STOCHASTIC MODELS AND TRANSITIVITY IN COMPLEX NETWORKS

Tiago Manuel Louro Machado De Simas

Submitted to the faculty of the University Graduate School

in partial fulfillment of the requirements

for the degree

Doctor of Philosophy

in Cognitive Science

Indiana University

May 2012

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirement for degree of Doctor of Philosophy.

Doctoral Committee

Luis Mateus Rocha - Chair, Ph.D.

Olaf Sporns - Co-Chair, Ph.D.

Alessandro Flammini

Johan Bollen

April 26, 2012

I dedicate this thesis to my father Miguel Machado de Simas, my mother Maria de Lurdes de Simas and my wonderful wife Ana Claudia Guerra, who supported me in each step of the way with love and understanding.
To all my brothers and sisters, Ana Maria, Miguel Jose, Maria Paula, Joao Pedro, Maria da Luz and Marianinha.

Acknowledgements

I would especially like to express my gratitude to my adviser Professor Luis Mateus Rocha, for his support at all levels of my academic and personal life, and for his attention to detail and enormous persistence on editing this thesis. I would like to thank Professors Olaf Sporns, Alessandro Flammini, Johan Bollen and Alessandro Vespignani for their collaboration, and for their enthusiastic support and availability. I also express my gratitude to Professor Rita Almeida Ribeiro, for her scholarship support, help in editing this thesis, and friendship. Finally, I thank my Ph.D colleges Alfredo Pereira and Artemy Kolchinsky for their support, discussions about this thesis and editing.

Tiago Manuel Louro Machado De Simas

Stochastic Models and Transitivity in Complex Networks

Networks are typically described as graphs: a set of nodes (vertices) related by links (edges), where a link can either represent the presence of a connection or the strength of the connection. When the connection between vertices is weighted we call these networks weighted graphs. Graphs are a useful abstraction for representing social connections, structural brain connections, functional brain connections, and others. The first contribution is a novel stochastic model that explains analytically, the cut-off behavior of real-world scale-free networks previously modeled computationally by Amaral et al. and others. This new mathematical model explains several existing computational scale-free network generation models, and yields a novel theoretical basis for understanding cut-off behavior in complex networks, previously only analyzed with simulations using distinct models - this contribution unifies the existing literature on cut-off behavior in scale-free networks. Further, as an analytical mathematical model, it allows us to study properties of the growth of scale-free networks that are not possible to study with current computational models. The second contribution is related to the transitive closure of fuzzy graphs, which is used to calculate the strongest connection between vertices via indirect paths, when edges weights denote proximity or similarity. In the field of complex networks, the Dijkstra algorithm is a

well-established algorithm to calculate the shortest path between any two vertices, where edge weights denote distance. Here, the transitive closure in fuzzy graphs is shown to be a generalization of Dijkstra algorithm in distance. This result bridges the theory of fuzzy graphs and the theory of complex networks. Finally, we propose a new methodology to analyze complex networks, based on the semi-metric behavior. We apply this methodology to the analysis of the small-world phenomenon in real-world complex networks.

The case studies included the US Airport Network, the Human Cortex Network, fMRI Human Brain network, the Scientific Collaboration Network, Astrophysics Collaborations Network and the High-Energy Theory Collaborations Network.

Luis Mateus Rocha - Chair, Ph.D.

Olaf Sporns - Co-Chair, Ph.D.

Alessandro Flammini

Johan Bollen

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Motivation

1.1 Towards Improving Mathematical Representations of Complex Networks

The main motivation for this thesis is improving current mathematical descriptions of real-world complex networks. Complex networks are often modeled using the theory of graphs [1] and graphs are based on binary relations. These can be undirected, directed and weighted, as presented in figures 1.1, 1.2 and 1.3. Graphs are the easiest way to model organization amongst elements of a set; for instance, they can model interactions, associations, similarity, distance, etc. Indeed, the binary relation (graph) is the simplest form of a general system in the effort to model the organization of complex systems [52].

[1] From now on we will describe networks as graphs

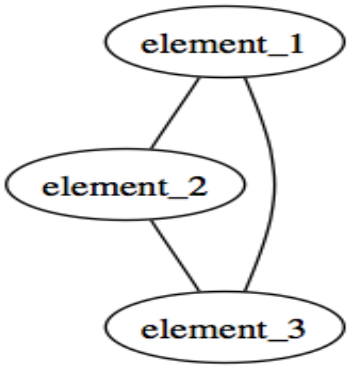

Figure 1.1: Undirected crisp graph.

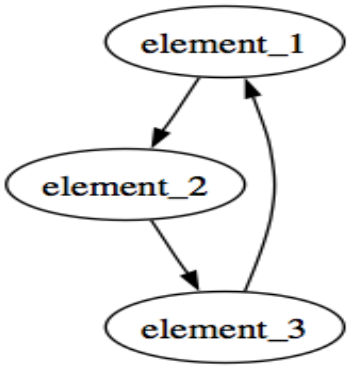

Figure 1.2: Directed crisp graph.

The first question we approach deals with models that explain the cut-off patterns found in the power-law degree[2] distribution of real-world networks. There are several models to explain how networks grow with a power law degree distribution, such as the seminal work of Barabasi and Albert [11]. Barabasi and Albert showed that many real world networks have a degree distribution which follows approximately a power-law, mathematically de-

[2]Degree is the number of edges (links) to and out of a given vertex (node) in a graph. Chapter 2 provides the necessary mathematical background for the treatment of networks.

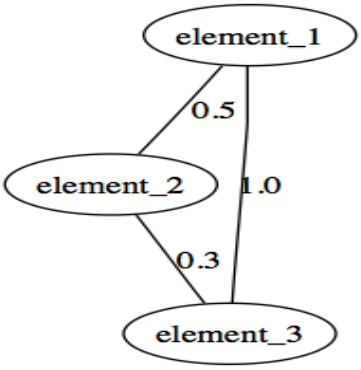

Figure 1.3: Weighted undirected graph.

scribed by

$$P(k) = ak^{-\gamma}$$

where k is the degree of a given vertex, a a constant and γ the characteristic distribution exponent. A probability distribution of this form is also known in Statistics as the Pareto distribution [65]. Graphs characterized by power-law degree distributions have a completely different topology than what is expected of random graphs. The latter have a Poisson degree distribution, thus there is a mean value for the degree distribution that is characteristic of this kind of graphs. In contrast, studies of real-world networks observed that they follow power-law degree distributions, and thus have no characteristic mean value degree and the variance converges to infinity with the size of the network. In other ways, graphs which are characterized by a power-law degree distribution have no scale. In networks that grow according to a power-law distribution we will always observe that most of the vertices have a small degree and there will be only a few with a high degree, while the

connectivity/degree of the network has no characteristic average degree.

The Barabasi and Albert model is concerned with the mechanisms behind the ubiquitous organization of complex networks. They proposed two simple mechanisms or axioms to generate scale-free behavior: *Growth* and *Preferential Attachment*. Together, they give rise to networks with power law degree distribution, particularly with the power law exponent equal to $\gamma = 3$. This mathematical model (described in chapter 2), is also known as "the rich get richer".

Many extensions of the work of Barabasi and Albert have been proposed, among them the work of Amaral et al. [6]. In this work real-world networks, which do not have a perfect power-law degree distribution, as they present a truncation or a cut-off for large degree vertices, are more accurately described. These truncations were identified as an exponential decay associated with the power-law. This is important because it allows the statistical moments to converge with the size of the network. Amaral et al. proposed an extended model of the Barabasi and Albert model by adding a third axiom; the *Aging of Vertices*. This axiom assumes that in the growing process the vertices of the network have a probability p of becoming inactive. The inactive vertices remain in the network but no longer accept new links from new vertices . With this new axiom, network organization exhibits an exponential decay or power law truncation of the degree distribution – which is a more realistic characterization of real networks.. Moreover, this truncation can be adjusted by setting the parameter p.

One issue with the work of Amaral et al. is that it was based only on simulations. Similarly, other works give an alternative explanation for the cut-offs in power law degree distributions. But these models, which are reviewed in chapters 2 and 3 explain the cut-off behavior of scale free networks without providing a mathematical explanation. My first motivation was to create an analytical explanation, which unifies all these contributions.

Another open question concerns the calculation of shortest paths in weighted graphs, where weights represent distances. To calculate shortest paths in a distance graph it is usual to use the Dijkstra algorithm [29] integrated in the All Pairs Shortest Paths (APSP) Johnson's algorithm [27]. Weighted graphs are equivalent to fuzzy graphs [53], for which there are many ways to calculate transitive closure (TC) [53], which is a means to calculate the strongest indirect associations between edges, where weights denote proximity or similarity. In the fuzzy graph literature transitive closure is computed using t-norm and t-conorm operators [53]. Here an open question arises as to how is the Dijkstra algorithm related to the transitive closure in fuzzy graphs? More generally, is it useful to transfer broadly known mathematical methods from the fuzzy graphs to the complex networks field? We show that this transfer is fruitful for both fields. First, we relate fuzzy transitive closure to the Dijkstra algorithm via two theorems, and derive some practical advice for choosing appropriate closure and shortest path algorithms. Second, we use this knowledge to study the small-world phenomenon in weighted graphs, an open question in complex networks.

In 1998, Watts and Strogatz published ground breaking work on the Small World phenomena in social networks [97]. Watts and Strogatz had as an initial goal the study of firefly synchronization. In their research they found a very interesting experiment of a sociologist, Milgram, who studied the small world phenomenon in 1963 [90]. Stanley Milgram was interested in the following problem: if we imagine a population as a social network, what is the average path length between any two random selected vertices, or people, in this network? To answer this question he performed the following experiment: selecting people or individuals from cities, which were geographically and socially distant he sent, from the original city a letter addressed to a target individual in the distant city. These letters contained only the basic information of the contact person in the target city, without providing any address. The idea was whether the individuals who received such letters knew the target persons or not; if yes they were instructed to send the letter directly to the target, otherwise to another person who, they thought would, probably know the target. With this, Milgram wanted to assess the average path length between the origin and the target. The cities chosen were Omaha and Wichita as origin and Boston as target, in the United States of America. Stanley Milgram found the surprising fact that an average path length of six individuals between the origin and the target, which we now colloquially call the *six degrees of separation*. This means that between any two random vertices in a social network such as the US population, on average, we know a person who knows a person, ..., who knows the target person, with six peo-

ple in between. Watts and Strogatz noticed that in a general description of networks, random networks and lattice networks were two extreme positions on organization defined by two parameters: *shortest path* and *clustering coefficient*. Random graphs, were well established by the work of Erdos and Reyni [21], two mathematicians who created the field. In this kind of network representations, the edges between any two vertices are chosen randomly, so that the degree distribution follows a Poisson distribution and the average path length is relatively small, compared with the size of the network. On the other extreme we have lattice networks, which are extremely well organized; that is, each vertex is connected only to a subset of nearest neighbors. Such networks have a relatively large average path length compared with the size of the network. Watts and Strogatz realized that in addition to the average path length, there is another important property behind these two kinds of networks: the *clustering coefficient*, which gives us a measure of the structure , or more precisely, the (immediate) level of transitivity in a network. In other words, the clustering coefficient is one means to describe communities in these networks. The clustering coefficient (or immediate transitivity) for a given vertex is measured as the the proportion of observed triangles with the direct neighbors, given the total possible number of such triangles. The clustering coefficient for random graphs is minimum, i.e. too small, showing non existing community structure. On the other extreme the clustering coefficient is maximum for lattice networks. Watts and Strogatz, found that many real networks such as the Milgram network fall somewhere in between

these two extremes. They have a relatively small average path length but a relatively high clustering coefficient. Watts and Strogatz, categorized this type of networks as *small world networks*.

The clustering coefficient for weighted networks has been generalized and analyzed (see Chapter 2 for an overview). However, the characteristic path length of weighted graphs is not well understood, because we can define many (infinite) different measures of shortest path or transitive closures in weighted graphs. This poses us the question: is there a preferred way to measure shortest paths in weighted graphs as models of real complex networks? By using some concepts from fuzzy graph theory, we produce a more comprehensive understanding of possible closures and consequently understand better the small-world phenomena in weighted networks.

Related to shortest path length is the notion of *semi-metric behavior*, which measures the violations of transitivity in a weighted graph, Rocha [72] defined and described several measures to measure semi-metric behavior. One of our motivations is to analyze the semi-metric behavior of weighted graphs with the semi-metric ratios (see Chapter 2 for an overview) to uncover which edges are more semi-metric. Highly semi-metric edges identify vertices (nodes) that are more indirectly (by some path) related than expected from a metric relationship on the original network.

One common approach when analyzing weighted networks is the application of a specific threshold to the edge weights of the graph resulting a binary (crisp) subgraph retaining only the edges above the threshold. The analysis

consists then of applying well known techniques from binary graphs to the resulting subgraph. This approach does not consider the semi-metricity of the weighted graph which is a problem we explore in this dissertation.

An alternative way to study weighted networks is proposed in this thesis. We introduce the semi-metric behavior to weighted graphs, by establishing a semi-metric threshold to the weighted graph. This semi-metric threshold is set based on the shortest paths between vertices. We calculate the shortest paths between any two vertices, then we set the semi-metric threshold and apply it to all direct connections of the original weighted graph. If the shortest path between two vertices that have a directed connection is bellow the semi-metric threshold the corresponding direct edge weight is removed otherwise is preserved with the original weight. We end up with a weighted subgraph of the original weighted graph. With this approach we preserve all weak edges, which have a strong indirect path but remove all weak edges that also have a weak indirect path.

With the approach described above and the generalization of the clustering coefficient and average path length we are able to introduce a new way to characterize the small-world phenomenon in weighted networks.

To support the theoretical results we analyzed six real networks: the US Airport network, the Human Cortex network, the fMRI Human Brain network, the Scientific Collaboration network, the Astrophysics Collaborations network and the High-Energy Theory Collaborations network. This analysis consists of the study of shortest paths, clustering coefficients and semi-metric

behavior in real and random networks.

1.2 Summary of Contributions

The contributions in this thesis are:

(a) Produce an analytical solution and integrative model of cut-offs in the power-law degree distribution , which gives us the ability to better predict the organization of complex networks.

(b) Relate a mathematical treatment of transitive closure in fuzzy graphs to the Dijkstra algorithm [29] in weighted graphs. This result bridges the gap between complex networks and fuzzy graphs and gives an insight about how we measure the shortest paths between any two vertices in a weighted graph, since there are no unique way to perform this measurement.

(c) Propose a new methodology to analyze complex networks and study properties such as: the average path length, semi-metric behavior and clustering coefficient in weighted graphs. This helps us to characterize more effectively the small-world phenomena in weighted networks.

1.3 Dissertation Outline

Chapter 2 provides the relevant necessary formal background on complex networks and fuzzy graphs. In Chapter 3 we propose a stochastic model for cut-offs in complex networks. In chapters 4, we relate weighted graphs

with fuzzy graphs, and generalize Transitive Closure to complex networks. Chapter 5 we discussed the performance of metric closure on recommender systems. Chapter 6 we discuss the small-world phenomena in weighted complex networks as a model of real networks. Finally, in chapter 7 we discuss the implications of this work and propose future work. Figure 1.4 summarizes the dissertation outline.

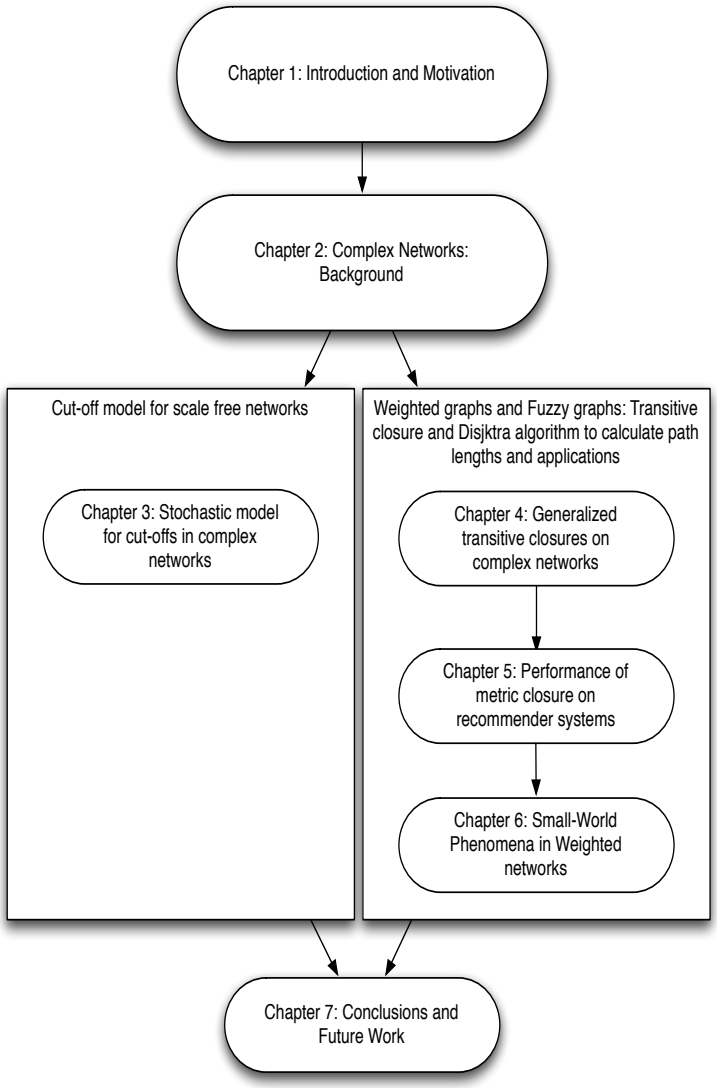

Figure 1.4: Logical structure of the thesis.

Chapter 2

Complex Networks: Background

In this chapter we present the relevant background to this thesis. The basic concepts of complex networks and the recents advances in theory and applications.

2.1 Introduction

In the last decade, much work has been done to understand the general mechanisms that influence the growth dynamics of complex networks. Complex networks have been studied by mathematicians, social scientists, physicists and others. Perhaps the two most influential contributions to this field are: *small world phenomenon* introduced by Watts and Strogatz [97] and the

preferential attachment mechanism behind *scale-free networks*, proposed by Barabasi and Albert [11]. Because of the pervasiveness of both the small-world phenomenon and scale-free networks in nature and society, there is an intense interest in the study of their structure and dynamics [68] [32][22]. Applications of complex networks can be found across disparate fields: understanding the Internet[68], the World Wide Web [71], protein-protein interaction networks[26], metabolic networks[45][94] and other natural networks.

Amaral et al. [6] proposed a computational model that explains the cut-offs in power-law degree distributions of real-world scale-free networks. Other models such as [20], [42], [61], based on finite size effects, have been proposed as well. However, the Amaral et al. model is among the simplest and therefore amenable to analysis.

In these studies, networks were described by graphs. A (crisp[1]) graph $G = (V, E)$ is characterized by a set of vertices (nodes) and a set of edges (links) E. A graph is defined by its connectivity matrix E, where elements $e_{i,j} \in \{0, 1\}$ denote the existence or abcence of an edge between vertex v_i and v_j. Recently, weighted graphs, have received much interest in the literature of complex networks [67] [14] [13] [7]. Weighted graph is a graph where the edges are characterized by having real values: $e_{i,j} \in \Re$. A good overview on the latest developments on complex networks can be found in [19]. To understand the small-world phenomena and the scale-free properties of complex networks represented as weighted graphs. We have to generalize concepts

[1]in the terminology of Fuzzy Graphs.

such as clustering coefficient, average path length, and node degree. Several definitions have been proposed in order to generalize the clustering coefficient to weighted networks [67][14]. Also the degree distribution has been generalized to the strength degree of a node and well studied in [14]. However, an important property, the average path length, has still not been well studied for weighted networks, but it plays an important role in classifying, for example, a small-world network.

2.2 Relations, graphs and weighted graphs

2.2.1 Relations

In the mathematical formalization of nature, relations play an important role by allowing us to associate objects of the same or different nature. Relations can give us insights about the structure of problems we intend to explore. Binary relations, which associate elements of set X with itself $R(X, X)$, can be seen as graphs, where R characterizes the nature of edges (e.g. weights).

Crisp relations represent a crisp association or interaction between the elements of two or more sets of objects. These associations or interactions are either present or absent, that is, the relation assumes values on the set $\{0, 1\}$.

A crisp relation among sets $X_1, X_2, ..., X_n$ is a subset of the cartesian product $X_1 \times X_2 \times, ..., \times X_n$ and is represented by $R(X_1, X_2, ..., X_n)$. More specifically we can define a *characteristic function* which assigns to each tuple

in the relation a value of 1.

$$R(x_1, x_2, ..., x_n) = \begin{cases} 1 & \text{iff } < x_1, x_2, ..., x_n > \in E \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

The elements of the relation are n-tuples $< x_1, x_2, ..., x_n >$. Another way to define: it is as a n-dimensional membership array (or tensor) $R = [r_{i_1,i_2,...,i_n}]$,

$$r_{i_1,i_2,...,i_n} = \begin{cases} 1 & \text{iff } < r_1, r_2, ..., r_n > \in R \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

A fuzzy relation is when $r_{i_1,i_2,...,i_n} \in [0, 1]$, where the values typically denote strength of the association. In this case, $r_{i_1,i_2,...,i_n}$ is known as a membership degree.

A graph is a special type of a binary relation E on the same set V. It is a subset of the cartesian product $V \times V$ and is represented by the set of pairs (2-tuples) $< v_i, v_j >$, which can be denoted on $e_{i,j}$. In this case the membership array is the same E, whose elements $e_{i,j}$ denote the presence or strength of edges between vertices v_i and $v_j \in V$, and where $i, j \in 1, ..., |V|$. A binary relation on 2 distinct sets, V_1, V_2 is a bipartite graph.

A graph $E(V, V)$ is *null* iff $E(v, v) = 0$, for all $v \in V$. A graph $E(V, V)$ is *reflexive* iff $E(v, v) = 1$, for all $v \in V$, otherwise, $E(V, V)$ is *irreflexive*. If $E(v, v) = 0$, for all $v \in V$ the graph is called *antireflexive*. Moreover $E(V, V)$ is *symmetric* iff $E(v_i, v_j) = E(v_j, v_i)$ for all $v_i, v_j \in V$, otherwise, the graph

is *asymmetric*. A fuzzy relation that is reflexive and symmetric is denoted as a *proximity*, *compatibility or tolerance* relation.

Graphs are special case of relations. The concepts derived for graphs can be generalized for relations, see [47] [60][53].

