation system in the world by number of stations. It has 487 stations and 439 connections and opened in 1904. Finally, Boston subway consisting of 124 stations and 125 connections had the first subway line opened in the United States in 1897.

Each node stands for a station, edges for direct railway connection between stations. Networks are created collecting latitude and longitude coordinates about stations locations and converting them into x, y coordinates using Miller cylindrical projection [77] (Mercator projection might be another technique to use). We finally normalize them in such a way every couple $(x_i, y_i) \in [0, 1]$. We also consider the US airline transportation system in which nodes represent airports and edges are non-stop flights. The US airline network has 235 airports and 1296 non-stop flights. Datasets of transportation networks are freely available on the net.

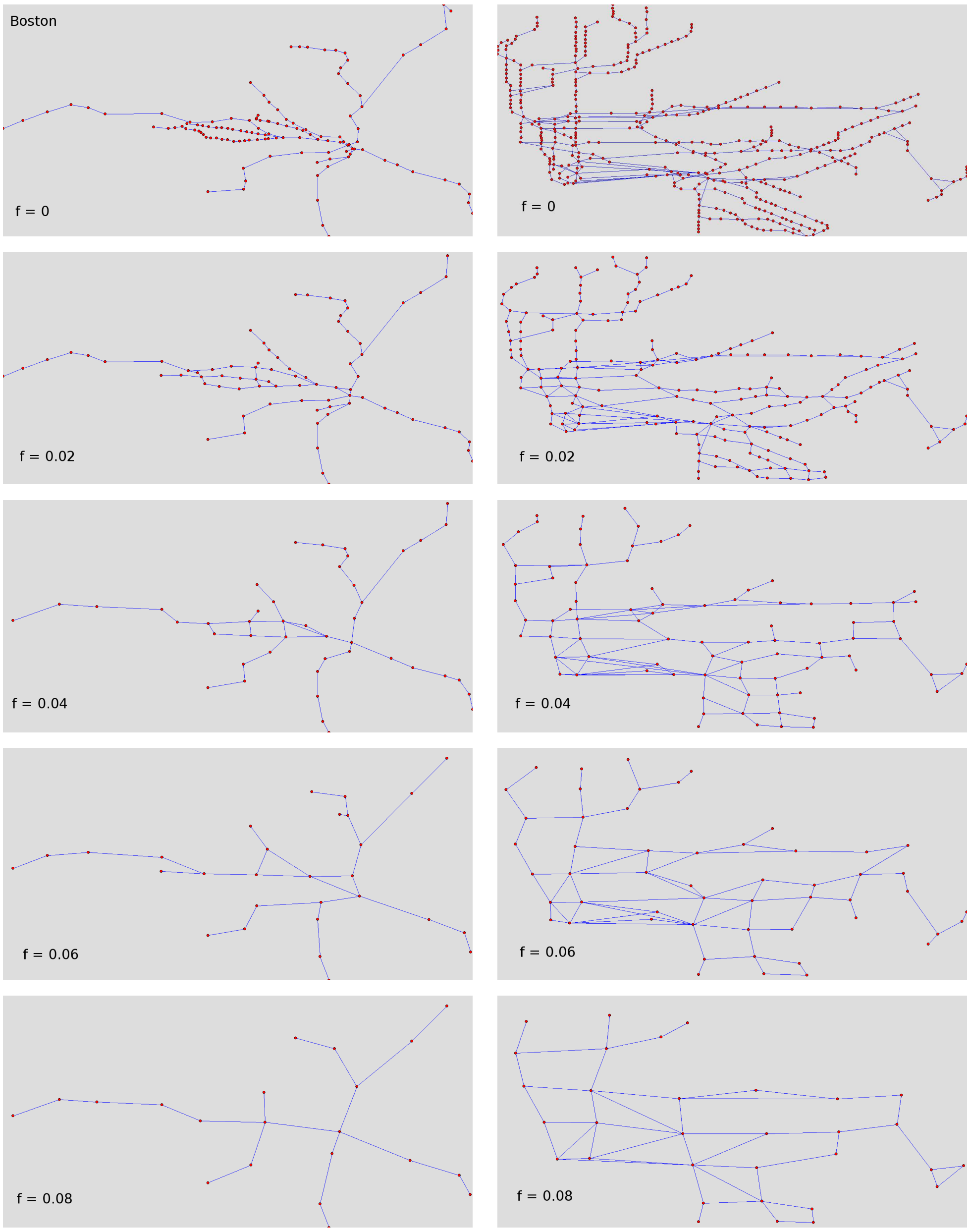
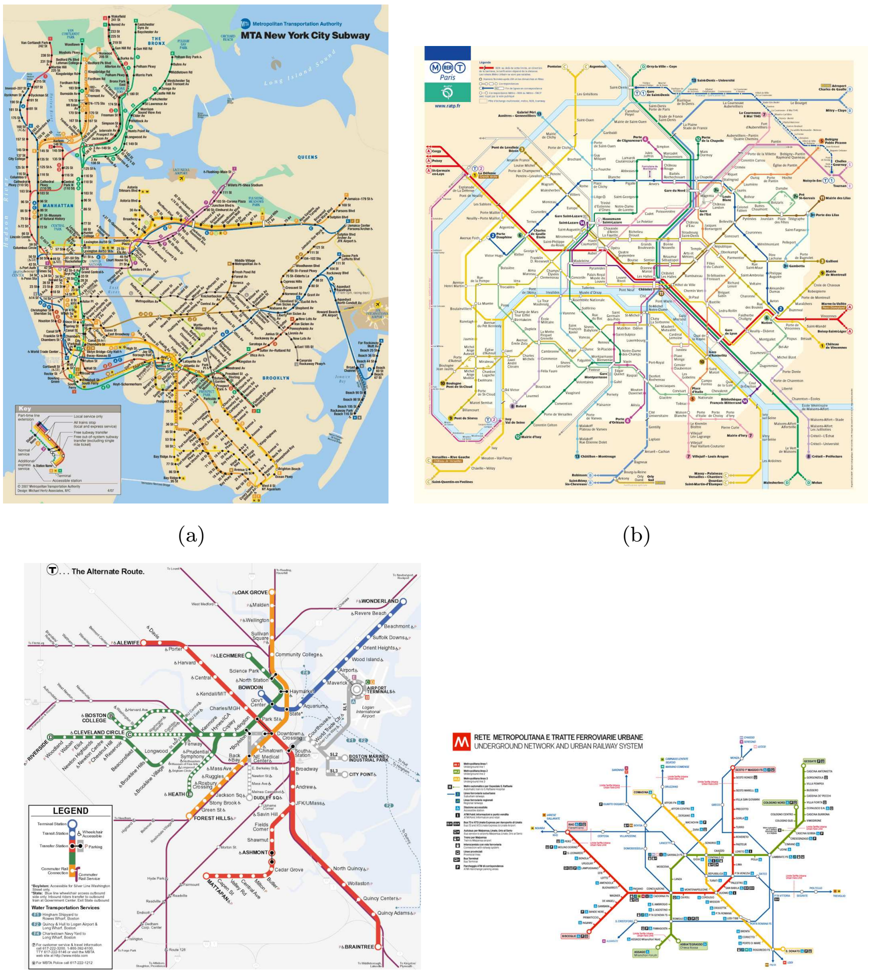


Figure 4.7: How Boston (leftmost panels) and New York (rightmost panels) subway networks vary according to spatial informative axis at increasing values of fuzziness. From the panel, it is clear that the spatial structure of the system remain unchanged in the first steps of the abstraction process.


MTA New York City Subway
Metropolitan Transportation Authority
Key
. . . The Alternate Route.
LEGEND
RETE METROPOLITANA E TRATTE FERROVIARIE URBANE
UNDERGROUND NETWORK AND URBAN RAILWAY SYSTEM

(a)

(b)

(c)

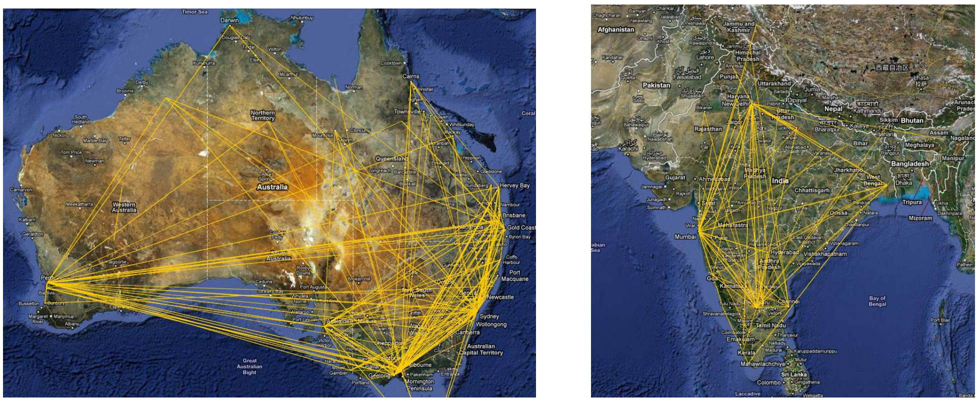
(d)

Figure 4.8: Maps of the subway networks we will consider in this Thesis: New York City (a), Paris (b), Boston (c) and Milan (d) ordered by number of stations.

To investigate the effect of this novel analysis to other than transportation networks, we also consider on line social networks. In particular, we use VirtualTourist[2] dataset. VirtualTourist (in the following abbreviated as VT) is an on-line tourist guide in which users share their travel experiences, suggest and review hotels, write comments and opinions on VT forums, find a place to visit, share photos and videos: it is a community of people that love traveling around the world. Users can meet new people and create a network of virtual friendships, like in many other social networks.

We explored the VirtualTourist social network through web crawling [44] all the publicly available profiles, and for each anonymized user we collected the following attributes: gender, birth date, subscription date and living location. We filter out users with empty location or unreliable fields (for example those whose format is not compliant).

Since VT locations span more than 150 countries, we decided to consider only those countries with the highest number of users such as Australia, the Netherlands, India, the United Kingdom and Italy and analyze them individually, leaving for future work the analysis of the whole worldwide dataset.

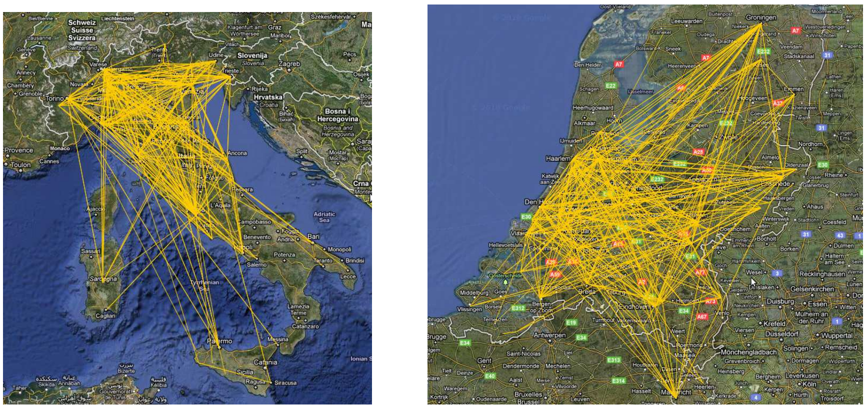


(a)

(b)

Figure 4.9: Example of city-based networks created from VirtualTourist on line community of Australia and India users. Lines (yellow) represent edges of the network connecting cities that share at least one friend.

Applying the telescopic algorithm to these networks may induce an unexpected increase in the number of collapsed nodes starting immediately at small values of f. This is caused by the presence of many users at the same location (for example when they live in big cities). In order to overcome this issue, we decided to transform these on line social networks into city-based networks, where nodes stands for cities



(a) (b)

Figure 4.10: Example of city-based networks created from VirtualTourist on line community of Italy and the Netherlands users. Lines (yellow) represent edges of the network connecting cities that share at least one friend.

(in which lives at least one VT user) and links express friendship relations between users of those cities. These networks now describe friendship relations at the level of cities instead of the users. The GPS coordinates of the cities were gathered from Geonames open source web service[1] and edge weights will be the Euclidean normalized distance between cities (see figure 4.9 and 4.10). As the time of gathering data, Italy network has 85 cities and 270 social ties, Australia has 76 cities and 183 links, the Netherlands has 106 cities and 340 edges, India has 46 cities and 81 edges and finally the United Kingdom has 446 cities and 1322 social ties. Both transportation and city-based networks are undirected because people can move either in both directions of the transportation line and friendships relations in VT are bidirectional. Table 4.1 reports statistics of the datasets we used in this section of the Thesis. We calculated the most important statistical properties such as the number of nodes n, edges m, maximum degree k_{max}, average degree $\langle k \rangle$, standard deviation of the degree σ_k, degree correlation ρ, diameters, local and global efficiencies and costs (see § 2.1.1 for the definitions). We indeed reported (table 4.2) on the same statistical quantities for randomized transportation and social-based networks.

Among these datasets, subway networks are not scale-free nor small-world because the diameter D_t do not scale as $\log(n)$, the average shortest path L is high, the clustering coefficient is low (like in random networks) and efficiency is also low (see table 4.1 and classification [11]). On average, these networks have low degree nodes, i.e. the majority of stations are not interchange points where user can switch to

	Boston	Milan	New York	Paris	US airline	Social IT	Social AU	Social NL	Social IN	Social UK
n	124	80	439	295	235	85	76	106	46	446
m	125	81	487	346	1296	270	183	340	81	1322
k_{max}	4	4	8	8	130	40	31	28	22	188
$\langle k \rangle$	2.02	2.03	2.22	2.35	11.03	6.35	4.82	6.42	3.52	5.93
σ_k	0.54	0.57	0.81	1.06	17.98	7.82	6.81	7.68	4.71	12.66
ρ	0.32	-0.08	0.14	-0.07	-0.36	-0.26	-0.43	-0.07	-0.44	-0.19
D_t	43	34	68	38	4	5	5	6	6	7
D_p	1.08	1.12	1.15	1.17	1.26	1.08	1.01	1.07	1.00	1.07
D_m	1.54	1.47	1.93	1.36	1.86	2.00	2.47	2.29	2.16	2.11
E^t_{glob}	0.11	0.14	0.07	0.11	0.46	0.41	0.41	0.36	0.44	0.34
E^m_{glob}	0.65	0.76	0.62	0.75	0.65	0.49	0.52	0.44	0.65	0.45
E^t_{loc}	$5.38\times^{-03}$	0.00e+00	3.17e-02	2.00e-02	6.97e-01	4.19e-01	4.50e-01	3.66e-01	2.64e-01	2.80e-01
E^m_{loc}	9.60e-05	0.00e+00	1.90e-03	8.83e-04	1.36e-01	7.62e-02	6.03e-02	8.09e-02	5.61e-02	4.81e-02
C_t	1.64e-02	2.56e-02	5.07e-03	7.98e-03	4.71e-02	7.56e-02	6.42e-02	6.11e-02	7.83e-02	1.33e-02
C_m	1.89e-03	3.19e-03	3.63e-04	7.96e-04	3.46e-02	5.53e-02	5.58e-02	5.43e-02	6.27e-02	1.15e-02
C_t/E_t	1.50e-01	1.80e-01	6.00e-02	7.00e-02	1.00e-01	1.80e-01	1.50e-01	1.60e-01	1.70e-01	3.00e-02
C_m/E_m	0.00e+00	0.00e+00	0.00e+00	0.00e+00	5.00e-02	1.10e-01	1.00e-01	1.20e-01	9.00e-02	2.00e-02

Table 4.1: Datasets statistics of subways, the US airline and city-based social networks: number of nodes n and edges m of the graphs, maximum degree k_{max} and average node degree $\langle k \rangle$, standard deviation of the degree σ_k, assortativity mixing by degree ρ, physical, topological and metrical diameter, global and local efficiency E_{glob}, E_{loc}, costs and C/E property (defined as the ratio between cost and global efficiency). Both *topological* and *metrical* versions are calculated of the latter three indicators.

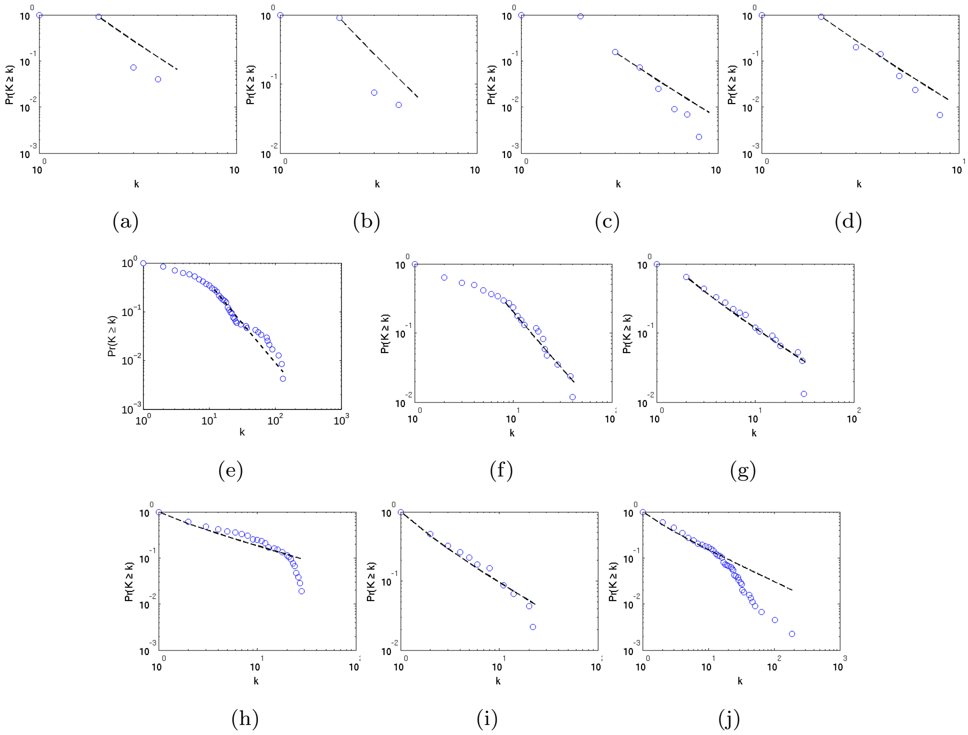
Rnd:	Boston	Milan	New York	Paris	US airline	Social IT	Social AU	Social NL	Social IN	Social UK
n	124	80	439	295	235	85	76	106	46	446
m	125	81	487	346	1296	270	183	340	81	1322
k_{max}	6.95	5.85	7.45	7.65	20.80	13.25	10.30	13.75	7.80	14.55
$\langle k \rangle$	2.02	2.03	2.22	2.35	11.03	6.35	4.82	6.42	3.52	5.93
σ_k	1.45	1.38	1.46	1.51	3.22	2.53	2.11	2.55	1.77	2.40
ρ	-0.05	-0.04	-0.01	-0.04	-0.00	-0.02	-0.01	-0.04	-0.03	-0.00
D_t	13.55	13.05	16.60	14.55	4.10	4.90	5.80	5.15	6.50	7.00
D_p	1.28	1.24	1.35	1.33	1.32	1.26	1.25	1.28	1.21	1.35
D_m	6.97	6.68	9.03	7.78	1.89	2.36	2.97	2.58	3.32	3.57
E^t_{glob}	0.14	0.16	0.12	0.14	0.43	0.43	0.39	0.41	0.38	0.30
E^m_{glob}	0.11	0.13	0.09	0.11	0.44	0.41	0.34	0.39	0.32	0.26
C_t	0.02	0.03	0.01	0.01	0.05	0.08	0.06	0.06	0.08	0.01
C_m	0.02	0.03	0.01	0.01	0.05	0.08	0.07	0.06	0.08	0.01
C_t/E_t	0.11	0.15	0.04	0.05	0.11	0.17	0.16	0.14	0.20	0.04
C_m/E_m	0.14	0.19	0.05	0.07	0.10	0.18	0.19	0.15	0.24	0.05

Table 4.2: Statistical properties of randomized transportation and city-based networks. The perturbation that is applied consists of placing nodes in random positions and rewiring edges randomly. n and m corresponds to the number of nodes and edges of the graph, k_{max}, $\langle k \rangle$ to the maximum and average node degree, σ_k to the standard deviation of the degree, ρ to the assortativity by degree, D_p, D_t and D_m to the physical, topological and metrical diameter respectively, E_t and E_m to the global efficiency (topological and metrical) and finally C_t and C_m to the topological and metrical costs. C/E property is the ratio between cost and efficiency (topological and metrical versions are considered).

other lines. The maximum degree is 4, or 8 for the biggest subways (they can not be considered *hubs* as in scale-free networks though) and they are assortative or eventually uncorrelated. However, by considering weighted version of the networks, we found that subways are very efficient both locally and globally ($E^m_{glob} > 0.65$). This observation is also confirmed by previous studies [66] [32] in which authors tested small world property on Boston and worldwide subways.

On the other hand, city-based and the US airline transportation networks have a different connection pattern. We found that two randomly nodes are connected by means of less than ten edges and the clustering coefficient is high compared to the randomized versions (see table 4.2). k_{max} and σ_k are high compared to subway networks and the degree distributions all displays long right tails (see figure 4.11 letter e to j) that is evidence for the presence of hubs. Indeed, high efficiency and low diameter are detected. On average, high degree nodes tend to be connected to low degree nodes (ρ is always negative) like in technological, neural and protein-to-protein interactions networks[84]. These systems could be classified as small world scale-free [31].

In addition to the transportation and city-based networks presented above, we also consider artificial random networks (null models). These systems are created by starting with the original network and randomize it through the iteration of same fundamental rewiring move that alters the topology or by applying perturbation on nodes' positions but keeping fixed the number of nodes and edges. We considered separately random perturbation on nodes position and connectivity in order to clearly distinguish two randomness effects on telescopic analysis. Moreover, we studied another important connectivity manipulation based on rewiring the edges in a way that the final topological structure results in a scale-free pattern (see section 4.4).



(a) (b) (c) (d)

(e) (f) (g)

(h) (i) (j)

Figure 4.11: The log-log plots of the cumulative degree distributions $P_{cum}(k)$ of subways (Boston, Milan, New York, Paris, a to d), the US airline (e) and city-based social networks (letter f to j) of Italy, Australia, The Netherlands, India and the United Kingdom. The distributions are characterized by exponents γ of $P(k) \sim k^{-\gamma}$ that is one plus the slope of $P_{cum}(k)$ (in a log-log plot), i.e. $\gamma = 1+\gamma_{cum}$. The coefficient is $\gamma = 3.5$ for subways networks, 2.6 for the US airline, 1.85 for Indian city-based social network, 1.68 for the United Kingdom, 2.61 for Italy, 1.94 for Australia and 1.61 for the Netherlands. The coefficients for subways might not be precise due to the small dimension of the networks.

4.3.2 Results

In this section, we report on results obtained by applying the telescopic framework to real world and artificial networks and we show how this novel approach based on modifications of spatial axis on complex systems is effectively a robust tool that precisely describes networks at different detail levels. A key feature of our study was to compare the results with null models in order to investigate the role played by topological structure and by nodes' positions (in conjunction and separately).

In all the experiments we made in this Thesis, we randomly shift 10^4 times the position of the square boxes to limit the bias in the grid displacement (the entire set of boxes will be shifted, not the single boxes individually, see section § 4.2) and we finally take averaged results.

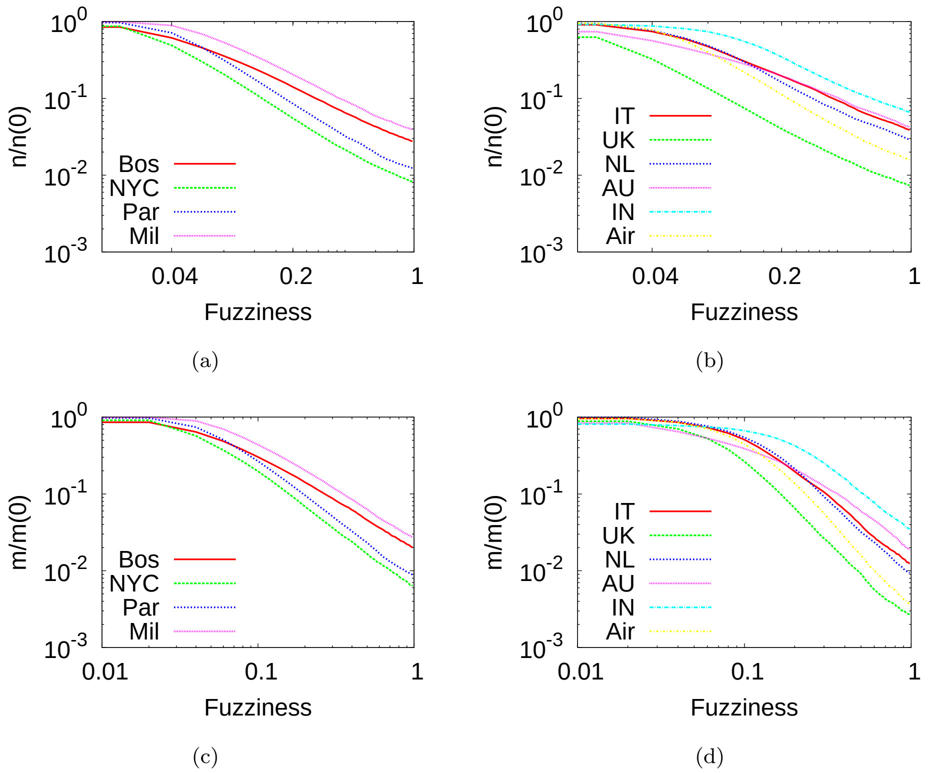
The telescopic process creates a family of graphs $\mathcal{F}_G = \{G_1, G_2, \ldots, G_k\}$ where $G_i = t(G, f_i)$, $i = 1, \ldots k$. We selected $k = 100$ as the granularity of the telescopic spectrum therefore the fuzziness will be increased (linearly) by 0.01 units at each step.

The telescopic process we are describing starts at low values of fuzziness (i.e., finer detail level) and by aggregating nodes together at every step, it generates networks that are more and more abstracted. This means that the networks given as input to the telescopic algorithm must have a high detail level. Since it is not possible to gather networks with arbitrary detail level, we decided to use as inputs the networks presented in section § 4.3.1. This choice not necessarily limits the challenges of the abstraction analysis we presented. Instead, we focused on the subset of the spectrum that starts as the same detail level of the networks used.

In the plots that follow, as a matter of clarity, we decided not to consider the most abstracted networks (at $f = 1$) since it contains only one node and consequently the metrics will get trivial results. Furthermore, some plots contain relative quantities. This means that the value obtained with fuzziness f, say v_f, will be divided by $v_{f=0}$, that is, the value obtained with no abstraction at all. This helps to depict the increase or decrease relative to the baseline.

Figure 4.12 shows how nodes and edges are merged together as a function of fuzziness f. It is interesting to note that the overall behavior of collapsing nodes and edges is similar over the same type of network (transportation and city-based networks). However, the rate with which vertices and links are merged depends on several factors such as the size of the system (instead of the history), the physical position and the structure of the network itself. Bigger networks, for instance New York or Paris have a faster merging rate.

Figure 4.13 shows how diameters metrics vary as a function of fuzziness f. The three versions of these statistical quantities accounts for three different characteristic



(a)

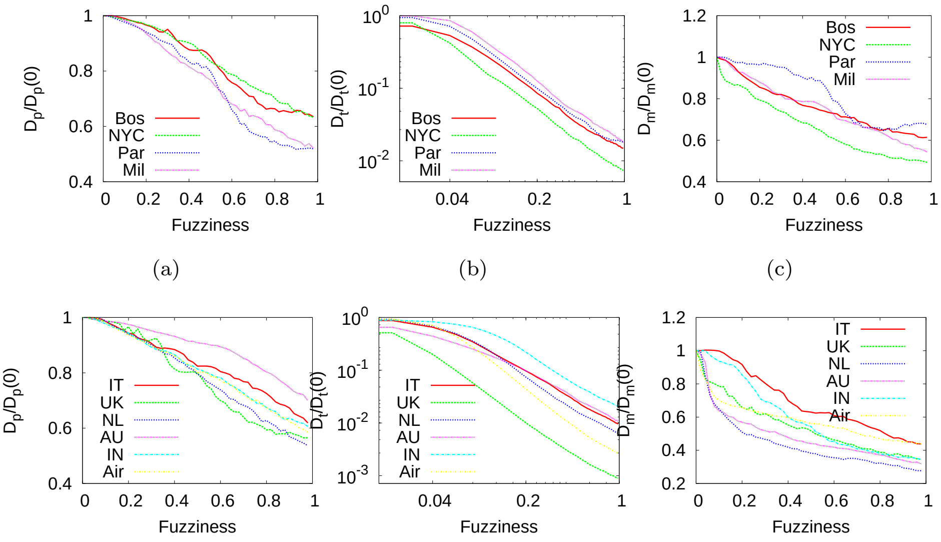
(b)

(c)

(d)

Figure 4.12: Number of collapsed nodes n and edges m as a function of f in log-log axes. The values are normalized by the baseline values $n(0)$ and $m(0)$ respectively, obtained at $f = 0$ (i.e., no abstraction applied). The leftmost panels refers to subway networks whereas the rightmost refers city-based and the US airline networks. The decrease of n and m is clearly exponential, even though the rate is influenced by many factors like network size and nodes' positions.

of the networks: (i) the maximum physical extension of nodes in a 1 × 1 unit square box, (ii) the maximum topological extension on the shortest path and (iii) the maximum metrical extension on the shortest weighted path. The first one, as expected, decreases linearly, mainly because of the linear increase of the fuzziness. The second one decreases exponentially and this is evidence that the telescopic process creates the right shortcuts links that decrease faster the diameter.



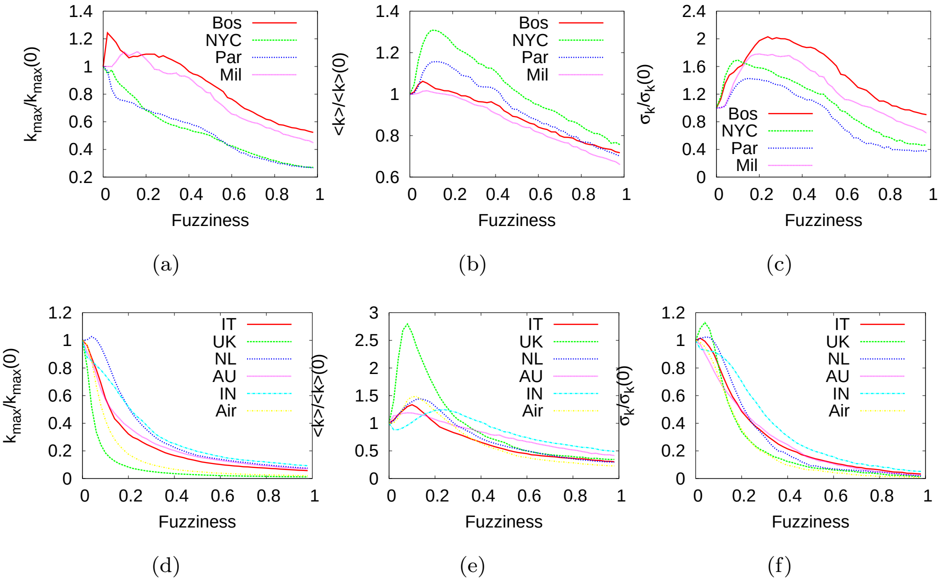
(a) (b) (c)

(d) (e) (f)

Figure 4.13: Effect of the telescopic abstraction for the physical D_p, topological D_t and metrical D_m diameter as a function of fuzziness f. All the values were normalized by the baseline values at $f = 0$ (i.e., no abstraction is applied). The top panels contain results of subways, the bottom ones of city-based and the US airline networks.