2.2.2 Graphs

A relation $E(V, V)$ among entities in the set V, more specifically $e_{i,j}$, can represent: association or interaction, proximity or similarity, dissimilarity, and distance between entities. In the first case we have a crisp relation with values in $\{0, 1\}$, which we denote as crisp graph or simply, graph. In the second and third cases we have a weighted relation which takes values between $e_{i,j} \in [a, b]$, where a and b are real values (Weighted Graph). In the fourth case our relation $e_{i,j} \in [0, \infty]$ has values in the extended positive real set $[0, \infty]$ (Distance Graph).Moreover, we can transform a dissimilarity into a similarity and vice versa, by means of linear functions, as well as normalize the values $[a, b]$ into the unit interval $[0, 1]$; i.e., after normalization we can use the linear function $S = 1 - D$ to transform a dissimilarity into a similarity, or vice versa. Similarity or dissimilary graphs normalized in the unit interval $[0, 1]$ are called Fuzzy Graphs [47] [60][53]. From all of the above we can conclude that fuzzy graphs and weighted graphs are equivalent; we overview fuzzy graphs below. Nonetheless, it is important to point out that a distance graph cannot be normalized into a fuzzy graph by means of a linear function. This conversion, is the study of the present work.

2.2.3 Fuzzy Graphs

Since the seminal paper of Zadeh [99] the Theory of Fuzzy Sets has expanded into several fields in Mathematics. Graph Theory is one of the fields where the concepts of Fuzzy Sets can be applied.

Definition 2.1. (Fuzzy Set) *Given a relevant universal set X of elements x, a fuzzy set A is defined by a membership function:*

$$\mu : X \to [0, 1]$$

We can see from this definition that a fuzzy set generalizes crisp set by endowing each element, with a degree of membership in a set.

Example 1 (Fuzzy sets versus Crisp sets) Consider the universal set $X = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$. An example of a crisp set is the set

$$H = \{1, 2, 3, 4\}$$

using the fuzzy set notation, all members on this set have membership *one* and the members $\{0, 5, 6, 7, 8, 9\}$ have membership *zero*.

An example of a fuzzy set is the set

$$A = \{(1, 0.5), (2, 1.0), (5, 0.1)\}$$

in this example 1 belongs to the fuzzy set A with a membership of 0.5, 2 with membership 1.0 and 5 with membership 0.1. The remaining elements from X have membership *zero*.

Definition 2.2. (t-norm) *A triangular norm (t-norm for short) is a binary operation T on the unit interval $[0,1]$, i.e., a function $T : [0,1]^2 \to [0,1]$, such that for all $x, y, z \in [0,1]$ the following four axioms are satisfied:*

(T1) $T(x,y) = T(y,x)$.

(T2) $T(x, T(y,z)) = T(T(x,y), z)$.

(T3) $T(x,y) \leq T(x,z)$ *wherever* $y \leq z$.

(T4) $T(x,1) = x$.

A *t-norm* is a generalization of intersection in set theory and conjunction in logic. It was first defined in the context of probabilistic metric spaces [80].

Definition 2.3. (t-conorm) *A triangular conorm (t-conorm for short) is a binary operation S on the unit interval $[0,1]$, i.e., a function $S : [0,1]^2 \to [0,1]$, such that for all $x, y, z \in [0,1]$, satisfies (T1)-(T3) and*

(S4) $S(x,0) = x$.

A *t-conorm* is a generalization of union in set theory and disjunction in logic.

There is an innumerable number of t-norms and t-conorms. In the following examples [51] we present the four basic t-norms and t-conorms.

Example 2 (t-norms) The following are the four basic t-norms T_M, T_P, T_L and T_D given by, respectively:

$T_M(x,y) = min(x,y)$ (minimum),

$T_P(x,y) = x \cdot y$ (product),

$T_L(x,y) = max(x+y-1,0)$ (Lukasiewicz t-norm),

$$T_D(x,y) = \begin{cases} 0, & \text{if } (x,y) \in [0,1[^2; \\ min(x,y), & \text{otherwise.} \end{cases}$$

(drastic product)

These t-norms cover the range for t-norms, from the strongest t-norm T_M to the weakest t-norm T_D. There are other t-norms, namely *parametric t-norms*, which range the spectrum of all possible t-norms. Examples of these t-norms are the *Dombi* and *Hammaker* t-norms, which we are going to use in this dissertation.

Definition 2.4. (Dombi)(t-norm) *The definition of Dombi* t-norm *is the following:*

$$DT_{\wedge}^{\lambda}(a,b) = \left\{ 1 + \left[\left(\frac{1}{a} - 1 \right)^{\lambda} + \left(\frac{1}{b} - 1 \right)^{\lambda} \right]^{\frac{1}{\lambda}} \right\}^{-1}$$

Where the parameter $\lambda \in]0, +\infty[$.

Definition 2.5. (Hamacher)(t-norm) *The definition of Hamacher* t-norm *is the following:*

$$H_{\wedge}^{r}(a,b) = \frac{ab}{r + (1-r)(a+b-ab)}$$

Where the parameter $r \in]0, +\infty[$.

Example 3(t-conorms) The following are the four basic t-conorms S_M, S_P, S_L and S_D given by, respectively:

$S_M(x,y) = max(x,y)$ (maximum),

$S_P(x,y) = x + y - x \cdot y$ (probabilistic sum),

$S_L(x,y) = min(x+y,1)$ (Lukasiewicz t-conorm),

$$S_D(x,y) = \begin{cases} 1, & \text{if } (x,y) \in [0,1[^2; \\ max(x,y), & \text{otherwise.} \end{cases}$$

(drastic sum)

These t-conorms define the specific range of t-conorms, from the strongest t-conorm S_D to the weakest t-norm S_M.

Definition 2.6. (Dombi)(t-conorm) *The definition of Dombi* t-conorm *is the following:*

$$DT^{\lambda}_{\vee}(a,b) = \left\{ 1 + \left[\left(\frac{1}{a} - 1 \right)^{\lambda} + \left(\frac{1}{b} - 1 \right)^{\lambda} \right]^{-\frac{1}{\lambda}} \right\}^{-1}$$

Where the parameter $\lambda \in]0, +\infty[$.

Definition 2.7. (Hamacher)(t-conorm) *The definition of Hamacher* t-conorm *is the following:*

$$H^{r}_{\vee}(a,b) = \frac{a + b + (r-2)ab}{r + (r-1)ab}$$

Where the parameter $r \in]0, +\infty[$.

Now we are able to define the transitivity property of a fuzzy relation.

Definition 2.8. (Transitivity) *A fuzzy relation $R(X, X)$ is transitive if*

$$R(x, y) \geq t - norm(R(x, z), R(z, y))$$

is satisfied $\forall x, y, z \in X$.

Definition 2.8 entails that transitivity depends on the pairs t-norm chosen.

Definition 2.9. (Fuzzy Complement) *A complement c of a fuzzy set satisfies the following axioms:*

(c1) $c(0) = 1$ $c(1) = 0$ (boundary conditions).

(c2) $\forall a, b \in [0, 1]$ if $a \leq b$, then $c(a) \geq c(b)$ (monotonicity).

The *Complement* of a fuzzy set measures the degree to which a given element of the fuzzy set does not belong to the fuzzy set. Two most desirable requirements, which are usually among of fuzzy complements are:

Definition 2.10. (Fuzzy Complement)(cont) *A complement c of a fuzzy set satisfies the following axioms:*

(c3) c is a continuous function.

(c4) c is involutive, which means that $c(c(a)) = a$ for each $a \in [0, 1]$.

In classical set theory, the operations of intersection and union are dual with respect to the complement in the sense that they satisfy the De Morgan laws. It is desirable that this duality be satified for fuzzy set as well. We

say that a t-norm T and a t-conorm S are dual with respect to a fuzzy complement c if and only if

$$c(T(a,b)) = S(c(a), c(b))$$

and

$$c(S(a,b)) = T(c(a), c(b)).$$

Examples of dual t-norms and t-conorms with respect to the complement $c_s(a) = (1-a)^s$ are:

$$< min(a,b), max(a,b), c_s >$$

$$< DT(a,b), DS(a,b), c_s >$$

$$< HT(a,b), HS(a,b), c_s > .$$

We can have weaker complements, which only obey to the first two axioms in definition 4.8 to allow other t-norm and t-conorm operators.

Next we follow with composition of fuzzy relations.

Definition 2.11. (Relation Composition) *Consider two binary fuzzy relations, $P(X,Z)$ and $Q(Z,Y)$ with a common set of Z. The standard composition of these relations, which is denoted by $P(X,Z) \circ Q(Z,Y)$ produces a*

binary fuzzy relation $R(X,Y)$ *on* $X \times Y$ *defined by*

$$R(X,Y) = [P \circ Q] = t - conorm(t - norm[P(x,z), Q(z,y)]),$$

$\forall x \in X$ *and* $\forall y \in Y$ *and* $\forall z \in Z$.

Algorithm 2.1. (Transitive Closure) *Given a binary fuzzy relation* $R(X,X)$*, its transitive closure* $R_T(X,X)$ *can be determined by a simple algorithm that consists of the following three steps:*

(1) $R' = R \cup (R \circ R)$.

(2) if $R' \neq R$*, make* $R = R'$ *and go to Step 1.*

(3) Stop: $R^T = R$.

The union in step 1 must be in accordance with the t-conorm defined in the relation composition. The resulting relation in step 3 is transitive with respect to the t-norm, t-conorm operations used. Moreover, given the last algorithm, a fuzzy graph is transitive if the algorithm stops at the first step. A reflexive, symmetric and transitive fuzzy relation is denominated as a *Similarity* or *Equivalence* relation.

As mentioned before, a graph can be defined as a relation $E(V,V)$ where V represents the set of vertices and E a relation among vertices on V. A Fuzzy graph, or weighted graph, can also be defined as a fuzzy relation $E : (v_i, v_j) \rightarrow [0,1]$ of a fuzzy set V with it self. If the vertex set V of a fuzzy graph $G = (V,E)$ is a fuzzy set, then the vertices are also weighted with a membership function $\mu : V \rightarrow [0,1]$. If V is a crisp set, then the

vertices are not weighted, and are simply present or not, i.e. $\mu : V \to \{0, 1\}$. In either case, the edges E of G are weighted with the relation weights $E : (v_i, v_j) \to [0, 1]$. In this dissertation we work with a crisp vertex set V without loss of generality. In summary a fuzzy graph $G = (V, E)$ is defined by a fuzzy set or a crisp set of vertices V, and a fuzzy set of edges given by the relation E.

2.2.4 More about t-norms and t-conorms

In this subsection we give a more detailed description of t-norms and t-conorms, fundamental to this thesis.

t-norms

The intersection of two fuzzy sets A and B is performed by a binary operation closed on the unit interval. There are an infinite number of t-norms from definition 2.2. One important class is that of *Archimedean t-norms*, see [53]. Before we introduce one of the fundamental theorems of t-norms, which provides us a method for generating Archimedean t-norms we introduce the following definitions:

Definition 2.12. *A decreasing generator φ is a continuous decreasing function from the unit interval $[0, 1]$ into the real extended interval $[-\infty, +\infty]$.*

Definition 2.13. *The pseudo-inverse of a decreasing generator φ is defined*

by

$$\varphi^{(-1)}(a) = \begin{cases} 1 & for \quad a \in (-\infty, 0) \\ \varphi^{-1}(a) & for \quad a \in [0, \varphi(0)] \\ 0 & for \quad a \in (\varphi(0), \infty) \end{cases}$$

Where φ^{-1} is the inverse function of φ.

Theorem 2.1. (Characterization Theorem of t-norms) *Let i be a binary operation closed on the unit interval. Then, i is an Archimidean t-norm iff there exists a decreasing generator φ such*

$$i(a, b) = \varphi^{(-1)}(\varphi(a) + \varphi(b))$$

for all $a, b \in [0, 1]$.

With this last theorem we can generate an infinite class of t-norms. Among many decreasing generators is the *Dombi t-norm generator*, (see definition 2.4):

$$\varphi(x) = \left(\frac{1-x}{x}\right)^{\lambda}$$

Parameter λ allow us to obtain the range from the T_D t-norm ($\lambda \to 0$) to the T_M t-norm ($\lambda \to +\infty$). For many other decreasing generators, see [51].

t-conorms

Set unions are generalized by the t-conorms in definition 2.3. There are an infinite number of t-conorms and ways to generate new t-conorms. One

important class of t-conorms is the *Archimedean t-conorms*, see [53].

Definition 2.14. *A increasing generator θ is a continuous increasing function from the unit interval $[0,1]$ into the real extended interval $[-\infty,+\infty]$.*

Definition 2.15. *The pseudo-inverse of a increasing generator θ is defined by*

$$\theta^{(-1)}(a) = \begin{cases} 0 & for \quad a \in (-\infty, 0) \\ \theta^{-1}(a) & for \quad a \in [0, \theta(0)] \\ 1 & for \quad a \in (\theta(0), \infty) \end{cases}$$

Where θ^{-1} is the inverse function of θ.

Theorem 2.2. (Characterization Theorem of t-conorms) *Let u be a binary operation closed on the unit interval. Then, u is an Archimidean t-conorm iff there exists an increasing generator θ such*

$$u(a,b) = \theta^{(-1)}(\theta(a) + \theta(b))$$

for all $a, b \in [0,1]$.

With this last theorem we can generate an infinite class of t-conorms. Among many increasing generators is the Dombi t-conorm generator:

$$\theta(x) = \left(\frac{x}{1-x}\right)^{\lambda}$$

Parameter λ allow us to obtain the range from the S_M t-conorm ($\lambda \to 0$) to S_D t-conorm ($\lambda \to +\infty$). For many other decreasing generators the reader

can see, [51].

2.2.5 Distance Graphs

A *distance graph* is a particular kind of weighted graph defined on the extended real line. In other words, the relation which defines the edge weights is a distance function

$$d : (X, X) \to [0, \infty]$$

such that

$$d(x, x) = 0$$

$$d(x, y) = d(y, x).$$

Distance graphs are very intuitive formalism to optimize associations among entities. Indeed, modelers who work with weighted graphs (dissimilarity or proximity) usually calculate distances between vertices. However, to convert a dissimilarity or proximity graph, where the edge weight are defined in the unit interval, into a distance graph, where the edge weight are defined in the real line, they have to apply a nonlinear function. Then to get the shortest distance between any two vertices in the weighted graph they typically, apply the All-Pairs Shortest Paths (APSP) Johnson's algorithm [23] by calling the Dijkstra algorithm n times (for simplicity of language we will call this algorithm the APSP Dijkstra algorithm), we will introduce this later in this chapter.

In general, distance graphs obtained from real-world data violate the triangle inequality, and therefore, are semi-metric. That is, the binary relation which defines the distance graph is symmetric and anti-reflexive but, there are indirect edges that breaks the triangle inequality, see [72] for more details. In order to extract some properties from these graphs we have to embed the semi-metric graph in a metric space using the APSP Dijkstra algorithm so as to close all transitivities in the graph. This procedure causes some distortion to the original semi-metric graph; which can be problematic because it has strong effects in the topology of these graphs.

In summary, we can always use a linear function to convert a fuzzy graph into another, e.g. a proximity to a dissimilarity graph. Same for any weighted graph where edge weights are bounded by finite limits, can be linearly mapped onto $[0, 1]$. But because distance graphs rely on edges defined on the extended positive real line, the conversion to a fuzzy graph and vice versa, must rely on a nonlinear function.

2.2.6 Properties of Fuzzy Graphs

The *order* or *size* of a graph $G = (V, E)$ is equal to the number of vertices $v_i \in V$ in the graph and it is written as $|V|$, the number of edges is written as $\|E\|$. A graph $G = (V, E)$ is *finite* or *infinite* according to its order. Unless otherwise stated, the graphs or weighted graphs we consider are all finite. Two vertices v_i and v_j are *adjacent* or *neighbors*, if $e_{i,j}$ or $e_{j,i}$ is an edge of G. Two edges $e_{i,j} \neq e_{k,l}$ are *adjacent* if they have a vertex in common $j = l$

or $i = k$, but not both conditions simultaneously.

Two $G_1 = \{V_1, E_1\}$ and $G_2 = \{V_2, E_2\}$ are *isomorphic* if there is an edge-preserving bijective vertex mapping $\varphi : V_1 \to V_2$, i.e. a bijection φ with

$$\forall u, v \in V_1 : e_{u,v} \in E_1 \Leftrightarrow e_{\varphi(u),\varphi(v)} \in E_2.$$

If $V_1 \subseteq V_2$ and $E_1 \subseteq E_2$, then G_1 is a *subgraph* of G_2 (and G_2 is a *supergraph* of G_1), written as $G_1 \subseteq G_2$.

The vertex *strength* s_i is defined by:

$$s_i = \sum_{j \in \nu(i)} e_{i,j} \tag{2.3}$$

where $\nu(i)$ is the set of neighbors of vertex i, and $e_{i,j}$ is the weight of edge between vertex i and vertex j.

A *path* in a Graph $G = (V, E)$ is a sequence of distinct adjacent edges $P = e_{0,1}, ..., e_{n,m}$. The length of the path is given to be the sum of all edges on the path, and makes more sense in the context of distance graphs.

The *path length* can be defined in various ways, such as the sum of all weights in path or the smallest weight in the path. In many applications we are interested in determining the shortest path between any two vertices in a fuzzy graph.

2.3 Complex Networks

We conceptualize a network as a graph $G = (V, E)$ where V is a set of *vertices* (or nodes) v_i and E is a set of *edges* $e_{i,j}$ which represent a connection or association between vertices v_i and v_j; If the graph is directed $e_{j,i}$ is not necessary equal to $e_{i,j}$. The *degree* (or valency) k_i of a vertex v_i is the number of connected vertices (incident edges) to v_i. In a directed graph, the *indegree* k_i^+ of a vertex v_i is the number of edges $e_{j,i}$ terminating at v_i, and the *outdegree* k_i^- of a vertex v_i is the number of edges $e_{i,j}$ originating at v_i. From this point on, unless otherwise specified, in the case of directed graphs, we will use degree (k_i) to mean indegree (k_i^+).

It is useful to characterize large graphs by their *degree distribution*, which is the distribution of the probability that the degree of a randomly chosen vertex is k [68]. A *power law distribution* is a distribution that follows the relation

$$P(k) \simeq ak^{-\gamma}$$

where γ and a are constants. Newman [64] defines a *scale-free* network as a graph whose degree distribution follows a power law.

2.3.1 The Barabasi-Albert Model

Given an initial connected network (or graph) G with n_0 vertices, generally a small network, the Barabasi-Albert model (BAM) [11] is based on the following axioms:

Axiom 1 (Growth). *A new vertex v_g is added to G at each time step;*

Axiom 2 (Preferential attachment). *An edge $e_{g,i}$ between v_g and $m \leq n_0$ vertices v_i is created at each time step with probability:*

$$\Pi(e_{g,i}) = \frac{k_i}{\sum_j k_j},$$

where k_i is the degree of vertex v_i in the previous time step and $\sum_j k_j$ is the total sum of the degree of every vertex in the network in the previous time step. In other words, the preferential attachment axiom, biases the generation of new edges towards vertices with higher degree. With these considerations and the evolution equation,

$$\frac{\partial k_i}{\partial t} = m \cdot \Pi(e_{g,i}) \tag{2.4}$$

where, in our case, the constant m is the rate of edges we are introducing each time step, Barabasi and Albert [11] have shown that the model generates a power law distribution, which is independent of time:

$$P(k) \propto k^{-3}$$

The growth and preferential attachment axioms implement the mechanism known as "the rich gets richer". This mechanism can be generalized in many ways, which are beyond the scope of this thesis, for an overview see [68].

2.3.2 The Amaral et al. cut-offs Model

Amaral et al [6] noticed that in several real networks the power law describing the degree distribution is truncated (or cut-off) for vertices with large degrees. In other words, the number of highly connected vertices is smaller than expected from the preferential attachment model. Several mechanisms can be behind this behavior. Strictly speaking, Amaral et al proposed two alternative mechanisms, which interact with the two axioms of the BAM: aging of the vertices and cost of adding edges to vertices. Both mechanisms produce a power law truncation, i.e., a cut-off in the power law degree distribution. Each alternative mechanism proposed by Amaral et al. [6] is defined by an additional axiom to the BAM axioms:

Axiom 3a 1 (Aging of the vertices). *at each time step every vertex may become inactive with a constant probability of aging p.*

An inactive vertex and its edges are still present in the network but it is not allowed to receive more edges. The other mechanism, cost of adding edges to vertices, is similarly implemented by an alternative third axiom:

Axiom 3b 1 (Cost of adding edges to vertices). *each vertex has a limit capacity k_c of edges that it can support. After this threshold a vertex becomes inactive.*

Axiom 3*b* leads to networks whose degree distribution follows the power laws obtain via BAM, except that it observes a spike at k_c, followed by an

abrupt and unrealistic cut-off. Therefore, the model obtained by axiom $3b$ is not as realistic and interesting as the one obtained via axiom $3a$ aging of the vertices, which is the only one we will discuss from now on.

Amaral et al. [6] have shown, with simulations that the BAM with axiom $3a$ leads to a truncation or a cut-off of the expected power law degree distribution for several probabilities p—the behavior observed in many real networks. This truncation is more prominent with higher values of p. However, the simulations of Amaral et al. do not allow us to determine precisely the ranges of values of p which allow the network to grow. We also, for instance,do not have a precise notion of how the vertices become inactive, or how many vertices are expected to be active in the network at a given time for various values of p. The model we propose in Chapter 3 proposes an analytical solution to these questions.

2.3.3 Small-World

Watts and Strogatz [97] introduced a model to describe the Small-World phenomena in complex networks. Small-World networks are characterized by a high clustering coefficient and a short average path length.

The Clustering Coefficient measures the local group cohesiveness [97]. According to Watts the clustering coefficient $C(i)$ for a given vertex i is defined by the following equation:

$$C(i) = \frac{e_i}{k_i(k_i - 1)/2} \tag{2.5}$$

where k_i is the degree of vertex i and ei the edges between vertex i and the neighbors of vertex i. The average clustering coefficient is defined by the equation:

$$< C >= \frac{1}{N} \sum_i C(i) \tag{2.6}$$

There are other definitions of the clustering coefficient such the one by Wasserman and Faust [96], where the clustering coefficient measures the number of local transitivities in the graph. In this manner the clustering coefficient is defined in the following equation:

$$C = \frac{3 \times numberoffullyconnectedtriples}{numberoftriples} \tag{2.7}$$

The *average shortest path* is the average of all shortest paths between all vertices in the graph.

Watts and Strogatz [97] studied several real networks and found they have a high clustering coefficient and a low average path length. In contrast, random networks studied by Erdos and Renyi [21], have a small average path length but low clustering coefficient and circular regular lattice networks, have a high average path length and a high clustering coefficient. Watts and Strogatz showed that real networks are in between these two kinds of networks. Watts and Strogatz classify the networks which have small av-

erage path length and high clustering coefficient as Small-Word networks. The average path length of small-world networks are characterized by small diameters, small average path lengths, which scale with the logarithm of the size of the network $< l > \approx log(N)$.

Watts and Strogatz [97] also proposed a generative model to explain the apperance of Small-World networks. This model starts with a circular regular lattice, then we start rewiring the network by rewiring one random edge to a random vertex in the network until we reach a completely random networks, figure 2.1. With this model we obtain all kinds of networks from regular lattice, small-world to random, according to the degree of rewirement.

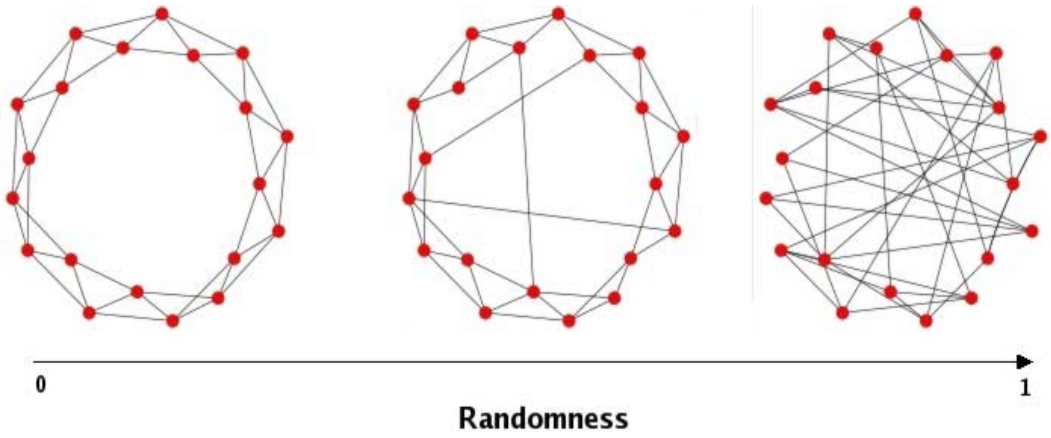

Figure 2.1: Regular Lattice (left), Small-World (midd), Random (right) Networks.

2.3.4 Clustering Coefficient in weighted graphs

There are many generalizations of clustering coefficients for weighted graphs [12][15] [86][67]. According to Barrat et al. [12] the *weighted clustering coefficient* for a given vertex i in a undirected graph is defined as:

$$C^w(i) = \frac{1}{s_i(k_i - 1)} \sum_{j,h} \frac{e_{i,j} + e_{i,h}}{2} a_{ij} a_{ih} a_{jh} \tag{2.8}$$

The term $\sum_{j,h} a_{ij} a_{ih} a_{jh}$ counts the number of triples in the neighborhood of vertex i. $\frac{e_{i,j}+e_{i,h}}{2}$ is the weight of two participant edges of vertex i for each triple in the neighborhood of vertex i. $s_i(k_i - 1)$ is a normalization factor to ensure that $C^w(i)$ is between 0 and 1, where s_i is the strength of vertex i and k_i the degree.

A definition for direct and undirected graphs is proposed by Onnela et al. [67]. They renormalize the clustering coefficient of equation 2.6 of the equivalent crisp network. The renormalization factors, *intensity* and *coherence* are defined in the following way:

$$I(h) = \left(\prod_{(i,j) \in l_h} e_{i,j} \right)^{\frac{1}{|l_h|}} \tag{2.9}$$

$$Q(h) = \frac{I(h) \times |l_h|}{\sum_{(i,j) \in l_h} e_{i,j}} \tag{2.10}$$

where $I(h)$ is the Intensity for vertex h, $e_{i,j}$ the weight between vertex i and j, l_h is the number of edges for vertex h and $Q(h)$ is the coherence for vertex

h. The *average intensity* and the *average coherence* are defined as:

$$< I(h) >= \frac{1}{e_h} \sum_{i \in N(h)} I(i) \tag{2.11}$$

$$< Q(h) >= \frac{1}{e_h} \sum_{i \in N(h)} Q(i) \tag{2.12}$$

where $N(h)$ denotes the neighborhood of vertex h and e_h the edges among vertex h neighbors. The clustering coefficient for vertex h in a weighted network is now defined as:

$$C^w(h) =< I(h) > \times C(h) \tag{2.13}$$

$$C^w(h) =< Q(h) > \times C(h) \tag{2.14}$$

where $C(h)$ is the clustering coefficient defined in equation 2.6 for the binary contra-part of the weighted graph. The clustering coefficient for the network is now defined by the renormalization as:

$$< C^w >= \frac{\sum_i < I(i) > \times C(i)}{\sum_i < I(i) >} \tag{2.15}$$

$$< C^w >= \frac{\sum_i < Q(i) > \times C(i)}{\sum_i < Q(i) >} \tag{2.16}$$

2.3.5 Statistical properties of networks

Besides the characteristic average path length, clustering coefficient and degree distribution, there are other important statistical properties to characterize networks such as the *betweenness distribution* and *degree correlations*.

Betweenness is a centrality measure. The shortest path betweenness centrality of a vertex v or edge e between two vertices s and t is based on the proportion of shortest paths that contain the vertex v or edge e and the number of shortest paths between s and t [23].

$$\delta_{st}(v) = \frac{\sigma_{st}(v)}{\sigma_{st}}$$

Then the total betweenness centrality $c_B(v)$ of a vertex v is given by [23]:

$$c_B(v) = \sum_{s \neq v \in V} \sum_{t \neq v \in V} \delta_{st}(v)$$

For edge betweenness we substitute in formulas above v for e.

Betweenness centrality gives us the importance of a given vertex or edge in a network. It became popular with the work of Girvan and Newman [38] on community detection.

The betweenness distribution is the probability distribution $P_{c_B}(c_B)$ that a vertex has betweenness c_B [15].

Another statistical property of complex networks is the *degree correlations*

[15]:

$$k_{nn}(k) = \sum_{k'} k' P(k'|k)$$

If the degrees of neighboring vertices are uncorrelated then $k_{nn}(k)$ is a constant. If there are correlations and if we can identify correlations increasing or decreasing with k, they are Assortative or Disassortative, respectively. Assortative networks, i.e., vertices with high degree have a larger probability of being connected with large degree vertices. Disassortative networks, high degree vertices, have a majority of neighbors with low degree vertices [15].

2.4 Shortest Paths and the APSP Dijkstra algorithm

The computation of the shortest-path distances between one specific vertex, called the source, and all other vertices is a classical algorithmic problem, known as Single Source Shortest Path (SSSP) problem [23]. The problem of computing the shortest path distances between all vertex pairs is called All-Pairs Shortest Paths problem (APSP) [23].

Dijkstra [29] [23] provided the first polynomial-time algorithm for SSSP for graphs with non-negative edge weights. Figure 2.2 presents the SSSP Dijkstra algorithm.

The APSP on a weighted graph is achieved by the Johnson's algorithm [23]. It first calculates the distances from an artificial source to all vertices

Algorithm Dijkstra's SSSP algorithm

Input: Graph $G = (V, E)$, edge weights $\omega : E \to \mathbb{R}$, source vertex $s \in V$
Output: Shortest path distances $d(s, v)$ to all $v \in V$

$P = \emptyset, T = V$
$d(s, v) = \infty$ for all $v \in V, d(s, s) = 0, pred(s) = 0$
while $P \neq V$ **do**
    $v = \text{argmin}\{d(s, v) | v \in T\}$
    $P := P \cup v, T := T \setminus v$
    for $w \in N(v)$ **do**
        if $d(s, w) > d(s, v) + \omega(v, w)$ **then**
            $d(s, w) := d(s, v) + \omega(v, w)$
            $pred(w) = v$

Figure 2.2: SSSP Dijkstra Algorithm from [23]

in the graph using the Bellman-Ford algorithm [23] and then determines the distances of all pairs by calling the Dijkstra algorithm n times. This algorithm of APSP has a runtime of $\circ(n^2 log(n) + nm)$ [23]. In this thesis we call the Johnson's algorithm or the APSP Dijkstra algorithm.

2.5 Semi-metric behavior and closures

A *distance function* is a non-negative function that obeys the following axioms and defines a distance between elements of a given set X, where $d : X \times X \to \Re$.

1. $d(x, y) = 0$ if and only if $x = y$ (anti-reflexive)

2. $d(x, y) = d(y, x)$ (symmetry)

3. $d(x, z) \leq d(x, y) + d(y, z)$ (triangle inequality)

We say that a distance function is metric iff obeys axioms 1-3.

We further say that a distance function is *ultra-metric* iff follows axioms 1-2 and stronger triangle inequality:

4. $d(x,z) \leq max(d(x,y), d(y,z))$ (ultra-metric inequality)

which is a stronger restriction than 3. Naturally, if a distance function is ultra-metric is also metric.

A distance function is *semi-metric* if it follows axioms 1 and 2, but violates the triangle inequality (axiom 3). We can generalize the triangle inequality:

5. $d(x,z) \leq \rho(d(x,y) + d(y,z))$ (ρ-relaxed triangle inequality)

A distance graph is semi-metric if the distances defined by its edges obey axioms (1 and 2) but there is at least one edge, for some indirect path that violates the triangle inequality. More specifically, semi-metric behavior [72], arises when we measure the lengths of indirect paths between vertices in a distance graph and these are smaller than the direct edges length. The direct and indirect distance between two vertices v_x and v_z is given by $d(x,z)$ and $d_{shortest}(x,z) = d(x,y_1) + d(y_1,y_2) + ... + d(y_n,z)$ respectively, and from the ρ-relaxed triangle inequality we have:

$$\frac{d(x,z)}{d_{shortest}(x,z)} \leq \rho$$

where, if $\rho \leq 1$ for all pairs of vertices, the weighted graph is metric. If $\rho > 1$ than the weighted graph is semi-metric.

Rocha [72] defines several coefficients to measure the semi-metric behavior in a distance graph.

The *semi-metric ratio* is defined by:

$$s(v_i, v_j) = \frac{d_{direct}(v_i, v_j)}{d_{shortest}(v_i, v_j)} \tag{2.17}$$

Where $d_{shortest}$ is the shortest distance between vertices in the distance graph, and s is positive ≥ 1 for semi-metric pairs.

The *relative semi-metric ratio* is defined by:

$$rs(v_i, v_j) = \frac{d_{direct}(v_i, v_j) - d_{shortest}(v_i, v_j)}{d_{max} - d_{min}} \tag{2.18}$$

rs compares the semi-metric distance reduction to the maximum possible distance reduction in the distance graph, d_{max} is the largest distance in the graph, and $d_{min} = 0$ is the shortest distance in the entire graph. This ratio varies between 0 and 1 for semi-metric pairs, and negative for metric pairs.

The *semi-metric below ratio* is defined by:

$$b(v_i, v_j) = \frac{< d_{v_i} >}{d_{shortest}(v_i, v_j)} \tag{2.19}$$

where $< d_{v_i} >$ represents the average direct distance from v_i to all v_j for edges where $d_{direct}(v_i, v_j) > 0$. b is only applied to semi-metric pairs of vertices (v_i, v_j) where $d_{shortest}(v_i, v_j) < d_{direct}(v_i, v_j)$ and it measures how much the shortest indirect distance between v_i and v_j falls bellow the average distance of all vertices v_i directly connects v_j.

We define the *semi-metric percentage* (SM), as,

$$SM = \frac{\sum_{i,j} (s(v_i, v_j) > 1)}{|E|} \tag{2.20}$$

where $|E|$ is the total number of direct edges.

2.6 Community detection in graphs

In this section we will give a brief overview on community detection in complex networks, since this dissertation does not make any contribution or development into community detection.

Community detection or clustering in graphs or weighted graphs mean the same. It was introduced in 2002 by Girvan and Newman [38] as the term community detection in graphs, however the term clustering in graphs and weighted graphs were already used for graph partitioning [23].

There are several clustering algorithms for graphs and weighted graphs [23]. Most of these algorithms are based on the intra-cluster density in the cluster versus the inter-cluster sparsity between clusters. There are algorithms such the one by Girvan and Newman [38], which partitioning the graph by the use of centrality measure (edge betweenness).

In order to see if a given graph partition is good most of the time it is used an optimization process is applied using a utility function that maximize the intra-cluster density and the inter-cluster sparsity, so as to reach a better clustering or graph partition.

In Brandes et al [23] there is a good overview of the clustering algorithms until the year 2005 and the Fortunato review paper [35] (to be appear in Physics Reports) complements this overview.

2.7 Dynamics in Complex networks

In this section we just have made only some brief comments on the dynamics in complex networks, since this is not the subject of this dissertation.

Dynamics in complex networks is a vast area. Using techniques from Statistical Physics, Critical Phenomena and Mean-Field approximation solves several problems such as Phase Transitions. Resilience and robustness as well as Epidemic spreading in complex networks, etc. Most of this can be found in the book of Barrat et al. [15].

Chapter 3

Stochastic model for cut-offs in complex networks

We propose and analyze a novel stochastic model which explains, analytically, the cut-off behavior of real scale-free networks previously modeled computationally by Amaral et al. [6] and others. We present a mathematical model that can explain several existing computational scale-free network generation models. This yields a novel theoretical basis to understand cut-off behavior in complex networks, previously treated only with simulations using distinct models. Therefore, ours is an integrative approach that unifies the existing literature on cut-off behavior in scale-free networks. Furthermore, our mathematical model allows us to reach conclusions not hitherto possible with computational models: the ability to predict the equilibrium point of active vertices and to relate the growth of networks with the probability of

aging. We also discuss how our model introduces a novel and useful way to classify scale free behavior of complex networks.

3.1 Preferential Attachment with Vertex Aging

3.1.1 Stochastic Model

As we discussed in the previous section, the *Preferential Attachment with Vertex Aging* (PAVA) model of Amaral's et al. [6], is based on three axioms: growth, preferential attachment and aging of the vertices. In this section we propose a novel *Stochastic Theoretical Model* (STM) to study PAVA analytically.

Let us first analyze how the vertices (nodes) become inactive. This is a fundamental piece of the analysis. We start with a core, fully connected network of x_0 vertices. Notice that at each time step, axiom $3a$, is equivalent to computing $x(t)$ Bernoulli trials, one for each vertex, where $x(t)$ is the number of vertices at time t, and p is the probability that a vertex becomes inactive. The probability of l vertices remaining active after $x(t)$ independent Bernoulli trials is given by the binomial probability distribution:

$$P(l,t) = \begin{pmatrix} x(t) \\ l \end{pmatrix} \cdot (1-p)^l \cdot p^{(x(t)-l)} \tag{3.1}$$

Therefore, the dynamics of a network can be expressed by the following stochastic map, where for convenience $x(t) = x_t$ now represents the *mean* number of vertices at time t:

$$x_{t+1} = x_t + \alpha - px_t \tag{3.2}$$

where α the number of vertices we introduce at each time step. Because at each time step we perform x_t Bernoulli trials, and introduce α new vertices, there are $x_t + \alpha$ vertices in the next time step minus the ones that become inactive; the mean value of which for the binomial distribution is px_t [92]. We can rearrange the terms and write the map in the following way:

$$x_{t+1} = (1 - p)x_t + \alpha \tag{3.3}$$

The equilibrium points for this map, which refer to the situations when the network retains the same mean number of vertices from iteration to iteration, can be identified by solving the equation:

$$(1 - p)x_t + \alpha = x_t \tag{3.4}$$

which results in the unique equilibrium point $\bar{x}_t = \frac{\alpha}{p}$, that is asymptotically stable when $\|f'(\bar{x})\| < 1$, where the first derivative of the map is given by

$$f(x) = (1-p)x + \alpha \Rightarrow f'(x) = 1 - p$$

Therefore, the unique equilibrium point $\bar{x_t} = \frac{\alpha}{p}$ is asymptotically stable for $p > 0$. Interestingly, when $p = 0$ our stochastic map yields the pure BAM of section 2.3.1. In this case the dynamical system does not have an equilibrium point and it diverges, i.e. the network keeps growing in size. When $p = 1$ the system, of course, does not grow since all vertices become immediately inactive. Finally, when $0 < p < 1$ the system observes the single asymptotically stable equilibrium point $\bar{x_t} = \frac{\alpha}{p}$, which depends on the value of p.

The equilibrium behavior of our STM can be better appreciated when we look at the solution of our stochastic map (eq. 3.2) [55].

$$x_t = \begin{cases} x_0 + \alpha t & if \quad p = 0 \\ \left(x_0 - \frac{\alpha}{p}\right) \cdot (1-p)^t + \frac{\alpha}{p} & if \quad p \neq 0 \end{cases} \tag{3.5}$$
$$t = 0, 1, 2, \ldots$$

Because $(1 - p) < 1$ for the second condition ($p \neq 0$) we can see that the dynamical system converges to the asymptotically stable equilibrium point $\bar{x} = \frac{\alpha}{p}$. In other words, after a transient the dynamical system converges to a network with a fixed mean number of active vertices and the system

remains in that state forever. This transient can be estimated as the time it takes for $\left(x_0 - \frac{\alpha}{p}\right) \cdot (1-p)^t$ to become sufficiently small, which can be better appreciated with a little manipulation of this expression:

$$\left(x_0 - \frac{\alpha}{p}\right) \cdot (1-p)^t = A \cdot e^{t \ln(1-p)} = A \cdot e^{-\frac{t}{t_0}} \tag{3.6}$$

where,

$$A \equiv \left(x_0 - \frac{\alpha}{p}\right)$$
$$t_0 \equiv -\frac{1}{\ln(1-p)}$$

Now, because the map in our model is stochastic there is variation about the equilibrium point. In our model (equation 3.2), x_t is a binomial variable and for large enough t we can approximate it by a normal distribution [92] and study its variation. We assume the ergodic hypothesis is true, therefore the statistical mean of x_t, which we denote as $< x_t >$ is equal to $\bar{x_t} = \frac{\alpha}{p}$, the equilibrium point given by equation 3.5. In our simulation section we validate this assumption. The variations can then be studied by solving the equation for variance:

$$\sigma^2 =< x_t^2 > - < x_t >^2 \tag{3.7}$$

From equation 3.2 and 3.5 and from the ergodic hypothesis we have for

the statistical mean

$$\mu =< x_{t+1} >=< x_t >= \frac{\alpha}{p} \tag{3.8}$$

Extending x_t in equation 3.2 to real values to make the approximation to the normal distribution feasible, and substituting this equation and 4.9 into 3.7,

$$\sigma^2 =< ((1-p)x + \alpha)^2 > -\frac{\alpha^2}{p^2} \tag{3.9}$$

and assuming a normal distribution as discussed above we reach the following expression:

$$\sigma^2 = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} ((1-p)x + \alpha)^2 e^{\frac{-(x-\mu)^2}{2\sigma^2}}\, dx - \frac{\alpha^2}{p^2} \tag{3.10}$$

solving equation 3.10 in order of σ for $t > t_0$ (equation 3.6) we obtain,

$$\sigma = \sqrt{\frac{1}{1-(1-p)^2}} \tag{3.11}$$

For large t equations 4.9 and 3.11 define our stochastic variable x_t, as a normal variable.

3.1.2 Exponential decay

Let's now take a closer look at axiom $3a$. According to this axiom a vertex at each time step may become inactive with a constant probability p. Each

vertex follows a binomial distribution which is equivalent to a random walk process. In this case, it is as if each vertex is trying to give z steps in a maximum of r steps all in the same direction—where r can be interpreted as the maximum iterations t. If a vertex changes the direction of its step it becomes inactive. This can be expressed by the following probability function,

$$P(Z = z) = \frac{\begin{pmatrix} r - z \\ qr - z \end{pmatrix}}{\begin{pmatrix} r \\ qr \end{pmatrix}} \tag{3.12}$$

where $q = 1 - p$ is the probability that the vertex succeeds and remains active for the next step. This equation can be simplified in the following form:

$$\begin{aligned} P(Z = z) &= \frac{(r - z)!(qr)!(r - qr)!}{(qr - z)!(r - qr)!r!} \\ P(Z = z) &= \frac{\prod_{k=0}^{z-1}(qr - k)}{\prod_{k=0}^{z-1}(r - k)} \end{aligned} \tag{3.13}$$

For a given $r = 10,000$ and $p = 0.1$, we obtain the following distribution presented in figure 3.1.

We can see that the respective exponential decay is independent of r, by taking the limit when $r \gg 1$.

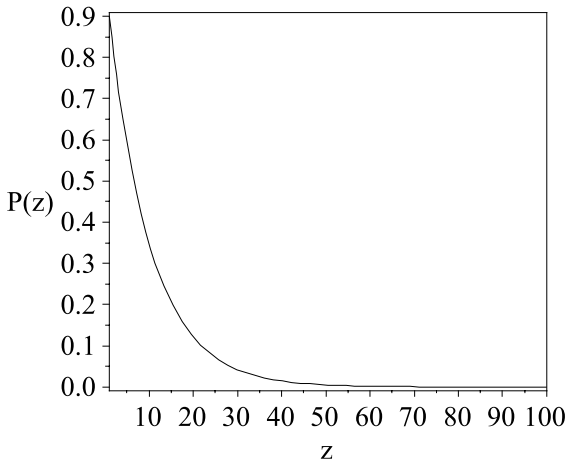

Figure 3.1: $P(Z = z)$ distribution

$$P(Z=z) = \frac{\prod_{k=0}^{z-1}(qr-k)}{\prod_{k=0}^{z-1}(r-k)} \underset{r\gg 1}{\cong} \frac{(qr)^{z-1}}{r^{z-1}} = q^{z-1} \tag{3.14}$$

In this last equation we can see that the binomial experiment of a vertex gets inactive is an exponential decay.

3.1.3 Exponential decay for the degree distribution

It was found in [87] that for networks where the number of active vertices is a subset of the total of vertices in the network, the power law degree distribution presents a truncation,

$$P(K=k) = C(a_m + b_m)k^{-(a_m+1)}e^{-b_m k} \tag{3.15}$$

Where m is the number of active vertices and a_m, b_m are parameters in function of m, and k the degree.