Figure 4.14 shows the effect of the telescopic abstraction on k_{max}, $\langle k \rangle$ and standard deviation of the degree. The explanation of the observed behavior is not so trivial even though a couple of observations can be made. First, we note a clear distinction on result between the two types of networks (first and second row). For instance, in subway networks, k_{max} decrease almost linearly and a joint observation that takes into account both $\langle k \rangle$ and σ_k suggests that telescopic analysis triggers an increase of the average degree (even though this is not necessarily evidence that in some part of the telescopic spectrum, the analysis produces networks with hubs). Conversely, in city based and the US airline networks (bottom panels), the effect of abstraction on the degree is more prominent. Where in subways the decrease on relative k_{max}

were almost linear, in these networks the rate with which maximum degree decrease is exponential. This is mostly caused by the presence of hubs that will be collapsed almost immediately as fuzziness f increases.

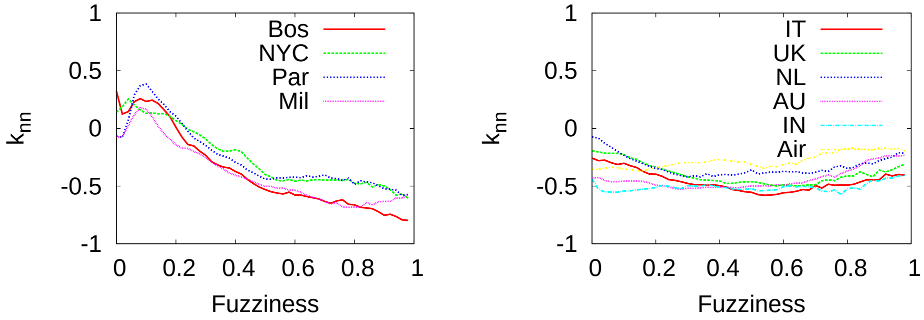


(a) (b) (c)

(d) (e) (f)

Figure 4.14: Effect of the telescopic analysis on the degree: maximum degree k_{max} (leftmost column), mean degree $\langle k \rangle$ (centermost column) and standard deviation σ_k (rightmost column) for subways (top panels), the US airline and social-based networks (bottom panels). All values were normalized relatively to the baseline value at $f = 0$ (where no abstraction is applied). The explanation of the results obtained is not so trivial. In general, the degree properties of the networks will be drastically modified as fuzziness increases. The degree tend to decrease linearly in subway networks whereas in airline and social-based the telescopic effect results in an exponential decrease.

Degree correlations k_{nn} on (unweighted) networks measure the level of interdependence between nodes. From table 4.1 we identified two different connection pattern as we consider subways or city-based and the US airline. The first class of networks is neutral whereas the second one is negatively correlated (that is, nodes with high degree link to small degree nodes). Figure 4.15 summarize the degree correlations in the telescopic spectrum. We note that the abstraction process yields disassortative correlation at high values of fuzziness regardless of the network considered. As a consequence, the initial topological structure of subway networks will be significantly changed toward a completely different configuration whereas city based and airline remain relatively stable in the entire spectrum (at least they remain disassortative).



(a)

(b)

Figure 4.15: Impact of the telescopic analysis on degree correlations ρ as a function of f for subway networks (left panel), the US airline and city-based networks (right panel). It is worth noting that the telescopic process yields disassortative networks regardless of the network. This means that in subways, the topological structure will be drastically changed whereas in the other networks the degree correlation tends to remain stable (at least will have the same sign).

Figure 4.16 shows how global efficiency E_{glob} is influenced by the granularity level with which a network is described. In particular, metrical and topological versions are considered. One of the aims of this study is to verify whether the detail scale with which networks are described affects network efficiency. Different observations can be made for topological and metrical quantities. Firstly, we clearly note (top panels) that the efficiency is strongly influenced by the current fuzziness value, regardless of the networks considered. In particular, at micro level (that is, when the network structure is highly detailed) the efficiency is smaller compared to the macro level. This is an important element suggesting that the abstraction process is effectively a useful methodology to simplify a system (in fact the number of nodes and edges decrease, see figure 4.12) by selecting the substructure of the network that works best and that is most efficient. One element that distinguishes the two different types of networks is the connection pattern, reflected at micro level. We clearly note that as the fuzziness increases the two classes of networks tend to gradually be similar, smoothing away the initial differences.

Secondly, we notice how metrical E_{glob} of subways networks (bottom left panel) is reasonable stable over telescopic spectrum meaning that metrical features will be preserved during the abstraction process. However, this finding holds only in exponential networks like subways where the metrical element plays an important role during network creation and evolution. Conversely, in networks embedded in Euclidean space but where physical constraints on edges are relaxed (like in the US airline and cities-based networks, right panels) the E_{glob} (both topological and metrical) is not universal in the spectrum and again strongly depends on fuzziness.

Generally speaking, this finding is evidence that under our telescopic hypothesis, results of analysis on scale-free small-world networks to date, refers only to a specific resolution level (fuzziness) and can not be considered as general results.

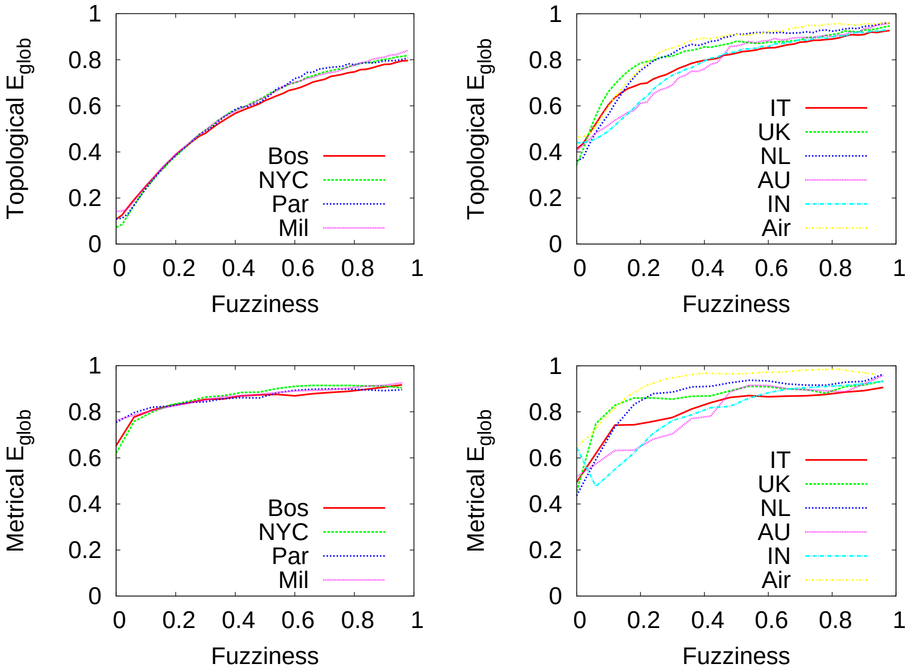


Figure 4.16: Effect of the telescopic process on subways (leftmost column), the US airline and city-based networks (rightmost column) as a function of f. The statistical properties considered in these panels are topological and metrical E_{glob}. The abstraction process does not preserves the topological E_{glob} (top panels) while varying f. In particular, regardless of the network considered, the networks viewed at macro level are simpler and more efficient compared to micro view. Conversely, the situation is slightly different for metrical E_{glob} (bottom panels). In this case, the connection pattern of the system considered alters significantly the outcome of the abstraction process. In fact, we detected that the structure of subway networks allow a good preservation of the metrical efficiency in the spectrum whereas in city-based networks this feature is absent.

E_{glob} is a quantity that accounts for the global system flowing of information along the paths of the networks. Conversely, with the formalization of the local efficiency E_{loc} (see section § 2.1.1), it is possible to detect how efficiently a system exchange information in the nodes' neighborhood.

Figure 4.17 shows how the telescopic analysis affects local efficiency as a f increases. At micro scale, we distinguish a completely different local connection pattern between the two types of networks (when no abstraction is applied). In fact, we note that subways are poorly connected locally because of the intrinsic physical and economic constraints that govern growth. In the US airline and city based networks, that are almost free from constraints (at least in the way nodes are linked), will have more redundant edges that increase local efficiency.

We observe again that the overall behavior of the metric considered is strongly influenced by the type of dataset involved. The main difference between the two types are not in the order of magnitude with which E_{loc} increase (in fact in both cases the quantity will raise) but instead in the detected behavior over the spectrum. We noted that E_{loc} is much more stable in subway networks compared to airline and city-based networks. This is probably caused by the characteristic of redundancy in the topological structure that makes E_{loc} more variable in the telescopic spectrum.

Even though local and global quantities are essential to characterize a network, the cost is another factor that has to be considered in order to get a better understanding of the entire system. Figure 4.18 reveals how metrical and topological costs behave in the telescopic spectrum. Regardless of the networks, we clearly see that as fuzziness f increases, the overall network cost will raise.

Although it seems counterintuitive because abstracted, i.e. simple, networks should be cheap, it is an expected effect because (as figures 4.16 and 4.17 show) these are also very efficient and as such, very expensive. All the curves are monotonically increasing functions, but subways result in smaller increase compared to city-based networks. This represents evidence that these networks have an economic inborn principle that is maintained also during the abstraction process.

Networks are defined as economic when they have low cost and high efficiency, i.e. whenever the ratio C/E tends to zero. Figure 4.19 shows how this variable changes in the telescopic spectrum. We clearly see that detailed networks are more economic than coarse-grained ones.

Figure 4.20 shows how the degree distribution $P(k)$ changes by increasing the fuzziness of 0.1 units at each step. We find that when decreasing the detail level, networks tend to lose their original topological structure and every node is likely to have the same degree. Therefore, hubs disappear and they became like low degree nodes. This is an expected result because it follows from the definition of network

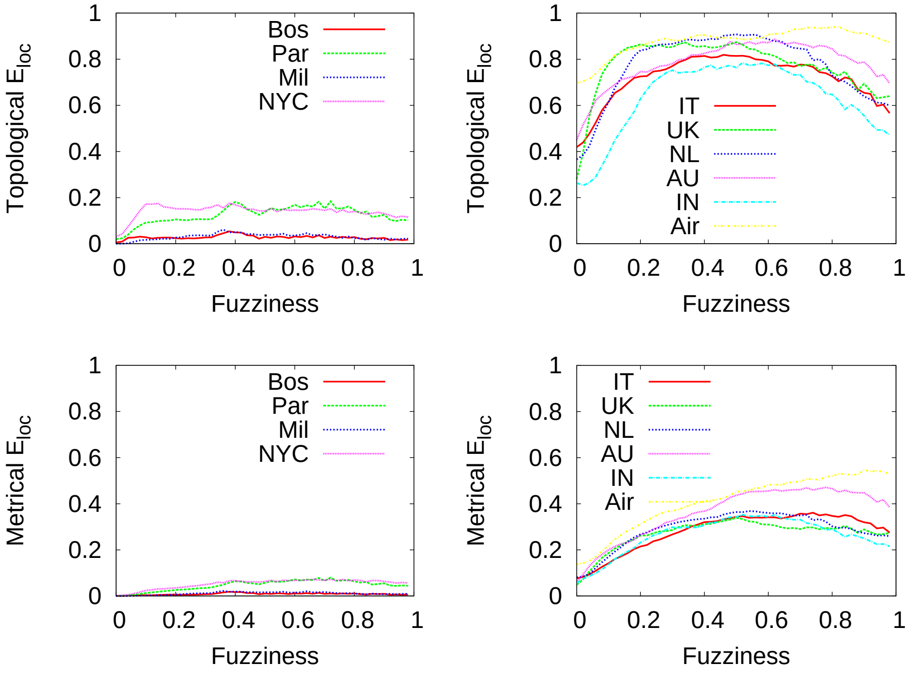


Figure 4.17: Effect of the telescopic analysis on topological and metrical E_{loc} as a function of f in subways (leftmost panels), the US airline and city-based networks (rightmost panels). The left most panels show that E_{loc} is almost stable in the spectrum meaning that the local properties of the subway networks are preserved by the analysis. However, in networks with heterogeneous topological structure, the telescopic process will further increase E_{loc} resulting in creating systems that are densely connected at local level.

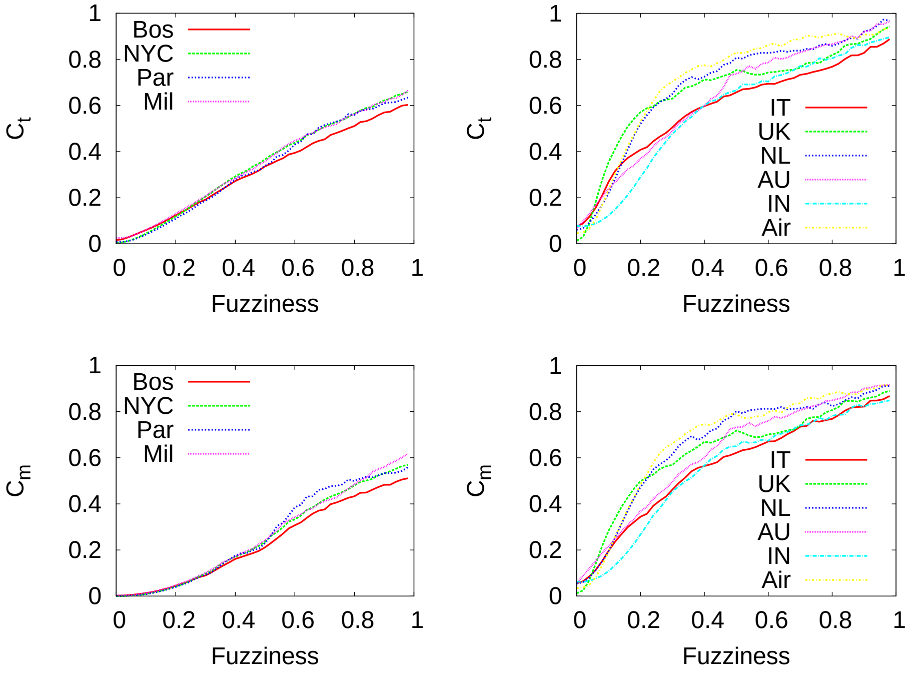


Figure 4.18: Effect of the telescopic analysis on topological and metrical cost (c_t and c_m) as a function of f for subways (leftmost column), the US airline and city-based networks (rightmost column). We note that our coarse graining process produces networks more expensive than detailed ones. This effect might be caused by the creation of redundant structures in macro level systems so that the whole cost will be higher. Even though both curves are positively correlated to f, the slope in subways networks is smaller compared to city-based networks. To verify whether this effect is not trivially caused by a low efficiency value, we will consider C/E_{glob} index (see figure 4.19).

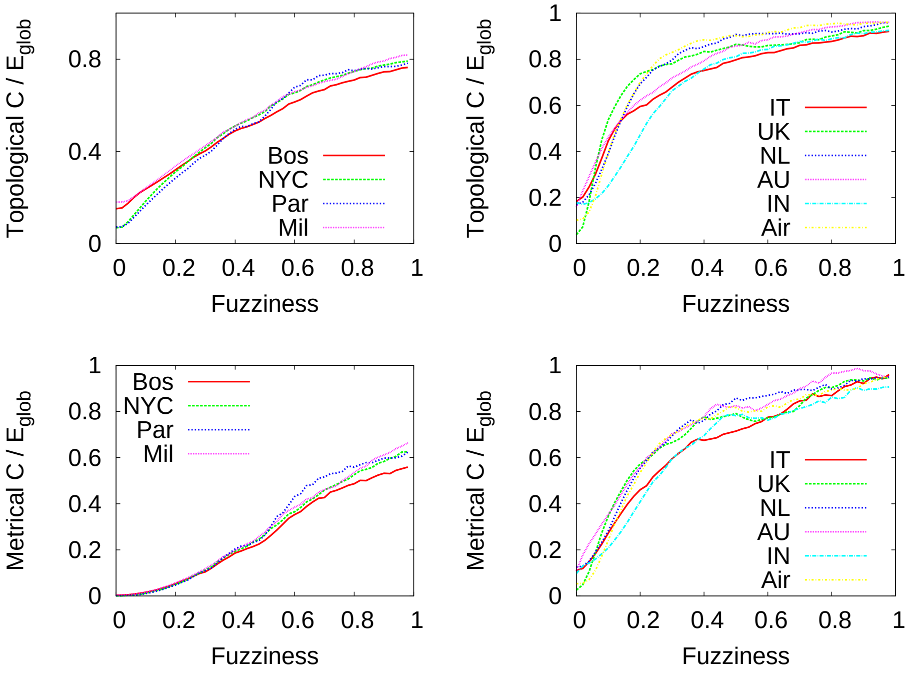
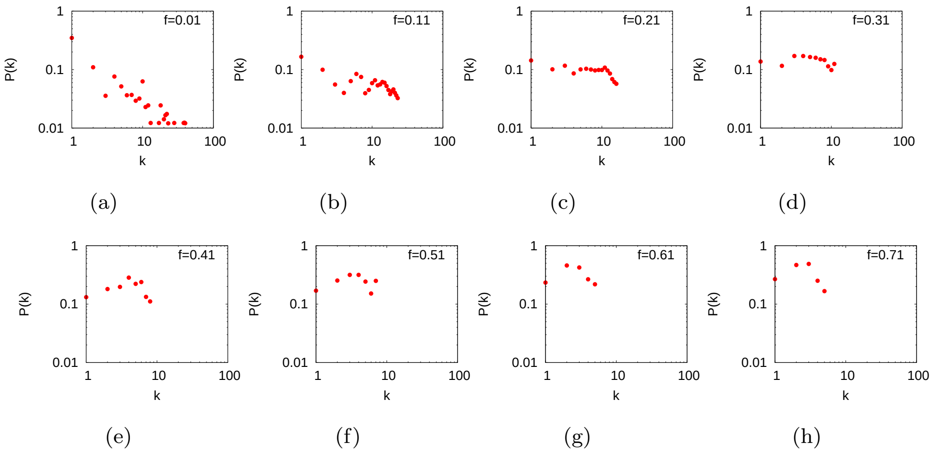


Figure 4.19: Effect of the telescopic analysis on topological and metrical normalized cost over efficiency for subways (leftmost column), the US airline and city-based networks (rightmost column). By dividing the cost of the networks by the global efficiency (that ranges between 0 and 1), we verified that subway networks are cheaper as well as very efficient, more than city based networks. This is evidence that subway network have an economic inborn principle that is maintained during the telescopic abstraction process.

abstraction.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 4.20: Effect of the abstraction process on the degree distribution $P(k)$ for increasing values of fuzziness f for the Netherlands city-based network. We detected that the behavior starts from a small world scale-free configuration and is ideally maintained for $f < 0.11$. When f increases, it changes to uniform and finally to random (when f is maximum).

4.4 Network perturbations

Here we report on the application of the telescopic algorithm to synthetic networks. The null models were created by iteratively applying different perturbations to the original real transportation and city-based networks presented in section 4.3.1. In this way, we precisely characterized the effect of randomization on the perception of networks at different abstraction levels. We proposed four variants of null models (for simplicity identified by a plus sign and a letter), namely:

- **+n**: Denotes a perturbation that consist on randomizing nodes' positions leaving unaltered the degree distribution. However, edge weights will be altered and will reflect the new Euclidean distance between nodes in the network.

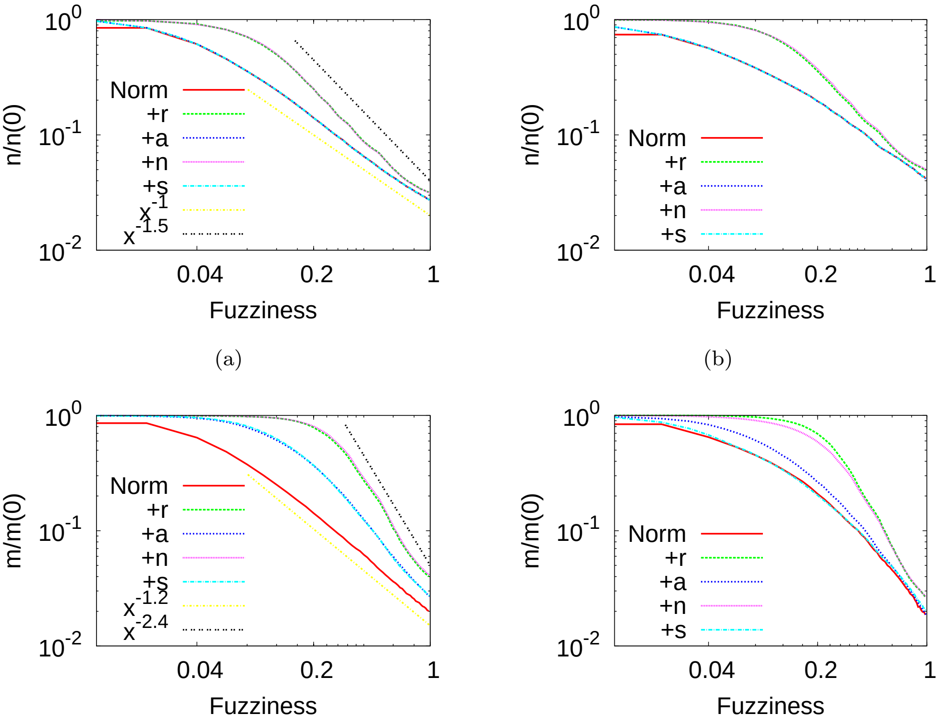
- **+a**: This modification consists of randomly rewire the edges of the network maintaining nodes' real positions. As a consequence, the degree distribution will not be preserved and edges weights will be modified accordingly to the new distances between nodes.

- **+r**: This perturbation simply represent the union of the previous two alterations and is useful in order to understand the joined effect of edges and nodes shuffling.

- **+s**: In this perturbation, we aim at changing the connection pattern of the network, switching to a scale-free structure, but maintaining the spatial layout of the nodes. There are many models for small-world scale-free network in the literature. However, no one allows matching the exact same number of nodes and edges of the referring networks. Therefore, we decided to use the Barabási-Albert scale-free model [15] because it best approximates $n = |V|$ and $m = |E|$ and because of its simplicity of setting up the parameters (i.e. starting with the number of nodes m_0, the number of links m that connect younger to older nodes). The alteration in the network topological structure results in edge weight modification though.

To limit the noise due to the randomness introduced by +r, +a and +n perturbations, we repeated the shuffling process generating one hundred random networks. Finally, averaged results are then considered. We first start by presenting the results obtained for randomized Boston subway and Australia city-based social network. We selected these two systems as the most representative of the two categories. However, appendix A includes the results for the other networks omitted in this section.

Figure 4.21 shows how the number of nodes n and edges m varies as a function of f and the type of perturbation selected for Boston subway (leftmost panels) and



(a)

(b)

(c)

(d)

Figure 4.21: Number of collapsed nodes n and edges m as a function of f and applying different perturbations. Curves are normalized by the baseline value obtained at $f = 0$. Straight lines refer to telescopic analysis without perturbations (also identified by +Norm). The leftmost column outlines results of Boston subway network whereas the rightmost column about Australia city-based social network (the other networks' results are presented in appendix A).

Australia city-based network (rightmost panels). All the values were normalized by $n(0)$ and $m(0)$ respectively, i.e. by the values at micro scale (at $f = 0$).

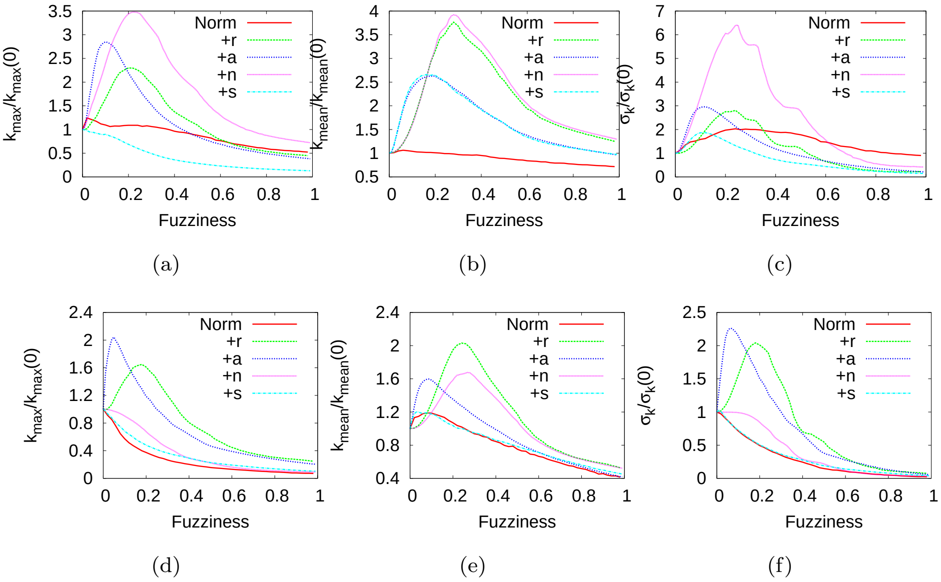
We note that the slopes of the curves are very similar regardless of the networks considered, even tough the underlying structure and node positioning is very different.

Indeed, the panels allow us to understand the effect of the single perturbations on the outcome of telescopic algorithm. In general, every alteration we considered slows down the collapsing rate of nodes and edges (compared to the baseline, i.e., the straight line). However, nodes' shuffling (+n) and combined nodes-edges (+r) are the one that mostly slow down the collapsing rate, leading to bigger y coefficient of the fitting k^{-y} curve from -1 to -1.5 (with $f < 1$). This is in accordance with the intuition that scattered nodes are more likely to be distinguishable during the abstraction process compared to a layout consisting of close elements. Similar effect was detected for edges. In fact, long-range edges will be collapsed later in the abstraction process.

From the panel it is interesting to note that the alteration of the topological structure (like with +a and +s) does not influence the collapsing pattern apart from subway networks where homogeneous spatial distribution of edges has the same effect of scattered nodes discussed above.

In figure 4.22 we show the maximum degree k_{max}, average degree $\langle k \rangle$ and standard deviation σ_k for Boston subway and Australia city-based networks as a function of f and perturbation rule selected. Although is far from trivial to explain the telescopic effect on network degree, we can make an interesting observation though. The panels show that both k_{max} and $\langle k \rangle$ increase for small value of f and, since the standard deviation also increase, we conclude that at least in the initial part of the spectrum, the abstraction process creates heterogeneous degree nodes. This effect is amplified in subways where there are no hubs compared to city-based networks. However, it deserves more research effort in order to assert that nodes with high degree produced by the abstraction process are effectively hubs for the network (it might be useful to look at the degree distribution) and affirm that the abstraction process might create small world scale-free networks at the beginning of the telescopic spectrum.

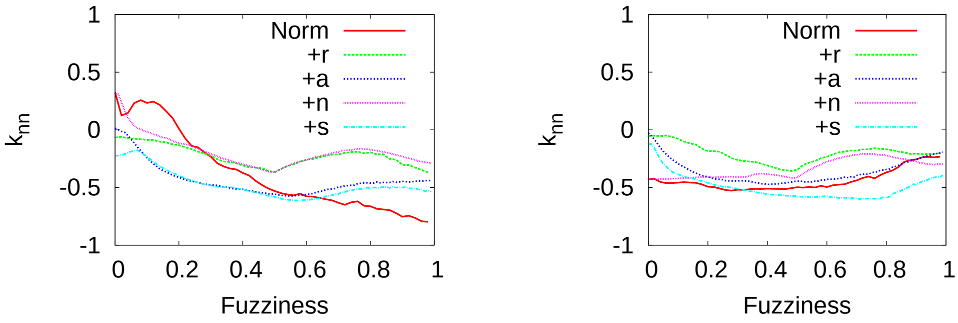
We see in figure 4.23 how k_{nn} varies as a function of fuzziness, for +r, +n, +a, +s null models. By focusing on the effect of perturbations on k_{nn} (i.e., the leftmost points in the x axis) we note that almost all perturbations tend to shift the initial k_{nn} toward neutral configurations. This behavior is due to the randomness introduced into the system by the different null models that destroy degree correlation. +n, as expected, is the only exception because as a perturbation, it leaves unchanged the topological structure and alters the nodes' positions only. Indeed, we see that the abstracted networks created by the telescopic process will be disassortative (or



(a) (b) (c)

(d) (e) (f)

Figure 4.22: Effect of the telescopic analysis on various degree based indexes: maximum degree k_{max}, average degree $\langle k\rangle$ and standard deviation σ_k for Boston subway (top panels) and Australia city-based network (bottom panels). As for previous panels, the discussion of the results for the omitted networks, were presented in the appendix A. All curves are normalized relatively to the baseline (i.e. the value at $f = 0$). The panel clearly shows that the telescopic abstraction on random perturbed networks (+r, +n and +a) drastically changes the connection pattern of the entire network.



(a)

(b)

Figure 4.23: Assortative mixing by degree ρ as a function of fuzziness f for Boston subway (leftmost panel) and Australia city-based network (rightmost panel). The randomization introduced by the perturbation process (see the leftmost point on the x axis) makes disassortative networks even at macro scale, regardless of the type of networks considered.

uncorrelated) regardless of the null model considered. This means that when networks are assortative at micro scale, they will become disassortative at macro scale. Vice versa, networks that start disassortative at micro scale (for instance, city-based or airline) will keep the same structure at macro level as well.

Figure 4.24 shows how the networks computed by the null models are globally perceived at micro/macro level in the context of the telescopic framework. We initially focused on the starting points of the curves (leftmost part of the x axis) and compare them with the baseline (starting points of the curve titled "Norm"). In this way we were able to precisely quantify the effect of +r, +a, +n, +s respectively. Every curve represents the results of the telescopic algorithm in which varying degree of randomness are applied to the networks.

In general, we note that after going beyond a certain fuzziness threshold, the alteration produced by the perturbations is smoothed away and all the networks, that initially were different, perform the same. This means that increasing the obfuscating value in diverse networks effectively produce output networks that are perceived as similar from an observer. This confirms the effectiveness of our abstraction process.

Figure 4.25 shows the effect of telescopic algorithm on different null models. Many observations can be made. By considering the leftmost panels (i.e., Boston subway network), we see that not all the perturbations by themselves alter the local connectivity both topological and metrical. In fact, all the curves start approximately from the same point. However, strong differences are then detected in the telescopic spec-

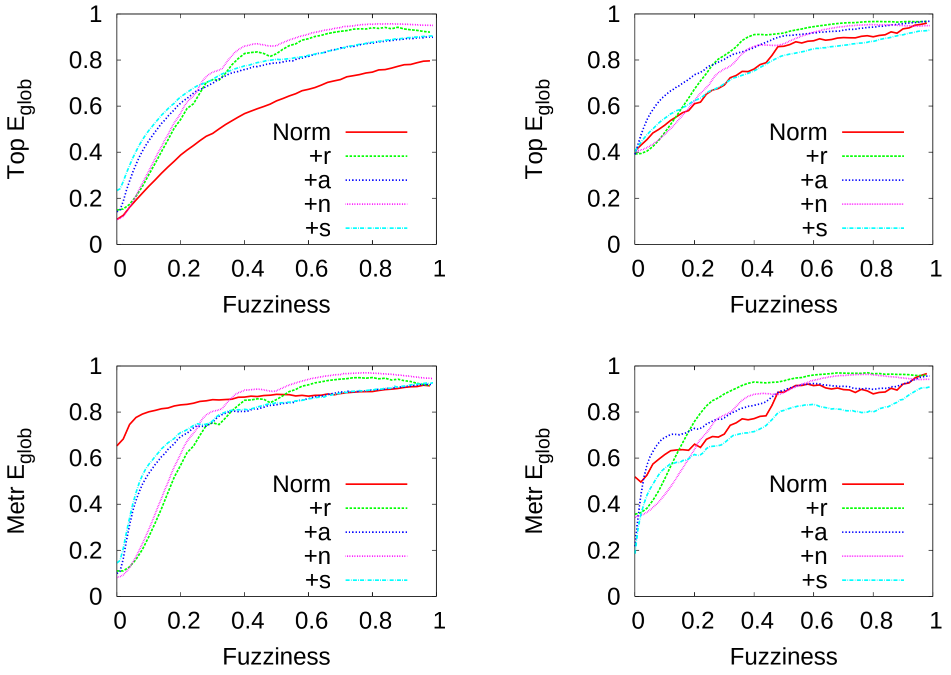


Figure 4.24: Effect of the telescopic process on Boston subway (leftmost column) and Australia city-based network (rightmost column) perturbed by different degree of randomness and topological modification (curve titled “Norm” represents the baseline, i.e., the telescopic analysis applied to the original network). Other networks perform similarly and are shown in the Appendix A. The statistical properties considered in these panels are topological and metrical E_{glob} as a function of f. From the plots, we clearly note that at macro level, every network produced by the different null models becomes simpler (will have fewer nodes and edges) and is very efficient compared to micro level. Indeed, another important point is the consequence of perturbations on the starting point (leftmost point in the x axis). In general, every perturbation alters the E_{glob} of the networks by slightly increasing it (for topological) or drastically decreasing it (for metrical).

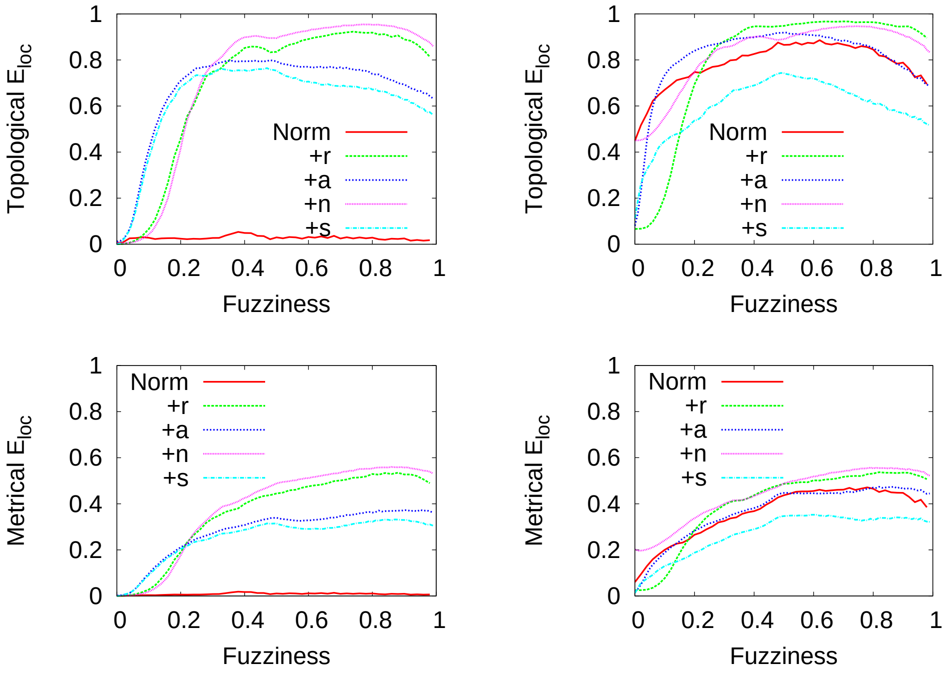


Figure 4.25: Effect of the telescopic analysis on topological and metrical E_{loc} as a function of f for Boston subways (leftmost column) and Australia city-based network (rightmost column). Other networks perform similarly and therefore are shown in the appendix A. By focusing on the left most points of the x axis, we clearly note that perturbations on Boston subway do not alter considerably the local efficiency. Conversely, in homogeneous networks like city-based social networks, perturbations influence considerably the local pattern and consequently the starting point will be different.

trum by comparing it with the baseline. This means that even though the original networks and null models might have the same locality pattern at micro scale, they can have very different connectivity at macro scale.

The situation is slightly different in small world scale free networks (rightmost panels). In fact, since +r, +a and +s modify the local connectivity introducing noise in the networks, the starting point of the curves will be altered toward a lower value. However, the differences from the baseline will be smoothed away as the fuzziness value increase and the abstracted networks will be similar.

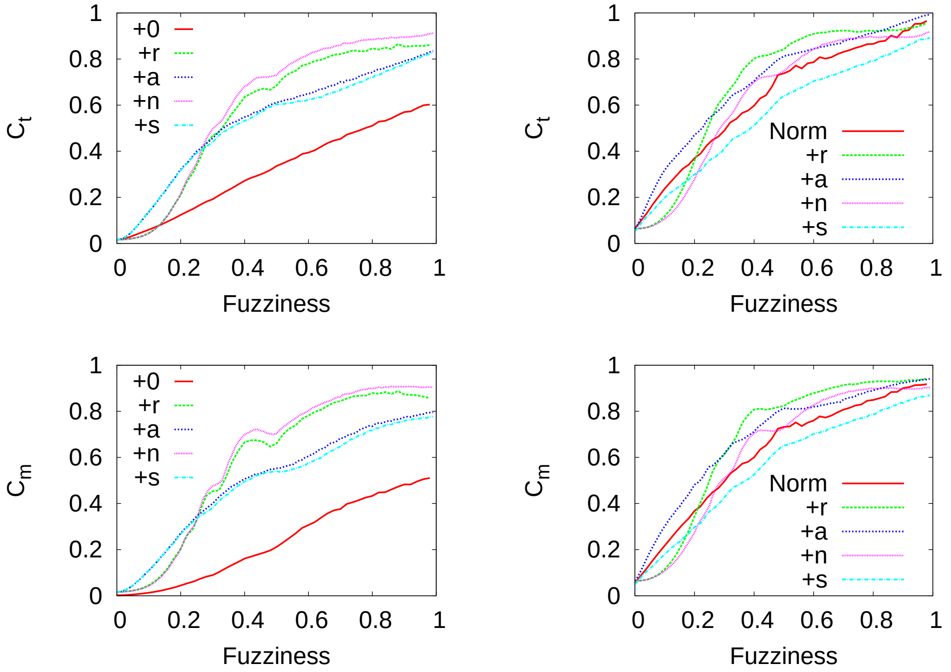


Figure 4.26: Effect of the telescopic analysis on topological and metrical cost (c_t and c_m) as a function of f for Boston subway (leftmost panels) and Australia city-based network (rightmost panels). Many perturbations of the original networks are considered. The plots for the other networks are presented in appendix A. As detected for real networks, the cost increases as networks become obfuscated. However, two different situations take places. In subway networks, the cost of null models is always bigger than the baseline (Norm curve) regardless of the perturbations. Conversely, for small world scale-free networks we found that the behavior is very similar to randomized networks. This is evidence that small world scale-free, in our telescopic context, share some characteristic with random networks.

In figure 4.26 we see how perturbations affect telescopic curves for metrical and topological costs. The effect of perturbations could be detected at the starting point of the curves (leftmost part of the x axis) and during the whole spectrum. In the former, no perturbation alters the cost of the networks. In the latter, the general telescopic behavior is strongly dependent on the class of the considered networks. Perturbations on small-world scale-free do not modify considerably the telescopic trend, as it grows. Conversely, on subway networks we detected that the cost always increase as the network will be more obfuscated. However, the increase is always lower than that of perturbation. In all cases, networks at micro scale are very cost efficient, whereas in macro scale they turn very expensive (even though they are very efficient, see figure 4.24).

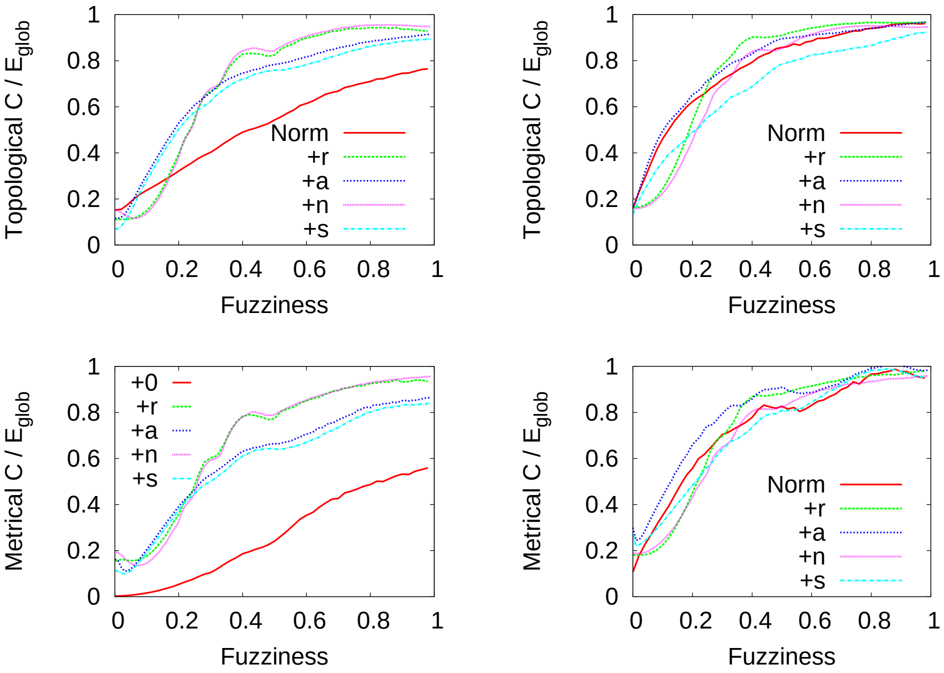


Figure 4.27: Effect of the telescopic analysis on topological and metrical normalized cost over efficiency for Boston subway (leftmost panels) and Australia city-based network (rightmost panels). By dividing the cost of network by the global efficiency (that ranges in $[0, 1]$), we verified that subway networks are cheaper, as well as more efficient, than city based networks. This is evidence that subway networks have an economic inborn principle (property) that is maintained during the telescopic abstraction process.

4.4.1 Null models as a function of networks' size

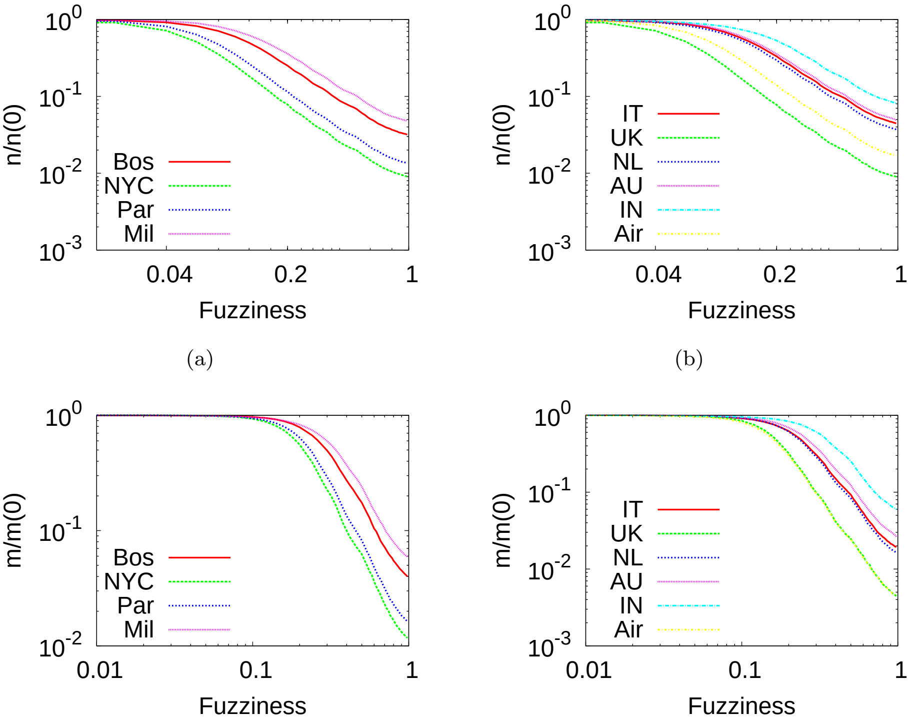
The previous section was devoted to describe how different null models affect the perception of networks at different abstraction scales. We now focus our attention to study how the size of networks' null models yields changes in the way networks are reconstructed at different detail scales. Between all null models proposed, we selected +r as it is less constrained by the original network features. This model starts with the original network and randomizes it through the iteration of rewiring moves and shuffling nodes' positions, altering both topological and spatial structure.

Figure 4.28 shows log-log plots of the number of nodes n and edges m as a function of the fuzziness. The values are normalized by the values obtained at $f = 0$. The null models have the same n and m as of subway (leftmost column) and airline and city-based networks (rightmost column). Clearly, the rate with which the nodes and edges collapse is influenced by the size itself. The more nodes (or edges) a network has, the faster they will be collapsed because the likelihood of belonging to the same square box is greater.

Figure 4.29 shows the effect of the abstraction on null models and in particular on diameter metrics. Physical diameter (leftmost column) linearly decreases as f increases. This is expected as the metric is based on the maximal extension of the network without considering any topological or metrical underlying structure. Conversely, topological and metrical diameter (centermost and rightmost column) decrease exponentially, meaning that in this case the abstraction process immediately creates the right shortcuts links that drop the maximum length that separates nodes.

On figure 4.30 are plotted the degree based metrics such as maximum k_{max}, average degree $\langle k \rangle$ and standard deviation σ_k. We note that big networks exhibit a peak approximately located at $f \sim 0.2$ regardless of the feature. This means that, at least initially, the telescopic analysis modify null models in such a way that high degree nodes appear, changing the network structure toward a more organized one (and similar to small world).

Topological and metrical efficiency E_{glob} as a function of fuzziness is pictured in figure 4.31. We clearly distinguish two very different situations at the extreme of the spectrum, similarly to what we have found for real world datasets. It is worth noting that, in accordance with figures 4.30 and 4.32, at approximately $f \sim 0.2$, the networks were abstracted in a way that both E_{glob} and E_{loc} are high and nodes degree tend to be homogeneous, evidence for a small world scale-free structure. However, a deeper investigation that include also the degree distribution as a function of fuzziness, might confirm this conjecture. Indeed, networks size influence the global efficiency with reference to the abstraction process as bigger networks will be more efficient



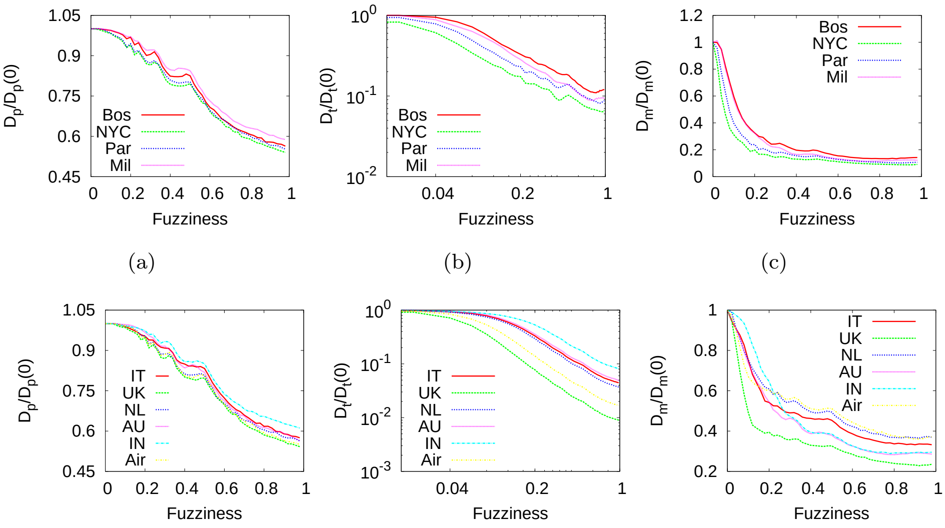
(a)

(b)

(c)

(d)

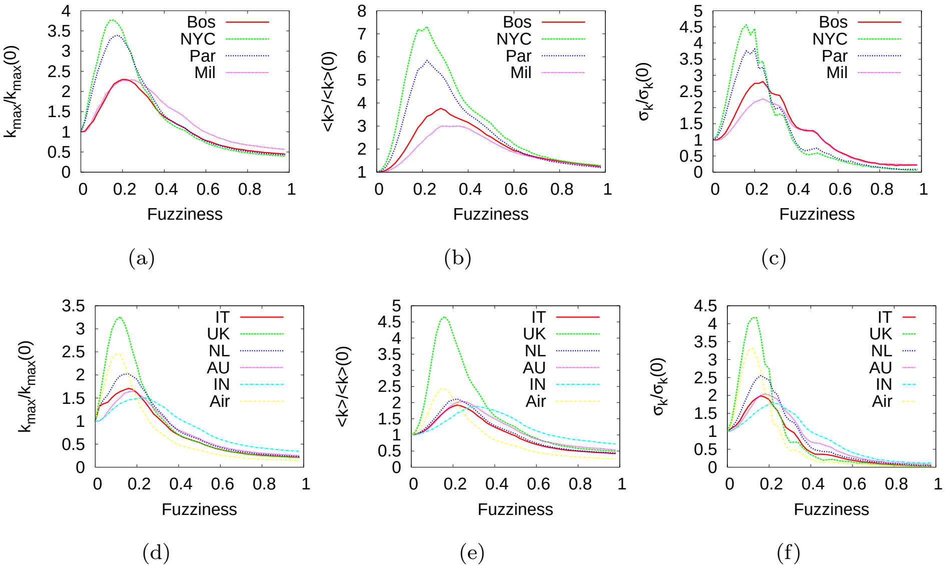
Figure 4.28: Number of collapsed nodes n (top panels) and edges m (bottom panels) as a function of f. The values are normalized by the first value obtained at $f = 0$ and denoted as $n(0)$ and $m(0)$ respectively. Leftmost column refers to +r null models created with the same number of nodes and edges of subway whereas the rightmost refers to the airline and city-based social networks. The plots clearly show that bigger networks tend to collapse nodes and edges faster than smaller ones. This effect is also due to the density of nodes in the 1×1 square box.



(a) (b) (c)

(d) (e) (f)

Figure 4.29: Effect of the telescopic abstraction on physical (leftmost column), topological (centermost column) and metrical (rightmost column) diameter for +r null models with different sizes created by randomizing nodes' positions and rewiring links of subways (topmost panels), airline and city-based social networks (bottom panels). Physical diameter linearly decreases as a consequence of how the metric is defined, i.e., the maximum distance between nodes without considering the underlying structure. Metrical and topological diameters decrease exponentially. This clearly means that the telescopic algorithm find the right shortcuts links that drop the maximum length.



(a) (b) (c)

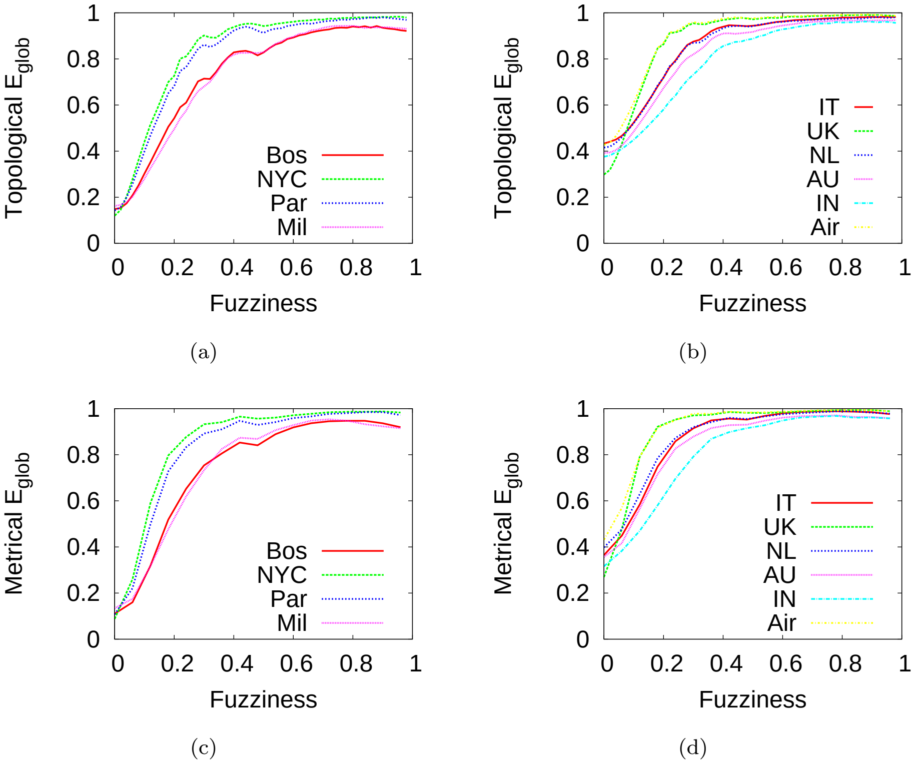
(d) (e) (f)

Figure 4.30: Effect of the telescopic abstraction on degree related metrics for +r null model created by perturbation of nodes' positions and rewiring links of subways (top panels), airline and city-based social networks (bottom panels).

compared to the small ones.

Figure 4.32 show the local efficiency E_{loc} as a function of fuzziness. Null models, under the telescopic framework, will have two different structures when compared at micro and macro level. At micro scale, networks are poorly connected, mainly due to the randomness with which networks were created. At macro scale, the local efficiency explodes for every networks size. This means that telescopic algorithm is able to create efficient networks even with null models (that are comparable to random networks).

Figure 4.33 represent topological and metrical cost in the telescopic spectrum as a function of fuzziness. We note that at macro scale, the cost level is high. The same situation was found in real world networks. This clearly means that at high values of fuzziness networks that originally are very different become comparable.



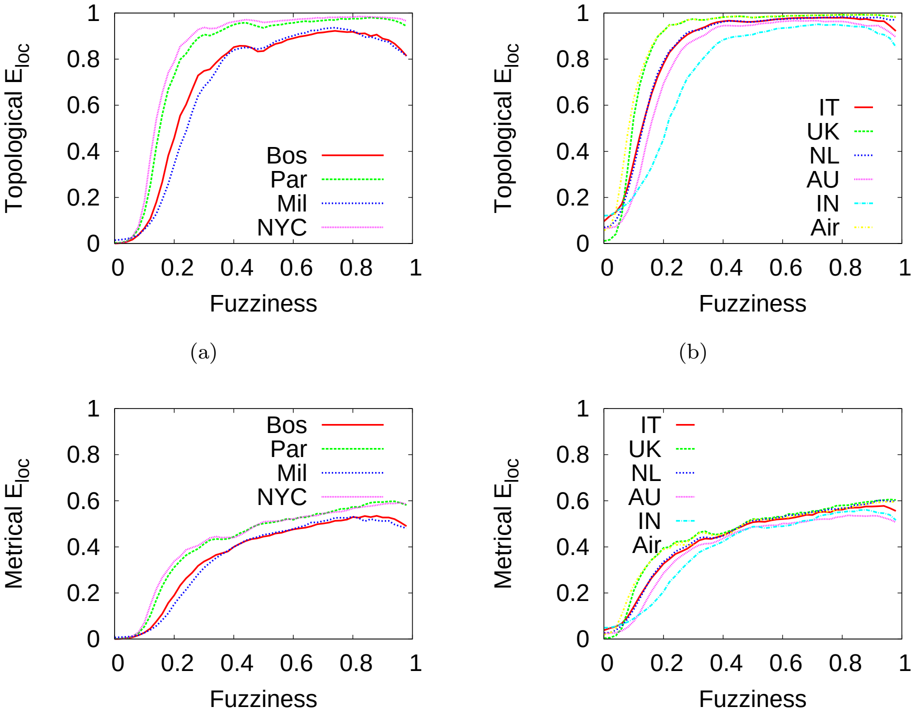
(a)

(b)

(c)

(d)

Figure 4.31: Effect of the telescopic abstraction on topological and metrical global efficiency E_{glob} for null model created by perturbation nodes' positions and rewiring links. Models will have the same number of nodes and edges of subways (leftmost panels) and the US airline and city-based social networks (rightmost panel).



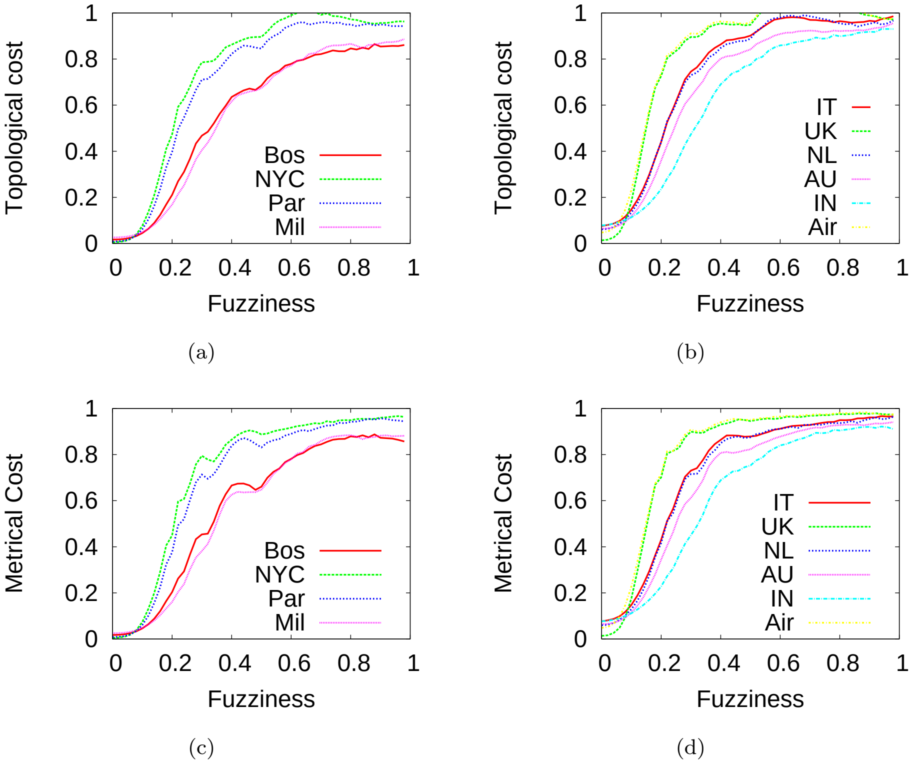
(a)

(b)

(c)

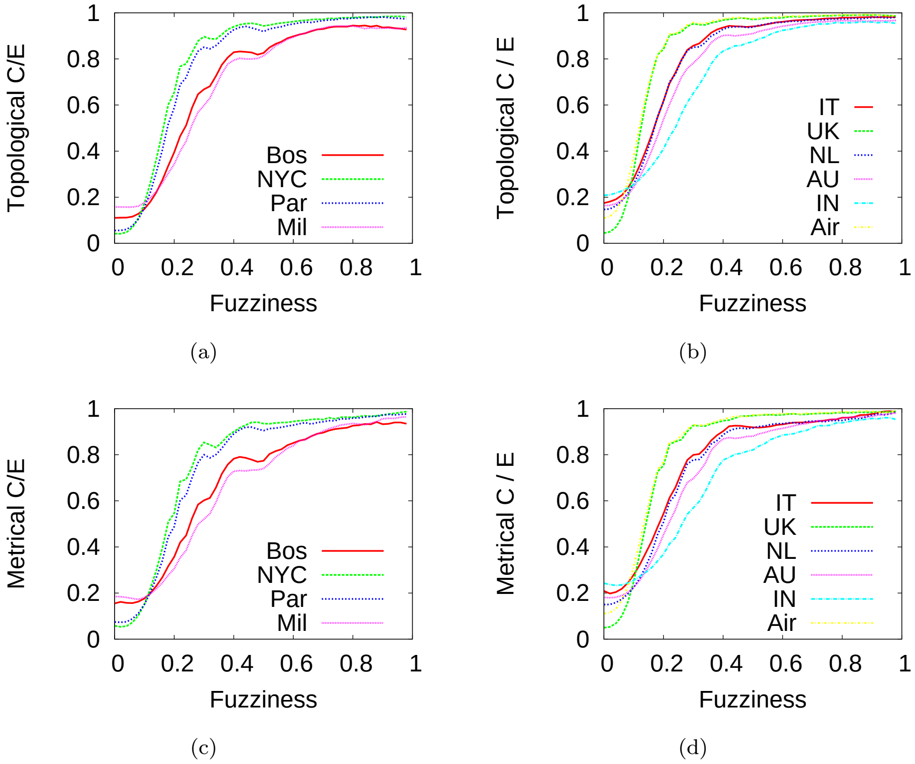
(d)

Figure 4.32: How local efficiency E_{loc} is affected by different sizes of null models (+r). Leftmost panels refers to subway networks size whereas the rightmost to the city-based social networks. Efficiency starts low at micro view (in accordance with many works that reported weak clustering coefficient in random networks [21]) and increases as f exceeds 0.1 reaching high values at macro view. As for global efficiency, E_{loc} is higher in bigger networks compared to small ones.



(a) (b) (c) (d)

Figure 4.33: The effect of the telescopic analysis on topological and metrical cost in randomized subways (leftmost column), the US airline and city-based social networks (rightmost column). In the topological version, the cost (also known as *density*) is defined as $\frac{|E|}{K_{|V|}}$ whereas in metrical one the formula accounts for weighted edges (see § 2.1.1). The cost is always increasing regardless of network size even though small networks tend to be cheaper than bigger ones (mainly at macro scale). This clearly states that, under the telescopic framework, abstracted networks are very expensive (but very efficient as well).



(a)

(b)

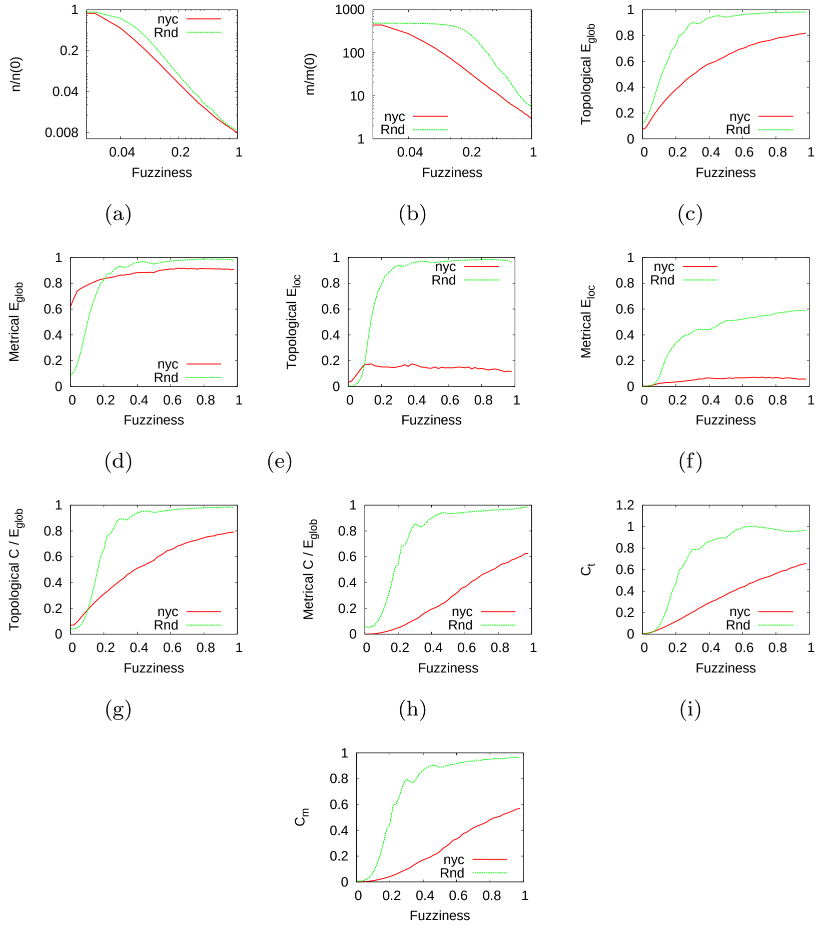
(c)

(d)

Figure 4.34: Topological and metrical C/E_{glob} in randomized subways (leftmost column), the US airline and city-based networks (rightmost column). This property measures the trade off between cost and efficiency. Small values indicate economic networks with high efficiency at low cost.