In our case we are in that situation, where the mean value of vertices active can be represented by m. In the next subsection we show how the STM can be validated by simulations of the PAVA model.

3.1.4 Network stop growing estimation

Because the STM is stochastic there is a probability that a network eventually will stop growing. From the results in sections; *Stochastic Model* and *Exponential decay*, we can estimate the probability for which a network will not grow. We have seen that we can approximate x_t as a normal distribution with mean and standard deviation given respectively by equations 4.9, 3.11. The next equation estimates the probability of variations (fluctuations) on x such that the network stops growing.

$$P(x \leq 0.5) = \Phi\left(\frac{0.5 - \mu}{\sigma}\right). \tag{3.16}$$

Where Φ is the cumulative probability function of a normal variable with mean μ and standard deviation σ—which depends exclusively on the values of α and p. We have chosen $x \leq 0.5$ in order to compensate the discrete extension of x_t to real values x. Moreover, this probability is the probability that a given network will die in t_{die} steps. This is exactly the idea behind equation 3.14, but now instead of a given vertex survival probability, we are

considering the probability of the network survival given by $w = 1 - P(x \leq 0.5)$, after t time steps and s trials, and for large s we have as before;

$$P(T = t) = \frac{\prod_{k=0}^{t-1}(ws - k)}{\prod_{k=0}^{t-1}(s - k)} \underset{r \gg 1}{\cong} \frac{(ws)^{t-1}}{s^{t-1}} = w^{t-1} = Be^{-\frac{t}{k_c}} \tag{3.17}$$

with,

$$B \equiv \left(\frac{1}{w}\right)$$
$$k_c \equiv -\frac{1}{\ln(w)}$$

In figure 3.2 we plot the probability for which a network will stop growing after t steps versus the probability of a vertex getting inactive. Most of network trials will stop growing after k_c. With this we can estimate the number of steps t_{die} for which the network stop growing. For example, for a probability of inactiveness $p = 0.2$ we have $P(x \leq 0.5) = 0.0035$, which means we will have $k_c = 288$. In this case with $p = 0.2$ most of the networks will stop growing after $t > 288$.

3.1.5 Simulations

In table 3.1, we can see the results that we have obtained comparing PAVA simulations with our analytic STM. First a confirmation of the system's

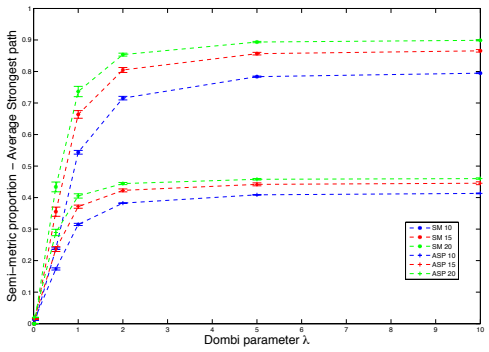

Figure 3.2: $P(x \leq 0.5)$ versus p (probability of a vertex getting inactive) for $\alpha = 1$.

Table 3.1: Comparison between PAVA and STM models for $\alpha = 1$ and $t = 10,000$. (*) measured after a transient period t_0 calculated after 11 simulations

	PAVA (*)	STM
p	$\bar{x} \pm \sigma$	$\bar{x} \pm \sigma$
0.1	10.01 ± 2.18	10.00 ± 2.29
0.08	12.49 ± 2.47	12.50 ± 2.55
0.06	16.62 ± 2.84	16.67 ± 2.93
0.05	19.50 ± 3.10	20.00 ± 3.20
0.03	33.56 ± 4.07	33.33 ± 4.11
0.01	99.75 ± 7.01	100.00 ± 7.09

stable equilibrium point for each p. For the PAVA model we performed eleven simulations for each probability p, from $p = 0.1$ to $= 0.01$, for a network with $10,000$ vertices and an $\alpha = 1$ — eleven simulations is usually considered the minimum experiments. In any case as it can be seen in table 3.1 eleven simulation yielded quite accurate results. The choice for p was based on results obtained in figure 3.2, where we can see for values of $p \leq 0.1$ the probability of the network not growing is essential null.

In table 3.1 we see that $\bar{x}$ is extremely well predicted by the STM, since values follow the same tendency as well as the standard deviation obtained by the PAVA simulations.

In figure 3.3 we can see that the standard deviation stabilizes after a transient period of time as was predicted by equation 3.6.

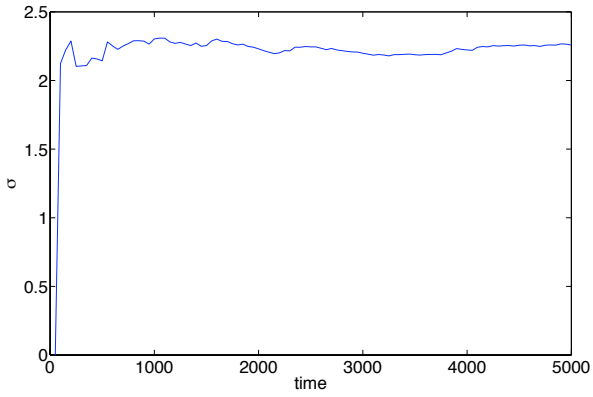

Figure 3.3: Evolution of the standard deviation of x_t for $p = 0.1$ and $\alpha = 1$ with time

In table 3.2 we compare the number of iterations for which most of the networks stop growing for both PAVA simulations and the STM. The choice for p was based on results obtained in figure 3.2. The STM values were estimated by using equation 3.17. To estimate the cut-off point of the exponential decay in PAVA we have fit with a 95% confidence an exponential function ae^{-bt}, to the experimental data. The estimated parameter b, allow us to compute $t_c = \frac{1}{b}$, which is the exponential cut-off. The comparison between PAVA and STM is made between t_c and k_c. There are some fluctuations but both follow the same tendency, which shows that eventually the

network will stop growing after t_c or k_c iterations.

From the results summarized in table 3.1 and table 3.2 it can be concluded that the number of active vertices observed by PAVA simulations follows the process described by the dynamic map inherent in the STM; in another words the aging of vertices process is a Binomial random process.

The PAVA and STM probability distribution $P(Z = z)$ for the number of steps a vertex succeeds without getting inactive is shown in figure 3.4 for $r = 10,000$ vertices and $p = 0.1$. In these figures we see the experimental probability distribution (dots) and the STM distribution fits well the equation 3.14, q^{z-1}, where $q = 1 - p = 0.9$. The results of the regression are an estimated $q = 0.89$ for 95% of confidence with $SSE = 0.04$ and $R^2 = 0.99$ and $RMSE = 0.02$.

In figure 3.5 we see that the degree distribution cut-off of the PAVA simulations does not change with the size as already observed by Amaral et. al. [6]. Similar results are observed for other p. The power law exponent γ does not change significantly; in the case of figure 3.5, the γ values are

Table 3.2: All vertex getting inactive after t iterations according to the probability p with $\alpha = 1$. t_c is the time step for which the PAVA network stop growing, $k_c = -\frac{1}{log(1-P(x\leq 0.5))}$ is the cut-off point and $P(x \leq 0.5)$ the theoretical probability for which a network will stop growing

	$PAVA$	STM
p	t_c	k_c
0.2	374	288
0.3	33	46
0.4	11	18

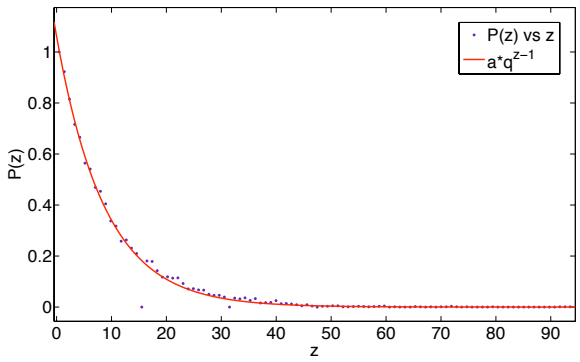

Figure 3.4: The simulation results for $P(Z = z)$ and curve fitting with and exponetial aq^{z-1}

around 2.6.

In figure 3.6, also as expected, the degree distribution cut-off point of the PAVA simulations decreases inversely with the probability of inactiveness p as already observed by Amaral et al. [6]. Also, the the power law exponent γ does not change significantly with p.

These results show that: (a) the network after a transient period of time t_0 reaches an equilibrium number of active vertices; (b) the network will eventually stop growing after t iterations according to the probability of inactiveness. According to all these results we can conclude that our STM is a good analytical model of the generation processes of PAVA, namely the aging of vertices process.

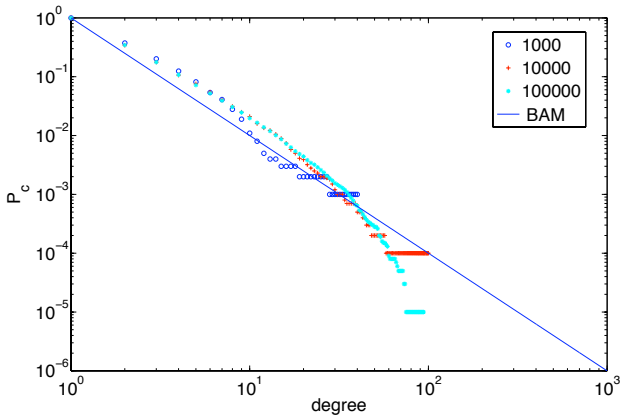

Figure 3.5: Cumulative degree distribution for network sizes: 1000, 10000 and 100000, with $\alpha = 1$ and $p = 0.05$. Also in solid we plot the BAM.

3.2 Discussion

We have seen in the previous sections that the PAVA model, follows a Binomial distribution and can be described by a discrete dynamical system. This dynamical system has an stable equilibrium point and two extreme behaviors, the pure random walk, when there is only one active vertex, and at the other extreme the Barabasi-Albert behavior, when we have all vertices active. Moreover, besides the simulations results made by Amaral et al. [6] (PAVA) our STM was also able to: (a) predict the equilibrium point of active vertices; (b) relate the growth of the network (size) with the probability of aging.

The PAVA simulations are an extension of the BAM and is useful for the study of several real complex networks, such the Actor network, Scientific Citations networks [24]. These networks are characterized by vertex dying

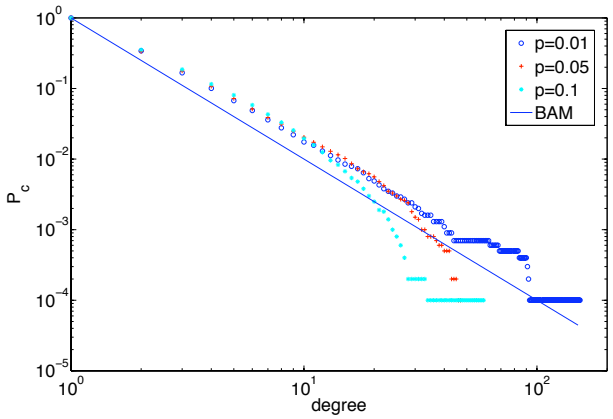

Figure 3.6: Cumulative degree distribution for probability of inactiveness $p = 0.1$, $p = 0.05$ and $p = 0.01$, with $\alpha = 1$ and size of the network$= 10,000$. Also in solid we plot the BAM.

over time. In the case of the actors network, an actor during a period of time plays in several movies with other actors and then they become inactive (die or retired). However, the influence of inactive actors still participates on the statistics of the network without removing the actor from the network. The same kind of reasoning can be applied to the network of citations, where a given paper becomes inactive (obsolete) after a given period of time but remains on the network.

Another way of interpreting Amaral's et al. PAVA simulations [6] and observe the relation between this generation process with other in the literature, is if we look at the network from the perspective of a new vertex. In PAVA a vertex gets inactive with a probability p as a result of a binomial process. Probabilistically equivalent to this is if a new vertex, at each time step, just sees a limited number of random vertices, the active vertices, which

are themselves limited by the equilibrium point reached by the network after a transient period of time. This can be seen as an information filtering done by each new vertex regarding the whole network. It is not possible for a new vertex to have a full knowledge of the entire network, it just has a partial knowledge. Therefore we can say the stochastic map defined in equation 3.5, represents the knowledge that a new vertex has about the entire network.

If $p = 0$ a new vertex has full knowledge of the network and in this case we are in one extreme, the purely Barabasi-Albert model. In the other extreme if we have a certain maximum $p_{max} < 1$ where the equilibrium point is $\bar{x} = 1$, only one vertex is active. In this case the new vertex does not have any knowledge at all; it just connects to the other vertices in a random way. The intermediate case happens when, $p < p_{max}$ and $\bar{x} > 1$. In this case each new vertex has partial knowledge of the network. Therefore, some scale-free networks range between two situations; absolute knowledge of the network, Barabasi-Albert model [11], and complete ignorance of the network, pure random process. The Amaral's et al. PAVA simulations [6] and obviously its STM formalization, seems to be a simpler model that could explain, as a first approximation, the general dynamic mechanisms behind scale-free networks between these two extremes. The parameter p measures how each vertex has complete knowledge of the network or complete ignorance.

3.3 Conclusions

In this work we have presented a stochastic theoretical model as a mathematical explanation of the Amaral's et al. PAVA model [6].

We believe this work can provide a simple explanation for the dynamics of some scale-free networks and through this knowledge, obtain a better understanding of how these scale-free networks can emerge. As we have seen at the introduction, the field of complex networks is an interdisciplinary field. Therefore a better understanding of the mechanisms behind complex networks can improve the understanding behind certain problems in areas like the Internet, World Wide Web, Neural Networks, Chemical Networks, Social Networks and so on.

Chapter 4

Generalized transitive closures on complex networks

Part of the work described in this chapter was published in [93], [76] and submitted for publication [84].

4.1 Introduction

Humans are unique in their ability to generate and utilize knowledge. In a nutshell, knowledge is valuable because it gives us ways to access new resources and to obtain them from others. Given the ubiquitous presence of the Web in our lives, much of our knowledge and social foraging is done online. Recently, thanks to computers and the Web infrastructure of systems for massive knowledge storage and distribution, we have crossed an impor-

tant threshold in our ability to collect, process and interpret quantitative data about the way knowledge and social interaction are organized [46]. We have just started to construct quantitative, large-scale characterizations of the geometry and dynamics of information exchanges and social interactions in human societies, but already there has been a breakthrough in our understanding of these interactions. Indeed, most interactions in knowledge and social space, as in other technological and biological systems, can be well described in terms of *complex networks* [5]. These windows into human organization are based on quantitative empirical observations that allow hypotheses to be better tested and falsified. This constitutes enormous progress relative to past standards.

The majority of research on complex networks treats interactions as binary edges in graphs, even though interactions in real networks exhibit a wide range of intensities or strengths. The varying strength nature of many, if not most, real networks has lead towards a more recent drive to study complex networks as weighted graphs [63, 14, 95, 39]. Certainly this shift towards weighted graphs as models of complex networks is welcomed. However, there is still much to do to bring decades of research on weighted graphs to bear on the field of complex networks. One field, in particular, that has accumulated substantial knowledge about weighted graphs is the field of Fuzzy Set Theory [53].

While the Fuzzy Set community has focused extensively on the mathematical characteristics of weighted graphs and how to compute them [60], it

has not focused much on the structure and dynamics of real networks obtained from and tested on empirical data. For instance, we show that the most common form of transitive closure in Fuzzy Graphs destroys the scale-free structure of complex networks for all networks we treated (see chapter 6). Conversely, the complex networks community has paid relatively little attention to the mathematics of weighted graphs. For instance, the very intuitive metric closure of distance graphs, related to the shortest-path Dijkstra's algorithm [29], has undesirable axiomatic features (see below).

We argue that in the field of complex networks there is still much to do to understand the *axiomatic* characteristics of various ways to compute transitive closures of weighted graphs obtained from real data. There is also ample need to study the effect of various forms of *transitivity* not only on the structure but also on the dynamics of complex networks modeled as weighted graphs. This is all the more relevant, as we discuss below, when we use *associative networks* as *knowledge representations* in social data mining, text mining and information retrieval. Indeed, the concept of transitive closure is important because it allows us to identify indirectly related items–which are potentially relevant and may possess an higher probability of direct co-occurrence in the future [72][76]. This has useful applications in recommender systems, text and literature mining, information retrieval, prediction of online social behavior, and social network modeling at large. However, unlike standard crisp graphs, in weighted graphs there is an infinite number of ways to compute transitive closures. Therefore, we need to understand which ones

preserve important characteristics of real networks as well as observe good axiomatic requirements. We show in this chapter that for a specific family of parametric logical connectives (the Dombi t-norm and co-norm family), it is possible to choose a transitive closure with optimal axiomatic and intuitive characteristics. However, there is a need to test which ones lead to better performance in information retrieval and recommender systems tasks (see chapter 5). Our mathematical insights lead to useful applications for extracting information from real knowledge, biological and social networks.

Our methodology is based on the extraction of networks (as weighted graphs) from large collections of documents, such as repositories of scientific articles. These networks have been used to build recommender systems for digital libraries [75, 72, 77, 78, 76], as we will see later in chapter 5.

4.2 Proximity Networks

Our approach is typically based on a specific proximity measure computed from fuzzy binary relations between any two sets of items (e.g. keywords and documents). This measure is a natural weighted extension [73] [75] [69] of the Jaccard similarity measure [40], which has been used extensively in computational intelligence [62] [74] [72] [76]. Given a generic binary relation R between sets X (of n elements x) and Y (of m elements y), we extract two complementary proximity graphs: XYP and YXP. $xyp(x_i, x_j)$ is the probability that both x_i and x_j are related via R to the same elements $y \in Y$

(and only those). Conversely, $yxp(y_i, y_j)$ is the probability that both y_i and y_j are related via R to the same elements $x \in X$ (and only those). In short, these measures equate proximity with co-occurrence; the respective formulas are:

$$xyp(x_i, x_j) = \frac{\sum_{k=1}^{m}(r_{ik} \wedge r_{kj})}{\sum_{k=1}^{m}(r_{ik} \vee r_{kj})}; \qquad yxp(y_i, y_j) = \frac{\sum_{k=1}^{n}(r_{ki} \wedge r_{kj})}{\sum_{k=1}^{n}(r_{ki} \vee r_{kj})} \tag{4.1}$$

Other co-occurrence measures can be used to capture a degree of association or closeness between elements of two sets in a binary relation. In information retrieval, in addition to variations of the Jaccard measure, it is common to use the cosine [9], Euclidean [89] and even mutual information measures [91]. For characterizing closeness in relations, we prefer our weighted Jaccard proximity measure because it possesses several desirable characteristics. The Euclidean measure is a similarity measure in the sense defined above (it is transitive for most commonly used criteria), but it generates non-sparse matrices, since all finite elements of the relation R lead to similarity greater than zero. This makes it impractical for very large data sets. The cosine proximity measure (which is not transitive for most commonly used criteria) is scale-invariant which makes it very appealing for text documents of varying size, but may be problematic in other domains. The weighted Jaccard measure has aspects of both the Euclidean and the cosine

measures [89], and leads to sparse matrices. We use our weighted extension of Jaccard proximity measure (eq. 4.1) in several applications. However, all of the theoretical work we propose below applies to any proximity graph (as defined above), independently of the measure used to obtain it from specific data sets.

4.3 Representing Knowledge in proximity networks

Proximity relations are fuzzy graphs which represent the closeness of elements in associative networks (e.g. terms extracted from documents, or users of a social networking web site). We derive our proximity networks using the proximity measures of formulae 4.1 computed from binary relations extracted from large collections of documents, websites, or records stored in databases. Such proximity graphs should be seen as *associative knowledge networks* that represent how often elements co-occur in a large set of documents [72, 78]. As with any other co-occurrence method, the assumption is that items that frequently co-occur are associated with a common concept, theme, or social community understood by the community of users and writers of the documents.

Notice that a proximity graph allows us to capture network associations rather than just pair-wise co-occurrence. In other words, we expect concepts or social communities to be organized in more interconnected sub-graphs, or

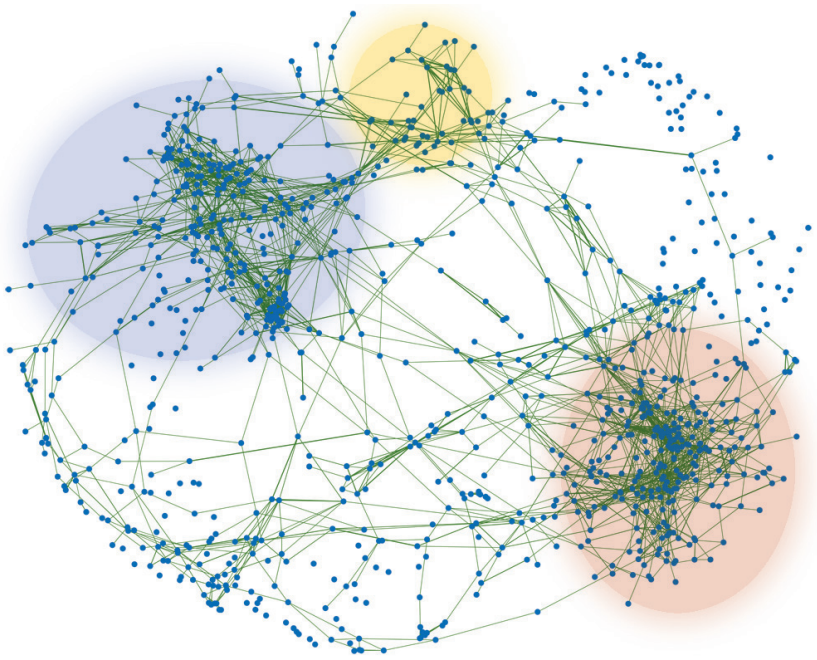

Figure 4.1: Social communities discovered in the proximity network of journals accessed by users of the *MyLibrary@LANL* recommender system [76]. In this proximity network, journals are closer to one another, if they tend to co-occur in the same user profile, and only in those. Drawn using the Fruchterman-Reingold algorithm in Pajek [16]. Figure reprinted from [76].

clusters of items in the proximity networks. Indeed, our group has successfully used proximity networks in several knowledge extraction and literature mining applications, such as several *BioCreAtIvE* Text Mining competition (Critical Assessment of Information Extraction in Biology) [93] [4] [54], as well as working recommender systems we have produced [77, 78, 76]. Figure 4.1 depicts a proximity network extracted from the recommender system we developed for the MyLibrary service of the digital library at the Los Alamos National Laboratory (LANL)[76]. The elements in this network are scientific

journals, and the proximity weights computed from co-occurrence of journals in user profiles. The figure clearly shows two main clusters of journals in the network. The Principal Component Analysis (PCA) analysis of this network revealed that the two first eigen-vectors (components) are very correlated with the two main clusters identified. The first component refers to a set of journals related to "Chemistry, Materials science and Physics" (left, blue). The second component refers to a set of journals related to "Computer Science and Applied Mathematics" (right, organge). However, these groups are further separated and refined into more specific clusters as we consider more components. A smaller third cluster in the figure refers to "Bioinformatics and Computational Biology" (top, yellow) [76].

The main clusters discovered in this network highlight the reality of the research pursued at LANL. Indeed, being a nuclear weapons laboratory, much of its research is concerned with Materials Science and Physics on the one hand, and Simulation and Computer Science on the other. Thus, the journal proximity network, produced from user profiles, captured the main communities of scientists (the users of MyLibray) at Los Alamos, as well as the knowledge associated with these communities (characterized by the journals in the respective components). Our user tests of the quality of recommendation based on this network were quite good [76]. Furthermore, the results of our proximity network approach on the Biocreative text mining competition [43], [70], [1], were among the very best results in the tasks we participated in [93]. This exemplifies how proximity networks can be seen as effective,

knowledge and social structure representations.

4.4 Semi-metric networks

A high value of proximity means that two items from one set (e.g. words) tend to co-occur frequently in another set of objects (e.g. web pages). But what about items that do not co-occur frequently with one another, but do occur frequently with the same *other* elements? In other words, even if two items do not co-occur much, they may occur very frequently with a third item (or more). Should we infer that the two items are related via indirect associations, that is, from *transitivity*?

In this work we study transitivity as a general topological phenomenon in weighted graphs such as proximity networks—where it can be computed in different ways. While the Fuzzy Set community has focused extensively on the mathematical characteristics of various possible conjunction/disjunction (T-Norm/T-Conorm) pairs to compute transitivity [48, 50, 58], it has not focused much on the structure and dynamics of real networks obtained from empirical data. Indeed, there is very little work on the identification of the most intuitive and appropriate forms of transitivity for information retrieval, text mining, or network analysis in general. Conversely, while the last decade witnessed a tremendous amount of scientific production towards understanding the structure of complex networks, including weighted networks to model Web processes (e.g. [12]), the complex networks community has paid little

attention to the effect of various forms of transitivity on network structure and dynamics.

To build up a more intuitive understanding of transitivity in weighted graphs, we convert our proximity graphs to distance graphs. Distance can be seen intuitively as the opposite of proximity. Various functions can be used to convert one into the other. Perhaps the most common way is to use a proximity-to-distance conversion function φ: $distance = \frac{1}{proximity} - 1$ [89], which is the Dombi t-norm generator with $\lambda = 1$ (see [51]). From the generic proximity measures XYP and YXP, obtained from a relation R between sets X and Y using formulae 4.1, we can compute generic distance functions among the elements of X or Y:

$$d_X(x_i, x_j) = \frac{1}{xyp(x_i, x_j)} - 1, \quad d_Y(y_i, y_j) = \frac{1}{yxp(y_i, y_j)} - 1 \tag{4.2}$$

where d_X and d_Y are distance functions because they are nonnegative, symmetric, real-valued functions such that $d(x, x) = 0$ (anti-reflexive) [37]. They define weighted graphs D_X and D_Y, which we refer to as *distance graphs*, whose vertices x_i or y_i are the elements of X or Y, and the edges are the values $d_X(x_i, x_j)$ and $d_Y(y_i, y_j)$, respectively. A small distance between elements implies a strong association between them.

In general, these distance graphs are not metric because, for some pair of elements x_1 and x_2, the triangle inequality may be violated: $d(x_1, x_2) \geq$

$d(x_1, x_3) + d(x_3, x_2)$ for some element x_3. This means that the shortest distance between two elements in D_X is not necessarily the direct edge but rather an indirect path. Distance functions that violate the triangle inequality are referred to as *semi-metrics* [37].

Rocha has compiled evidence [72] that pairs of elements with larger semi-metric behavior (those which possess at least one indirect path between them whose distance is much shorter than the direct link) denote a type of *latent association*. That is, an association which is not grounded on direct evidence provided by the relation R, but rather indirectly implied by the overall network of associations extracted from this relation. More formally, when $d(x_i, x_j) \gg d(x_i, x_k) + \cdots + d(x_l, x_m) + \cdots + d(x_p, x_j)$, then the edge (x_i, x_j) possesses a strong semi-metric or latent association in distance graph D_X. Clearly, semi-metric behavior is a question of degree: some semi-metric shortcuts are much shorter than others depending on how much the triangle inequality is violated. Thus, to measure a degree of semi-metric behavior we can use the *semi-metric* and *below average ratios* [72]:

$$s(x_i, x_j) = \frac{d(x_i, x_j)}{\underline{d}(x_i, x_j)}, \quad b(x_i, x_j) = \frac{\overline{d_{x_i}}}{\underline{d}(x_i, x_j)} \tag{4.3}$$

where $\underline{d}(x_i, x_j)$ is the shortest indirect distance between x_i and x_j in distance graph D_X, and $\overline{d_{x_i}}$ is the mean direct distance from x_i to all other $x_k \in X$ such that $d(x_i, x_k)$ is finite. $s > 1$ for semi-metric pair of elements. b is only applicable to semi-metric pairs of elements (x_i, x_j) where $0 < \underline{d}(x_i, x_j) <$

$d(x_i, x_j)$ and it measures how much the shortest indirect distance between x_i and x_j falls below the average distance of x_i to all its directly associated elements x_k. The below average ratio is designed to capture semi-metric behavior of pairs (x_i, x_j) which do not have a finite direct distance $d(x_i, x_j)$. Note that $b(x_i, x_j) \neq b(x_j, x_i)$. $b > 1$ denotes a below average distance reduction (see [72] for more details).

Rocha has proposed that in proximity graphs of keywords extracted from documents, a latent association identified by large values of s and b (eq. 4.3), implies novelty and can be used to identify trends [72]. We have also used and tested this idea, with good results, in a recommender system that was implemented at LANL's digital library [76]. In the case of this service, a strong semi-metric association in the journal network (figure 4.1) identifies a pair of journals that hardly co-occur in user profiles, but which are nonetheless very strongly implied via other journals which do co-occur with the pair.

Rocha in collaboration with Luis Bettencourt at LANL, have also tested our method on social networks of scientists, they started working with network of collaboration of scientists working on the field of Feynman diagrams from 1949 to 1956 [18]. While this is a very small network, it is validated by historical evidence compiled by David Kaiser at MIT. In ongoing, yet unpublished work, they have computed *co-collaboration* and *co-acknowledgement* networks using our proximity measure of eq. 4.1. The first network associates authors who tend to write papers with many of the same other au-

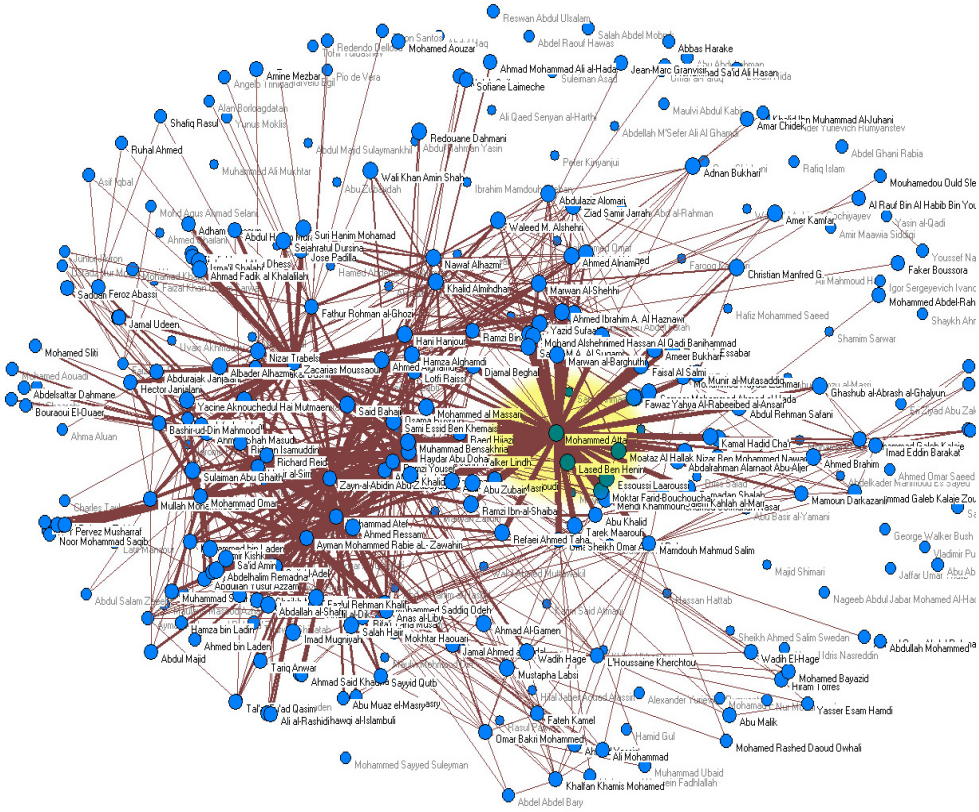

Figure 4.2: Terrorist proximity network obtained from intelligence data related to the 9/11 terrorist attacks on New York city and Washington DC [74]; strongly semi-metric edges, computed with parameter s in formulae (4.3), shown with thicker lines. The node for Mohammed Atta is highlighted. The strong links out of this node, denote potential terrorist associations not identified in intelligence data, but highly possible. Drawn using the Fruchterman-Reingold algorithm in Pajek [16]

thors, whereas the second associates authors who tend to acknowledge many of the same other authors. This preliminary study shows that a strong semi-metric behavior in the co-acknowledgement network, is highly correlated with a future association in the co-collaboration network and is also a very good predictor of a future direct collaboration. Figure 4.3 depicts a subset of the

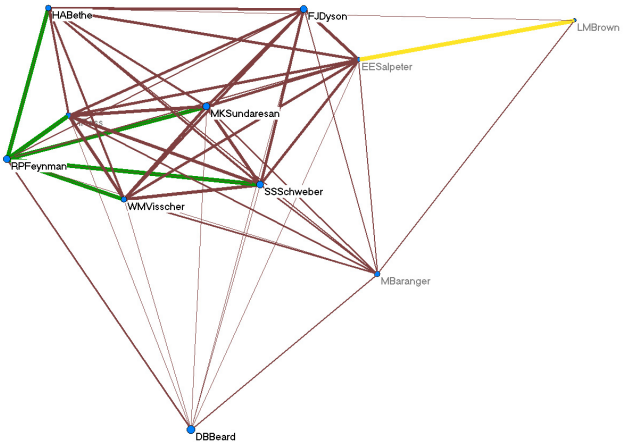

Figure 4.3: subset of the co-collaboration network (red edges) near Feynman, with superposed edges discovered with the semi-metric analysis of the co-acknowledgment network: green edges discovered with parameter s and yellow with parameter b of formulae (4.3).

co-collaboration network (red edges) near Feynman, with superposed edges discovered with the semi-metric analysis of the co-acknowledgment network: green edges discovered with parameter s and yellow with parameter b of formulae 4.3. All the superposed semi-metric edges, were found to indicate a future collaboration. We will see later in chapter 5 some evidence between highly semi-metric pairs and causality.

4.5 Computing Semi-metric pairs: metric closure

From a practical standpoint, one is naturally interested in identifying the specific pairs of elements that are most semi-metric. These pairs are useful

to issue recommendations in web services [76, 88, 59], to identify keywords appropriate to classify biological entities [93, 4], or social interactions that have a higher chance of occurring [74]. These pairs of items are associated not by direct co-occurrence in the data, but are rather implied (as a global property) by the transitivity of the proximity networks obtained from the same data. To obtain these pairs, we compute the metric closure of the relevant distance graph. By *metric closure* we mean that we calculate the shortest distance between any pair of elements in a distance graph D. To do this we compute a $(min, +)$ matrix composition of D with itself (distance product [100]) until closure is achieved—until the composition does not yield any changes. The final matrix obtained with this process (which, as discussed below, is related to Dijkstra's method [29]) is the metric closure D^{mc} of graph D:

Algorithm 4.1. *Metric Closure (AKA distance product [100])*

1. $D' = (D \circ D)$
2. *If* $D' \neq D$, *make* $D = D'$ *and go back to step 1.*
3. *Stop:* $D^{mc} = D'$

where $[D \circ D]_{i,j} = \min_k(d_{ik} + d_{kj}) = d'_{ij}$ is the distance product (see below).

Using D and D^{mc}, we identify all semi-metric pairs (edges) in D, which are the edges for which $d_{ij} > d^{mc}_{ij}$. Further, we choose the most semi-metric pairs (edges) in D, using the semi-metric ratios of formulae 4.3, see figure

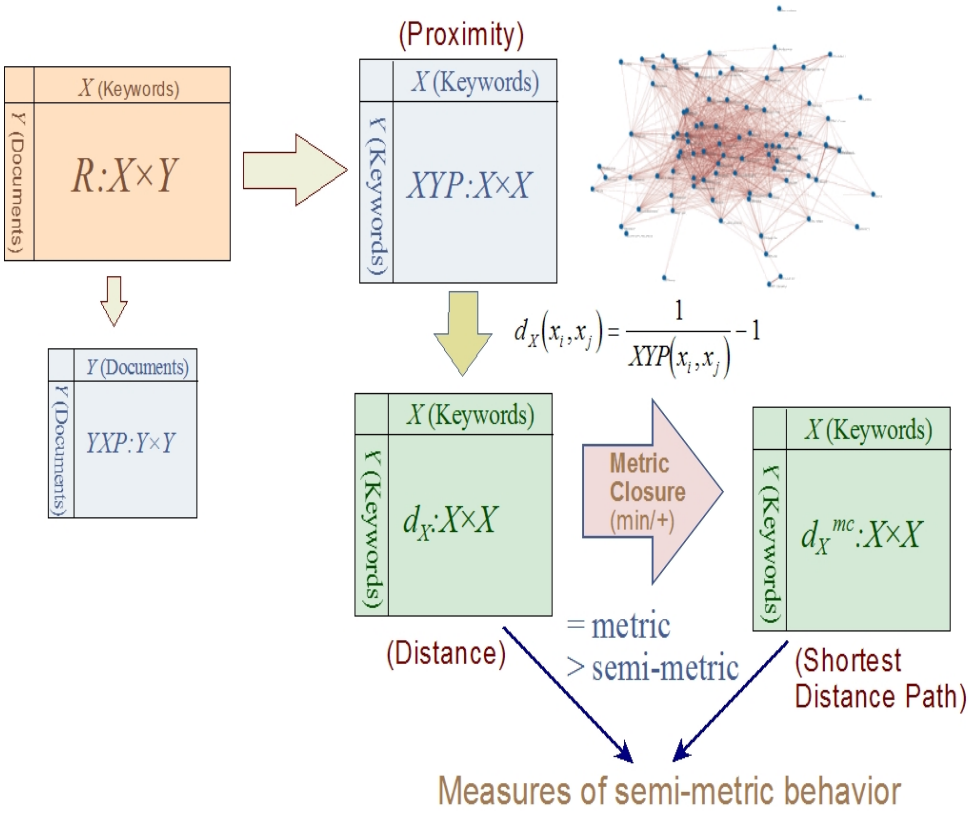

Figure 4.4: Computing semi-metric behavior. From a relation R between sets X and Y, keywords and documents in the example, two proximity matrices/graphs are produced: XYP (keywords) and YXP (documents). Distance closure is exemplified for XYP only. First, a distance matrix/graph D_X is computed using formulae 4.2. Then, the metric closure of this matrix, D^{mc}, is computed using $(min, +)$ composition. Semi-metric pairs are then identified via formulae 4.3.

4.2 for an example with terrorist networks. Figure 4.4, depicts the general process of computing semi-metric behavior given the proximity-to-distance isomorphism of formulae 4.2.

The *All Pairs Shortest Paths* (APSP) problem is one of the most fundamental algorithmic graph problems for computing shortest paths (distances)

among all pairs of vertices in (weighted) distance graphs [100]: it is equivalent to the metric closure of such graphs. The complexity faced by the fastest known algorithm for solving the APSP problem for weighted graphs is $O(mn + n^2 log\ n)$, where n and m are respectively the number of vertices and edges [23]. The APSP uses the Bellman-Ford algorithm for removing all negative cycles and then determines the distances of all pairs by calling the *Single-Source Shortest-Path* (SSSP) Dijkstra algorithm n times, [23]. However, for positive sparse weighted graphs the Johnson algorithm reduces to a time complexity $O(n^2 log\ n)$, [83]. In this work we refer to this algorithm as the APSP Dijkstra algorithm. There are other approaches to solving the APSP problem, such as Floyd-Warshall algorithm [23][83], but all of them fall in the $O(n^3)$ complexity range [100].

The *Distance Product*, or metric closure (Algorithm 4.1) has complexity $O(n^{2.575})$, and it is another approach to solving the APSP problem [100]. The distance product algorithm is based on matrix operations as we can see in algorithm 4.1 is a special case of transitive closure on Fuzzy graphs as we will see in the next sections. In the following sections, we define a general *Distance Closure*, which is a generalization of the metric closure and is equivalent to the transitive closure in Fuzzy, Proximity graphs.

4.6 Fuzzy Shortest paths

A generalization of shortest paths was first introduced by Dubois and Prade in [33] and studied in the last years among the Fuzzy Sets community, [10, 28, 57, 17]. Two main approaches have been proposed to give a degree of fuzziness to the graph edges, [57] [28]. The classical fuzzy shortest path problem the length of a given edge in the graph is attributed a fuzzy number. The second approach the length of a given path is a fuzzy number and each edge in the graph has a membership value. The search for the fuzzy shortest path in the graph can be done using several approaches such as: using a dynamic programming formulation, [57], methods based on the defuzzification of the fuzzy weights [57] and others [33, 10, 17].

In this work we tackle a slightly different problem, related of course to fuzzy shortest path. We study the shortest path problem in weighted graphs (Fuzzy graphs) as been classically studied. We relate it with the various possibilities of transitive closure, and look into the impact of different closures, have in the complex networks analysis.

4.7 General distance closure

Transitive Closure is a well established algorithm in the theory of Fuzzy Graphs, and used to calculate a transitive graph, whose edge weights are not smaller than every indirect path between the same edge vertices. Transitive closure is also behind many definitions and theorems in the theory of Fuzzy

Graphs [60]. All theories based on weighted graphs, such as the small-world phenomena, can profit from concepts already stablished in Fuzzy Graph Theory, such as transitive closure and its relation to the All Pairs Shortest Paths (APSP) problem, typically based on the Dijkstra algorithm [29].

The relation between *Transitive Closure* and *Distance Closure* can be seen in the mathematical framework of algebraic structures [44, 8, 3]. In fact the transitive closure can be seen as a closed semi-ring $I = \{S = [0,1], \vee, \wedge, tc, 0, 1\}$ [1], where $\vee, \wedge : S \times S \to S$ are two binary operations and $tc : S \to [0,1]$ is the closure. Likewise, the distance closure can be seen as a closed semi-ring, $II = \{S' = [0, +\infty], f, g, dc, 0, +\infty\}$, where $f, g : S' \times S' \to S'$ are two binary operations and $dc : S' \to [0, +\infty]$ is the closure. Here, we relate the set of conditions $(\varphi, \wedge, \vee, f, g)$ for the equivalence of transitive closure with a pair of t-norm $\wedge$, t-conorm $\vee$ in semi-ring I and it distance closure with a pair of binary operations f and g in semi-ring II, as constrained on isomorphism φ that maps between S and S'.

Of particular interest is the relationship between the metric closure (special case of distance closure, typically obtain via APSP Dijkstra algorithm or the distance product) and the transitive closure of Fuzzy Graphs. We provide a general mathematical framework to this problem, which is particularly targeted to the practice of complex networks and information retrieval from empirical data. In summary, if we have a proximity graph we use the

[1] A closed semi-ring is a semi-ring with two additional properties: (1) if $a_1, a_2, ..., a_n, \ldots$ is a countable sequence of elements of S then $a_1 \vee a_2 \vee ... \vee a_n \vee \ldots$ exists and is unique; (2) the operation $\wedge$ distributes over countably infinite $\vee$'s as well as finite $\vee$'s.

transitive closure to calculate a similarity graph, which is transitive [53]. If, instead, we have a distance graph, we can use the distance closure to compute the smallest possible distance between vertices. There are isomorphisms that map proximity graphs, (edges in $[0, 1]$), into distance graphs (edges in $[0, +\infty]$), such as the function of formula 4.2. However, there are no linear functions for this map, even though there is an infinity of non-linear functions that instantiate; this poses us with a problem of degeneracy of solutions to the metric closure in weighted graphs. Therefore, if we want to understand and make appropriate inferences from the metric or distance closure for a given weighted graph, we have to take a closer look at the space of non-linear functions that instantiate this isomorphism, which we do below. In fact, this non-linear isomorphism enforces a particular topological distortion of the original proximity graph used to construct the distance graph, which ultimately determines the way we compute shortest paths.

We also show that the isomorphism chosen is a generator of a *t-norm*, (see [51]). The concept of *t-norms* was introduced by Karl Menger to generalize transitivity in *probabilistic metric spaces* [80]. The results of Menger and his followers were then applied to the theory of Fuzzy Sets to generalize the concept of Conjunctions (Unions) and Disjunctions (Intersections) in Fuzzy logic (Sets) [53]. Transitive closure is a generalization of the APSP problem, and, as it is well known in the fuzzy set community, there are infinite solutions to this problem, [53]. Nonetheless, different *t-norms* provide lower and upper bounds of the strength of transitivity, where the strongest *t-norm*

is the *minimum* function and the weakest *t-norm* is the *drastic product* [51] [53]. The ability to sweep the transitivity space that results from the $t-norm$ bounds allows us to control and understand the topological distortion imposed on proximity graphs when, via a non-linear isomorphism, we convert to a distance graph to be fed to the APSP or metric closure.

We now determine the general constrains on closure imposed by the non-linear isomorphism between the space of proximity graphs and the space of distance graphs. Figure 4.5 shows the general picture of the problem. Suppose we have a proximity graph G_P (a fuzzy symmetric and reflexive graph), a t-norm $\wedge$, a t-conorm $\vee$, both acting on G_P, and two binary operations f , g acting on G_D (a distance symmetric and anti-reflexive graph), and on isomorphism φ (our distance function). We want to characterize what are the constrains that φ imposes on the closures computed from these graphs, under various algebraic operators. This isomorphism φ can only be a non-linear function because it maps the unit interval $[0, 1]$ into the positive real line $[0, +\infty]$.