4.4.2 Null model and real networks comparison

The aim of the previous section was to investigate the effect of networks' size on networks abstraction, with particular emphasis to null models created by shuffling nodes' positions and randomly rewiring edges. The following figures 4.35 and 4.36 compare the results obtained in real world networks (such as Nyc subway and city-based social network of the Netherlands) and null models. We observed two completely different behaviors: subway network have less redundancy and randomness so, in general, the abstracted system will not be comparable to the outcome achieved with null models, even at macro scale. Conversely, small world scale-free networks, being more fault tolerant, will have a similar behavior in the telescopic context.



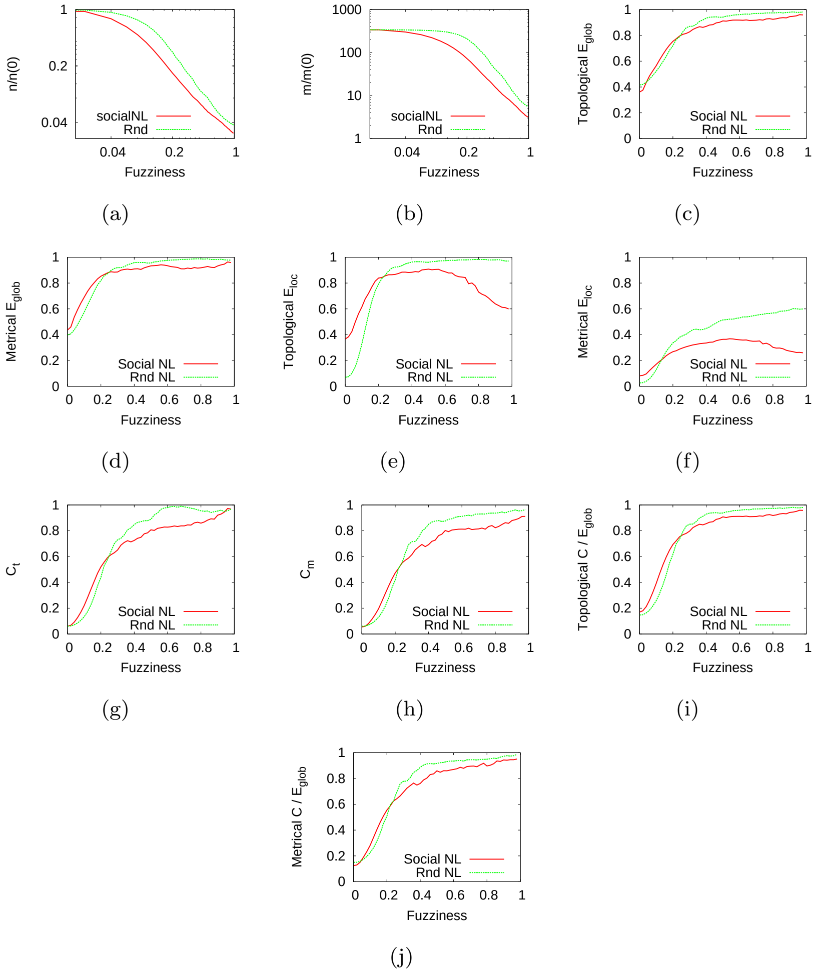
(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j)

Figure 4.35: Complete overview of the randomization effect on Nyc subway network. Comparison between results of telescopic analysis in terms of number of nodes n, edges m, efficiency, cost and cost over efficiency. Nodes and edges were normalized by the first value obtained at $f = 0$, defined as $n(0)$ and $m(0)$ respectively.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j)

Figure 4.36: Comparison between original and randomized city-based network of the Netherlands. The others' city-based networks were not plotted because the results were similar. All the most important statistics were used, such as number of nodes, edge, topological and metrical efficiency, cost and cost over efficiency. Nodes and edges were normalized by the value at $f = 0$, denoted as $n(0)$ and $m(0)$ respectively. It is interesting to note that the effect of abstraction process has strong similarities with that obtained in city-based networks.

Chapter 5

The time dimension

5.1 Introduction

In the previous chapter, we studied how abstraction modifies networks along with their spatial informational axes. We now focus on the other informative axis: time. This analysis is accomplished by keeping fixed the spatial axes and investigating how networks evolve as time pass by. Our analysis is designed to evaluate on line social systems. However, we think that the simplicity of our models can be easily fit to other complex networks.

Studies [57][112] suggest that on line interaction is driven by the same needs as face-to-face interaction, and should not be regarded as a separate arena but as an integrated part of modern social life [112]. Thus, communicative actions taken by members of on line communities can be expected to share many features with the web of human acquaintances and romances in the social off line world. Indeed, for many people in contemporary western societies, interaction on the Internet is as real as any other interaction [111]. For this reasons, the formation, the dynamics and the general evolution over time of the social networks in an Internet community can provide important information for enhancing our understanding of social networks. This is confirmed by latest works [12][57][24] that shed new light on the hidden rules behind social mechanism like creation, maintenance, dissolution, and reconstitution of interpersonal ties.

Researchers that study these communities use heavily the digital traces that virtual users leave while using social networks. Even though these datasets are usually owned by private companies, we're increasingly seeing new trends toward openness of data, for example by letting people to access sampled subset of the entire dataset (Twitter for example allow to get $1-2\%$ of all users' tweets, friendship relations, etc) or by releasing anonymized data for mining or edge prediction challenges. This benefits the network scientific community that is able, firstly, to work in this new challenging area and secondly to propose researches that are not limited by the number of persons involved in the experiments, as it was for the off-line social studies.

The most studied social mechanisms include random wiring, triadic closure and preferential attachment (the latter usually used to describe other than social networks evolution). The first model assumes that the social networks evolve without being constrained by exogenous or endogenous factors like geographical proximity, socio-demographics, belonging to topic-specific sub communities (homophily), technological differences, etc. Every couple of people is linked in a random fashion. The second one, also known as the friend of a friend's rule (or triadic closure), states that two friends of a person are more likely to know each other compared to two randomly chosen persons. In the latter model, new nodes are injected into the system and form links with existing nodes proportional to their current connectivity. Because of that, the older nodes will have a higher degree compared to the younger.

However, the main goal of the study of temporal dimension of complex network was to identify a new instinct, found in particular in on line social network, which drives network evolution. This interesting feature is detected by using special nodes (the so-called sirens) that, when added into the system, trigger a boost in network connectivity, spreading utilization to the entire community. This interesting phenomena was studied by trying to predict the best configuration in term of number of sirens, attractiveness (i.e., the ability to create new links), and length of time these nodes are effective.

In the following sections we first describe more formally the characteristics of the evolution models and in particular network growth with sirens. Finally, we report on the encouraging results obtained by simulating two on line social networks (VirtualTourist.com and Communities.com) and show the benefit that bring special nodes in enhancing growth.

5.2 Network growth models

In this section, we describe the details of network's growth models we investigated in this Thesis. We start from an empty network and by repeatedly apply rules at local level, the network evolves and (eventually) reaches a mature state. We already know that many real world systems such as power grids, communication networks, biochemical interaction as well as social networks can be modeled as graphs [7]. Based on graph theory notation introduced in § 2.1, we will consider on line social networks as unweighted undirected graphs $G = (V, E)$. Nodes $u_i \in V$ represent users while edges $(u_i, u_j) \in E$ mutual friendship relations between them. The evolution of a graph $G = (V, E)$ is conceptually represented by a series of graphs $G_1, \cdots, G_t$, so that $G_i = (V, E_i)$ is the graph at step i. Since $G_1, \cdots, G_t$ represent different snapshots of the same graph, we have $E_i \subseteq E$. The simulations that we will consider are constrained in the sense that the final graph G_t has to be equal to the referring networks. In this phase of our research, we will not consider links or nodes removal, therefore $E_1 \subseteq E_2 \subseteq \cdots \subseteq E_t$. This choice is motivated by the observation that in on line social networks, removals are fewer than users and friendships relations [49].

Simple network dynamics simulations need at least two parameters. The first is the definition of the order with which edges will be inserted into the network (the order will influence the overall connection pattern) and the second one defines how many edges will be added at each step. Network connectivity evolves according to the following three rules:

- *Random* order. Every edge will have the same probability $p = \frac{1}{|E|-|E_i|}$ to be selected during the growth process. This rule, as many studies showed, is far from real. However, it is a good candidate for a baseline.

- *Aristocratic* order. This rule is based on the preferential attachment process [15][92] where older nodes have an higher probability of attracting new links. The process selects edges by choosing a source node, according to degree, and a target node, randomly chosen on available neighbors' nodes. By randomly choosing target nodes, low degree nodes can acquire new links as well. More formally, the probability of a node u to be selected is the following:

$$p_u = \frac{1 + deg(u) \cdot \alpha}{\sum_{j \in V} (1 + deg(j) \cdot \alpha)}$$

  where α is a scale factor that increase or decrease the influence of degree on the final probability value, $deg(u)$ is the node degree.

- *Social* order. This rule is inspired by the local clustering of small world networks (also known as *triadic closure*), and in particular from the observation that two friends of a person is likely that know each other [51]. This rule will consider more likely to be selected those edges that close triangles. Edges that make more than one triadic closure will be inserted sooner into the network than others. More formally, the probability of edge (u, v) of being selected is the following:

$$p_{u,v} = \frac{1 + soc(u, v) \cdot \alpha}{\sum_{j \neq k \in V} (1 + soc(j, k) \cdot \alpha)}$$

where $soc(i, j)$ is the number of times edge (i, j) closes triangles (see for example figure 5.1). As for the previous rule, α tunes the effect of triadic closures on final probability value.

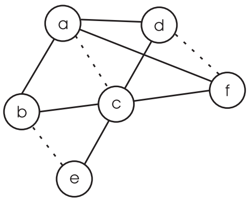


Figure 5.1: Example of social rule. The figure represents a hypothetical snapshot of graph G at time t during network evolution. Straight links indicate already existing edges whereas the dashed lines the ones that will be added in the following steps. Edge (a, c) closes three triades $(a, d)(d, c)$, $(a, b)(b, c)$ and $(a, f)(f, c)$ whereas (d, f) and (b, e) only close two and one triangle respectively. Therefore, the probability of been selected at time $t + 1$ is 0.5, 0.33 and 0.16 respectively.

5.2.1 Evolutionary models: serial and parallel

Although the previous three rules are sufficient to define the order with which nodes will be connected, it is also necessary to define how many edges are inserted into the network at every time. One trivial solution is to add *serially* (alternatively called inertial) an edge every time slot so, in a network composed by m edges, the simulation will lasts m simulated time units. This represents the baseline in our experiments (see section 5.4.2) and is crucial for reporting which rules achieve best when the system behavior will be unfolded.

However, since the previous dynamics is realistic but only in specific situations (for instance in the initial part of a network evolution), we also consider simulations where more than one edge are allowed to be inserted at the same time. We assume that the number of edges added changes as a function of network efficiency E_{glob}. This model, that we call *parallel* (alternatively called accelerated), is described in Algorithm 1.

The algorithm accepts as input: (i) a graph $G = (V, E)$ and (ii) a rule $r_n = \{random|aristocratic|social\}$. It starts from a graph G' that has the same nodes as G and no edges. The algorithm deals with parallel edges creation by selecting at each time a subset F such that $F \subseteq E$ to be added to G'. Since the edges can be added into the network only once, E will be updated adequately with the remaining edges.

The number of connections selected varies according to the following formula:

$$e = 1 + \left\lfloor C \cdot \frac{E(\mathbf{G}_{t-1})}{E(\mathbf{G}_{ideal})} \cdot (nar_{t-1} - 1) \right\rfloor \tag{5.1}$$

where G_{ideal} is the ideal network in which all edges exist $K_{|V|}$, nar_{i-1} is the number of edges that has to be inserted into the network, C is a constant factor and $E(G_{t-1})$ is the global efficiency (see § 2.1.1) of the network G at step $t-1$. At the beginning, few nodes will be inserted because of low efficiency and, as soon as the network grows and many people are involved in the network, more edge will be chosen and added concurrently. The one factor at the beginning of formula 5.1 allows to pick at least one edge at each step. The C factor is used to tune the effect of the efficiency in the number of chosen edges. We studied the effect of C on the network growth and we found out that it only expands ($C < 1$) or shrinks ($C > 1$) the time needed to get target efficiency without altering considerably the curve behavior (see figure 5.2). For this reason, we decided to use $C = 1$ in all our experiments.

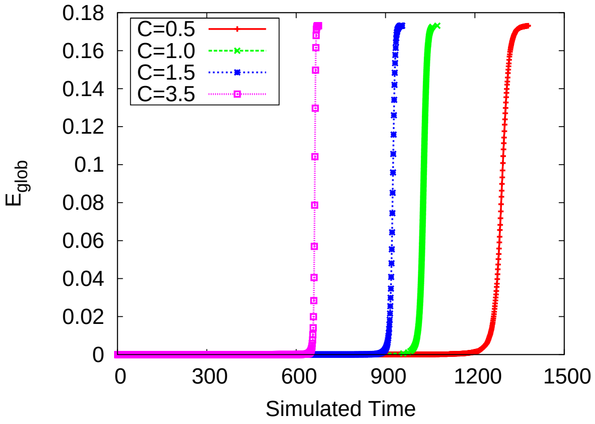


Figure 5.2: How the scale parameter C influences the global efficiency curve. It directly impacts the time span needed to get the referring network.

Algorithm 1: Parallel networks simulation

```
input : $G = (V, E)$, $r_n$
$G' = (V, E' = \varnothing)$
while $E \neq \varnothing$ do
    F = Choose $e$ edges from $E$ with method $r_n$
    $E = E - F$
    $E' = E' \cup F$
    calculates statistics on $G'$
    e = Number of edges as a function of $E_{glob}(G')$
end
```

5.3 Sirens

The aim of this section is to present a method we developed that is able to detect a new instinct, typical of social networks, that drives network evolution and boost network connectivity. This technique make use of external nodes (in the sense that they are new to the network) $V^s = \{s_1, s_2, \dots, s_m\}$, that we called *sirens*, whose goal is to interact with normal users and to stimulate overall network utilization, i.e. people engagement on on line social networks (see figure 5.3) and increase efficiency.

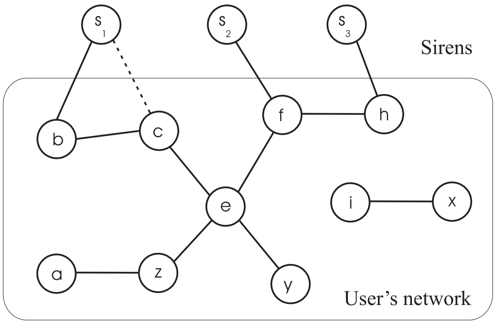


Figure 5.3: Example of on line social network with sirens $V^s = \{s_1, s_2, s_3\}$. Nodes inside the rounded rectangle belong to the users' graph. Sirens aim to connect to nodes of users' network in order to increase overall utilization (edge (s_1, c)).

The whole graph now becomes $G^s = (V \cup V^s, E \cup E^s)$, the total number of nodes $|V \cup V^s| = n + m$ and the number of edges $|E \cup E^s|$ depends on the following parameters:

- m, specifies the number of sirens,

- a, the sirens' ability of attracting new edges (i.e. generate interest in the community),
- d, the operating timespan of sirens.

In particular, we define a *configuration* as a tuple $c_i = (m, a, d)$ formed by the previous three parameters. Furthermore, we define a cost of a configuration, C_s, proportional to the previous parameters, that is:

$$C_s = m \cdot a \cdot d \tag{5.2}$$

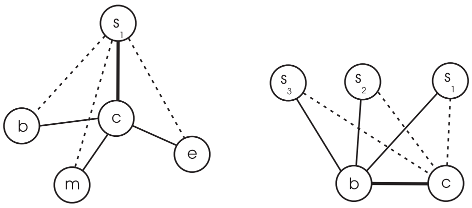
Combinations of these parameters lead to different costs and ideally to different growth behaviors. Another important goal of this Thesis is to understand how the overall network evolution changes as a function of c_i and in particular to test whether increasing the investment on the sirens (that is, C_s) yields a proportional benefit in the global efficiency. Furthermore, we study which configuration's parameter attains the best performance in the same cost configurations (see § 5.4.2). In this context, for the remainder of the thesis, we make the following assumptions:

- during the network evolution, edges between sirens $\{(s_i, s_j) | s_i \in V^s, s_j \in V^s\}$, are not allowed
- sirens $\{s_1, s_2, \ldots, s_m\}$ are active at the beginning of simulation only, i.e. from time t_0 to t_d. Improvements of this study might consider relaxing this constraint and studying the effect of the activation in different time periods.

The rules by which edges $\{(s, u) | s \in V^s, u \in V\}$ will be added are identical as for normal users, namely *random*, *aristocratic* and *social* (see previous section). In general, sirens' edges evolve independently of users' edges. However, an exception still exists. In fact, a new link in users' subnetwork could trigger a modification in the likelihood of sirens' edges of being selected, specifically with the social rule. Figure 5.4 pictures some examples. In particular, figure 5.4(b) shows what happens when a new link (b, c) in users' network is added: edges (s_1, c),(s_2, c),(s_3, c) will be more likely to be selected in the following steps because of triadic closure rule.

Algorithm 2 describes how simulations are made. It uses two sets E and E^s from which the edges will be selected, two rules r_n and r_s (that specify which edge to choose in the users' and sirens' subnetworks) and a configuration c_i. The algorithm has a main loop (line 4) in which two distinct phases are executed and each one manages users' and sirens' selection of edges (line 5 and 9) according to r_n and r_s respectively.

The number of edges selected in the first process (line 5) is calculated similarly as in algorithm 1 (using equation 5.1) apart from the input graph that now become



(a) (b)

Figure 5.4: Examples of word of mouth model (a-b). $s_i \in V^s$ and $\{b, c, e, m\} \in V$. Highlighted links indicate the edges that have just been added to the network, straight line are edges inserted in the previous steps and dashed ones represents the possible options for new links (and that have an higher probability to be chosen).

$G' = (V \cup V^s, E')$ instead of $G = (V, E)$. In the second phase (line 9), the number of selected edges is fixed and equal to $|V^s|\cdot|V|\cdot a$. This means that the total number of links between sirens and users can be estimated in advance as the following: $E^s = |V^s|\cdot|V|\cdot a\cdot d$ (this number will be reached at time t_d and will not change afterwards). The effect of sirens will be limited to the first d iterations, after that the sirens will be deactivated (line 8) and the system will evolve independently by itself.

Algorithm 2: Accelerated networks simulation with sirens

```
   input : $E$,$E^s$, $r_n$, $r_s$, m, a, d
1  $G' = (V \cup V^s, E' = \varnothing)$
2  $p \leftarrow 1$
3  $q \leftarrow a \cdot m \cdot n$
4  while $E \neq \varnothing$ do
5      F=Choose $p$ edges from $E$ with method $r_n$
6      $E = E - F$
7      $E' = E' \cup F$
8      if $s < d$ then
9          F=Choose $q$ edges from $E^s$ with method $r_s$
10         $E^s \leftarrow E^s - F$
11         $E' \leftarrow E' \cup F$
12         $s \leftarrow s + 1$
13     end
14     calculates statistics on $G'$
15     p = Number of edges as a function of $E_{glob}(G')$
16 end
```

The attractiveness parameter a quantifies how much a siren is able to promote the utilization (i.e. edges creation) of the on line community. It is defined as follow:

$$a(s) = \frac{q(s)}{\sum_{u \in V \cup V^s} q(u)} \tag{5.3}$$

where $q(s)$ is a weight function. Clearly, in order to meet the requirement that sirens have a better ability to establish new friendships, we assigned a doubled weight to them compared to normal nodes (see section § 5.4.2).

5.3.1 Network Cost

From a business point of view, every system has to deal with costs. In our context, the cost could be split into two parts. The first, that accounts for sirens cost (think of sirens as employees), and the second one that accounts for web site cost. More formally, the following formula gives an estimation of managing cost of setting up a network until it reaches a steady state:

$$f(C_s) = C_s + \beta \cdot T_{min}(C_s) \tag{5.4}$$

where C_s is the cost due by using a particular configuration (as defined in § 5.2), β is the cost by time units of the web site and T_{min} is the minimum timespan needed to the network to evolve toward a connection pattern with a specific global efficiency (given a configuration which costs C_s).

We are now interested in knowing the points in which the function f has minimums. For this reason, by calculating the first derivative of (5.4) and solve $f'(C_s) = 0$, we found that

$$f'(C_s) = 1 + \beta \cdot T'_{min}(C_s) = 0$$

and so the cost is minimum when:

$$T'_{min} = -\frac{1}{\beta} \tag{5.5}$$

We will return to this topic in section § 5.4.2 where we test how f's minimal points change versus hypothetical values of β.

5.4 Experimental Results

In this section, we report our experimental analysis. The simulation algorithms were implemented in Python and C programming languages. All the experiments were conducted on three Linux machines equipped with i5 Intel processor at 3.2Ghz and 8Gb of RAM.

5.4.1 Datasets

We conducted experiments on two real world datasets: Communities[1] and VirtualTourist [2] online social networks. Communities, considered the first social network in the world started in 1996, is similar to many other social networks like Facebook, Google Plus or LinkedIn where users meet new people, share photos and chat with friends. Communities is managed by users themselves that creates customized web pages in which they express passions, loves and friendships. Every user can keep track of friends in friends list, can use a guestbook, blog or use photo gallery. In Communities, users can establish virtual contacts, but unlike real world, these ties could be easily maintained over distance. This produces a network of virtual social ties that connects the entire world. Communities let members to create and join communities in order to easily find group of people with same interests. Joining a community means being able to chat with other members in community forums and chat rooms. User's locations in Communities are widely distributed and span over 185 nations.

VirtualTourist (in the following abbreviated as VT) is an on line travel-oriented community, started in 1998, in which users share their own travel experiences, suggest and review hotels, write comments and opinions on forums, find a places to visit, share photos and videos: it is a community of people that love traveling around the world. In VirtualTourist, users can meet new people and create a network of social virtual friendships as well.

Both analyzed networks were collected by crawling web pages of the sites as time of 2005 and 2006 [44][28]. Publicly available profiles and friendships were parsed and anonymized. At that time, there were approximately 700 thousands users in VT, 650 of which are singletons (92.4% of the total), i.e., users that have joined the service but have never made a connection with another user. Conversely, 57 thousands users have at least one friend (approximately 7.6% of the total). There were more that 200 thousands social ties at that time. The VT network has a giant component, a group of users who are pair wisely connected through paths in the social network, formed by 53.034 nodes (92% of the total nodes with degree greater than zero). The rest of the network is formed by 2077 small (less than 14 nodes each) isolated communities (also called middle region [65]) that are disconnected from the giant component.

In Communities, there were about 30 thousands registered users, 18 of which are singletons (60% of the total) and 12 have more than one friend (about 40%). There exists approximately 60 thousands friendship links. Apart from singletons, the vast majority of the nodes (about 12.131, 92.7% of the total) of the community belong to

[1] `www.communities.com`

[2] `www.virtualtourist.com`

Feature	Communities	VirtualTourist	Randomized CM	Randomized VT
$\|V\|$	12.479	57.639	12.479	57.639
$\|E\|$	60.209	211.415	60.209	211.415
$\langle k \rangle$	9.64	7.34	9.64	7.34
k_{max}	656	963	24	21
L	4.18	4.95	4.42	5.72
C	0.1067	0.04425	0.0006	0.0001
E_{glob}	0.238683	0.201822	0.23296	0.17817
E_{loc}	0.131466	0.056248	0.00074	0.00013
$Cost$ (density)	0.00077	0.00013	0.00077	0.00013
$Cost/E_{glob}$	0.00324	0.00063	0.00332	0.00073
γ	2.5	2.7	~ 0	~ 0
$\#CC$	161	2078	3	43
ρ	-0.027	-0.027	-0.002	0.00082

Table 5.1: Statistical feature of Communities.com and Virtualtourist.com on line social networks together with randomized versions of the same networks: number of nodes $|V|$, number of edges $|E|$, average node degree $\langle k \rangle$, maximum degree k_{max}, average shortest path L and average clustering coefficient C (for the largest connected component), global efficiency E_{glob}, local efficiency E_{loc}, cost, cost over efficiency, exponent of the cumulative degree distribution γ, number of connected clusters $\#CC$ and the correlation pattern ρ.

the giant component, whereas the rest to the middle region were small communities having less than 8 nodes each.

Since social ties are bidirectional in both systems, we mathematically tract those graphs as undirected. Table 5.1 summarizes the most important network statistical features. It also contains the metrics calculated on randomized versions of the same graphs. We note that both networks have small average shortest path L (less than 5 hops separates two randomly chosen nodes) and high clustering coefficient C (compared to the randomized versions, fourth and fifth columns). High E_{glob} and E_{loc} has been detected too. These facts are evidence for classifying them as small world[28]. Both networks are formed by many connected clusters, so average path length and clustering coefficient are calculated on the largest connected component (LLC), whereas global and local efficiency on the entire network (the latter two quantities work correctly even for disconnected networks, see § 2.1.1). Indeed, since cumulative degree distributions P_{cum} have tails that decays as power law with exponents equal to 2.5 and 2.7 (see figure 5.5) and maximum degree k_{max} is higher compared to the average $\langle k \rangle$, they could be classified as scale-free networks.

The plots on figure 5.6 and the assortativity values ρ on table 5.1 suggest that both networks are disassortative (Pearson correlation equal to -0.59 and -0.30 respectively). This means that, on average, users with many connections tend to connect to users with few friends (see classification on [84]). Many other studies found the

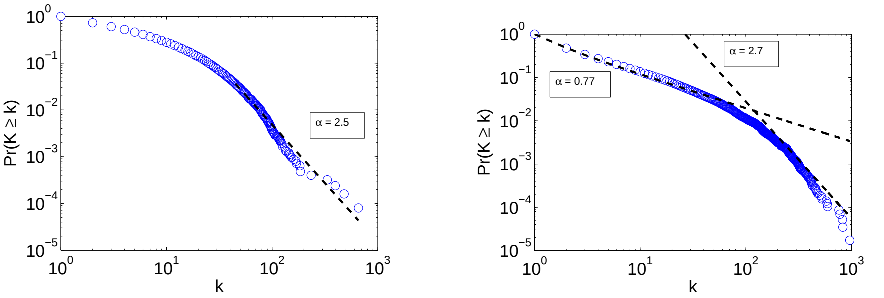


Figure 5.5: Cumulative degree distributions $P_{cum}(k)$ of Communities (left panel) and VirtualTourist (right panel). k is the degree, α is the coefficient of the fitting (dashed) line $k^{-\alpha}$. Figures clearly show a power law behavior in the degree and α is approximately equal to 2.5 and 2.7 respectively.

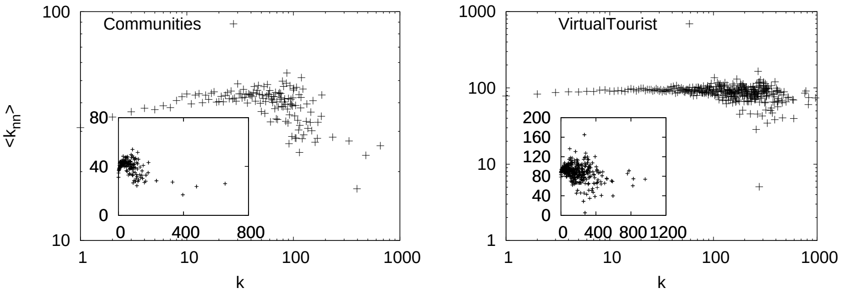


Figure 5.6: Node degree correlations in Communities.com (left panel) and VirtualTourist.com (right panel) on line social networks. $\langle k_{nn} \rangle$ is the average degree of first neighbors. Figures show a negative correlation. In fact, Pearson correlation is equal to -0.59 and -0.30 respectively. The inset graphs contain the same data but plotted in linear axes.

same correlation pattern in on line social networks, like for instance in Youtube[78], pussokram[57] or Cyworld[5] networks. Being elite in on line social networks simply means to have many connections and is just a matter of clicks[59]. However, this assortative pattern is the opposite compared to the real world where establishing and maintaining friendships require time and efforts and where many factors might influence the likelihood of being a friend of a person, such as cultural, economical factors.

$\|V\|$	m	$q(s)$	a_n	a_s
12479	6	10	0.000079751	0.000797512
12479	6	20	0.000079371	0.001587428
12479	12	10	0.000079371	0.000793714
12479	12	20	0.000078623	0.001572451
57639	6	10	0.000017331	0.000173313
57639	6	20	0.000017313	0.000346266
57639	12	10	0.000017313	0.000173133
57639	12	20	0.000017277	0.000345548

Table 5.2: Summarize of attractiveness values used during network simulations in Communities (first four rows) and VirtualTourist (last four rows). $|V|$ is the number of nodes, m the number of sirens, $q(s)$ is the weight assigned to sirens, a_n is the attractiveness of normal nodes, a_s is the attractiveness of sirens.

C_s	Configurations c_i		
600	(6,10,10)		
1200	(6,10,20)	(6,20,10)	(12,20,20)
2400	(12,10,20)	(12,20,10)	(6,20,20)
4800	(12,20,20)		

Table 5.3: List of the all possible configurations available with $m = 6, 12$, $a = 10, 20$ and $d = 10, 20$ with the corresponding costs.

5.4.2 Results

We now evaluate our models by simulating the network evolution with respect to suitable configurations c_i. We selected the global efficiency as the main statistical feature that has been tracked during the experiments. A configuration is defined as a tuple composed by (i) the number of sirens m used, (ii) the sirens' attractiveness a and (iii) the length of time d in which the sirens are active (starting from t_0). We decided to employ 6 or 12 sirens and to use those special nodes for the firsts 10 or 20 initial time units. To estimate the attractiveness as defined in equation 5.3, we use a weight function that is $q(u) = 1$ for $u \in V$ and $q(s) = 10$ or 20 for $s \in V^s$, in order that these special nodes acquire more links compared to normal nodes. Table 5.2 present the estimated attractiveness values of normal users a_n and sirens a_s, together with the variables used to calculate them. With the previous parameters, we created a set of 8 configurations and 4 levels of cost (listed in table 5.3).

The possible configurations that were planned do not exhaust all the axes along with our simulations are based. In fact, two more parameters are needed: the rule that selects edges between normal nodes and the rule that select edges between sirens and users. Since these dynamics are independent but the names of the rules are still the same, we define the sirens' rules as *Broadcast*, *Word of Mouth* and *Preferential* models in order to uniquely distinguish them from the users' rules.