Definition 4.1. *Two undirected weighted graphs* $G_1 = (V, E_1)$ *and* $G_2 = (V, E_2)$ *are isomorphic if there is a vertex-preserving bijective edge mapping* $\varphi : E_1 \to E_2$*, i.e. a bijection* φ *with*

$$\forall u, v \in V : e_{u,v} \in E_1 \Leftrightarrow \varphi(e_{u,v}) \in E_2$$

Let X be the set of vertices and P be the connectivity matrix of the

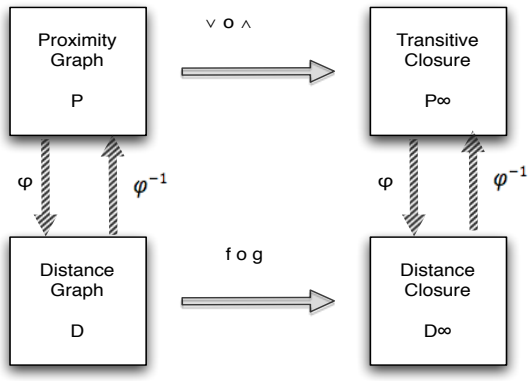

Figure 4.5: Transitive and Distance Closure space.

proximity edges of $G_P = (X, P)$. Let D be the connectivity matrix of the distance edges of $G_D = (X, D)$. P is reflexive, i.e. $p_{x,x} = 1$, and D is anti-reflexive, i.e. $d_{x,x} = 0$, $\forall x \in X$.

Definition 4.2. *Let $\varphi : [0, 1] \to [0, +\infty]$, $d_{x,y} = \varphi(p_{x,y})$, be a function that maps the edge weights $p_{x,y} \in [0, 1]$ of a fuzzy proximity graph $G_P = (X, P)$ into the edge weights $d_{x,y} \in [0, +\infty]$ of a distance graph $G_D = (X, D)$, $\forall x, y \in X$, let also $\Phi : [0, 1] \times [0, 1] \to [0, +\infty] \times [0, +\infty]$ be the graph function that maps the proximity connectivity matrix into the distance connectivity matrix, $D = \Phi(P)$. We define φ and Φ in the following way:*

(1) φ is strictly monotonic decreasing, $\forall a, b \in [0, 1] : a > b \Rightarrow \varphi(a) < \varphi(b)$;
(2) $\varphi(0) = \infty$ and $\varphi(1) = 0$;
(3) $\Phi(P) = [\varphi(p_{x,y})]$, $\forall x, y \in X$.

Because φ is a real valued function and it is strictly monotonic it is also

bijective, therefore the graphs G_P and G_D are isomorphic with Φ, with the same set of vertices X.

Theorem 4.1. *Let $G_P = (X, P)$ be a proximity (symmetric and reflexive) graph and Φ the graph distance function in definition 4.2, then $G_D = (X, D)$, where $D = \Phi(P)$ is symmetric and anti-reflexive.*

Next we define a set of binary operators, which operate on distance graphs.

Definition 4.3. (TD-norms and TD-conorms) *Let $f, g : [0, +\infty] \times [0, +\infty] \to [0, +\infty]$, such that for all $a, b, c \in [0, +\infty]$ the following four axioms are satisfied:*

(1) $f(a, b) = f(b, a)$, $g(a, b) = g(b, a)$ (commutativity).

(2) $f(a, f(b, c)) = f(f(a, b), c)$, $g(a, g(b, c)) = g(g(a, b), c)$ (associativity).

(3) $f(a, b) \leq f(a, c)$, $g(a, b) \leq g(a, c)$, whenever $b \leq c$ (monotonicity).

(4) $f(a, \infty) = a$, $g(a, 0) = a$, with $a \leq \infty$ (boundary conditions).

We call g a TD-norm and f a TD-conorm.

Theorem 4.2. *If φ is a distance function as in definition 4.2 and $\wedge$, $\vee$ a t-norm, t-conorm pair and f and g a TD-conorm and TD-norm pair as defined in 4.3, the following statements are true:*

(1) $\varphi(a \wedge b) = g(\varphi(a), \varphi(b))$;

(2) $\varphi(a \vee b) = f(\varphi(a), \varphi(b))$.

Where $a, b \in [0, 1]$.

Definition 4.4. *Let* $G_P = (X, P)$ *be a fuzzy proximity graph, with edges* $p_{x,y} \in [0, 1]$ $\forall x, y \in X$. *We define the n-power of* P *as*

$$P^n = \underbrace{P \circ P \circ \cdots \circ P}_{n},$$

where $(P \circ P)(x, y) = \underset{z}{\vee}(\wedge(p_{x,z}, p_{z,y}))$, $\forall x, y, z \in X$.

Definition 4.5. *The transitive closure of a fuzzy proximity graph is computed as:*

$$P^\infty = P \cup P^2 \cup P^3 \cup \cdots$$

Where $\cup$ *is some fuzzy union (t-conorm) of two sets. It is computed via Algorithm 2.1.*

Theorem 4.3. *If the graph is finite then, the power of* P *converges for a certain* k *(largest path in the graph), the transitive closure of* G_P, *is the graph* $G_P \equiv (X, P^{tc})$, *obtained by:*

$$P^\infty = P \cup P^2 \cup P^3 \cup \cdots \cup P^k = P^{tc}.$$

The proof of this theorem can be found in [53] [60].

Definition 4.6. (Distance Composition) *Let $G_D = (X, D)$ be the distance graph where edges are $d_{x,y} \in [0, +\infty]$, $\forall x, y \in X$. We define the n-power of D by*

$$D^n = \underbrace{D \circ D \circ \cdots \circ D}_{n}$$

where $(D \circ D)(x, y) = \underset{z}{f}\{g(d_{x,z}, d_{z,y})\} : \forall x, y, z \in X$. Where f and g are two binary functions defined in 4.3.

Definition 4.7. (Distance Closure) *The distance closure is given by:*

$$D^\infty = D\dot{\cap}D^2\dot{\cap}D^3\dot{\cap}\cdots$$

Where $(A\dot{\cap}B)(x, y) = f(a_{xy}, b_{xy})$, $\forall x, y \in A, B$, where f is a TD-conorm.

Theorem 4.4. *If the graph is finite then the power of D converges for a certain k, thus, the distance closure of G_D, is the graph $G_D \equiv (X, D^{dc})$, obtained by:*

$$D^\infty = D\dot{\cap}D^2\dot{\cap}D^3\dot{\cap}\cdots\dot{\cap}D^k \equiv D^{dc}.$$

This theorem can be easily proven from theorems 4.3, 4.5 and 4.6 (see below). Next we give a more general definition of the metric closure computed with algorithm 4.1.

Definition 4.8. (Metric Composition) *Same as definition 4.6, with $f \equiv min$ and $g \equiv +$.*

Definition 4.9. (Metric Closure) *Same as definition 4.7 with $f \equiv min$ and $g \equiv +$. The metric closure of G_D is the graph $G_D^{mc} \equiv (X, D^{mc})$. It is computed via Algorithm 4.1 (AKA distance product).*

Theorem 4.5. *if $G_P = (X, P)$ is a fuzzy proximity graph, φ the isomorphism (distance function) in definition 4.2 and $G_D = (X, D)$ is the respective distance graph, where $D = \Phi(P)$. The following is true with the distance composition:*

1) $\Phi(P) \dot{\supseteq} \Phi(P^2) \dot{\supseteq} \Phi(P^3) \dot{\supseteq} \cdots \supseteq \Phi(P^\infty)$;

2) $D \dot{\supseteq} D^2 \dot{\supseteq} D^3 \dot{\supseteq} \cdots \dot{\supseteq} D^\infty$.

where $\Phi(P^i) \dot{\supseteq} \Phi(P^{i+1})$ means that: $\forall x, y \in X : \varphi(p_{x,y}^i) \geq \varphi(p_{x,y}^{i+1})$.

Theorem 4.6. *Given a proximity graph $G_P = (X, P)$, a distance graph $G_D = (X, D)$, and isomorphism φ, Φ as defined in 4.2. For a t-norm $\wedge$ and t-conorm $\vee$ used to compute the transitive closure of P, then there exists a TD-conorm/TD-norm pairs f and g, to compute the distance closure of D, $\Phi(P^{tc}) = D^{dc}$, which obey the condition:*

$$\forall x, y \in X : \underset{z}{f}\{g(\varphi(p_{x,z}), \varphi(p_{z,y})\} = \varphi(\underset{z}{\vee}\{\wedge(p_{x,z}, p_{z,y})\})$$

(where Φ^{-1} is the inverse of Φ and φ^{-1} is the inverse function of φ). The same is true if we fix f (TD-conorm), g (TD-norm) and isomorphism φ, Φ, to obtain a pair of t-conorm, t-norm $\vee$ and $\wedge$.

The conditions of this theorem leads to the equations of theorem 4.2:

$$g(d_{x,z}, d_{z,y}) = \varphi(\wedge(\varphi^{-1}(d_{x,z}), \varphi^{-1}(d_{z,y})))$$

$$f(d_{x,z}, d_{z,y}) \equiv \varphi(\vee(\varphi^{-1}(d_{x,z}), \varphi^{-1}(d_{z,y})))$$

.

From these last equations we can also find $\vee$ and $\wedge$ given f, g and the isomorphism φ:

$$\vee(p_{x,z}, p_{z,y}) = \varphi^{-1}(f(\varphi(p_{x,z}), \varphi(p_{z,y})))$$

$$\wedge(p_{x,z}, p_{z,y}) = \varphi^{-1}(g(\varphi(p_{x,z}), \varphi(p_{z,y})))$$

4.8 Exploring the Proximity/Distance isomorphism space

The conditions of theorem 4.6, allows us to compute f, g given $\vee$, $\wedge$, and φ, as well as $\vee$ and $\wedge$ given f, g, and φ. This allows us to study several closure scenarios (e.g., Metric, Ultra-metric, dombi conjugate t-norm/t-conorm for $\lambda = 1$), which lead to different distortion of the original graphs.

Given this space of possible transitivity criteria, it is reasonable to ask several questions: for a given proximity-to-distance isomorphism φ, what is the equivalent of the fuzzy (max, min) closure for a distance graph? Per-

haps more interestingly, what is the fuzzy equivalent of the metric closure of a distance graph? Which closures preserve important characteristics of real complex networks and observe good axiomatic requirements? These questions are important because all the applications of complex networks that use transitivity produce different results depending on the connectives employed. Not only do we want intuitive connectives (e.g. leading to a metric closure), we want those that lead to best results in specific applications.

Example 4.1 (Ultra-Metric Closure) Let φ be any t-norm generator (see [53]), and $f(x, y) \equiv min(x, y)$ the min TD-conorm and $g(x, y) \equiv max(x, y)$ the max TD-norm. Where $a, b \in [0, 1]^2$, $a = \varphi^{-1}(x)$ and $b = \varphi^{-1}(y)$.

We know from theorem 4.6:

$$\vee(a, b) = \varphi^{-1}(f(\varphi(a), \varphi(b)))$$

it easy to show that

$$\vee(a, b) = max(a, b)$$

since φ is strictly monotonic decreasing.

We apply the same reasoning to $\wedge$:

$$\wedge(a, b) \equiv \varphi^{-1}(g(\varphi(a), \varphi(b)))$$

it is easy to show that,

$$\wedge(a, b) = min(a, b)$$

since φ is strictly monotonic decreasing. Therefore, the ultra-metric closure $(g, f) \equiv (max, min)$ become in the proximity space $(\wedge, \vee) \equiv (min, max)$.

Ding et al [30] have previously shown this relationship, which derives easily for any φ in our framework. In this case, the $(\vee \equiv max, \wedge \equiv min)$ closure of a fuzzy graph is equivalent to the *ultra-metric* closure of a distance graph $(f \equiv min, g \equiv max)$, where instead of the triangle inequality, a stronger inequality is enforced: $d_{ij} \geq max(d_{ik}, d_{kj})$. Ding et al further used this closure to compute cliques in protein interaction networks—a problem relevant for computational Biology.

Example 4.2 (Metric Closure) Let $\varphi(x) = \frac{1}{x} - 1$ (formula 4.2), which is also the Dombi t-norm generator with $\lambda = 1$ (see [53]), and $f(x, y) \equiv min(x, y)$ the TD-conorm and $g(x, y) \equiv +(x, y)$. Where $a, b \in [0, 1]^2$, $a = \varphi^{-1}(x)$ and $b = \varphi^{-1}(y)$.

We know from theorem 4.6:

$$\vee(a, b) = \varphi^{-1}(f(\varphi(a), \varphi(b)))$$

if $a \leq b$ then,

$$\vee(a, b) = \varphi^{-1}(min(\varphi(a), \varphi(b))) = b = max(a, b)$$

therefore,

$$\vee(a,b) = max(a,b)$$

We apply the same reasoning to $\wedge$:

$$\wedge(a,b) = \varphi{-}1(g(\varphi(a),\varphi(b)))$$

$$\wedge(a,b) = \varphi^{-1}(\varphi(a)+\varphi(b)) = \varphi^{-1}(\frac{a+b-2ab}{ab})$$

and since $\varphi^{-1}(x) = \frac{1}{x+1}$ we obtain,

$$\wedge(a,b) = \begin{cases} 0 & for \quad (a,b) = (0,0) \\ \frac{ab}{a+b-ab} & for \quad (a,b) \in]0,1]^2 \end{cases}$$

Therefore, the metric closure $(g,f) \equiv (+,min)$ in the proximity space is $(\wedge,\vee) = (DT^1_\wedge, max)$.

Figure 4.6 depicts the two examples 1 and 2 closures related to the questions above, for the proximity-to-distance isomorphism φ of formulae 4.2.

We will see below from an applied viewpoint, that the (max,min) closure of proximity graphs (or ultra-metric closure in distance graphs) is quite destructive, this means that every item becomes highly related to every other indirectly linked item, however far, resulting in very low performance in information retrieval applications, for instance. We already knew from our recommender systems [77, 76], as well as from Rocha's analysis of social and knowledge networks [72, 74, 93, 4], that the metric closure of distance graphs

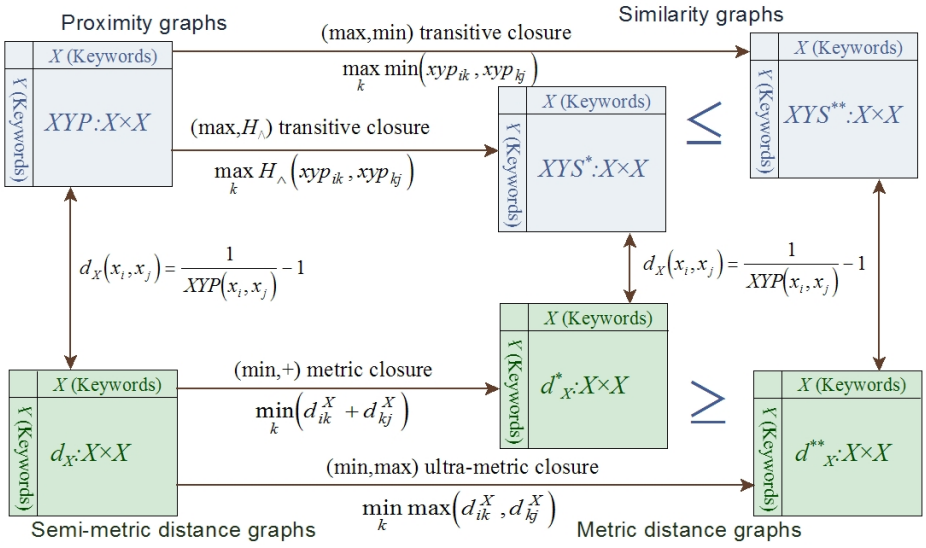

Figure 4.6: Metric and ultra-metric distance closures, and their fuzzy proximity graph counterparts for $\varphi : distance = \frac{1}{proximity} - 1$. The ultra-metric distance closure is equivalent to the (max, min) closure of a fuzzy graph. The metric closure is equivalent to the $(max, H_\wedge)$ closure of a fuzzy graph, where $H_\wedge$ is the base Hammacher conjunction [53].

produced better and more intuitive results than the (max, min) closure of proximity graphs—insofar as the search for relevant indirect associations is concerned. The metric closure is a weaker constraint than the ultra-metric (or $max - min$ of fuzzy graphs), which results in significantly fewer edges being altered in the original graph: only those indirect paths where every edge is very short.

Now let us go in the other direction, and start from a well-known $\wedge$ and $\vee$ in Fuzzy graphs.

Example 4.3 (Dombi Closure $\lambda = 1$) Let $\varphi(x) = \frac{1-x}{x}$ be the Dombi

t-norm generator with $\lambda = 1$ (see [53]), and $\wedge \equiv DT^1_\wedge(a,b) = \frac{ab}{a+b-ab}$ the Dombi t-norm with $\lambda = 1$ (see [53]) and $\vee \equiv DT^1_\vee(a,b) = \frac{a+b-2ab}{1-ab}$. Where $a, b \in [0,1]$, $a = \varphi^{-1}(x)$ and $b = \varphi^{-1}(y)$.

We know from theorem 4.6:

$$g(x,y) = \varphi(\wedge(\varphi^{-1}(x), \varphi^{-1}(y)))$$

It is easy to show that $\varphi^{-1}(x) = \frac{1}{x+1}$, thus from example 4.2:

$$g(x,y) \equiv x + y$$

.

We apply the same reasoning to f:

$$f(x,y) \equiv \varphi(\vee(\varphi^{-1}(x), \varphi^{-1}(y)))$$

$$f(x,y) = \frac{1 - DT^1_\vee(\varphi^{-1}(x), \varphi^{-1}(y))}{DT^1_\vee(\varphi^{-1}(x), \varphi^{-1}(y))}$$

$$f(x,y) = \frac{1 - DT^1_\vee(\frac{1}{x+1}, \frac{1}{y+1})}{DT^1_\vee(\frac{1}{x+1}, \frac{1}{y+1})}$$

we have,

$$f(x,y) = \begin{cases} y & for \quad x = +\infty \\ x & for \quad y = +\infty \\ \frac{xy}{x+y} & for \quad (x,y) \in]0,+\infty[^2 \\ 0 & for \quad (x,y) = (0,0) \end{cases}$$

The three examples above, are three well-known cases of closure. But let us pursue a more comprehensive explanation of possible closures.

One way to explore the isomorphism space is to constrain $f \equiv min$ and $g \equiv +$ to compute the metric closure of distance graph D.

Definition 4.10. *The pseudo-inverse of a decreasing generator φ is defined by*

$$\varphi^{(-1)}(a) = \begin{cases} 1 & for \quad a \in (-\infty, 0) \\ \varphi^{-1}(a) & for \quad a \in [0, \varphi(0)] \\ 0 & for \quad a \in (\varphi(0), \infty) \end{cases}$$

Theorem 4.7. (Characterization Theorem of t-norms) *Let $\wedge$ be a binary operation on the unit interval. Then, $\wedge$ is an Archimedean t-norm iff there exist a decreasing generator φ such*

$$\wedge(a,b) = \varphi^{(-1)}(\varphi(a) + \varphi(b))$$

for all $a, b \in [0,1]$.

Both definition 4.10 and theorem 4.7 are from [53]. The next corollary

follows from theorem 4.6 and generalizes the Metric Closure as we vary t-norm generator *varphi* (isomorphism).

Corollary 4.1. *According to the formulation of theorem 4.6, let* $f \equiv \min()$, $g \equiv +()$ *and* φ *a distance function, as defined in 4.2. If* $\vee = \max()$ *as t-conorm, then the t-norm operator* $\wedge$ *exists and* φ *is its generator function.*

Corollary 4.1 states that in the formulation of theorem 4.6, when we fix t-conorm $\vee = max()$ and $(f, g) = (min, +)$ operators, there exists a t-norm $\wedge$, which preserves the isomorphism between proximity and distance graphs, as well as their closures with the respective operators. Moreover, the isomorphism function φ is in fact the t-norm generator. This corollary has a strong impact when we convert a proximity graph into a distance graph and then calculate a distance closure. The specific isomorphism function we choose defines a t-norm, in the proximity graph space, which ultimately influences how we measure distances in the distance graphs. Sweeping the space of φ functions allows us to study their topological effects on measuring distance,that can be easily computed via APSP or distance product.

We can see from theorem 4.6 and corollary 4.1 that the transitive closure of weighted graphs entails a very wide space of possibilities. Each one leads to a distinct way of computing the distance closure or the APSP Dijkstra algorithm. Consequently these closures are not unique as already known in the theory of fuzzy graphs: for a given application, it is important to pay attention to the distortion created by the distance closure or transitive closure

chosen on the original proximity information extracted from relational data [53].

We define *distortion*, Δ (see below), as the difference between the edges in the original graph (proximity) and the edges obtained by a given closure. We show that the distortion is smaller with a metric closure than with a ultra-metric closure.

$$\Delta(P) = \sum_x \sum_y |p^c_{xy} - p_{xy}| \tag{4.4}$$

Theorem 4.8. *Given the isomorphism φ, if D^{mc} is the metric closure with $f \equiv min$ and $g_1 \equiv +$, and D^{um} is the ultra-metric closure with $f \equiv min$ and $g_2 \equiv max$ then $D^{mc} \dot{\supseteq} D^{um}$ is equivalent to $P^{mc} \subseteq P^{um}$, where $D^{mc} = \Phi(P^{mc})$ and $D^{um} = \Phi(P^{um})$.*

Theorem 4.7 shows that the transitive closure with (max,min), which is the ultra-metric distance closure, produces a larger distortion of the original graph, than what we get from the metric closure of the distance graph $\Delta_{um} \geq \Delta_{mc}$. These results are also shown in Figure 4.6.

In the next section we search for distance closures with good axiomatics, close to metric closure. Afterwards, using corollary 4.1, we proceed to study the range of t-norms generated by φ, a range bounded by the drastic product and the minimum t-norms: $\wedge \in [drastic\ product\ T_D, min]$.

4.9 Axiomatic characteristics of Distance Closure

In the Fuzzy logic community, considerable work has been produced to identify pairs of logical connectives and complements that satisfy desirable axiomatic characteristics (e.g. De Morgan's laws [31])[2]. These pairs of general (fuzzy) logic conjunction and disjunction operations are known as conjugate *t-norms* and *t-conorms* respectively [53, 49]. As discussed above in this work, each distinct conjunction/disjunction pair leads to a specific transitive closure of an initial proximity graph. Therefore, transitivity in weighted graphs depends on the particular logical connectives used. However, only some of these entail intuitive logical connectives: $\wedge$ and $\vee$. For instance, the (max, min) logical connectives, with the standard fuzzy complement ($\bar{x} = 1 - x$), follow De Morgan's laws. So do many other logical connectives and complements, see [53] for a good overview.

The metric closure of a distance graph, in answer to the second question above, corresponds to an interesting fuzzy logic conjunction/disjunction pair—in the case of the isomorphism φ of formulae 4.2 as shown in example 4.2. In the proximity space, the summation operation used in the distance graph metric closure becomes a bounded sum operation, which is a special case of the Hamacher [41] and the Dombi [31] conjunctions (or intersections or T-Norms) [53]:

[2]This section was done in collaboration of Bharat Dravid

$$H_\wedge \equiv DT^1_\wedge(a,b) = \frac{ab}{a+b-ab} \tag{4.5}$$

for any real a, b. Thus, the metric closure of distance graphs, corresponds to the $(max, DT_\wedge)$ transitive closure of a fuzzy graph for the most common isomorphism defined in formula 4.2. Unfortunately, this pair of fuzzy logic connectives, with any complement function, leads to a fuzzy algebra with very poor axiomatic characteristics. It can be easily seen that this pair of connectives does not satisfy De Morgan's laws, for instance (see proof in appendix). One could seek a different isomorphism φ to generate a t-norm with better logical characteristics, or we could seek a different fuzzy t-norm/t-conorm pair, in some sense close to the metric closure pair $(max, DT_\wedge)$, but with better axiomatics. However, these avenues need to be pursued with care since our choice of the proximity-to-distance isomorphism and the concept of metric closure (related to Dijkstra's algorithm) are quite intuitive, and indeed used ubiquitously.

Furthermore, the space of possible transitive closures leads to quite distinct results in performance of information retrieval systems and in the analysis of complex social and knowledge networks. For instance, in the recommendation system for the MyLibrary@LANL system (see section §4.3) we used the proximity network of scientific journals depicted in Figure 4.1. To recommend items that users might be interested in, we computed the metric closure of this network and identified the highly semi-metric edges (see

§4.5), which lead to very good performance in user tests [76]. When we used the more traditional (max, min) (ultra-metric) closure of fuzzy proximity graphs, many more irrelevant items are picked up, lowering the information retrieval measures of performance. The use of an appropriate closure is important for all other applications of weighted networks, be it for social analysis, literature mining of biological entities, or information retrieval and recommendation systems based on knowledge and social networks.

The field of fuzzy sets has been very concerned with studying conjunction/disjunction pairs that lead to good logical axiomatic constraints. For instance, it is reasonable to expect a complement to be involutive, so that $\bar{\bar{x}} = x$. It is also reasonable that disjunction, conjunction and complement follow De Morgan's laws: $\overline{a \vee b} = \bar{a} \wedge \bar{b}, \overline{a \wedge b} = \bar{a} \vee \bar{b}$. Such good axiomatic characteristics are also important for fuzzy graphs, especially when we use them to model knowledge networks. Indeed, when we use proximity networks as knowledge representations (§4.3), it may be useful to have an intuitive understanding of what is the complement of a given network, or better, be able to compute the conjunction and disjunctions of various networks obtained from distinct data sources. For instance, in the recommender system developed for MyLibrary@LANL [76], it may be useful to issue recommendations on a aggregate journal network built from a conjunction of two constituent networks (e.g. journal proximity obtained from user access data *and* journal proximity obtained from citation data). If, as we have shown (§4.3), proximity networks are good knowledge representations for many applications, we

need to be able to combine networks obtained from different data sources, to compute compound logical statements from the knowledge they store.

As we discussed above, the metric closure of distance graphs becomes the $(max, DT^1_\wedge)$ transitive closure in fuzzy graphs, using the isomorphism of formula 4.2. Since no involutive complement exists that can satisfy De Morgan's laws with the $(max, DT^1_\wedge)$ conjunction/disjunction pair, we now ask what are the closest conjunction/disjunction pairs to the metric closure, that with an involutive complement obey De Morgan's laws. This search is pursued here by inspecting the space of known t-norm/t-conorm families [53], using the Dombi family [31] for the simplest isomorphism φ of formula 4.2.

$$DT^\lambda_\vee(a,b) = \frac{1}{1+\left[\left(\frac{1}{a}-1\right)^{-\lambda}+\left(\frac{1}{b}-1\right)^{-\lambda}\right]^{-\frac{1}{\lambda}}} \tag{4.6}$$

From this general Dombi t-conorm formula we can find a parameter value for it which will lead to obeying De-Morgan's laws when using the Dombi t-norm used in the metric closure ($\lambda = 1$).

Let us investigate if De-Morgan's Laws work with complement $C_1(x) = 1 - x$;

$$\bar{a} \vee \bar{b}|\vee \equiv D^\lambda_\vee(\overline{a},\overline{b}) = \frac{1}{1+\left[\left(\frac{1}{a}-1\right)^{\lambda}+\left(\frac{1}{b}-1\right)^{\lambda}\right]^{-\frac{1}{\lambda}}}$$

$$\overline{a \wedge b}|\wedge \equiv \overline{D}^1_\wedge(a,b) = 1 - \frac{ab}{a+b-ab} = \frac{a+b-2ab}{a+b-ab}$$

For De-Morgan's Law to hold, $\overline{a \wedge b} = \overline{a} \vee \overline{b}$:

$$-ab\left[\left(\frac{1}{a}-1\right)^{\lambda}+\left(\frac{1}{b}-1\right)^{\lambda}\right]^{\frac{1}{\lambda}}+a+b-2ab=0$$

This equation has $\lambda = 1$ as a straightforward solution which is not surprising because for λ, the triple satisfies De-Morgan's Laws [53]. We can say that the left side of the equation is the *error* or *deviation* from a t-norm/t-conorm pair that obeys De-Morgans laws with standard complement. An integral of the left side of the above equation gives an estimate of the total deviation from ideal axiomatics over the entire domain of the function.Thus we can define the error function, F(λ) as:

$$F(\lambda)=\int_0^1\int_0^1\left(-xy\left[\left(\frac{1}{x}-1\right)^{\lambda}+\left(\frac{1}{y}-1\right)^{\lambda}\right]^{\frac{1}{\lambda}}+x+y-2xy\right)dxdy$$

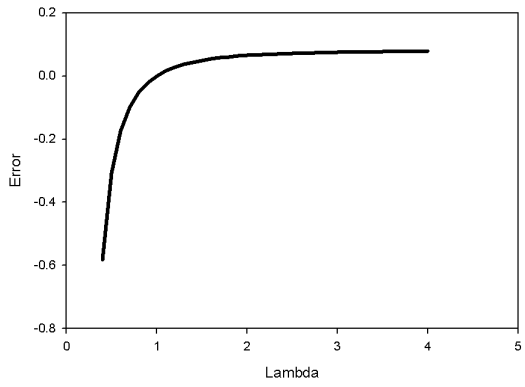

Figure 4.7: Error between the surface established by the desired axiomatic constraints, and $(DT_{\vee}^{\lambda}, DT_{\wedge}^{1})$ as λ varies.

Figure 4.7 shows the error (computed as the double integral, above) be-

tween the surface established by the desired axiomatic constraints.

While ($DT^1_\wedge$, max) does not possess good axiomatic characteristics, it not deviates too much from the axiomatically desirable ($DT^1_\vee, DT^1_\wedge$). Nonetheless, if one intends to use proximity or distance graphs as knowledge graphs, i.e., as representations of knowledge extracted from relational data, the ($DT^1_\vee, DT^1_\wedge$) t-norm/t-conorm pair is the only that preserves desirable characteristics with the most intuitive and simple isomorphism. It allows De Morgan and involution rules to be systematically applied without concern. In contrast, the t-norm/t-conorm pair used in metric closure ($max, DT^1_\wedge$) will result in the accumulation of errors. However,the relatively small error of the metric closure space ($\lambda \to +\infty$) may be acceptable in certain circumstances[3]. There are many other conjunction/disjunction families to explore, which can be studied as future work.

4.10 Exploring the isomorphism with the Dombi t-norm

In the previous section, we fixed the isomorphism φ to the simplest and most common distance function given by formula 4.2. This allowed us to search the t-norm/t-conorm pairs that better preserve logical axiomatics and are closed to our intuitive metric closure operator. Here, we fix the metric closure operators instead, and search the space of possible functions φ.

[3]The curve in Figure 4.7 asymptotically approaches 0.1 when $\lambda \to +\infty$

We have seen that we can apply an infinity of pairs of t-norms and t-conorms to calculate distance closure, and compute shortest paths in distance graphs. In this formulation (see corollary 4.1), we fix the t-conorm with $\vee \equiv max$, allowing us to explore many options for the t-norm $\wedge$. The t-norm is defined via the t-norm generator isomorphism φ (corollary 4.1). Then, using ($f \equiv min, g \equiv +$) binary operations for computing the metric closure, via the APSP Dijkstra, distance product or equivalent, we can sweep the space of possible t-norms, thus simultaneity exploring the range of possible isomorphisms. This poses us the following question: which t-norm/isomorphism is optimal, given a set of assumptions, for the shortest paths calculation? We answer this question here for the Dombi t-norm family, which provides the range of t-norms between the lower and upper bounds through the λ parameter. Recall that the Dombi t-norm generator is:

$$\varphi(x) = \left(\frac{1}{x} - 1\right)^{\lambda} \tag{4.7}$$

where λ is the sweeping parameter. The parameter λ in the t-norm generator takes values in $]0, +\infty[$: $\lambda \to 0$ lower bound (*drastic product*) and $\lambda \to \infty$ is the upper bound (*minimum*). The reason we choose this t-norm generator is because it yields the more commonly used isomorphism from proximity to distance; when $\lambda = 1$, [2] [89], the function 4.7, becomes isomorphism 4.2,

which we have used in the previous section:

$$\varphi(x) = \frac{1}{x} - 1.$$

We have seen that when t-norm and t-conorm $(\vee, \wedge)$ are fixed the transitive closure and the distance closure are equivalent via isomorphism φ.

For empirical analysis of complex networks it is desirable that properties of the graphs obtained via specific closures, such as *average shortest path*, be simultaneously characteristic in both spaces (proximity and distance). That is, the fluctuations of the mean, must be constrained on both spaces (average shortest path and average strongest path). In order to have a characteristic average path length, the shortest paths distribution must follow approximately a normal distribution. We want to find the best λ, using the Dombi t-norm generator, which guarantees these assumptions.

Assuming that the shortest path distribution of a distance graph follows a normal distribution, the probability density function for a normal random variable X here, the shortest path, is given by:

$$h_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{4.8}$$

where μ and σ are the mean and standard deviation of the normal distribution.

The mean of a random variable $Y = j(X)$, which is a monotonic function of X, where X is the random variable representing shortest path in a distance

graph, and Y the random variable representing the strongest path in the isomorphic distance graph, is given by:

$$< Y >= \int_0^\infty j(x) h_X(x) dx \tag{4.9}$$

In our case,

$$j(x) = \varphi^{-1}(x) = \frac{1}{x^{\frac{1}{\lambda}} + 1}$$

Therefore, the fluctuations to the mean, in the proximity space are given by:

$$CV_p = \frac{\sigma_p}{\mu_p} = \frac{\sqrt{< Y^2 > - < Y >^2}}{< Y >} \tag{4.10}$$

where CV_p is the *coefficient of variability*[4], and σ_p and μ_p are the standard deviation and mean of the strongest path in the proximity space and $< Y^2 >$ is given by:

$$< Y^2 >= \int_0^\infty j^2(x) h_X(x) dx \tag{4.11}$$

The fluctuations in the distance space of the shortest path, are given by the *coefficient of variability*, CV_d:

$$CV_d = \frac{\sigma}{\mu} \tag{4.12}$$

The dependence of CV_p on CV_d comes from equations 4.8, 4.9 and 4.11. In figure 4.8 we plot the theoretical relation between λ and CV_p for $\mu = 10$

[4]The coefficient of variability is scale invariant.

(average shortest path in distance space is normally distributed) and $CV_d = 0.2$, using equation 4.10; the shape is preserved for different parameter values. We can see from this figure that the coefficient of variability in the proximity space is minimum when λ converges to the *min* t-norm ($\lambda \to +\infty$); the ultra-metric closure. However, from our assumptions we require that $CV_p \approx CV_d = 0.2$, in this case. The marked point in the figure 4.8 shows the point where the assumptions are met. We observe that $\lambda \approx 1$ in this scenario.

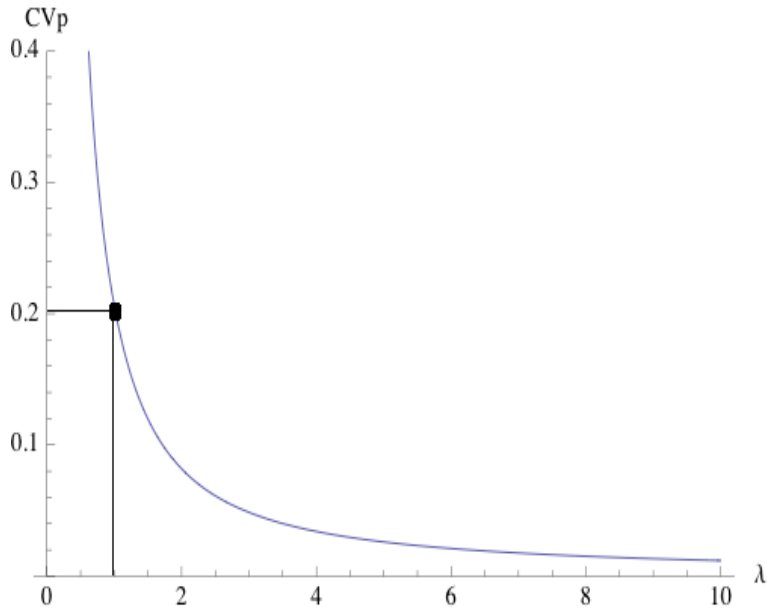

Figure 4.8: Study of the fluctuations in proximity space, CV_p as function of λ for $\mu = 10$ (average path length in distance space) with $CV_d = 0.2$.

To inspect in more detail the best value or values for λ, using the metric closure we plot, in figure 4.9 the theoretical λ versus μ (average shortest path), for several acceptable coefficients of variability in both spaces, assuming that the optimal value should share a controlled $CV_d \approx CV_p \leq 0.6$. The

results from this figure are obtained by finding the root (λ) of the equation:

$$CV_p^{theoretical}(\lambda) - CV_p = 0$$

$$CV_p^{theoretical}(\lambda) = \frac{\sqrt{< Y^2 > - < Y >^2}}{< Y >}$$

Where $< Y^2 >$ and $< Y >$ are given by equations 4.8, 4.9 and 4.11 with $j(x) = \frac{1}{x^{\frac{1}{\lambda}+1}}$ and we assume $h_X(\mu, \sigma)$ is normally distributed with $\sigma = \mu \times CV_d$ (μ is the average shortest path) with $CV_d \approx CV_p$ the real data fluctuations. We use *Mathematica 7* to find the roots of this equation. From this figure we can see that when we increase the coefficients of variability, λ also increases. However, λ remains contained in the interval $[0.8, 1.9]$. For small average shortest paths the best $\lambda \in [0.8, 1.2]$, where after a transient ($\mu \approx 25$), λ reaches an equilibrium, independent of scale factors (λ becomes invariant). The scale factor associated to the average shortest path length (characteristic for each network), depends mainly on the weights distribution. We can also observe that for very small fluctuations ($CV_d = CV_p = 0.1$), λ becomes invariant for values ≈ 1. $\lambda = 1$ is an optimal asymptotic value for small fluctuations, since $CV \geq 0$. In real data in order to guarantee a characteristic mean (average strongest path and average shortest path), in both spaces (proximity and distance), the fluctuations should be as small as possible. However in real data the shortest paths distribution only approximates to the normal distribution, which is one of our assumptions, resulting in higher fluctuations, for both spaces. For fluctuations $CV_d \approx CV_p \in [0, 0.4]$

we should use an isomorphism with $\lambda \in [0.8, 1.9]$. For $CV \approx 0$ the asymptotical optimal value is $\lambda = 1$ (see figure 4.9). This gives us a lower bound to calculate the desired metric closure in a distance graph to minimize fluctuations, λ should be larger or equal than 1 ($\lambda \geq 1$). To control fluctuations in both spaces (proximity, distance) we should choose λ according to the fluctuations obtained in the distance or proximity spaces (this can be seen as an optimization problem).

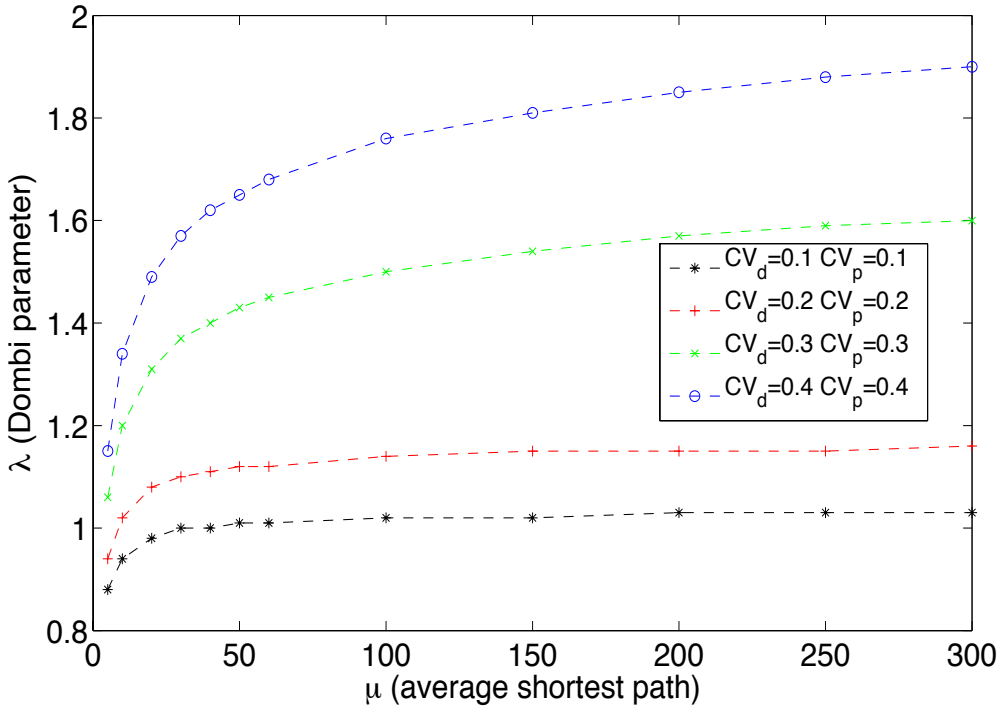

Figure 4.9: λ versus μ for several coefficients of variability CV_d and CV_p

In most computing applications, including the Complex Networks community researchers use mappings between proximity and distance spaces sim-

ilar to $\lambda = 1$, using isomorphisms $\varphi = \frac{1}{x}$ or $\varphi = \frac{1}{x} - 1$. We have to alert that the first choice $\varphi = \frac{1}{x}$ is not mathematically correct, since it maps $\varphi : [0, 1] \rightarrow [1, +\infty]$, which is not a distance space. $\lambda = 1$ leads to the more common φ and asymptotical optimal value, assuming small fluctuations. However, to constrain fluctuations we may want to use other values of $\lambda \geq 1$, depending on the level real data fluctuations.

4.11 Conclusions

In this thesis we have studied a theoretical framework to measure distances in weighted complex networks. From this study we reached the following conclusions:

First we formalized the relation between Transitive Closure in Fuzzy Sets and the Distance Closure through an isomorphism which maps a proximity graph into a distance graph. APSP Dijkstra algorithm or the Metric Closure in Computer Science is only one of the possible ways to calculate the distance closure. We saw, that it is possible to sweep different isomorphisms t-norm in the proximity space, given a fixed t-conorm (e.g., $\vee \equiv max$).

We also have studied the axiomatics that characterize this isomorphism we identified the prefered t-norm/t-conorm operations, which are closest to the APSP on metric closure in distance space, but are logically, consistent: ($DT^1_\wedge$, $DT^1_\vee$) this allows us to perform logical operations on distance graphs (via isomorphism to Fuzzy graphs), which can be useful in many areas of

research.

We explored the isomorphism with the Dombi t-norm generator. We estimated the best λ parameter in order to reduce fluctuations on the average shortest path and average strongest path. We saw that the asymptotical optimal λ occurs at $\lambda = 1$ assuming small fluctuations in real data, which is a lower bound and therefore λ must be $\lambda \geq 1$. It was observed also that λ is a scale invariant of the network. This scale depends mainly on the edge weights type of distribution.

While Complex Networks community have been suggesting the use of $\varphi = \frac{1}{x} \approx \frac{1}{x} - 1$ (Dombi generator with $\lambda = 1$) to perform the mapping between proximity and distance graphs, we have seen that this setting assumes fluctuations (average shortest path and average strongest path) close to zero in both spaces, which is not valid in real networks. However, $\lambda = 1$ is the asymptotical optimal value, which can justify this setting.

4.12 Appendix - Proofs of the Theorems

Theorem 4.1. *Let $G_P = (X, P)$ be a proximity (symmetric and reflexive) graph and Φ the graph distance function in definition 4.2, then $G_D = (X, D)$, where $D = \Phi(P)$ is symmetric and anti-reflexive.*

Proof. Since G_P is reflexive then $p_{x,x} = 1$ and from definition 4.2 we have $d_{x,x} = \varphi(p_{x,x}) = \varphi(1) = 0$, therefore G_D is anti-reflexive. Let x and y be two vertices of G_P, because a proximity graph is symmetric we have $p_{x,y} = p_{y,x}$, since φ is bijective $d_{x,y} = \varphi(p_{x,y}) = \varphi(p_{y,x}) = d_{y,x}$, therefore G_D is symmetric. □

Theorem 4.2. *If φ is a distance function as in definition 4.2 and $\wedge$, $\vee$ a t-norm, t-conorm pair and f and g a TD-conorm and TD-norm pair as defined in 4.3, the following statements are true:*

(1) $\varphi(a \wedge b) = g(\varphi(a), \varphi(b))$*;*
(2) $\varphi(a \vee b) = f(\varphi(a), \varphi(b))$*.*

Where $a, b \in [0, 1]$.