The simulations considered in this Thesis are constrained in the sense that every edge that is added must exists in the original network. We decided to use this approach because other techniques like for instance stochastic simulations (Construct, Link Probability Model) are not well suited to describe big social systems (incurring in computational issues) and because they usually require to set an high number of initial parameters (incurring in a not trivial initial setting).

In order to evaluate the effectiveness of our methods to detect new instincts in social systems and to verify whether they are valuable as incentive for network utilization, we test (i) how faster global efficiency will increase when using sirens and (ii) whether growth curve will be altered by using these special nodes. Before that, we start by looking at results obtained for serial analysis (see section 5.2.1) so as to understand the effect each rule has in the unfolded network evolution and subsequently, we consider the more realistic situation in which more than one edge can be added at the same time (called accelerated or parallel analysis). We are interested to uncover all this aspects of on line virtual communities by trying to answers the following questions:

Q0 Does each rule behaves equally in the inertial (serial) context? What happens in the accelerated context?

Q1 How the same cost configurations influence efficiency?

Q2 How do parameters variations influence global efficiency?

Q3 How much we have to invest in special nodes?

Q4 How beneficial is the adoption of sirens in on line social networks?

Before delving into the details of the answers for the previous questions, we can delineate a general discussion about the results for all the growth patterns. In fact, regardless of the configuration adopted, we found that S-shaped curve characterizes all the growth pattern of E_{glob}. It is well known that S-shaped curve is at the heart of many diffusion processes and is characteristic of a chain reaction, in which the number of people who adopt a new behavior follows a logistic-like function [104]: a slow growth in the initial stage, rapid growth for critical mass time, and a rapid flattening of the curve beyond this point. Because of that, our models and rules could be considered as good candidate for estimating the real network evolution.

Q0: Unfolded serial setting. Figure 5.7 shows the unfolded behavior of the systems

	CM			VT		
	rnd	soc	ari	rnd	soc	ari
Normal	1384	1333	1931	3130	2996	7505
Randomized	2585	2571	2294	13718	13704	12003

Table 5.4: Summarize of T_{min} for all accelerated simulations in Communities (CM), VirtualTourist (VT) and randomized version of both networks. Random, aristocratic and social rules are considered.

for the three proposed rules, namely: random, aristocratic and social. Each curve represents global efficiency E_{glob} of the temporal networks that have been created by adding one edge at time. The plots allow for interesting observations. First, we note that until one sixth of the complete spectrum, each rule produces an indistinguishable behavior probably due to weak network structure. After that point, we observe the cumulative effect of drawing edges in different ways. The behavior detected is super-linear for aristocratic rule meaning that preferential attachment is an effective way to boost network efficiency in networks. Conversely, with social rule we observe a weak sub-linear increase ideally meaning that triadic closure is not the only key ingredient for network evolution. Linear increase is then detected for random rule. To avoid the bias of randomness, we made 100 simulations and then averaged results are considered. Standard deviations are small are not plotted in favor of clearer plots.

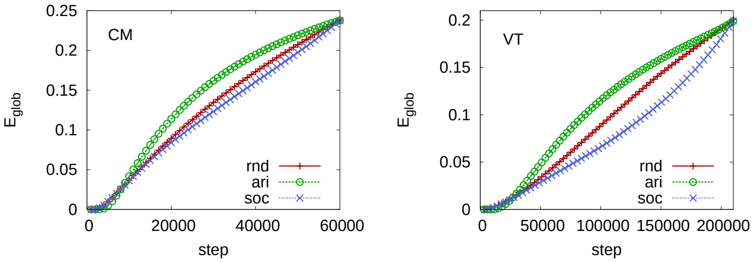


Figure 5.7: Effect of serial network simulation for Communities (left panel) and VirtualTourist (right panel). After an initial time span (approximately one sixth of the entire simulation time), the preferential attachment rule (aristocratic) outperforms the others.

Even though preferential attachment seems to perform better than other rules in serial evolution, this not necessarily holds even in accelerated (parallel) simulations. In fact, as figure 5.8 shows, random and social rule turn out to be 30% and 60% (CM and VT) faster in reaching the maximum E_{glob} compared to the preferential attachment (see table 5.4 that contains the minimum time needed (T_{min}) to get the

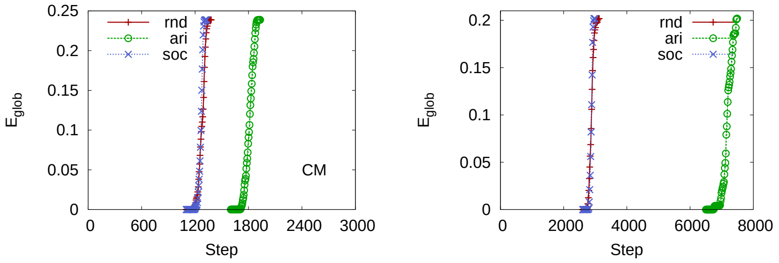


Figure 5.8: Effect of parallel network simulation for Communities (left panel) and VirtualTourist (right panel). These plots help us to understand how parallel links' creation modifies the dynamics of on line social systems. Surprisingly, preferential attachment, that outperforms other rules in inertial setting, is the slowest obtaining bad performance in terms of time needed to reach the target efficiency. Curves start at simulation time t_0, but we cropped the points for low values of E_{glob} for graphical clarity. Standard deviations are very small and are not plotted for graphical reasons.

original efficiency). This is probably due to a combined effect of topological structure and rule applied. In fact, on line social networks, like social networks in general[51], are formed by weak ties responsible of keeping together sub-communities and preserving the global reachability among nodes. According to preferential attachment, nodes with high degree are more likely to acquire new links. However, weak ties are not necessarily connected to hubs meaning that they will not be selected at the beginning, maintaining low the global efficiency.

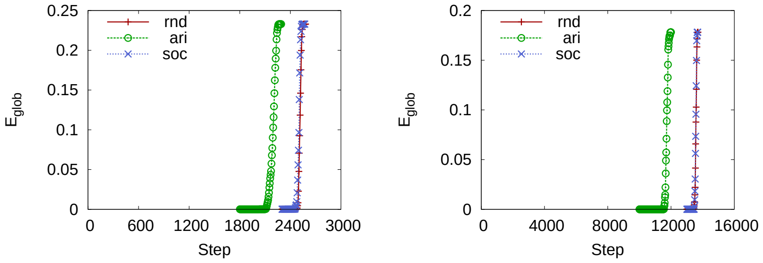


Figure 5.9: Effect of simultaneous network simulation in randomized version of Communities (left panel) and VirtualTourist (right panel). Curve starts at simulation time t_0, but we cropped the points for low values of E_{glob} for graphical clarity. Standard deviation is very small and is not plotted for graphical reasons.

In order to verify whether network topology affects the overall behavior, we applied the same rules on randomized version of the networks. Surprisingly, as figure 5.9 and table 5.4 show, the preferential attachment rule that previously was the slowest, now is 11% and 12% (CM and VT) faster that the others.

The explanation of this phenomenon is again dependent on the specific topological structure of the random networks. The main characteristic of these networks is that global connectivity is not based on the weak ties, but instead by scattered edges that connects randomly chosen nodes. As a consequence, the preferential attachment effect will be now strongly limited by the degree homogeneity of these networks and consequently the likelihood of selecting long-range edges is higher, bringing together far away substructures, yielding a fast increase of efficiency.

Q1: Same cost configurations. The following set of figures (from 5.10 to 5.16) represents how the same cost configurations affect the global efficiency. In particular, we consider the cost levels C_s that have at least two configurations c_i, namely 1200 and 2400. In table 5.3 collects all the possible configurations with a specific cost.

A single simulation's run needs three parameters: a configuration and two rules. The first rule that specifies the the dynamics of users and the second of the sirens'. In order to limit the bias due to the randomness of selecting the edges, we decided to repeat 100 times the same simulation and get the averaged results. However, the (simulated) timespan needed to get the target efficiency might vary in every run, making the calculation of averages not so straightforward. For this reason, we extended the timespan in order that every simulation fits to the longest. In this way, we were able to average the y values at fixed x intervals. We found that when the range of the timespan values is large (see for instance figure 5.13 (c)) and in particular with some C_s of VirtualTourist dataset, the method we use for averaging the results could create averaged behavior that seems like stepping functions. We think this issue could be easily figured out by increasing the number of simulations.

Figure 5.10 shows that only one specific configuration performs better, compared to the others and in particular the one that has the higher value of attractiveness. Surprisingly this result is also quite general because it holds no matter what cost level or rule selected and indeed regardless of sirens' dynamics chosen.

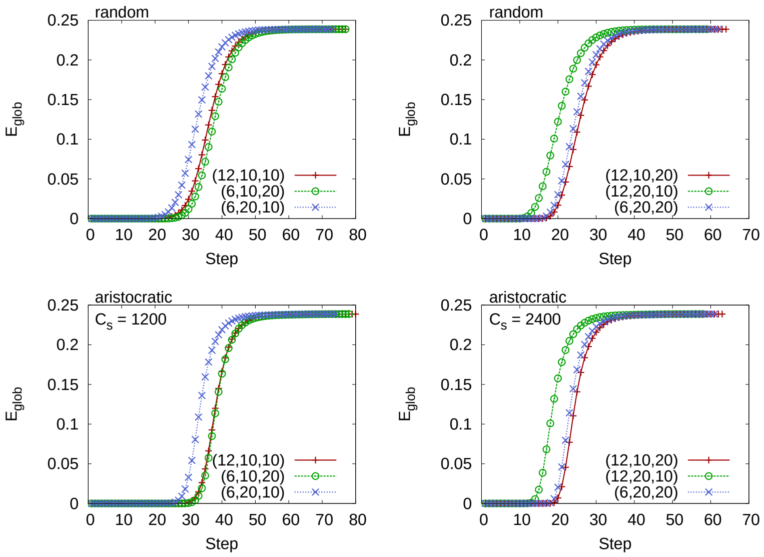


Figure 5.10: The figures describe the benefit of higher *attractiveness* for the same cost configurations of Communities on line social network. In particular, we selected *broadcast* model, random and aristocratic rules. Two cost levels have been considered: $C_s = 1200$ (left panels) and $C_s = 2400$ (right panels). Configurations $(6, 20, 10)$ and $(12, 20, 10)$ outperform the others and in this case network efficiency will start to increase earlier regardless of the growing rule of the users' network.

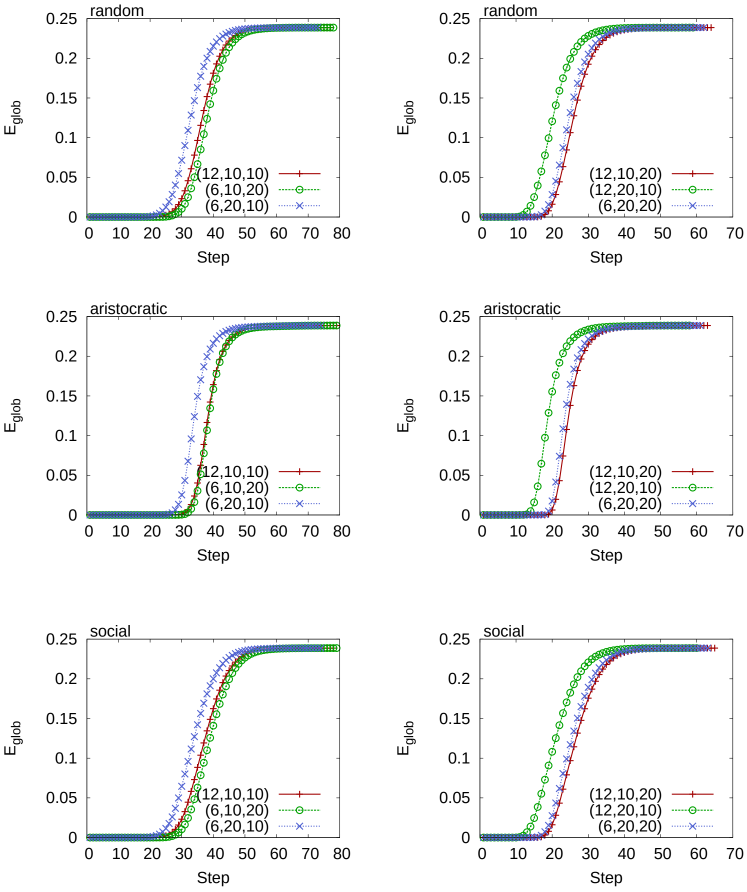


Figure 5.11: Comparison between same cost configurations of Communities on line social network. We consider cost $C_s = 1200$ (left panels), $C_s = 2400$ (right panels) and random, aristocratic and social rules. All plots refer to *word of mouth* model. We clearly see that network efficiency increases faster in configurations that have an higher value of attractiveness, no matter what cost level or rule has been selected.

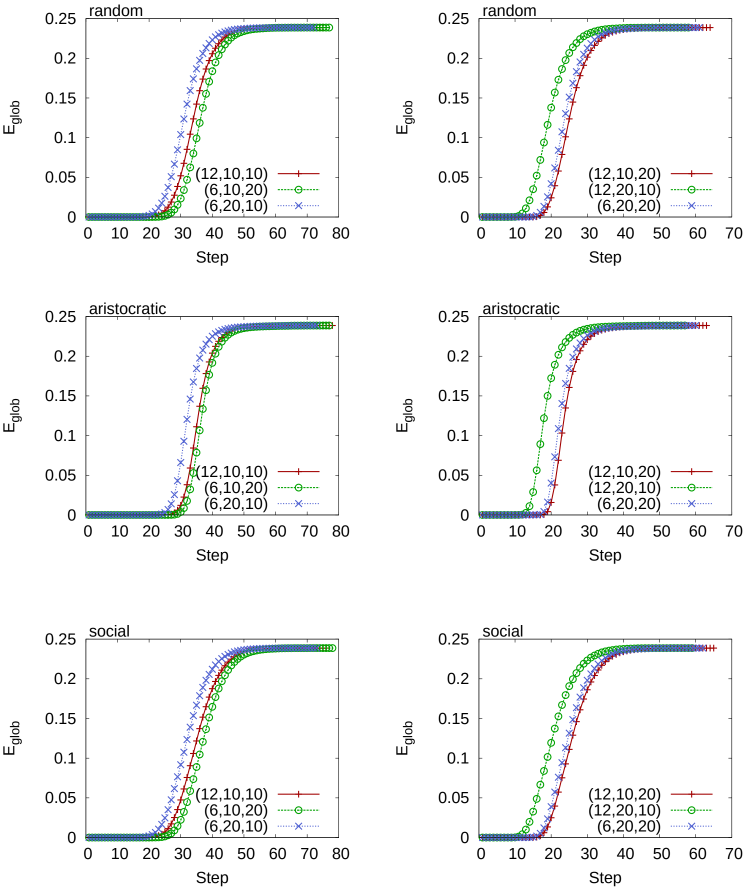
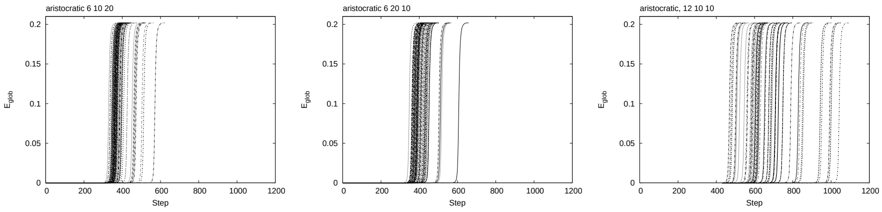
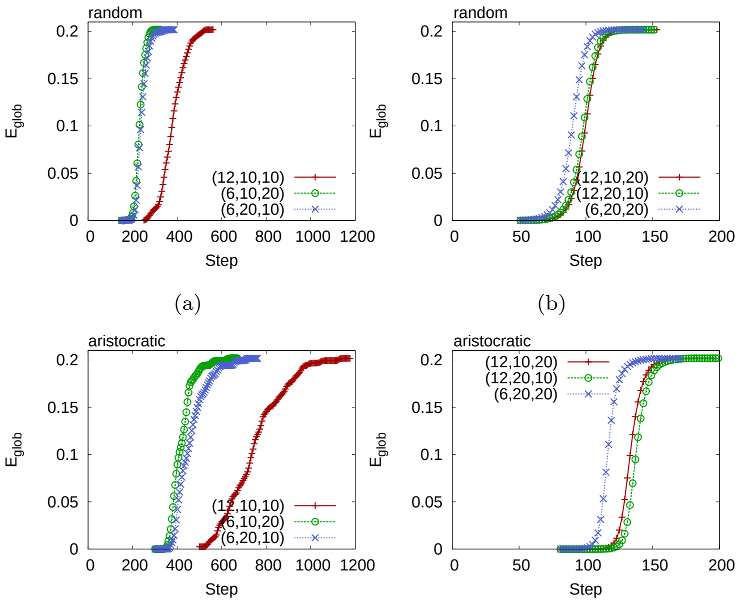


Figure 5.12: Accelerated analysis with sirens, random and aristocratic, social rules, *preferential*, for Communities social network with cost $C_s = 1200$ (left panels) and $C_s = 2400$ (right panels).



(a) (b) (c)

Figure 5.13: Effect of multiple runs of simulations on the same VT dataset.



(a)

(b)

(c)

(d)

Figure 5.14: Behavior of the network's E_{glob} with two different cost levels: $C_s = 1200$ (left panels) and $C_s = 2400$ (right panels) for VirtualTourist social network, *broadcast model*. In total, six configurations are considered. The one that has higher attractiveness is the favored one because can reach the efficiency of the original network faster than others.

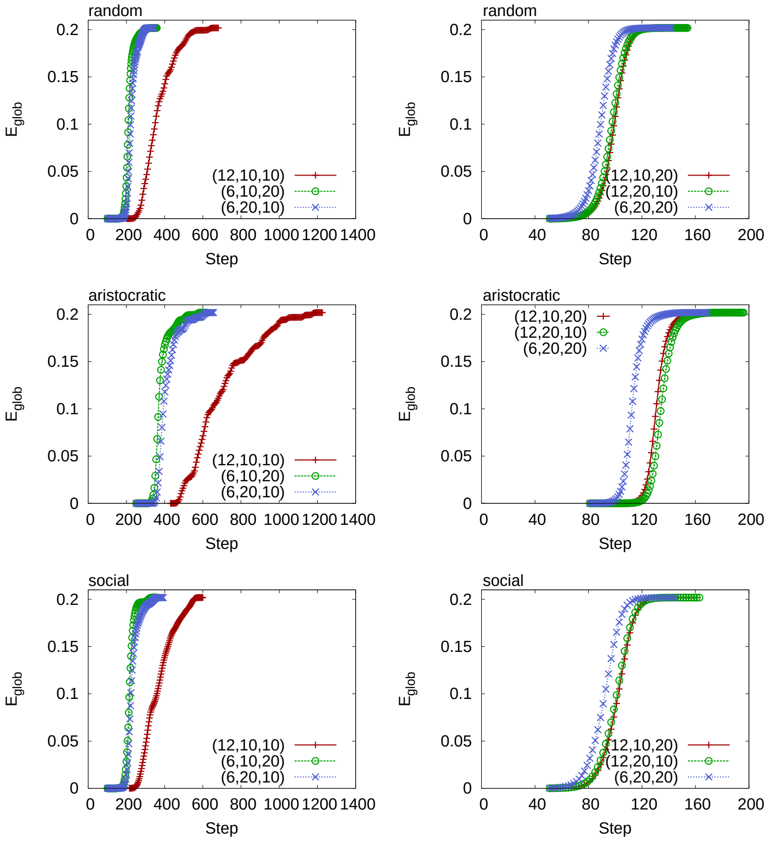


Figure 5.15: Same cost configurations for VirtualTourist, $C_s = 1200$ (left panels) and $C_s = 2400$ (right panels), *word of mouth model*. For each cost C_s, three configurations are then considered. In all experiments, the configuration that performs better is the one that has fewer sirens and higher attractiveness (or equivalently that last more). In accordance with the results of accelerated analysis with no sirens (figure 5.7), random and social rules attain the target efficiency in fewer steps than aristocratic.

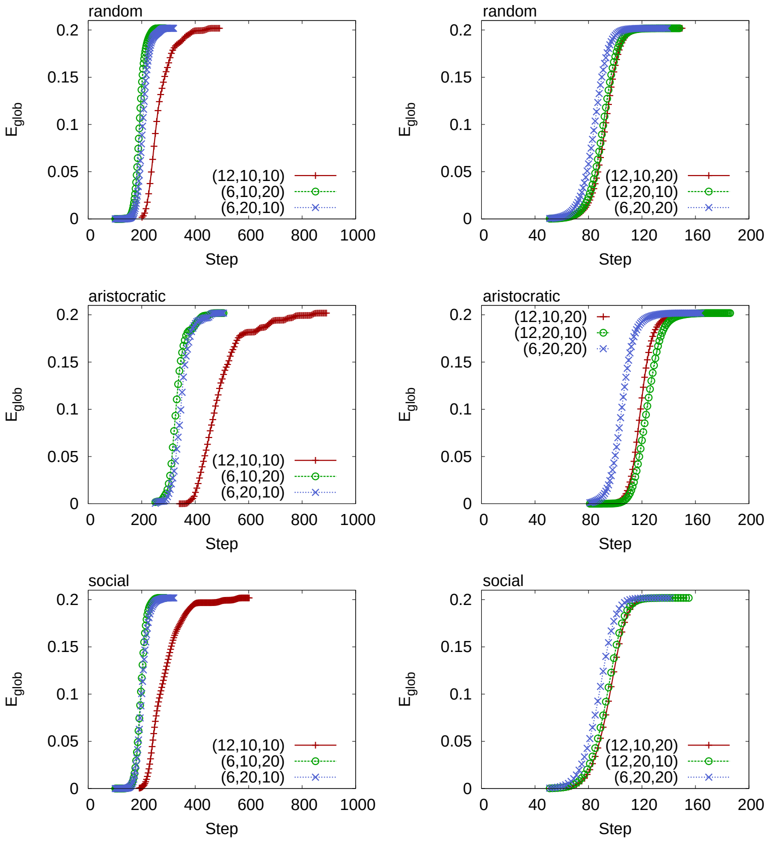


Figure 5.16: Accelerated analysis with sirens, random, aristocratic and social rules, *preferential* model, in VirtualTourist social network with cost equal to $C_s = 1200$ (left panels) and $C_s = 2400$ (right panels). We clearly see that the configurations that have higher attractiveness reach faster the target efficiency, regardless of the users' growing rules.

Q2: Parameters variation. In the following experiments, we investigate the effects of parameters' variation in configurations. In particular, we fixed the number of sirens ($m = 6$ and $m = 12$) and check the performance of other configurations compared to the baseline (that are $(6, 10, 10)$ and $(12, 10, 10)$ respectively). The plots presented are grouped by network: figure 5.17, 5.18 and 5.19 show the results obtained for Communities by varying the dynamics with which the sirens connect to users, namely broadcast, word of mouth and preferential sirens' dynamics. Conversely, figures 5.20, 5.21 and 5.22 show the results of VirtualTourist.

All the plots presented in this section clearly show that increasing C_s results in shrinking times to obtain the target efficiency. This is a very interesting result because it confirms the effectiveness of employing more sirens in order to boost network efficiency and in particular to lower the threshold after which the connectivity spreads all over the network. Indeed, other interesting observations could be made. We note that even though increasing C_s always triggers a broader connectivity distribution, the benefit is not proportional to C_s. For instance, quadrupling C_s, the simulated time shrinks less than four times. The question Q3 will account for quantitatively define this benefit. Remarkably, the observed effects are universal in the sense that they hold regardless of network and rule considered suggesting their validity in a wide class of social networks.

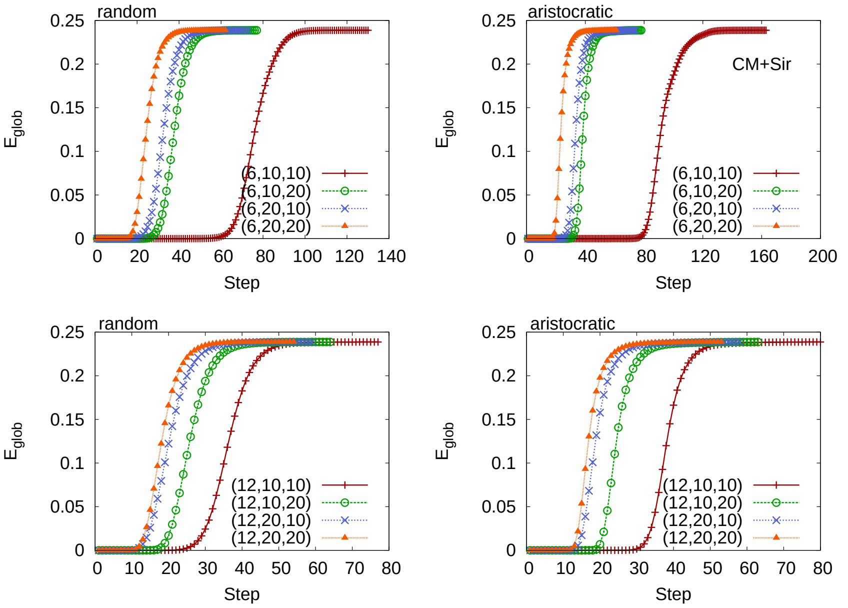


Figure 5.17: Accelerated analysis with sirens, *broadcast model* (for sirens' dynamics), Communities on line social network, fixing $m = 6$ (top panels) and $m = 12$ (bottom panels). Three cost levels are then considered for each plot.

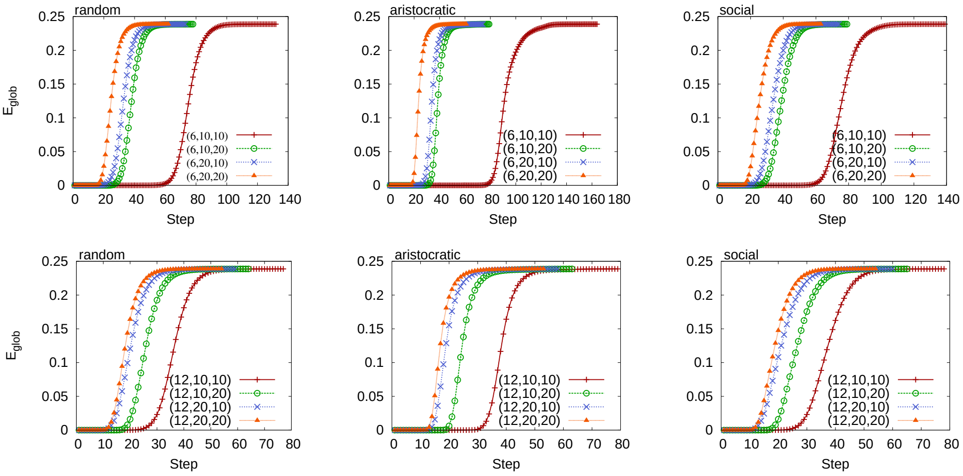


Figure 5.18: Accelerated analysis with sirens fixing $m = 6$ (top plots) and $m = 12$ (bottom plots), random, aristocratic and social rules, *word of mouth model* model (for sirens' dynamics). Communities on line social network.

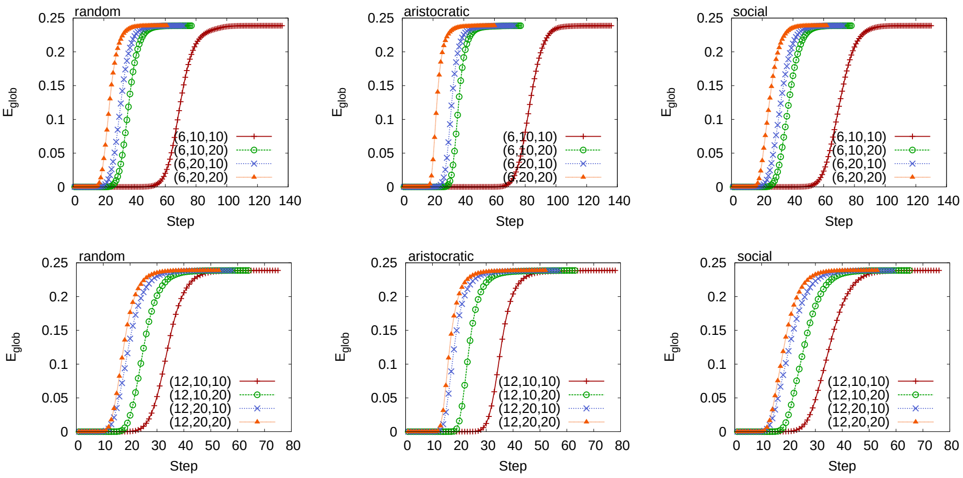


Figure 5.19: Accelerated analysis with sirens fixing the number of sirens to $m = 6$ and $m = 12$, random, aristocratic and social rules. The sirens' dynamics evolve according to the *preferential* model. Communities.com on line social network.

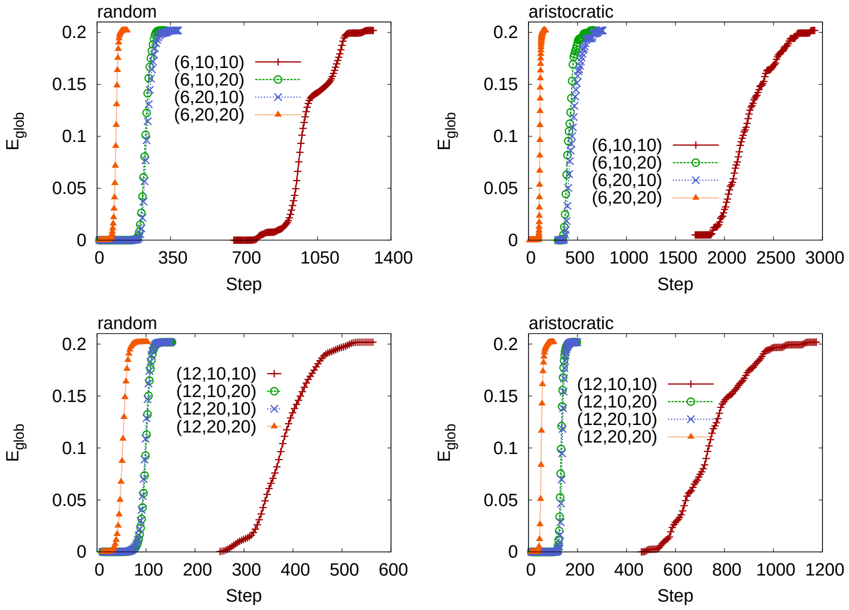


Figure 5.20: Effect on parameters' variation on the configurations fixing the number of sirens to $m = 6$ (top panels) and $m = 12$ (bottom panels). Four cost levels are then considered in each plot, from 600 to 4800. Broadcast model.

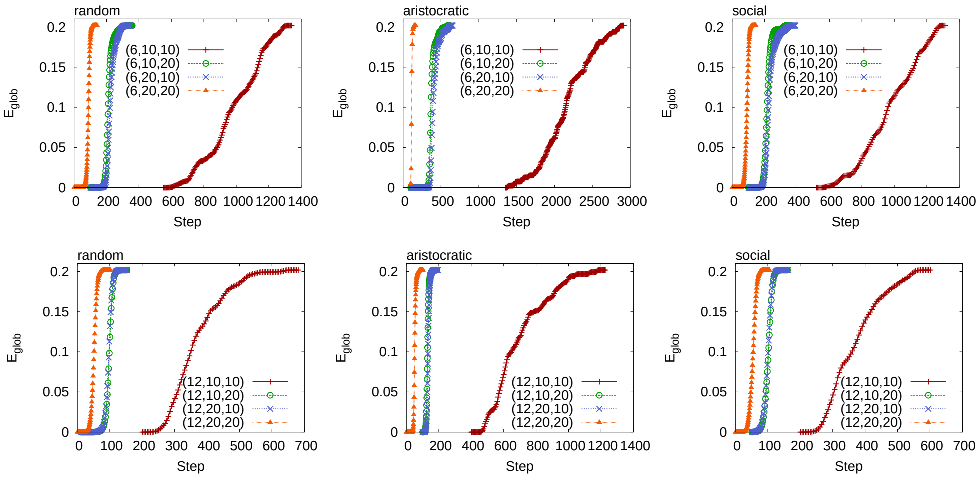


Figure 5.21: Accelerated analysis with sirens fixing $m = 6$ and $m = 12$, random, aristocratic and social rules, *word of mouth model* (for sirens' dynamics), Virtualtourist on line social network.

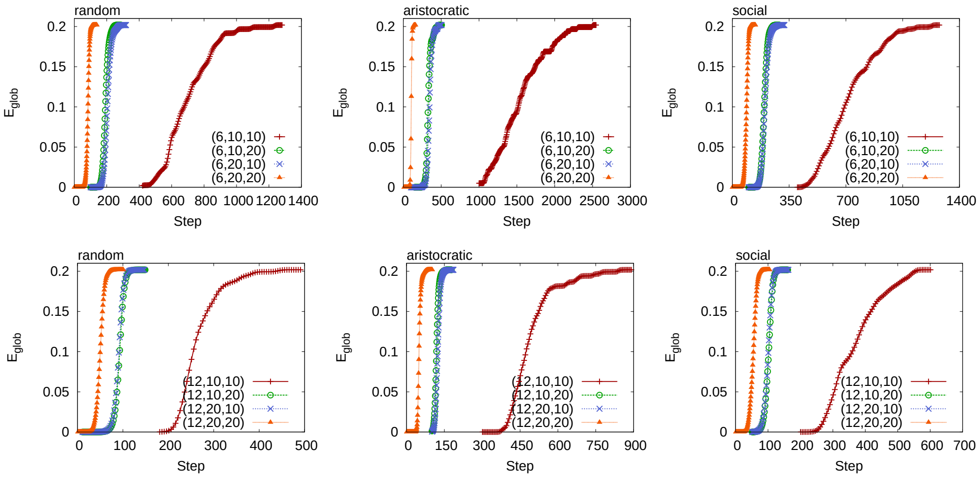


Figure 5.22: Accelerated analysis with sirens fixing $m = 6$ and $m = 12$, with random, aristocratic and social rules, *preferential model* (for sirens' dynamics), Virtualtourist social network.

Q3: Trade-off between the benefit of investing on sirens and the cost. In the plots presented previously, we described how the timespan needed to get the reference efficiency varies according to C_s. The figures 5.23 and 5.24 (rightmost panels) show T_{min} as a function of C_s and this allow to describe more quantitatively the benefit of investing on sirens. In fact, plots clearly show that this is not linear as one might guess, instead it is inversely proportional as C_s. We think this is probably due to system saturation. In other words, the network is not able to respond to high level of exogenous stimuli from sirens resulting in performance that are comparatively similar to those obtained with lower cost configurations.

In order to test whether this finding holds when considering threshold values of efficiency, we investigated the time needed to get half E_{glob} and one third E_{glob}. Surprisingly, as figures show (left most and centermost panels), the benefit is still inversely correlated to C_s. This means that the efficiency growth behavior is quite regular.

In section 5.3.1 we introduced the Eq. (5.4) that accounts for the cost of running a social network and we analytically found that $f(C_s)$ is minimum when Eq. (5.5) holds. Since in real contexts, β, i.e. the hypothetical cost of running a web service, is influenced by many factors, a unique value might not exist. For this reason, we tried many combinations of β that meet Eq. (5.5). To obtain those values we calculate the first discrete derivative of $T_{min}(C_s)$ (different threshold values are then considered). Table 5.7 lists all the values of β we examined in our experiments. Table 5.8 shows the total cost C_t as a function of C_s and for different threshold values. This is a very interesting finding because, once the cost per unit time of a web service is known, our method can estimate the C_s that accounts for the minimum C_t. Indeed, since many configurations can have the same cost, the one that has the higher value of attractiveness will be the one that reach faster the target efficiency (see question Q1).

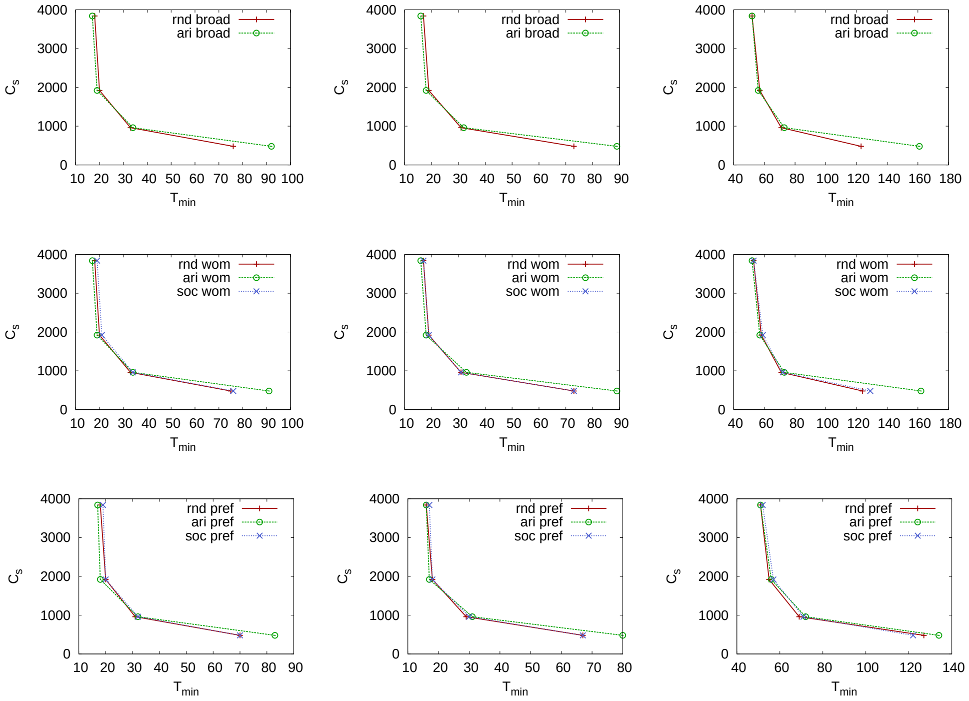


Figure 5.23: Scatter plots between cost C_s and T_{min} in Communities. T_{min} represents the minimum number of steps (in simulated time units) necessary to get target efficiency (E_{glob}). We consider three thresholds: half efficiency (leftmost column), one third (centermost column) and full (rightmost column). Every row represent a different sirens' model namely *broadcast*, *word of mouth* and *preferential*.

CM	bro rnd	bro ari	word rnd	word ari	word soc	pref rnd	pref ari	pref soc
(no sirens)	1381	1930	1381	1930	1328	1381	1930	1328
(6,10,10)	112.92	128.16	111.66	126.95	113.01	106.82	118.48	108.16
(6,10,20)	73.10	73.67	73.80	73.80	74.93	72.45	71.28	74.20
(6,20,10)	68.28	68.29	68.61	68.61	69.90	67.38	66.88	69.08
(6,20,20)	58.14	56.75	58.52	56.65	59.09	57.61	55.02	59.22
(12,10,10)	72.35	73.32	72.07	74.23	73.34	70.51	70.39	72.34
(12,10,20)	60.29	58.69	60.44	58.09	62.01	59.25	56.70	60.70
(12,20,10)	55.08	52.73	55.21	52.92	56.12	53.90	51.90	55.60
(12,20,20)	49.44	47.90	50.05	48.60	50.81	49.23	47.11	50.66

Table 5.5: Summarize of the average T_{min} in all combinations of the number of sirens (m), attractiveness (a) and length of time (d) of accelerated analysis with and without sirens, Communities social network.

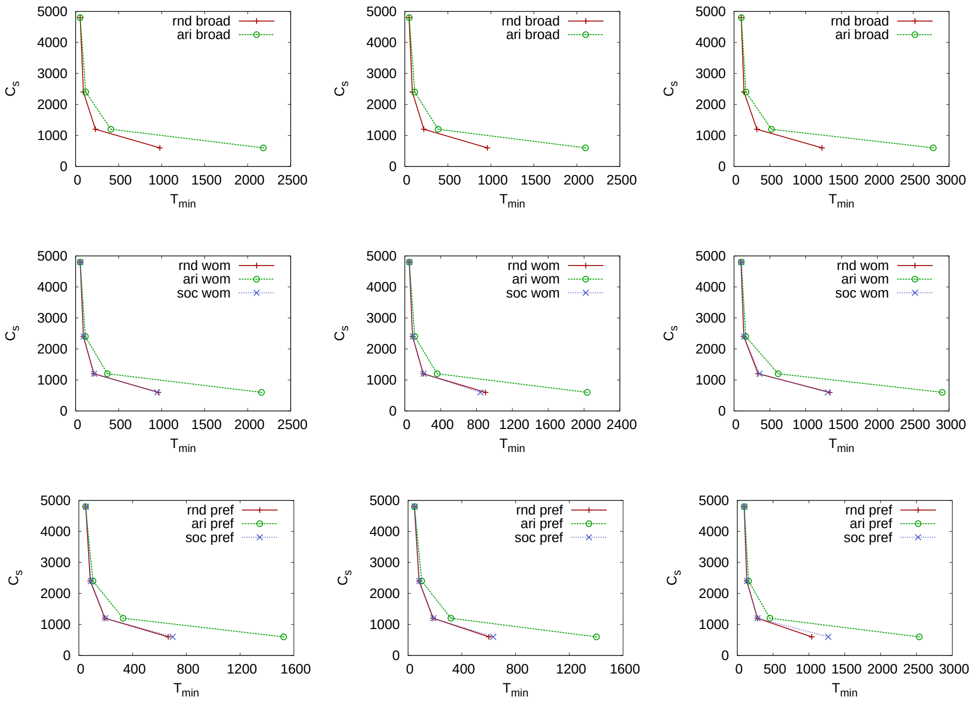


Figure 5.24: Scatter plots between cost C_s and T_{min} in VirtualTourist. T_{min} represent the minimum number of steps (in simulated time units) necessary to get the target efficiency (E_{glob}). We consider three thresholds: half (leftmost column), one third (centermost column) and full (rightmost column) efficiency. Every row represent a different sirens' model namely *broadcast*, *word of mouth* and *preferential*.

VT	bro rnd	bro ari	word rnd	word ari	word soc	pref rnd	pref ari	pref soc
(no sirens)	3120	7496	3120	7496	2987	3120	7496	2987
(6,10,10)	1051.37	2272.84	1041.35	2215.34	999.88	752.16	1607.41	771.30
(6,10,20)	283.12	461.48	262.40	428.12	264.58	243.19	375.16	242.17
(6,20,10)	293.92	512.32	276.52	454.18	276.97	253.02	392.34	245.84
(6,20,20)	137.42	163.84	134.42	156.91	136.71	130.49	148.74	132.45
(12,10,10)	431.78	807.28	432.89	725.23	415.61	311.19	521.40	333.13
(12,10,20)	147.04	180.44	144.94	174.75	146.64	138.53	163.60	140.48
(12,20,10)	145.96	184.40	143.78	178.64	146.17	137.33	168.59	138.82
(12,20,20)	98.46	98.70	96.63	95.89	98.95	94.74	93.39	96.72

Table 5.6: Summarize of the average T_{min} in all combinations of m, a and d of accelerated analysis with and without sirens, VirtualTourist social network.

	$T'_{min}(1200)$	β	$T'_{min}(2400)$	β	$T'_{min}(4800)$	β
$1/1 \cdot E(\mathbf{G})$	-0.05875	17.02	-0.01104	90.56	$-1.145e-3$	872.72
$1/2 \cdot E(\mathbf{G})$	-0.055	18.12	-0.0094	106.38	$-4.1\overline{6}e-4$	2400
$1/3 \cdot E(\mathbf{G})$	-0.054	18.23	-0.0092	108.10	-0.0003	3333

Table 5.7: Summarize of $T'_{min}(C_s)$ and β calculated for different threshold values of $E(\mathbf{G})$ and C_s.

(a)

C_s	C_t
600	**1.732**
1200	**1.732**
2400	2.728
4800	5.110

(b)

C_s	C_t
600	7.248
1200	**4.327**
2400	**4.327**
4800	6.621

(c)

C_s	C_t
600	150.600
1200	71.760
2400	**45.888**
4800	**45.888**

(d)

C_s	C_t
600	**1.700**
1200	**1.700**
2400	2.698
4800	5.085

(e)

C_s	C_t
600	7.126
1200	**4.168**
2400	**4.168**
4800	6.490

(f)

C_s	C_t
600	201.833
1200	92.733
2400	**56.933**
4800	**56.933**

(g)

C_s	C_t
600	**2.314**
1200	**2.314**
2400	3.289
4800	5.642

(h)

C_s	C_t
600	9.724
1200	**7.132**
2400	**7.132**
4800	9.283

(i)

C_s	C_t
600	88.527
1200	58.363
2400	**48.000**
4800	**48.000**

Table 5.8: Total cost $C_t = C_s + \beta \cdot T_{min}(C_s)$ as a function of C_s and different values of β. First row, from left to right: $\beta = 18.12$ (a), 106.38 (b) and 2.400 (c) for half E_{glob}. Second row: $\beta = 18.23$ (d), 108.23 (e) and 3.333 (f) for one third of the efficiency. Third row: $\beta = 17.02$ (h), 90.56 (h) e 872.72 (i) with no threshold at all. Once β is known, our method estimates the best C_s to obtain the minimum cost. For instance, suppose that the cost per unit time β is approximately equal to 90 (with no threshold on E_{glob}), the configurations that achieve the minimum cost are those with $C_s \in [1200, 2400]$. Indeed, since there are many configurations with the same cost level, the one that performs better will have the higher value of attractiveness.

Chapter 6

Conclusions and Future Directions

In this Thesis, we have introduced a novel network analysis that we called multidimensional. This new framework consists of studying the two most important informational axes along with a complex network evolve: space and time. To the best of our knowledge, this is the first work that examines complex networks in that way.

In order to achieve this goal we investigated singularly each dimension by keeping fixed the other. In this way, we addressed two problems: the first one was to understand how systems change their spatial structure when viewed at different abstraction scales (macro and micro detail levels). Conversely, by keeping fixed the space (x, y dimensions) we depicted how they evolve over time, and in particular shedding new lights on what are the basic instincts that trigger networks evolution.

The so-called telescopic analysis, inspired from the human eyes capability to distinguish two points when placed at some distance from a point of view, was devoted to propose a new method that arbitrarily models networks under different levels of abstraction. Its importance stems from the ability of changing nodes' spatial coordinates and connectivity according to some predefined rules. Doing so, we were able to understand what happens to the most important statistical network properties not only when the network detail is high (at micro level) or low (at macro level), but also in between these two extremes. At this point, we were concerned to answer a set of questions such as: which properties are safe to consider after abstracting a network? Which topological structure better preserves system attributes? Are the results of static analysis incomplete because they strongly rely on the detail level with which a network is constructed?

Our experiments are focused on many networks that are embedded in the space, whose evolution is constantly shaped by the surrounding environment. We considered rapid transportation networks (such as subways and airline) and city-based on line social networks. An important finding suggests that complex networks, when observed at finer or coarse-grained level of detail, exhibit statistical features that in many cases are different, meaning that networks characteristics are not stable under the telescopic (or abstraction) process. Because of that, many networks researches are confined to describe only one of all the possible configurations a network could take, showing results that might not be valid for the entire grained spectrum.

The second dimension of complex networks that we studied in this Thesis is the time. To date, many models of network growth have been proposed, especially in the context of social network. For instance, some researches observed that new social ties are driven by randomness. Because of that, some classical models are based on random wiring rule, whereas others are based on preferential attachment (i.e., the rate with which older nodes acquire new links is faster than new nodes) or on triadic closure rule (also known as friend's friend rule, that is, two friends of a person are more likely to know each other compared to two randomly chosen people).

Even though the previous classical growth models are well known and applied in social networks as well as in many other complex networks settings, we found that, there exists a new instinct, never studied before, that is a fundamental ingredient in the process of network evolution and that is also able to considerably boost connectivity. We detected this new phenomenon in the context of on line social communities. In particular, we were able to identify this important instinct by using a new set of special nodes, called "sirens", whose aim was to increase network utilization by establishing new links with existing nodes.

The main questions we raised in this Thesis are the following: Are the sirens beneficial as a way to widely spread the adoption of new on line social systems? How the global network behavior is shaped by employing special nodes? Which of serial (i.e., one edge added at time) and parallel (i.e., the number of edges added varies dynamically according to the current efficiency) model achieve the best performance? How same cost configurations influence network efficiency? How much does it cost using the sirens?

We systematically simulated two on line communities with different sizes and topics demonstrating the effectiveness of sirens to drive social evolution and boost community engagement in general. Indeed, simulations were performed as a function of the configurations. Configurations represent model's growth variables and are: number of sirens, attractiveness and time span of sirens' utilization respectively. We found that at the same cost, the configurations that attain the best results are those with high attractiveness, regardless of the on line social network considered.

Several other interesting features could be explored as a natural extension of this Thesis. The first one, with regard to telescopic analysis, is to expand the types of networks and the perturbations accounted for in the simulations. For instance, it will be interesting to know whether force-based layouts such as Kamada-Kawai [62] or Fruchterman-Reingold [48] can influence the overall outcome of the telescopic abstraction process. This could greatly increase the understanding of the telescopic effect on different settings. Secondly, in the context of time dimension, it could be of great importance to investigate whether the found instinct is present in other than on line social networks. Another future direction that might be important in the context of social network modeling with special nodes is to consider other statistical features such as local efficiency, assortativity, centrality, etc to develop a more complete insight on network evolution.

Here, we introduced a new approach that study singularly two important dimensions of network, making progress toward a network analysis framework that not only is able to understand the single dimensions when taken apart (minimalist approach), but also infer good results at global perspective. Interestingly, this can opens several new perspectives in the understanding of the general behavior of complex systems.

In conclusion, we believe that this novel approach of studying complex networks by investigating the two most important informative axes, time and space, has a great potential to impact the way the future complex system analysis will be.

Acknowledgments

The present Thesis describes the results of the research I performed at University of Padua from January 2010 to December 2012 under the supervision of Prof. Massimo Marchiori.

I would like first to express my acknowledgments to Massimo Marchiori. Thanks to his guidance throughout these three years, I could discover the pleasure of doing research on both theoretical and more applied problems. His availability and his patience have been extremely important for me and our frequent discussions have been an invaluable source of inspiration.

During this Thesis, I had the opportunity to collaborate with Filippo Menczer at Indiana University Bloomington and in particular with the Scholarometer team, Jasleen Kaur, Xiaoling Sun and Snehal Patil and many others at the School of Informatics and Computing. I never met such great people in my life, honest, professional and good friend. I could not omit to mention the foosball guys at IU.

I'd like to thank Jesús Gómez-Gardeñes who accepted to be the expert for this Thesis.

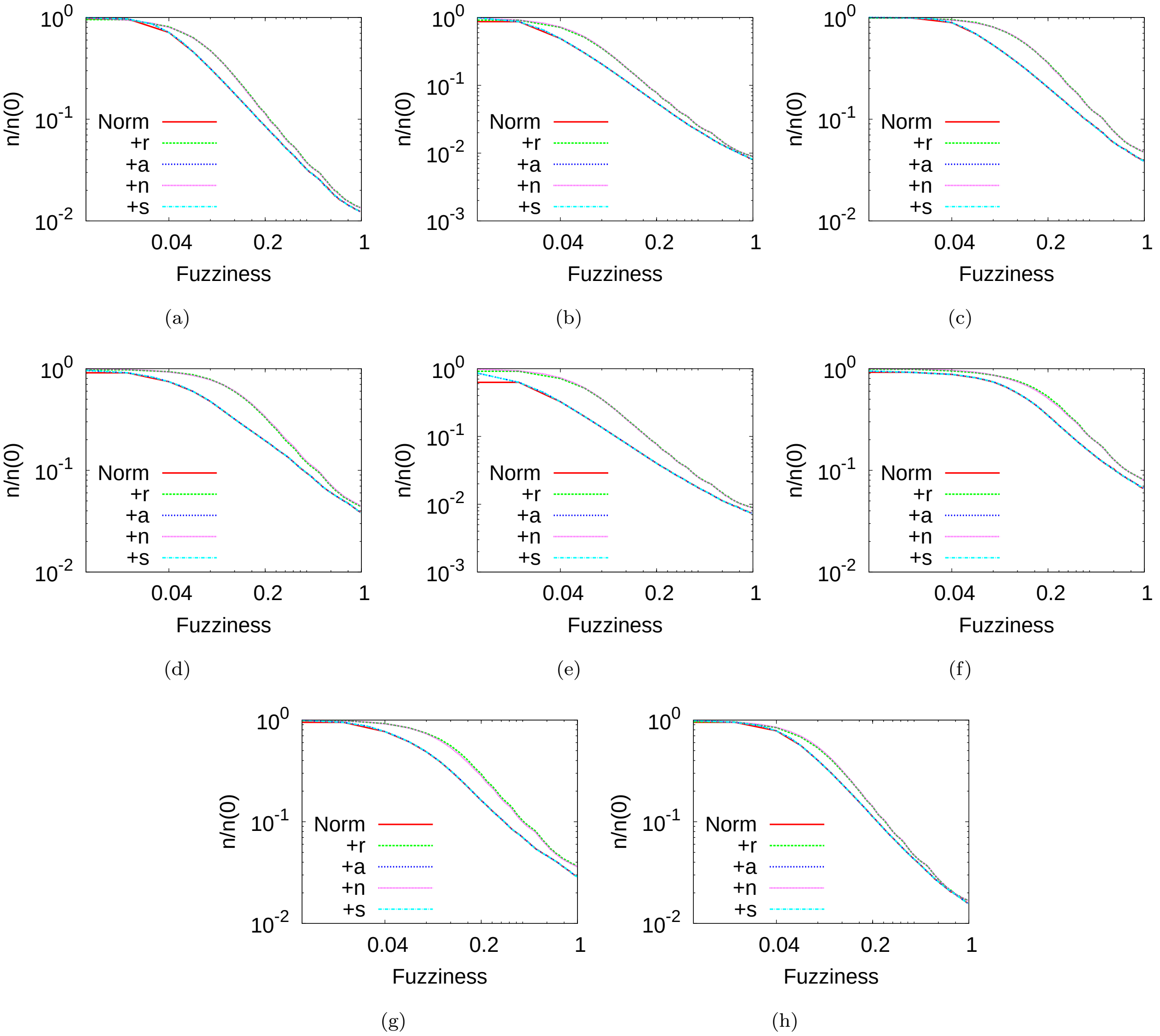
More personally, I want express my recognition to my parents that from the beginning trusted my thought and ideas; I could not have reached this stage in my life without their encouragement and patience. Thanks to all friends, colleagues and people that I met during this PhD program.

Appendix A

Supplementary material for space dimension

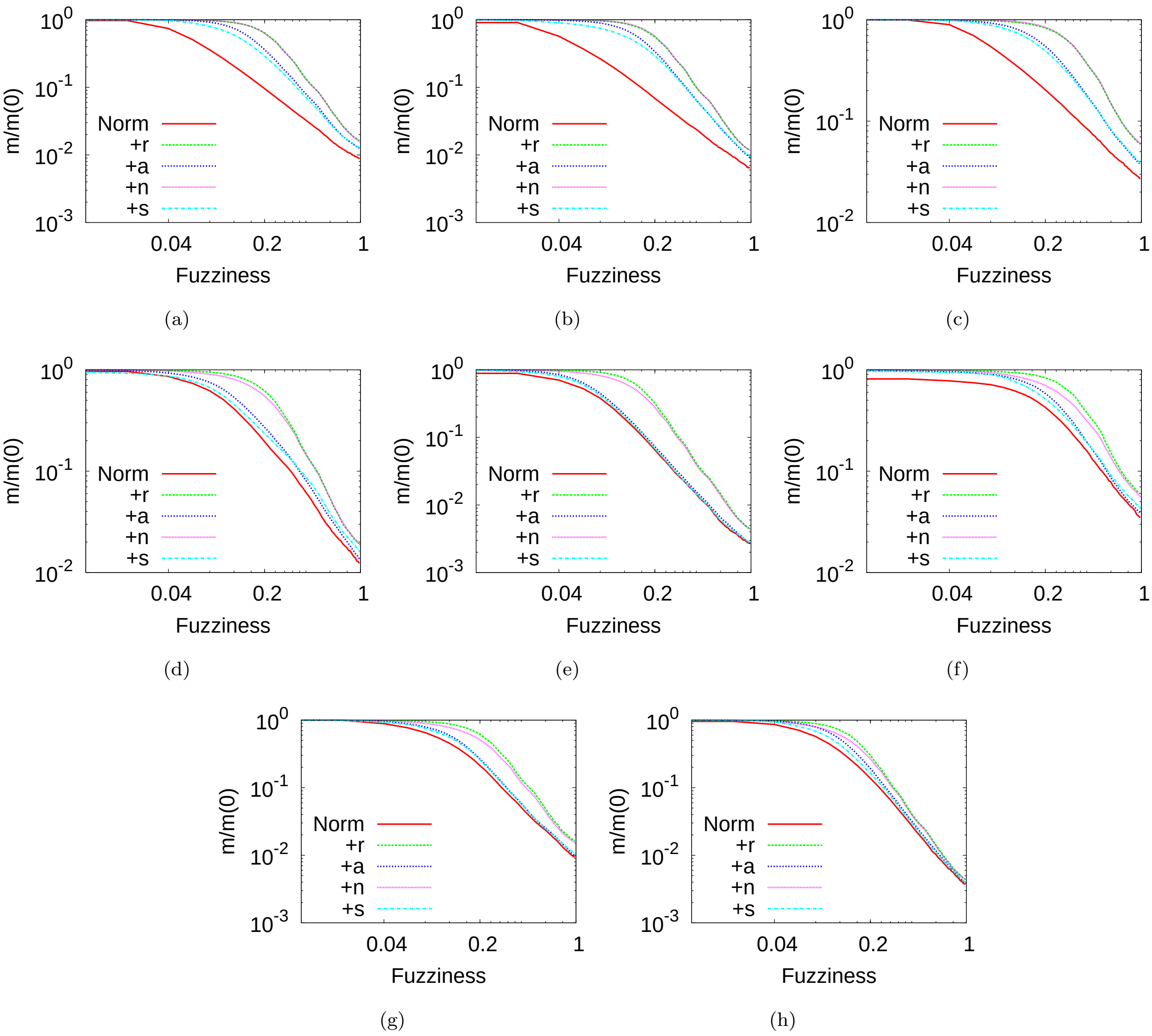
Our novel analysis, called multidimensional, consists of separately study how a complex network evolves along the two most important dimensions: time and space. We were able to identify fundamental facts that were not yet known. Telescopic analysis consists of fixing the time axes and study how networks' statistical properties varies while modifying the spatial structure. In particular, we were able to successfully implement a method that can create networks at different abstraction levels.

This section contains the outcome of the experiments of the telescopic analysis. In particular, we applied the algorithm to transportation networks such as subway and airline and to city-based online social networks. We considered four subways, such as Boston, Paris, New York and Milan; the US airline network formed by non-stop flights and airports; the VirtualTourist city-based online communities of Italy, United Kingdom, India, Australia and the Netherlands. In section § 5.4.2 we presented the results of Boston subway and Australia online community.



(a) (b) (c) (d) (e) (f) (g) (h)

Figure A.1: Log-log plots of the number of nodes (normalized by the baseline, i.e. at $f = 0$) as a function of f in subway networks such as Paris (a), Nyc (b), Milan (c) and in VirtualTourist city-based network of Italy (d), United Kingdom (e), India (f), the Netherlands (g) and airline (h). Straight lines represent the baseline, that is, no perturbation is applied; in the other cases, different perturbations on nodes (+n), edges (+a, +s) and on both (+r) were applied. In all networks we found a slower decrease in those specific perturbations that shuffle nodes (+r and +n). This is arguably due to the scattered nodes' positions in the space and this means that nodes will be visible in more abstraction level and will be collapsed later in the process.

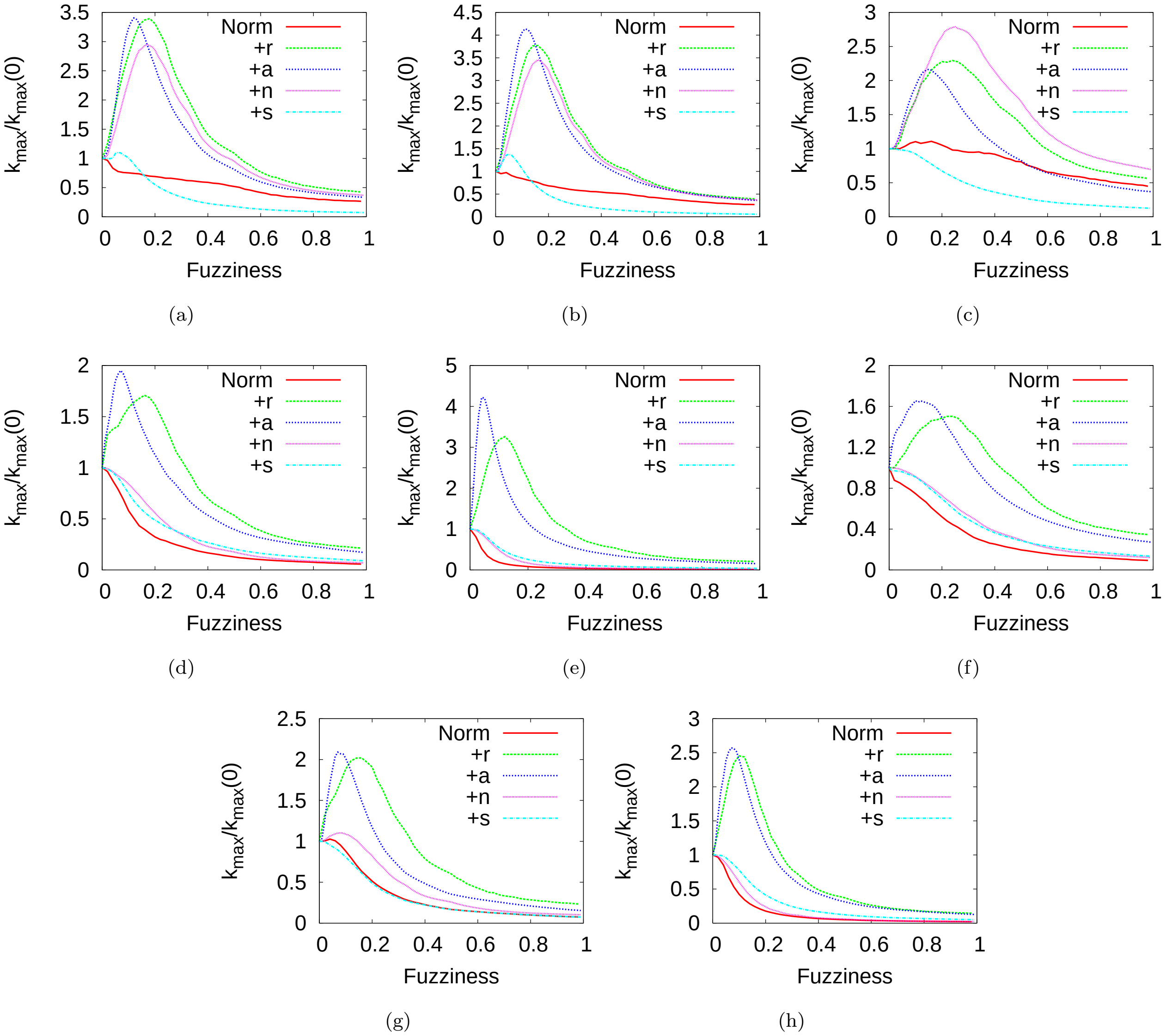


(a) (b) (c)

(d) (e) (f)

(g) (h)

Figure A.2: Log-log plots of edges (normalized by the baseline, i.e. at $f = 0$) as a function of f in rapid transportation systems of Paris (a), Nyc (b) and Milan (c) and in VirtualTourist city-based network of Italy (d), United Kingdom (e), India (f), the Netherlands (g) and airline (h). Straight lines represent the baseline, that is, no perturbation is applied. Four types of perturbations are applied on nodes' positions (+n, +r) and on edges (+a, +s). Clearly, without any perturbation, edges will be collapsed faster due for instance to local clustering (even though in subway networks is low) and in general to some sort of organization in wiring edges. Conversely, any other perturbations considered slow down the rate with which edges are collapsed mainly because of the long range edges created by the randomization process.

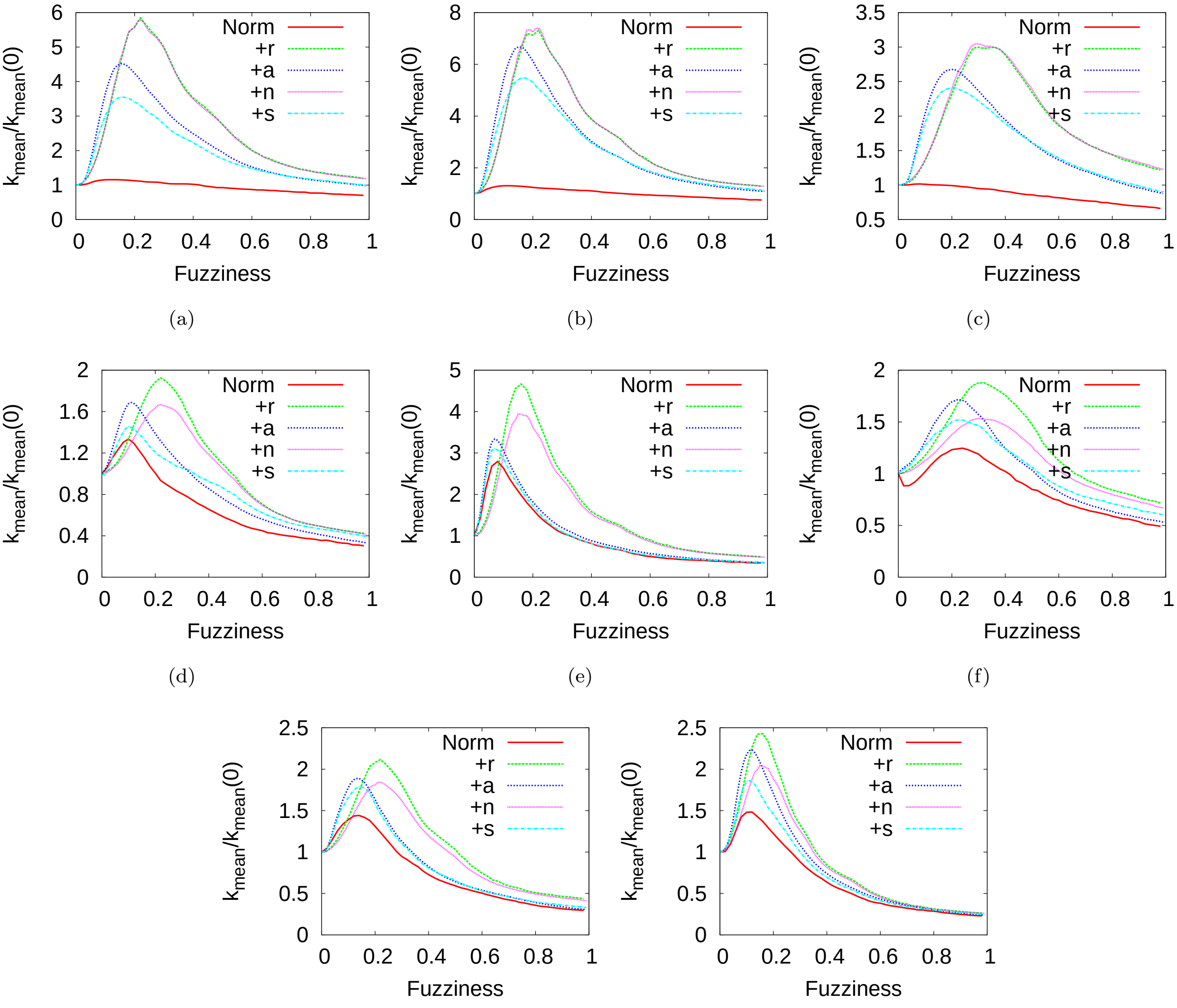


(a) (b) (c)

(d) (e) (f)

(g) (h)

Figure A.3: Normalized k_{max} as a function of f in subway networks such as Paris (a), Nyc (b), Milan (c) and in VirtualTourist city-based network of Italy (d), United Kingdom (e), India (f), the Netherlands (g) and airline (h). Values are normalized by the baseline (straight lines), i.e. at $f = 0$. Interestingly, only the randomized perturbations (+r, +n and +a) produce a peak in the maximum degree, approximately at $f/5$ of the telescopic spectrum. This is evidence that telescopic analysis on subway networks creates hubs (this is also confirmed by results on *kmean* and standard deviation in the figures that follow).



(a)

(b)

(c)

(d)

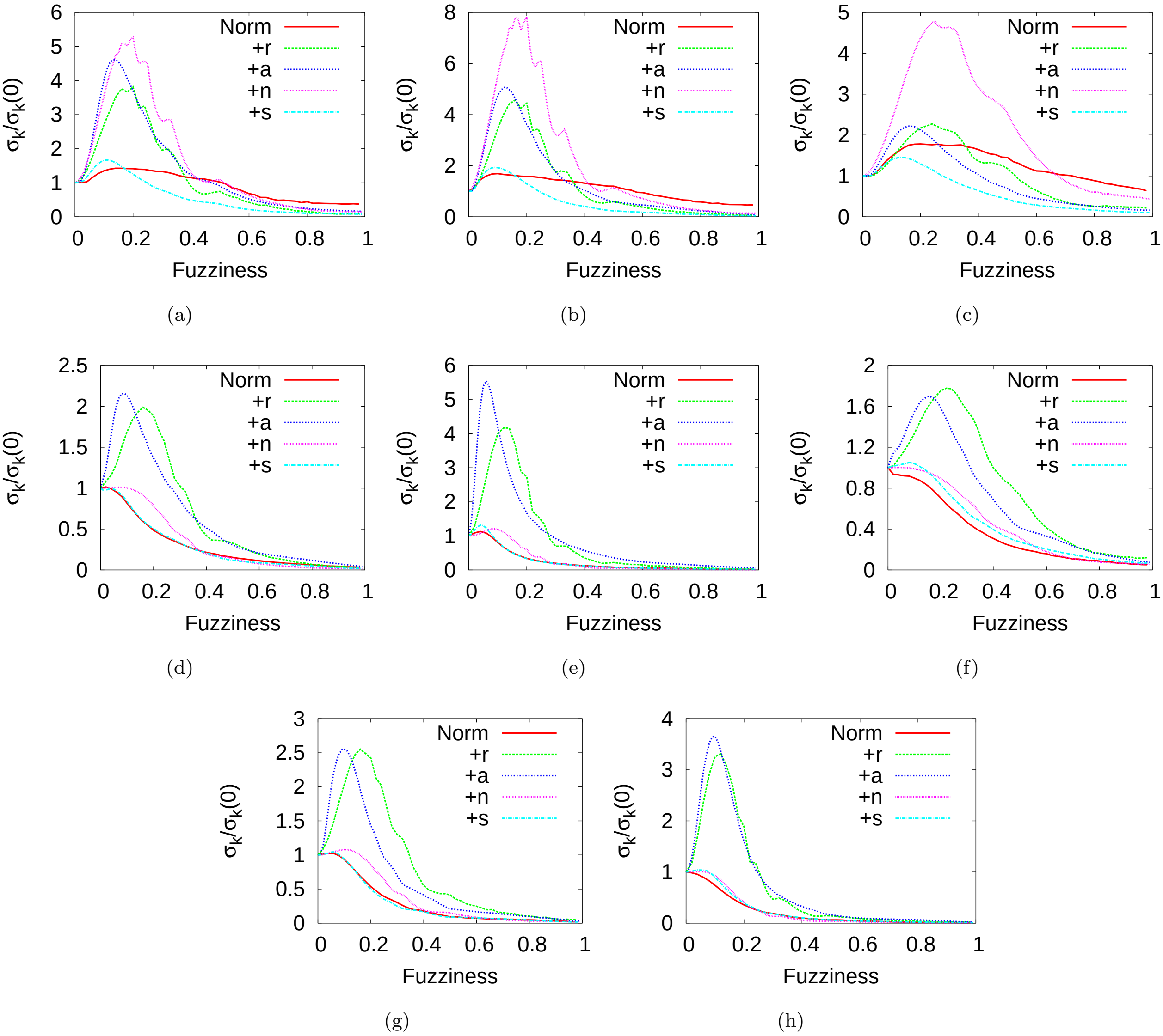
(e)

(f)

(g)

(h)

Figure A.4: Normalized k_{mean} as a function of f (the normalization is carried out by dividing the values by the baseline at $f = 0$) in Paris (a), Nyc (b), Milan (c) rapid transportation systems and in VirtualTourist city-based networks of Italy (d), United Kingdom (e), India (f), the Netherlands (g) and airline (h). Values are normalized by the baseline (straight lines), i.e. at $f = 0$. Many perturbations applied: those who randomize nodes' positions (+n and +r) and connectivity (+a, +r and +s). We note that any type of rewiring or node shuffling cause an increase immediately as the network is abstracted.

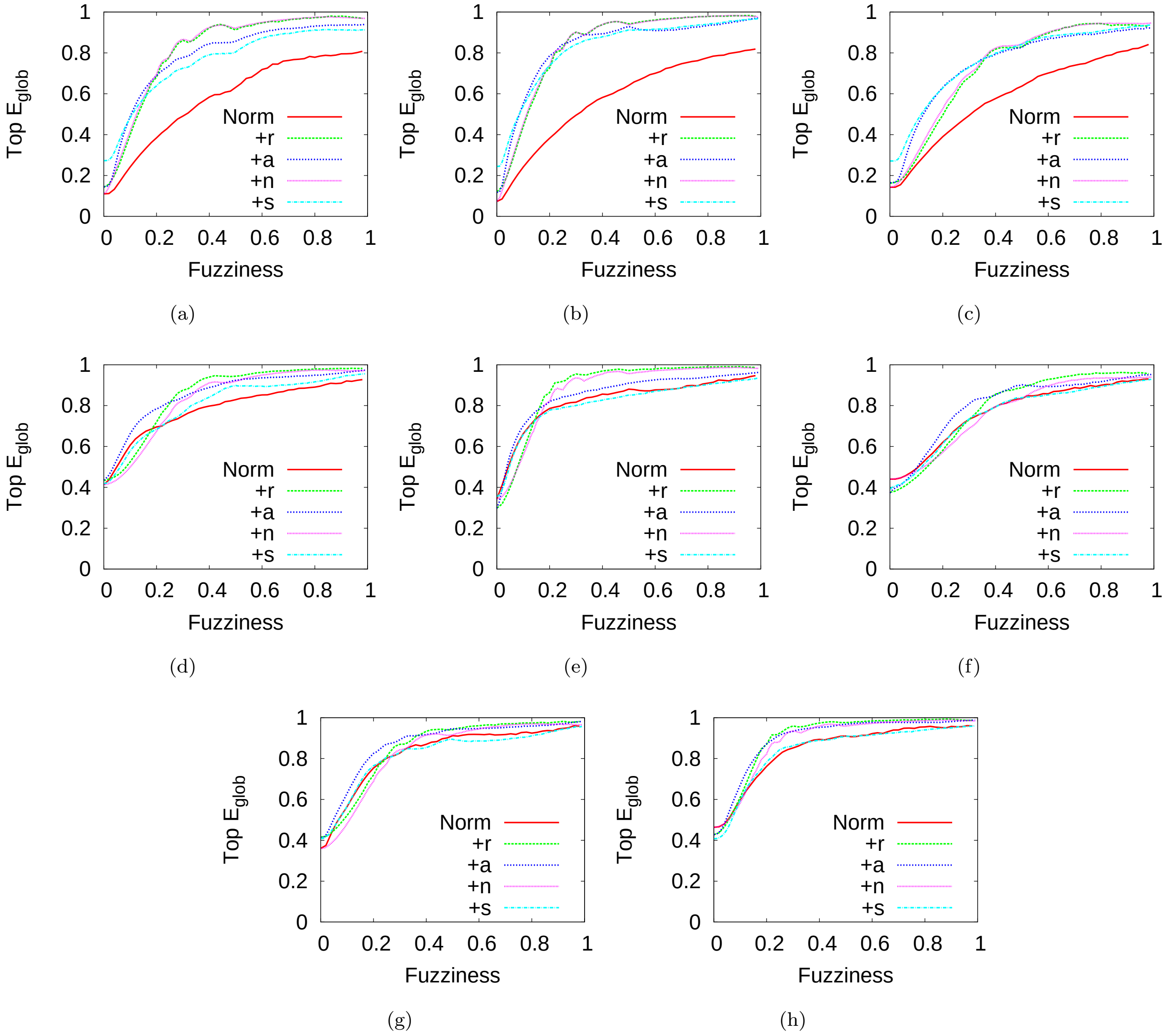


(a) (b) (c)

(d) (e) (f)

(g) (h)

Figure A.5: Standard deviation of the degree as a function of f in subway networks such as Paris (a), Nyc (b), Milan (c) and in VirtualTourist city-based networks of Italy (d), United Kingdom (e), India (f), the Netherlands (g) and airline (h). Straight lines represent the baseline, i.e., when no perturbations applied. This measure is very useful in conjunction with other degree related statistics in order to have a complete perspective on how nodes' degree change while abstracting networks.

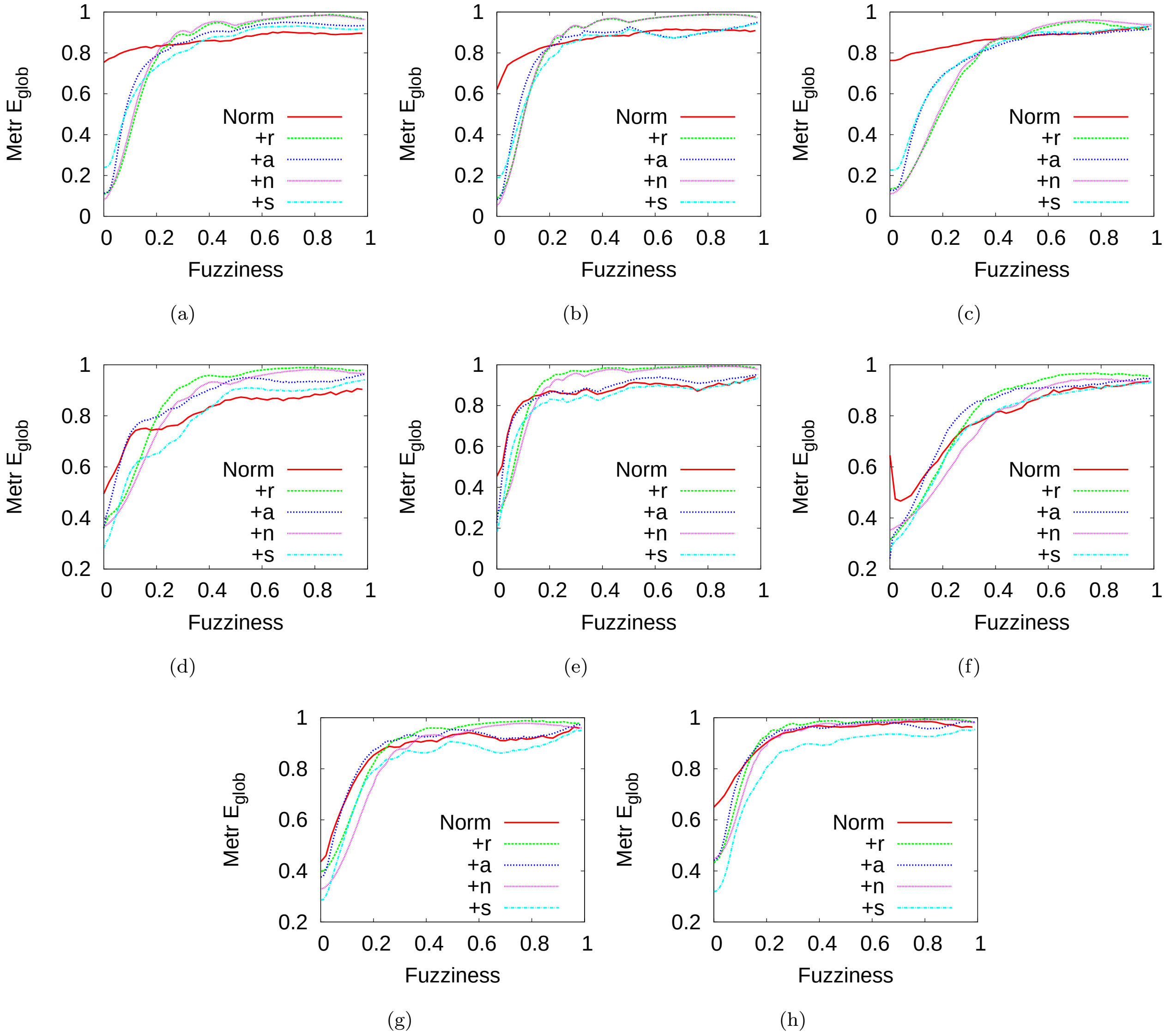


(a) (b) (c)

(d) (e) (f)

(g) (h)

Figure A.6: Topological E_{glob} as a function of f in rapid transportation systems of Paris (a), Nyc (b), Milan (c) and in VirtualTourist city-based networks of Italy (d), United Kingdom (e), India (f), the Netherlands (g) and airline (h). Many perturbations are then considered, +n that shuffles node positions, +a rewires in a random fashion the edges, +r that both shuffles and rewires and +s that creates a small world scale-free structure leaving unchanged the nodes' positions. The figures show an important characteristic of analyzing networks at different granularities. In fact, networks that are poorly connected at micro scale (leftmost points on the plots) become highly efficient at micro scale. Under our framework, we never found that an highly connected network at micro scale become poorly connected at macro scale.

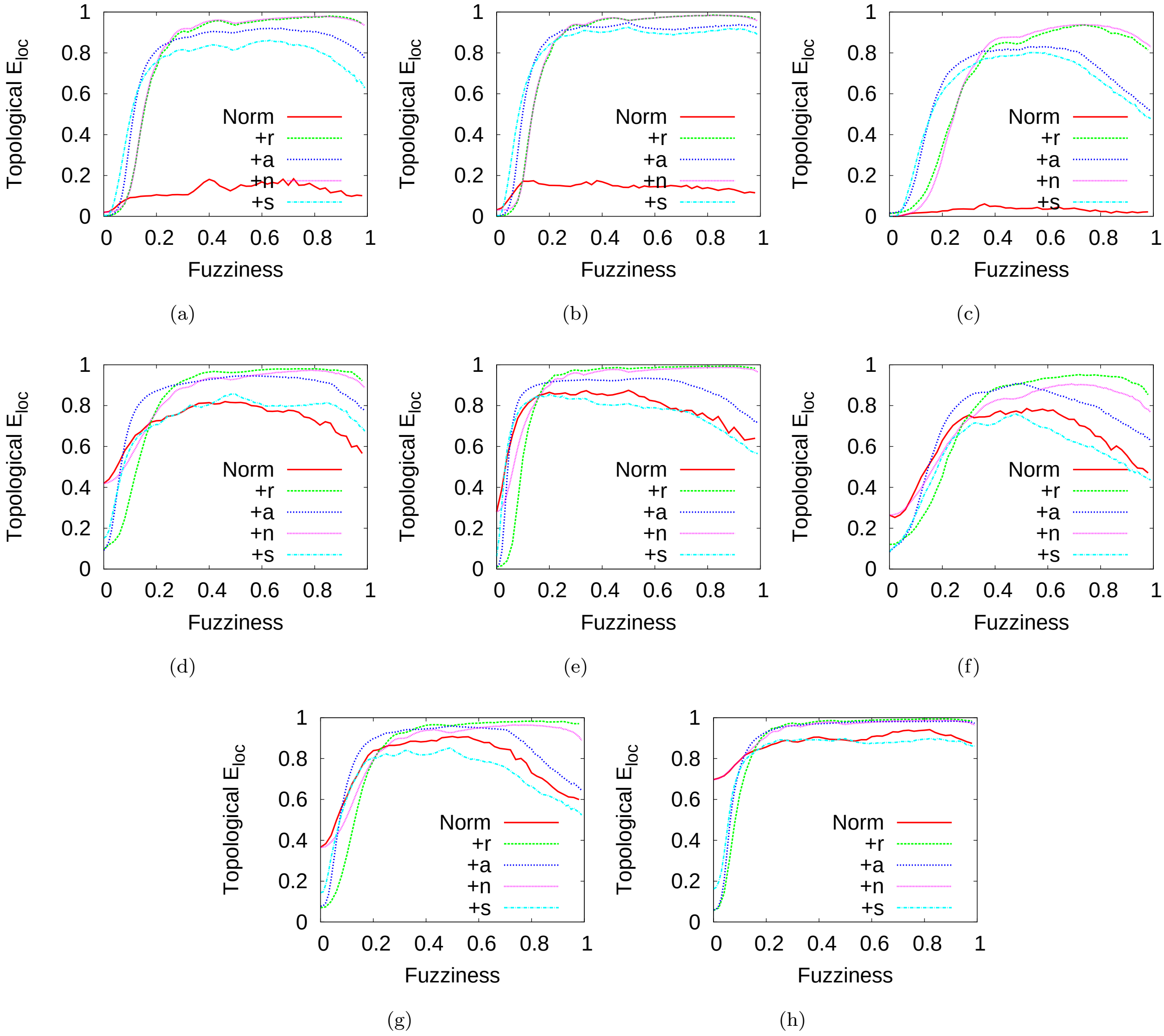


(a) (b) (c)

(d) (e) (f)

(g) (h)

Figure A.7: Metrical E_{glob} normalized as a function of f in rapid transportation systems such as Paris (a), Nyc (b) and Milan (c) and in VirtualTourist city-based networks of Italy (d), United Kingdom (e), India (f), the Netherlands (g) and airline (h). Straight lines represent the baseline, i.e., when no perturbations applied. Different perturbations are then considered in order to see how randomization affects the outcome of the telescopic analysis and compare them to the baseline (i.e. the straight lines, where no perturbations at all are applied). As expected, the randomization on all dimensions destroys the connectivity (leftmost points on the figures). However, on the other side of the telescopic spectrum, the original randomized networks are indistinguishable because of the high level of fuzziness causing high uncertainty on the network structure.



(a)

(b)

(c)

(d)

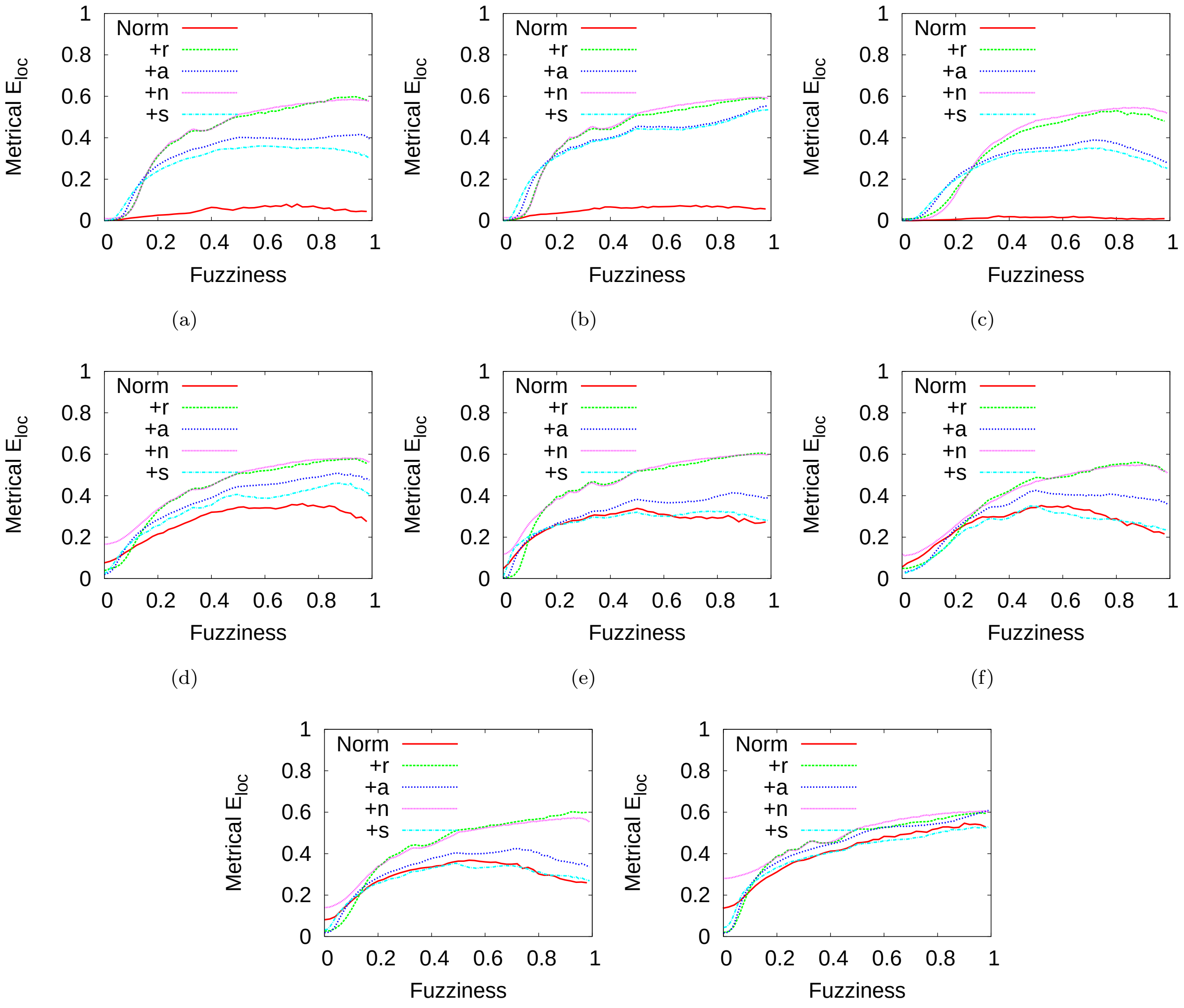
(e)

(f)

(g)

(h)

Figure A.8: Topological E_{loc} as a function of fuzziness in subway networks of Paris (a), Nyc (b), Milan (c) and in VirtualTourist city-based networks of Italy (d), United Kingdom (e), India (f), the Netherlands (g) and airline (h). Straight lines represent the baseline, i.e., when no perturbations applied. Randomization on nodes' positions (+n), link connectivity (+a, +s) and on both (+r) are then considered. Even though not all perturbations alter local efficiency at micro scale, the spatial and topological modifications affect considerably the way networks are perceived at macro scale. We note that small world/scale-free networks are more sensible to perturbations compared to subway systems. Conversely, when viewed at macro scale, they tend to look like perturbed networks.

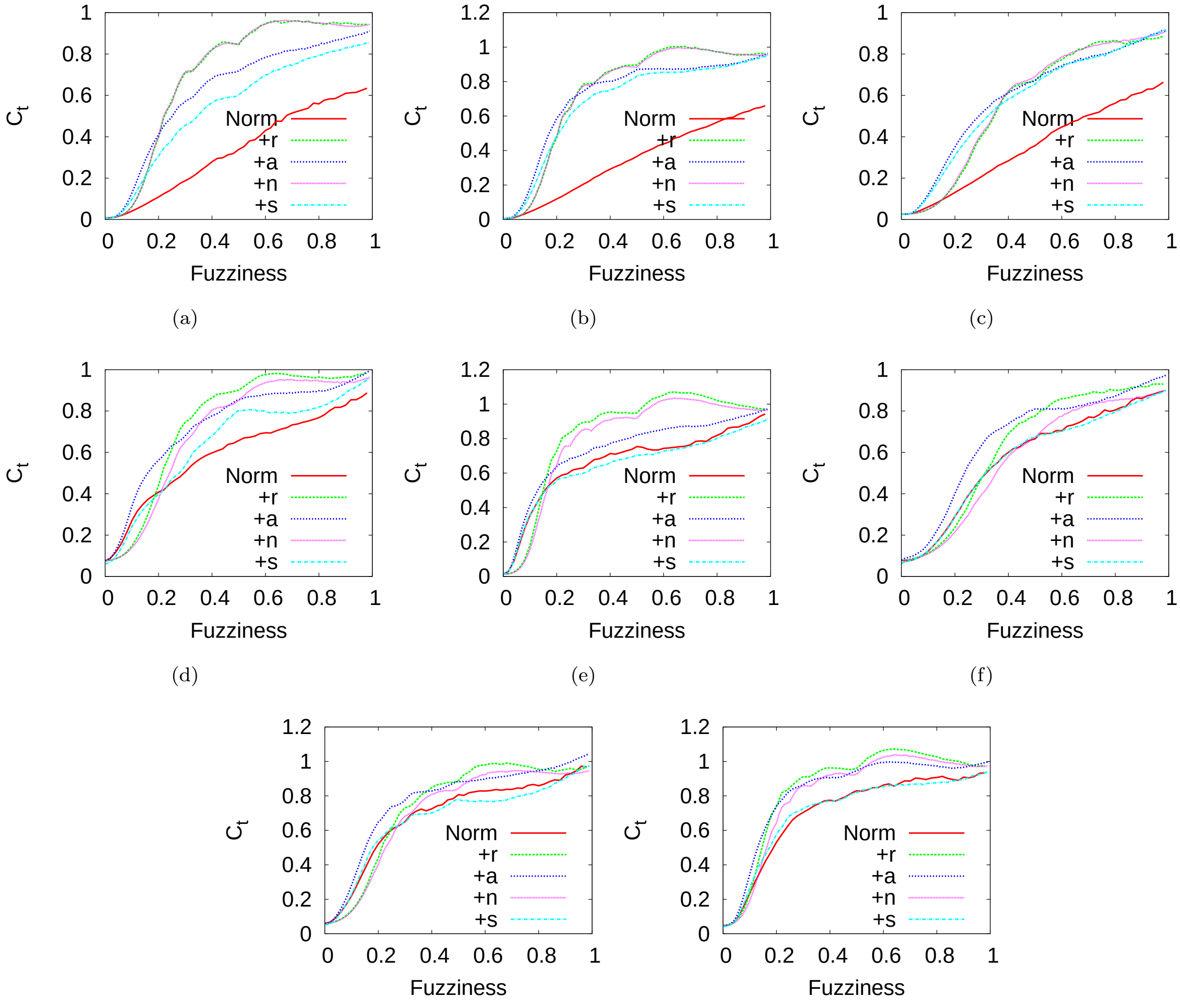


(a) (b) (c)

(d) (e) (f)

(g) (h)

Figure A.9: Metrical E_{loc} as a function of f in subway networks of Paris (a), Nyc (b), Milan (c) and in VirtualTourist city-based networks of Italy (d), United Kingdom (e), India (f), the Netherlands (g) and the US airline network (h). Straight lines represent the baselines. Perturbations on nodes' positions (+n,+r) and on edges (+a,+s) are then considered. Just like topological E_{loc}, at macro scale the effect of perturbation is modest and randomized and real networks perform the same. However, as the abstraction process starts, small world/scale-free networks do not exhibit a different behavior compared to randomized versions.



(a)

(b)

(c)

(d)

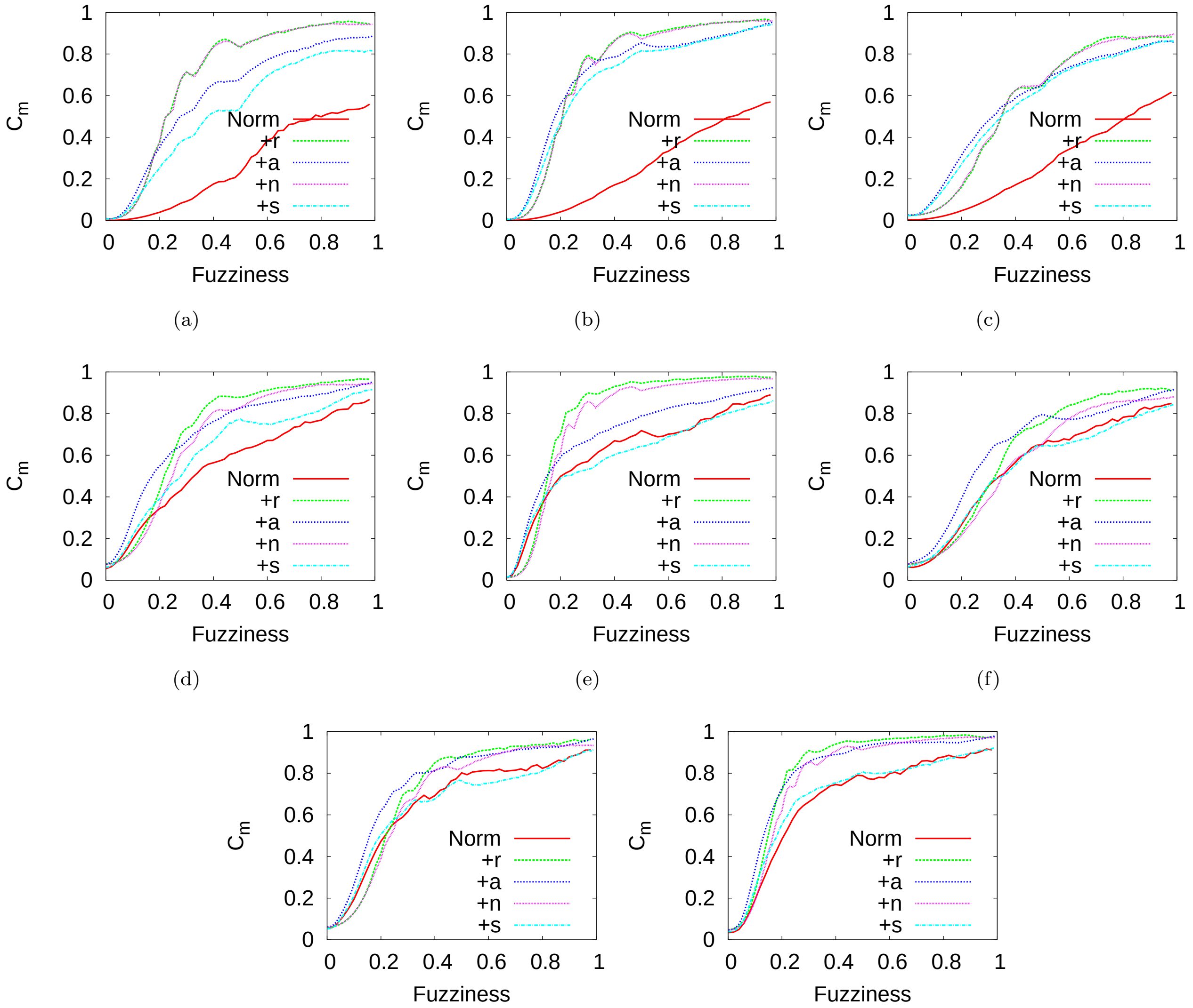
(e)

(f)

(g)

(h)

Figure A.10: Topological cost as a function of f in subway networks such as Paris (a), Nyc (b), Milan (c) and in VirtualTourist city-based networks of Italy (d), United Kingdom (e), India (f), the Netherlands (g) and the US airline network (h). Straight lines represent the baselines. Different perturbations on nodes' positions (+n, +r) and on edges (+a,+s) are then applied. Generally speaking, these plots suggest that topological cost is higher at macro level compared to micro. Indeed, c_t of small world scale-free networks in the telescopic spectrum is comparable to the randomized versions. However, subway networks have a different behavior in the telescopic spectrum, due to the high efficient topological structure that allow, at macro level, to have high E_{glob} and low cost.



(a)

(b)

(c)

(d)

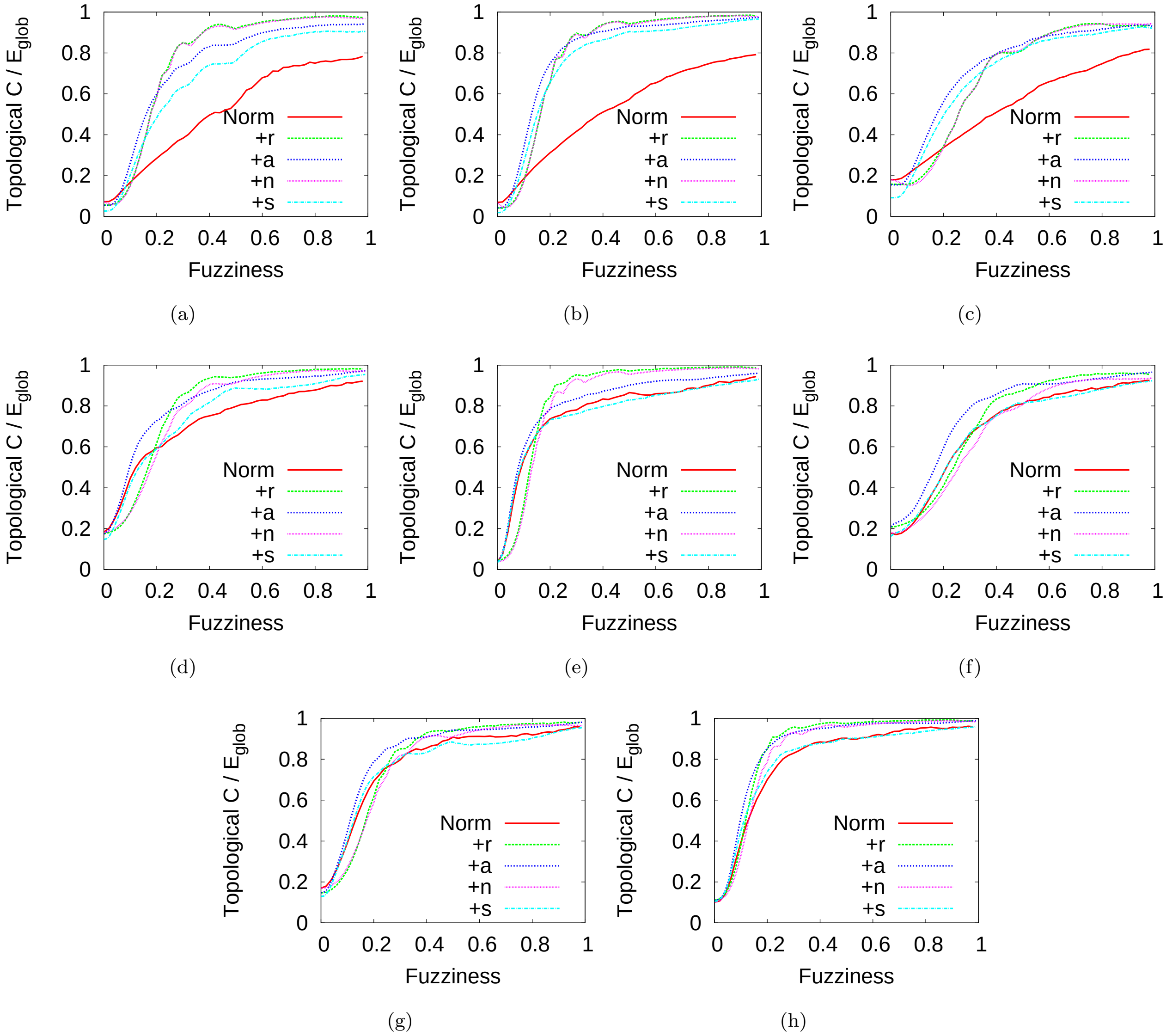
(e)

(f)

(g)

(h)

Figure A.11: Metrical cost as a function of f in subway networks such as Paris (a), Nyc (b), Milan (c) and in VirtualTourist city-based networks of Italy (d), United Kingdom (e), India (f), the Netherlands (g) and the US airline network (h). Straight lines represent the baselines. Different perturbations on nodes' positions (+n, +r) and on edges (+a,+s) are then applied.



(a)

(b)

(c)

(d)

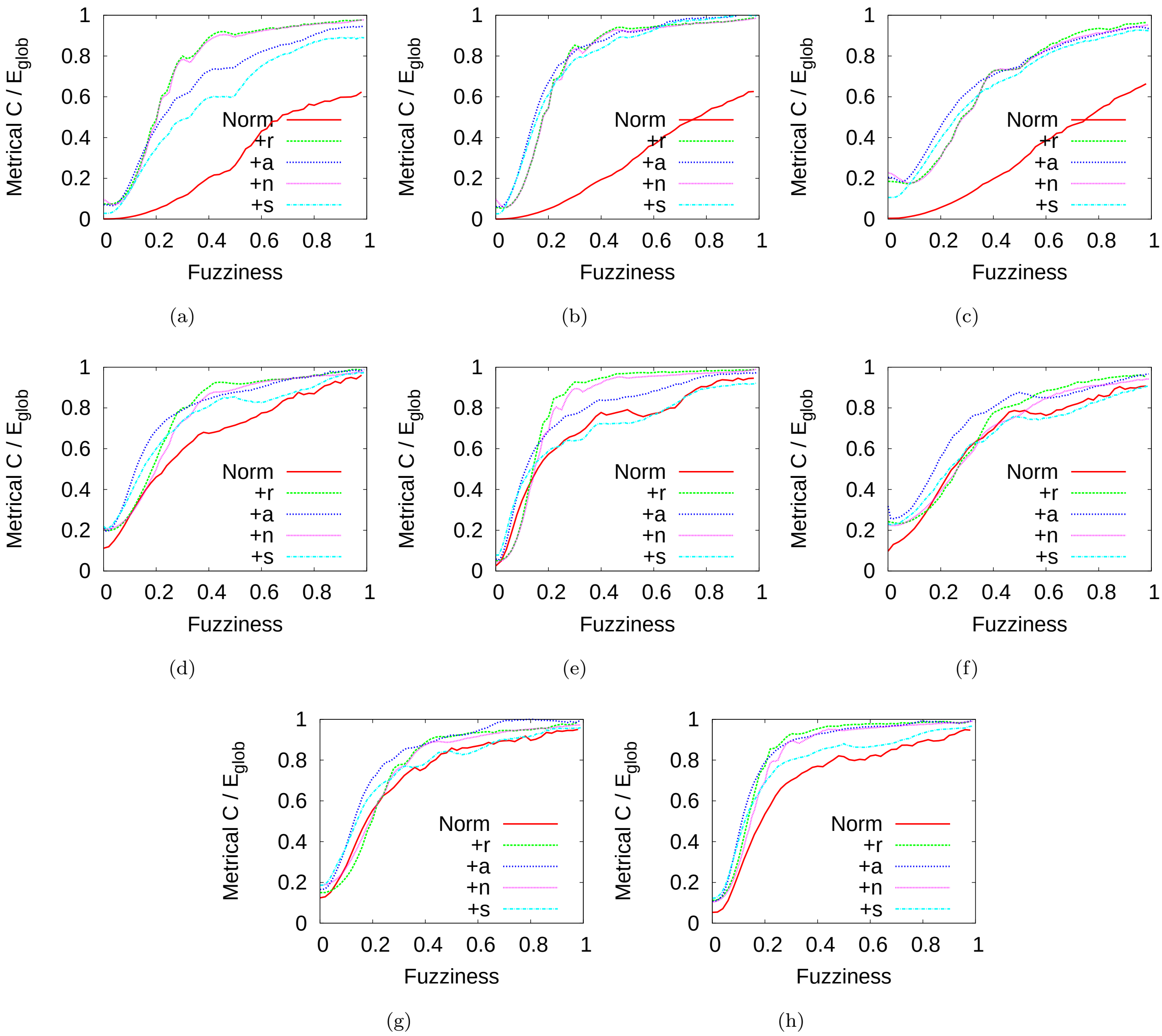
(e)

(f)

(g)

(h)

Figure A.12: Topological C_t/E_{glob} as a function of f in subway networks such as Paris (a), Nyc (b), Milan (c) and in VirtualTourist city-based networks of Italy (d), United Kingdom (e), India (f), the Netherlands (g) and the US airline network (h). Straight lines represent the baselines. Different perturbations on nodes' positions (+n, +r) and on edges (+a,+s) are then applied. Economic are networks cheap to create and with good local and global connectivity pattern. Cost over efficiency is an indicator of how much a network is economic. The best one is that with low C_t/E_{glob} value. Under our framework, we found that only at micro scale networks are very economic, whereas at macro the networks are very efficient but still very expensive.



(a) (b) (c)

(d) (e) (f)

(g) (h)

Figure A.13: Topological C_t/E_{glob} as a function of f in subway networks such as Paris (a), Nyc (b), Milan (c) and in VirtualTourist city-based networks of Italy (d), United Kingdom (e), India (f), the Netherlands (g) and the US airline network (h). Straight lines represent the baselines. Different perturbations on nodes' positions (+n, +r) and on edges (+a,+s) are then applied. Economic are networks cheap to create and with good local and global connectivity pattern. Cost over efficiency is an indicator of how much a network is economic. The best ones are those with low C_t/E_{glob} value. Under our framework, we found that only at micro scale networks are very economic, whereas at macro the networks are very efficient but still very expensive.

Bibliography

[1] Geonames, geographical database. `http://www.geonames.org`.

[2] Virtual tourist home page. `http://www.virtualtourist.com`.

[3] Sophie Achard, Chantal Delon-Martin, Petra E. Vértes, Félix Renard, Maleka Schenck, Francis Schneider, Christian Heinrich, Stéphane Kremer, and Edward T. Bullmore. Hubs of brain functional networks are radically reorganized in comatose patients. *Proceedings of the National Academy of Sciences*, 2012.

[4] Lada A. Adamic. The small world web. In *Proceedings of the ECDL*, pages 443–452, 1999.
`http://www.hpl.hp.com/research/idl/papers/smallworld/smallworld.ps`.

[5] Yong-Yeol Ahn, Seungyeop Han, Haewoon Kwak, Sue Moon, and Hawoong Jeong. Analysis of topological characteristics of huge online social networking services. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 835–844, New York, NY, USA, 2007. ACM.

[6] Luca M. Aiello, Alain Barrat, Ciro Cattuto, Rossano Schifanella, and Giancarlo Ruffo. Link creation and information spreading over social and communication ties in an interest-based online social network. *EPJ Data Science*, 1(1):12+, 2012.

[7] R. Albert and A.L. Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47–97, 2002.

[8] Reka Albert, Jeong Hawoong, and Albert-Laszlo Barabasi. Error and attack tolerance of complex networks. *Nature*, 406:378–382, 2000.

[9] B. Albert-László and E. Bonabeau. Scale-free networks. *Scientific American*, 288, 2003.

[10] E Almaas, B KovÃ¡cs, T Vicsek, Z N Oltvai, and A-L Barabasi. Global organization of metabolic fluxes in the bacterium escherichia coli. *Nature*, 427(6977):839–43, February 2004.

[11] L.A.N. Amaral, A. Scala, M. Barthélémy, and H.E. Stanley. Classes of small-world networks. *Proceedings of the National Academy of Sciences of the United States of America*, 97(21):11149, 2000.

[12] L. Backstrom, E. Sun, and C. Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World wide web*, pages 61–70. ACM, 2010.

[13] Per Bak, Kim Christensen, Leon Danon, and Tim Scanlon. Unified scaling law for earthquakes. *Phys. Rev. Lett.*, 88:178501, Apr 2002.

[14] Shweta Bansal, Shashank Khandelwal, and Lauren Meyers. Exploring biological network structure with clustered random networks. *BMC Bioinformatics*, 10(1):405+, 2009.

[15] Albert Barabasi and Reka Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[16] A. Barrat, M. Barthelemy, R. Pastor-Satorras, and A. Vespignani. The architecture of complex weighted networks. *PNAS*, 101(11):3747–3752, March 2004.

[17] A. Barrat and M. Weigt. On the properties of small-world network models. *The European Physical Journal B-Condensed Matter and Complex Systems*, 13(3):547–560, 2000.

[18] Alain Barrat, Marc Barthlemy, and Alessandro Vespignani. *Dynamical Processes on Complex Networks.* Cambridge University Press, New York, NY, USA, 2008.

[19] Marc Barthelemy and Nunes L. A. Amaral. Small-world networks: Evidence for a crossover picture. *Physical Review Letters*, 82:3180–3183, 1999.

[20] M. Biasiolo, M. Forcato, L. Possamai, F. Ferrari, L. Agnelli, M. Lionetti, K. Todoerti, A. Neri, M. Marchiori, S. Bortoluzzi, et al. Critical analysis of transcriptional and post-transcriptional regulatory networks in multiple myeloma. In *Pacific Symposium on Biocomputing.*, page 397, 2010.

[21] S. Boccaletti, Vito Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. Complex networks: Structure and dynamics. *Physics Report*, 424:175–308, 2006.

[22] B. Bollobas. *Random Graphs.* Accademic Press, London, 1985.

[23] II Michael J. Bommarito, Daniel Katz, and Jon Zelner. Law as a seamless web?: comparison of various network representations of the united states supreme court corpus (1791-2005). In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, ICAIL '09, pages 234–235, New York, NY, USA, 2009. ACM.

[24] Robert M. Bond, Christopher J. Fariss, Jason J. Jones, Adam D. I. Kramer, Cameron Marlow, Jaime E. Settle, and James H. Fowler. A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415):295–298, September 2012.

[25] Dan Braha, Blake Stacey, and Yaneer Bar-Yam. Corporate competition: A self-organized network. *Social Networks*, 33(3):219–230, 2011.

[26] G. Caldarelli. *Scale-Free Networks: Complex Webs in Nature and Technology.* Oxford Finance Series. Oxford University Press, USA, 2007.

[27] G. Caldarelli, A. Capocci, P. De Los Rios, and MA Munoz. Scale-free networks from varying vertex intrinsic fitness. *Physical review letters*, 89(25):258702, 2002.

[28] Alessandro Casagrande. Analisi di comunità virtuali: il caso Communities.com. Master's thesis, Università Ca' Foscari, Venezia, 2006.

[29] F. Chung and L. Lu. The diameter of sparse random graphs. *Advances in Applied Mathematics*, 26(4):257–279, 2001.

[30] R. Cohen, D. Ben-Avraham, and S. Havlin. Percolation critical exponents in scale-free networks. *Physical Review E*, 66(3):36113, 2002.

[31] 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.

[32] Regino Criado, Benito Hernandez-Bermejo, and Miguel Romance. Efficiency, vulnerability and cost: an overview with applications to subway networks worldwide. *I. J. Bifurcation and Chaos*, 17(7):2289–2301, 2007.

[33] P. Crucitti, V. Latora, M. Marchiori, and A. Rapisarda. Error and attack tolerance of complex networks. *Physica A: Statistical Mechanics and its Applications*, 340(1-3):388–394, 2004.

[34] P. Crucitti, V. Latora, and S. Porta. Centrality measures in spatial networks of urban streets. *PHYSICAL REVIEW E*, 73(3):036125, Part 2006.

[35] DJ Daley and DG Kendall. Epidemics and rumours. 1964.

[36] 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.

[37] G.F. Davis, M. Yoo, and W.E. Baker. The small world of the corporate elite. *Preprint, University of Michigan Business School*, 2001.

[38] D. J. de S. Price. A general theory of bibliometric and other cumulative advantage processes. *J. Amer. Soc. Inform. Sci.*, 27:292–306.

[39] A. Demers, D. Greene, C. Hauser, W. Irish, J. Larson, S. Shenker, H. Sturgis, D. Swinehart, and D. Terry. Epidemic algorithms for replicated database maintenance. In *Proceedings of the sixth annual ACM Symposium on Principles of distributed computing*, pages 1–12. ACM, 1987.

[40] Reinhard Diestel. *Graph Theory (Graduate Texts in Mathematics)*. Springer, August 2005.

[41] Victor M. Eguíluz, Dante R. Chialvo, Guillermo A. Cecchi, Marwan Baliki, and A. Vania Apkarian. Scale-free brain functional networks. *Phys. Rev. Lett.*, 94:018102, Jan 2005.

[42] P. Erdos and A. Renyi. On random graphs I. *Publ. Math. Debrecen*, 6(290-297):156, 1959.

[43] P. Erdos and A. Renyi. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5, 1960.

[44] Stefano Falco. Analisi di comunità virtuali: il caso Virtualtourist.com. Master's thesis, Università Ca' Foscari, Venezia, 2005.

[45] K.J. Falconer and J. Wiley. *Fractal geometry: mathematical foundations and applications*. Wiley New York, 2003.

[46] L.C. Freeman. Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239, 1979.

[47] Linton C. Freeman. A Set of Measures of Centrality Based on Betweenness. *Sociometry*, 40(1):35–41, March 1977.

[48] Thomas M. J. Fruchterman and Edward M. Reingold. Graph drawing by force-directed placement. *Softw. Pract. Exper.*, 21(11):1129–1164, November 1991.

[49] Lazaros K. Gallos, Diego Rybski, Fredrik Liljeros, Shlomo Havlin, and Hernán A. Makse. How people interact in evolving online affiliation networks. *Phys. Rev. X*, 2:031014, Aug 2012.

[50] A. V. Goltsev, S. N. Dorogovtsev, and J. F. F. Mendes. Percolation on correlated networks. *Phys. Rev. E*, 78:051105, Nov 2008.

[51] M. S. Granovetter. The Strength of Weak Ties. *The American Journal of Sociology*, 78(6):1360–1380, 1973.

[52] P. Grassberger. On the critical behavior of the general epidemic process and dimensional percolation. *Mathematical Bioscience*, 63:157–72, 1983.

[53] J.W. Grossman and P.D.F. Ion. On a portion of the well-known collaboration graph. *Congressus Numerantium*, pages 129–132, 1995.

[54] R. Guimera, S. Mossa, A. Turtschi, and L.A.N. Amaral. The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(22):7794, 2005.

[55] A. Gutfraind, L. Ancel Meyers, and I. Safro. Multiscale Network Generation. *ArXiv e-prints*, July 2012.

[56] H.W. Hethcote. The mathematics of infectious diseases. *SIAM review*, 42(4):599–653, 2000.

[57] Petter Holme, Christofer R. Edling, and Fredrik Liljeros. Structure and time evolution of an internet dating community. *Social Networks*, 26(2):155–174, 2004.

[58] Petter Holme and Jari Saramäki. Temporal networks. *CoRR*, abs/1108.1780, 2011.

[59] Haibo Hu and Xiaofan Wang. Evolution of a large online social network. *Physics Letters A*, 373(12–13):1105–1110, 2009.

[60] H. Jeong, S.P. Mason, A.L. Barabási, and Z.N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, 2001.

[61] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabasi. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, October 2000.

[62] Tomihisa Kamada and Satoru Kawai. An algorithm for drawing general undirected graphs. *Information Processing Letters*, 31(1):7–15, 1989.

[63] Beom Jun Kim. Geographical coarse graining of complex networks. *Phys. Rev. Lett.*, 93:168701, Oct 2004.

[64] P.L. Krapivsky, S. Redner, and F. Leyvraz. Connectivity of growing random networks. *Physical Review Letters*, 85(21):4629–4632, 2000.

[65] Ravi Kumar, Jasmine Novak, and Andrew Tomkins. Structure and evolution of online social networks. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pages 611–617, New York, NY, USA, 2006. ACM.

[66] V. Latora and M. Marchiori. Is the Boston subway a small-world network? arXiv:cond-mat/0202299, 2002.

[67] V. Latora and M. Marchiori. Economic small-world behavior in weighted networks. *The European Physical Journal B*, 32(2):249–263, 2003.

[68] Vito Latora and Massimo Marchiori. Efficient behaviour of small-world networks. *Physical Review Letters*, 87, 2001.

[69] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, KDD '05, pages 177–187, New York, NY, USA, 2005. ACM.

[70] David Levary, Jean-Pierre Eckmann, Elisha Moses, and Tsvi Tlusty. Loops and self-reference in the construction of dictionaries. *Phys. Rev. X*, 2:031018, Sep 2012.

[71] Priya Mahadevan, Dmitri Krioukov, Kevin Fall, and Amin Vahdat. Systematic topology analysis and generation using degree correlations. *SIGCOMM Comput. Commun. Rev.*, 36(4):135–146, August 2006.

[72] B. Mandelbrot. *The Fractal Geometry of Nature.* Freeman and Co., San Francisco, 1982.

[73] N.D. Martinez. Artifacts or atributes? Effects of resolution on the Little Rock Lake food web. *Ecological Monograph*, 61:367–392, 1991.

[74] Mirko S. Mega, Paolo Allegrini, Paolo Grigolini, Vito Latora, Luigi Palatella, Andrea Rapisarda, and Sergio Vinciguerra. Power-law time distribution of large earthquakes. *Phys. Rev. Lett.*, 90:188501, May 2003.

[75] M. Menezes, CF Moukarzel, and TJP Penna. First-order transition in small-world networks. *EPL (Europhysics Letters)*, 50:574, 2000.

[76] Stanley Milgram. The small world problem. *Psychology Today*, 22:61–67, 1967.

[77] O.M. Miller. Notes on a cylindrical world map projection. *Geograph. Rev*, (32):424–430, 1942.

[78] Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, IMC '07, pages 29–42, New York, NY, USA, 2007. ACM.

[79] R. Monasson. Diffusion, localization and dispersion relations on small-world lattices. *The European Physical Journal B-Condensed Matter and Complex Systems*, 12(4):555–567, 1999.

[80] J. M. Montoya and R. V. Solé. Small world patterns in food webs. *Journal of Theoretical Biology*, 214(3):405–412, 2002.

[81] Y. Moreno, JB Gómez, and AF Pacheco. Instability of scale-free networks under node-breaking avalanches. *EPL (Europhysics Letters)*, 58:630, 2002.

[82] M. E. J. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America*, 98(2):404–409, January 2001.

[83] M. E. J. Newman and Juyong Park. Why social networks are different from other types of networks. *Phys. Rev. E*, 68:036122, Sep 2003.

[84] M.E.J. Newman. Assortative mixing in networks. *Physical Review Letters*, 89(20):208701, 2002.

[85] M.E.J. Newman. Mixing patterns in networks. *Physical Review E*, 67(2):26126, 2003.

[86] MEJ Newman. Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46(5):323–351, 2005.

[87] Jae Dong Noh. Percolation transition in networks with degree-degree correlation. *Phys. Rev. E*, 76:026116, Aug 2007.

[88] J. Park and M.E.J. Newman. Statistical mechanics of networks. *Physical Review E*, 70(6):66117, 2004.

[89] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Physical review letters*, 86(14):3200–3203, 2001.

[90] Romualdo Pastor-Satorras, Alexei Vasquez, and Alessandro Vespignani. Dynamical and correlation properties of the internet. *Physical Review Letters*, 87:258701, 2001.

[91] Tobias Preis, Dror Y. Kenett, H. Eugene Stanley, Dirk Helbing, and Eshel Ben-Jacob. Quantifying the behavior of stock correlations under market stress. *Scientic Reports*, 2(752), 2012.

[92] Albert Réka and Albert-László Barabási. Topology of evolving networks: Local events and universality. *Physical Review Letters*, 85:5234, 1999.

[93] Garry Robins, Tom Snijders, Peng Wang, and Mark Handcock. Recent developments in exponential random graph (p*) models for social networks. *Social Networks*, 29:192–215, 2006.

[94] Camille Roth, Soong M. Kang, Michael Batty, and Marc Barthelemy. A long-time limit for world subway networks. *Journal of The Royal Society Interface*, May 2012.

[95] A. Saichev, Y. Malevergne, and D. Sornette. *Theory of Zipf's law and beyond.* Springer, 2009.

[96] M. Ángeles Serrano, Ana Maguitman, Marián Boguñá, Santo Fortunato, and Alessandro Vespignani. Decoding the structure of the www: A comparative analysis of web crawls. *ACM Trans. Web*, 1(2), August 2007.

[97] Chaoming Song, Shlomo Havlin, and Hernan A. Makse. Self-similarity of complex networks. *Nature*, 433(7024):392–395, January 2005.

[98] Animesh Srivastava, Bivas Mitra, Niloy Ganguly, and Fernando Peruani. Correlations in complex networks under attack. *Phys. Rev. E*, 86:036106, Sep 2012.

[99] H.E. Stanley. Introduction to phase transitions and critical phenomena, Clarendon, 1971.

[100] D. Stauffer and A. Aharony. *Introduction to percolation theory.* Taylor & Francis, 1992.

[101] Mark Steyvers and Joshua B. Tenenbaum. The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth. *Cognitive Science*, 29(1):41–78, 2005.

[102] S. Strogatz. *Sync: how order emerges from chaos in the universe, nature, and daily life.* Hyperion, 2004.

[103] P. Tieri, S. Valensin, V. Latora, G. C. Castellani, M. Marchiori, D. Remondini, and C. Franceschi. Quantifying the relevance of different mediators in the human immune cell network. *Bioinformatics*, 21(8):1639–1643, April 2005.

[104] A. Tsoularis. Analysis of logistic growth models. *Mathematical Biosciences*, 179(1):21–55, August 2002.

[105] Alexei Vazquez and Yamir Moreno. Resilience to damage of graphs with degree correlations. *Phys. Rev. E*, 67:015101, 2003.

[106] T. Victek. *The Fractal Growth Phenomena.* World Scientific, Singapore, 1992.

[107] W. Vogels, R. Van Renesse, and K. Birman. The power of epidemics: robust communication for large-scale distributed systems. *ACM SIGCOMM Computer Communication Review*, 33(1):131–135, 2003.

[108] I. Vragović, E. Louis, and A. Díaz-Guilera. Efficiency of informational transfer in regular and complex networks. *Phys. Rev. E*, 71:036122, Mar 2005.

[109] Stanley Wasserman and Katherine Faust. *Social Network Analysis. Methods and Applications.* 1994.

[110] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature (London)*, **393**:440–442, 1998.

[111] B. Wellman. Computer networks as social networks. *Science*, 293(5537):2031–2034, 2001.

[112] B. Wellman and C.A. Haythornthwaite. *The Internet in everyday life.* Blackwell Pub., 2002.

[113] D.B. West et al. *Introduction to graph theory*, volume 1. Prentice Hall Upper Saddle River, NJ, 2001.