Proof. Let us assume $a \leq b$.
(1) Suppose $\varphi(a \wedge b) > g(\varphi(a), \varphi(b))$, thus the inequality is true if the maxima of $\varphi(a \wedge b)$ (must be maximum) is bigger than the minimum of $g(\varphi(a), \varphi(b))$ (must be minimum). $\varphi(a \wedge b)$ is maximum for $\wedge \equiv T_D$ (drastic product, see

[51] [53]) and $g(\varphi(a), \varphi(b))$ is minimum for $\varphi(b) = 0$, thus for $\varphi(b) = 0$ we obtain, $\varphi(a) \leq g(\varphi(a), \varphi(b))$, from the other side $\varphi(a \wedge b) \leq \varphi(min(a, b)) = \varphi(a)$. Therefore, $\varphi(a \wedge b) \leq g(\varphi(a), \varphi(b))$.

Suppose $\varphi(a \wedge b) < g(\varphi(a), \varphi(b))$, thus $\varphi(a \wedge b)$ must be minimum and $g(\varphi(a), \varphi(b))$ must be maximum. $\varphi(a \wedge b)$ is minimum for $\wedge \equiv min$ and $g(\varphi(a), \varphi(b))$ is maximum for $a = 0$, thus for $a = 0$ we obtain, $g(\varphi(a), \varphi(b)) \leq \varphi(a)$, from the other side $\varphi(a \wedge b) \geq \varphi(min(a, b)) = \varphi(a)$. Therefore, $\varphi(a \wedge b) \geq g(\varphi(a), \varphi(b))$, and from above this implies $\varphi(a \wedge b) = g(\varphi(a), \varphi(b))$, which proves statement (1).

(2) Suppose $\varphi(a \vee b) > f(\varphi(a), \varphi(b))$, thus $\varphi(a \vee b)$ must be maximum and $f(\varphi(a), \varphi(b))$ must be minimum. $\varphi(a \vee b)$ is maximum for $\vee \equiv max$ and $f(\varphi(a), \varphi(b))$ is minimum for $\varphi(a) = 0$, thus for $\varphi(a) = 0$ we obtain, $f(\varphi(a), \varphi(b)) \geq 0$, from the other side $\varphi(a \vee b) \leq \varphi(max(1, b)) = \varphi(1) = 0$. Therefore, $\varphi(a \vee b) \leq f(\varphi(a), \varphi(b))$.

Suppose $\varphi(a \vee b) < f(\varphi(a), \varphi(b))$, thus $\varphi(a \vee b)$ must be minimum and $f(\varphi(a), \varphi(b))$ must be maximum. $\varphi(a \vee b)$ is minimum for $\vee \equiv S_D$ (drastic sum, see [51] [53]) and $f(\varphi(a), \varphi(b))$ is maximum for $b = 0$, thus for $b = 0$ we obtain, $f(\varphi(a), \varphi(b)) \leq \varphi(a)$, from the other side $\varphi(a \vee b) \geq \varphi(max(a, b)) = \varphi(a)$. Therefore, $\varphi(a \vee b) \geq f(\varphi(a), \varphi(b))$, and from above this implies $\varphi(a \vee b) = f(\varphi(a), \varphi(b))$, which proves statement (2). □

Theorem 4.5. *if $G_P = (X, P)$ is a fuzzy proximity graph, φ the isomorphism (distance function) in definition 4.2 and $G_D = (X, D)$ is the respective distance graph, where $D = \Phi(P)$. The following is true with the distance*

composition:

1) $\Phi(P)\dot{\supseteq}\Phi(P^2)\dot{\supseteq}\Phi(P^3)\dot{\supseteq}\cdots\supseteq\Phi(P^\infty)$ *;*

2) $D\dot{\supseteq}D^2\dot{\supseteq}D^3\dot{\supseteq}\cdots\dot{\supseteq}D^\infty$*.*

where $\Phi(P^i)\dot{\supseteq}\Phi(P^{i+1})$ means that: $\forall x,y\in X:\varphi(p^i_{x,y})\geq\varphi(p^{i+1}_{x,y})$.

Proof. 1) φ is a monotonic decreasing function and because P is reflexive, from [60] we have $P\subseteq P^2\subseteq P^3\subseteq\cdots\subseteq P^\infty\Rightarrow\Phi(P)\dot{\supseteq}\Phi(P^2)\dot{\supseteq}\Phi(P^3)\dot{\supseteq}\cdots\dot{\supseteq}\Phi(P^\infty)$ which proves the statement.

2) To prove the second statement we first need to prove that $D\dot{\supseteq}D^2$, which is equivalent to showing that, $\forall x,y,z\in X:d^2_{x,y}=\underset{z}{f}\{g(d_{x,z},d_{z,y})\}\leq d_{x,y}$. Lets prove by absurd this statement: suppose $d^2_{x,y}>d_{x,y}$ then the minimum of $\underset{z}{f}\{g(d_{x,z},d_{z,y})\}$ must be $>d_{x,y}$. $\underset{z}{f}\{g(d_{x,z},d_{z,y})\}$ is minimum if f and g are minimum. g is minimum if $d_{z,y}=0$ for all $z\in X-\{x\}$, then $g(d_{x,z},d_{z,y})\geq d_{x,z}$. f is minimum if $d_{x,z}\geq d_{x,y}$ for all $z\in X-\{y\}$ then $f(d_{x,y},d_{x,z})\leq f(d_{x,y},+\infty)\leq d_{x,y}$, which contradicts our assumption, $d^2_{,x,y}>d_{x,y}$. Therefore, $d^2_{x,y}\leq d_{x,y}$.

By induction we can prove the general result.

$\forall x,y,z\in X:d^{n+1}_{x,y}=\underset{z}{f}\{g(d^n_{x,z},d_{z,y})\}$ by hypothesis $d^n_{x,y}\leq d^{n-1}_{x,y}$, thus $d^{n+1}_{x,y}\leq\underset{z}{f}\{g(d^{n-1}_{x,z},d_{z,y})\}=d^n_{x,y}$, which proves the second statement. □

Theorem 4.6. *Given a proximity graph* $G_P=(X,P)$*, a distance graph* $G_D=(X,D)$*, and isomorphism* φ*,* Φ *as defined in 4.2. For a t-norm* $\wedge$ *and t-conorm* $\vee$ *used to compute the transitive closure of P, then there exists a TD-conorm/TD-norm pairs* f *and* g*, to compute the distance closure of D,*

$\Phi(P^{tc}) = D^{dc}$, *which obey the condition:*

$$\forall x, y, z \in X : \underset{z}{f}\{g(\varphi(p_{x,z}), \varphi(p_{z,y})\} = \varphi(\underset{z}{\vee}\{\wedge(p_{x,z}, p_{z,y})\})$$

(where Φ^{-1} *is the inverse of* Φ *and* φ^{-1} *is the inverse function of* φ*). The same is true if we fix* f *(TD-conorm),* g *(TD-norm) and isomorphism* φ*,* Φ*, to obtain a pair of t-conorm, t-norm* $\vee$ *and* $\wedge$*.*

Proof. The transitive closure of P is given by P^{k_1} and the distance closure of D by D^{k_2}, with k_1 and k_2 integers. Let $n = max(k_1, k_2)$, thus for $\Phi(P^n) = D^n$ to be true, the following must also be true:

$$\forall x, y, z \in X : \underset{z}{f}\{g(\varphi(p_{x,z}), \varphi(p_{z,y})\} = \varphi(\underset{z}{\vee}\{\wedge(p_{x,z}, p_{z,y})\})$$

We can prove by induction that $\Phi(P^n) = D^n$ is true if we assume that the condition in this theorem is true.
The condition in this theorem is equivalent to:

$$\Phi^{-1}(\Phi(P) \circ \Phi(P)) = P^2 = P \circ P$$

Where $\Phi(P) \circ \Phi(P)$ is the distance composition using f and g, and $P \circ P$ is the transitive composition using $\wedge$ and $\vee$. We also can define D^n in function of Φ and P.

$$D^n = \underbrace{D \circ \cdots \circ D}_{n} = \underbrace{\Phi(P) \circ \cdots \circ \Phi(P)}_{n}$$

Therefore, what we want to prove is:

$$\Phi^n(P) = \Phi(P^n)$$

given the condition on this theorem is true.
by induction:

(1) $\Phi(P) \circ \Phi(P) = \Phi(P^2)$ (Basis);
(2) $\Phi^n(P) = \Phi(P^n)$ (Hypothesis);
(3) $\Phi^{n+1}(P) = \Phi(P^{n+1})$ (Thesis).

Assuming the condition on this theorem $\Phi^{-1}(\Phi(P) \circ \Phi(P)) = P^2$ is true, then it is also true that $\Phi(P) \circ \Phi(P) = \Phi(P^2)$. Thus, $\Phi^{n+1}(P) = \Phi^n(P) \circ \Phi(P) = \Phi(P^n) \circ \Phi(P) = \Phi(P^{n+1})$ from statements (1) and (2), which proves the theorem.

Let us prove that there exist a pair of binary functions f and g as defined in 4.3. From theorem 4.2 we have

$$g(\varphi(p_{x,z}), \varphi(p_{z,y})) = \varphi(\wedge(p_{x,z}, p_{z,y}))$$

and from the condition in this theorem, we have

$$\underset{z}{f}\{g(\varphi(p_{x,z}), \varphi(p_{z,y}))\} = \varphi(\underset{z}{\vee}\{\wedge(p_{x,z}, p_{z,y})\})$$

$$\underset{z}{f}\{\varphi(\wedge(p_{x,z}, p_{z,y}))\} = \varphi(\underset{z}{\vee}\{\wedge(p_{x,z}, p_{z,y})\})$$

Therefore,

$$f(d_{x,z}, d_{z,y}) \equiv \varphi(\vee(\varphi^{-1}(d_{x,z}), \varphi^{-1}(d_{z,y})))$$

The conditions of this theorem leads to the equations of theorem 4.2:

$$g(d_{x,z}, d_{z,y}) = \varphi(\wedge(\varphi^{-1}(d_{x,z}), \varphi^{-1}(d_{z,y})))$$

$$f(d_{x,z}, d_{z,y}) \equiv \varphi(\vee(\varphi^{-1}(d_{x,z}), \varphi^{-1}(d_{z,y})))$$

.

From these last equations we can also find $\vee$ and $\wedge$ given f, g and the isomorphism φ:

$$\vee(p_{x,z}, p_{z,y}) = \varphi^{-1}(f(\varphi(p_{x,z}), \varphi(p_{z,y})))$$

$$\wedge(p_{x,z}, p_{z,y}) = \varphi^{-1}(g(\varphi(p_{x,z}), \varphi(p_{z,y}))) \qquad \square$$

Corollary 4.1. *According to the formulation of theorem 4.6, let* $f \equiv \min()$, $g \equiv +()$ *and* φ *a distance function, as defined in 4.2. If* $\vee = \max()$ *as t-conorm, then the t-norm operator* $\wedge$ *exists and* φ *is its generator function.*

Proof. We have seen in theorem 4.2 that $\varphi(x \wedge y) = g(\varphi(x), \varphi(y))$ therefore

$\forall x, y, z \in P$ and by theorem 4.6:

$$\varphi^{-1}(\underset{z}{min}\{\varphi(p_{x,z}) + \varphi(p_{z,y})\}) = \underset{z}{max}\{\wedge(p_{x,z}, p_{z,y})\}$$
$$\underset{z}{max}\{\varphi^{-1}(\varphi(p_{x,z}) + \varphi(p_{z,y}))\} = \underset{z}{max}\{\wedge(p_{x,z}, p_{z,y})\}$$
$$\Rightarrow$$
$$\varphi^{-1}(\varphi(p_{x,z}) + \varphi(p_{z,y})) = \wedge(p_{x,z}, p_{z,y})$$

This last result is the characterization function of t-norms, according to theorem 4.7 [53], which states that $\wedge$ is a t-norm and φ is the decreasing generator function (obeying definition 4.2). □

Theorem 4.7. *Given the isomorphism φ, if D^{mc} is the metric closure with $f \equiv min$ and $g_1 \equiv +$, and D^{um} is the ultra-metric closure with $f \equiv min$ and $g_2 \equiv max$ then $D^{mc} \dot{\supseteq} D^{um}$ is equivalent to $P^{mc} \subseteq P^{um}$, where $D^{mc} = \Phi(P^{mc})$ and $D^{um} = \Phi(P^{um})$.*

Proof. We can prove by induction that:

1) $D^2 \dot{\supseteq} \Phi(P^2)$;

2) $\begin{cases} H : D^n \dot{\supseteq} \Phi(P^n) \\ T : D^{n+1} \dot{\supseteq} \Phi(P^{n+1}) \end{cases}$

Let's prove 1)

$\forall x, y, z \in X : D^2_{mc} = \underset{z}{f}(d_{x,z} + d_{z,y}) = \underset{z}{f}(\varphi(p_{x,z}) + \varphi(p_{z,y})) \geq \underset{z}{f}(g_2(\varphi(p_{x,z}), \varphi(p_{z,y}))) = D^2_{um}$, therefore $D^2_{mc} \dot{\supseteq} D^2_{um}$.

2) by the hypothesis we know that $\forall x, y, z \in X : D^n \geq \Phi(P^n)$, then using this result we have $\forall x, y, z \in X : D^{n+1} = \underset{z}{f}\{d^n_{x,z} + d_{z,y}\} \geq \underset{z}{f}\{\varphi(p^n_{x,z}) + \varphi(p_{z,y})\}$

, because $\underset{z}{f}\{\varphi(p^n_{x,z}) + \varphi(p_{z,y})\} \geq \underset{z}{f}\{\varphi(p^n_{x,z}) \vee \varphi(p_{z,y})\}$ and using theorem 4.2, $\underset{z}{f}\{g_2(\varphi(p^n_{x,z}), \varphi(p_{z,y}))\} = \varphi(\underset{z}{\vee}\{p^n_{x,z} \wedge p_{z,y}\}) = \Phi(P^{n+1})$, so

$\forall x, y, z \in X : D^{n+1} \geq \Phi(P^{n+1})$, which proves that $D^{mc} \equiv D^n \dot{\supseteq} \Phi(P^n) \equiv D^{um}$. □

Theorem 4.8. *Given a fuzzy complement $c(x)$, a t-norm $DT^1_\wedge = \frac{ab}{a+b-ab}$ and a t-conorm $max(a, b)$, then the triple has no involutive complement, which satisfies the De Morgan's laws.*

Proof. A complement is involutive if $c(c(x)) = x$. If the complement $c(x)$ satisfies the De Morgan's laws we have:

$$\overline{a \vee b} = \bar{a} \wedge \bar{b}$$

$$c(max(a, b)) = \frac{c(a)c(b)}{c(a) + c(b) - c(a)c(b)}$$

without loss of generality let $a \geq b$

$$c(a) = \frac{c(a)c(b)}{c(a) + c(b) - c(a)c(b)}$$

$$c(a)(1 - c(b)) = 0$$

$$c(a) = 0 \vee c(b) = 1$$

the only function that satisfies this condition is the threshold function, which is not involutive [53]. □

Chapter 5

Performance of metric closure on recommender systems

In the previous chapter we related transitive and distance closure with the all pairs shortest path problem. The distance closure and more specifically the metric closure allows us to study semi-metric behavior in weighted graphs. In this chapter and next chapter we apply the metric closure to interpret the semi-metric behavior in complex networks. First, in this chapter we search for evidence between semi-metric behavior and prediction and second, next chapter, we relate semi-metric behavior with the study of the structure of complex networks.

5.1 Introduction

The search for strength of association between time events is inherent to many systems, such as: recommender systems, social behavior, functional brain interaction and many more. Recommender systems are a good example of prediction, since given the information from the past (e.g. the relation between users and items in collaborative filtering) recommend new useful items to the users in the future. In Rocha et al [76] we found some evidence on this relation in recommendation systems, between semi-metric behavior and prediction. We employed two different types of weighted graphs in our analysis and development: Proximity graphs, a type of Fuzzy Graphs based on a co-occurrence probability (Fuzzy Jaccard measure, see chapter 4), and (semi-metric) distance graphs, which do not necessarily observe the triangle inequality of Euclidean distances for all edges. Both types of graphs were used to develop intelligent recommendation and collaboration systems for the MyLibrary@LANL web service, a user-centered front-end to the Los Alamos National Laboratory's digital library collections and Web resources. Recommendation were issued using the semi-metric behavior of distance graphs derived from users access profile. The quality of recommendation was assessed using expert evaluations [76]. This assessment has shown that semi-metric recommendations were relevant. In this chapter, instead of using experts assessment we use a benchmark database from the group MovieLens [1], to

[1] http://movilens.umn.edu

test the accuracy of the recommendations, and compare with some previous work done on the same data. The advantage of this type of assessment is that we do not have the subjectivity of human experts. The disvantage is that the results are specific to the Movilens database on the topic of movies preferences only. There are other datasets such the one provided by Netflix[2], where we can test the evidence between semi-metric behavior and prediction. However, we do not intend in this chapter to perform an exhaustive study on recommender systems. We leave for future work a more detailed study on recommender systems where we intend to use the Netflix benchmark.

In collaborative filtering we manage a database that has information relating items with users. Each user gives a certain score to a specific item. The assumption in collaborative filtering is that if users share the same interests in the past, they will also have similar preferences in the future.

In the next section we show how we built our recommender system.

5.2 Collaborative Filtering Based Recommendation Systems

We developed and tested two types of collaborative filtering algorithms, proximity and semi-metric based. Collaborative filtering systems start with the relation between Items and Users. This relation consists on the history or assessment done by the users to items. Examples are: the relation between

[2]www.netflix.com

users and items such as in Amazon.com where users buy books and other items, MovieLens where users rate movies, etc.

Given this relation between users and items we can associate users or items to each other with some similarity or dissimilarity measures between users (user-based) or items (item-based), respectively.

We build our item-based and user-based proximity graph by applying the generalized Jaccard similarity measure (see chapter 4). Other measures can be used such as: cosine projection (vector based), mutual information, Pearson correlation and many more.

Given the proximity between users (user-based) or the proximity between items (item-based) a collaborative filtering system identifies, for a given user, a set of items to be recommended. In the case of user-based approaches we search for a given user a neighborhood (using e.g. nearest neighbors method, see [98]) of similar users and recommend the items more popular among the set of neighbors. In the case of item-based recommendation, for each user, we search for items similar to the ones that user have already consumed or rated.

Item-based systems are based on the proximities between items. Each user has associated a set of items, which is a subset of all items available to the users. Given the relation between items, we compute for each user a set of items which are similar to the items associated to that user.

The following algorithm describes an item-based collaborative filtering:

Algorithm 5.1. *Item-Based*

For each user:

1. *We retrieve the set of items from the training set (relation between users and items from the past $U \times I$) for this particular user and form a vector.*

2. *For each item in the proximity relation $I \times I$ (row vector), obtained using the Fuzzy Jaccard similarity measure (equation 4.1) on the relation $U \times I$, we identify the items from the previous step* 1. *in this row vector and calculate the average of the identified values. With this we get a score for each item in $I \times I$.*

3. *Each user is recommended the top n items.*

4. *Do the previous steps for all users.*

We tested the following item-based algorithms:

1. *Proximity Item-based algorithm (Prox-Item-based)*

   The simplest item-based algorithm follows algorithm 5.1 using the matrix $I \times I$ calculated from the proximity measure of equation 4.1.

2. *Semi-metric algorithm (SM-Item-based)*

   Here we calculated the metric closure from the proximity relation $I \times I$ using the isomorphism of equation 4.7 (Dombi t-norm generator with $\lambda = 1$). From the resulting matrix we identify the semi-metric pairs (edges) where the below average ratio is above a given threshold (equations 2.19 and 4.3), and insert the corresponding edges from transitive

closure of $I \times I$) into the proximity graph ($I \times I$). Finally use this proximity graph as input for item-based algorithm 5.1.

User-based systems are based in the similarities between users. For each user is determined a neighborhood of users, which are more similar to that user. It is recommended the set of items more popular in that neighborhood of users.

Algorithm 5.2. *User-Based*

For each user:

1. *determine the number of users n in $U \times U$, which form the neighborhood of the user according to a given alpha-cut (threshold);*
2. *For this set of users calculate the top m items more frequent among that neighborhood;*
3. *Recommend this set of items to the user;*
4. *Do the previous steps to all users.*

We tested the following user-based algorithms:

1. *Proximity User-based algorithm (Prox-User-based)*

   The simplest user-based algorithm follows algorithm 5.2 using the matrix $U \times U$ calculated from the proximity measure of equation 4.1.

2. *Semi-metric algorithm (SM-User-based)*

   Here we calculated the metric closure from the proximity relation $U \times U$

using the isomorphism of equation 4.7 (Dombi t-norm generator with $\lambda = 1$). From the resulting matrix we identify the semi-metric pairs (edges) where the below average ratio is above a given threshold (equations 2.19 and 4.3), and insert the corresponding edges from transitive closure of $U \times U$) into the proximity graph ($U \times U$). Finally use this proximity graph as input for item-based algorithm 5.2.

5.3 Experimental Evaluation

Data Sets

In this work we used the benchmark data set MovieLens. This data set is a collection of votes given by web users (943 users) in respect to a given movie (1682 movies), as a total of 100,000 ratings. The group Movilens provides a set of datasets. In these datasets were retained only users that had rated 20 or more movies (943 users). Each user gives his opinion (vote) in respect to a movie graded in a scale from one to five. These data sets are based in the full matrix items (movies) versus users votes and partitioned in sets of training and test. It was studied by the group [79] that the best partition is a training set with 80% of the votes and a test set with the remaining 20%. These data sets are divided in two major data sets one with a test set with the ten votes per user (about 10,000 ratings), while the training set contains the remainder of the ratings (about 90,000 ratings). None of the edges belongs both to the training and test sets.

In our experiment we are only interested on the relation between semi-metric behavior and prediction, i.e., based on the past watched movies from a given user what are the movies he/she will watch in the future. Therefore, we converted these data sets to binary votes: one or zero.

Evaluation Metrics

As evaluation metrics, we used Precision and Recall and F1 measure and a variant of the Somers D, the degree of agreement metric to assess the performance of the recommender system, [82]. Precision, recall, and the F1 measures are the traditional measures in information retrieval computed using unordered datasets. There are other assessment metrics for order datasets such as the Area Under Curve (AUC). However, this measure is difficult to implement on this particular dataset. Instead, we use the Somers D, which is easy to apply and already tested in a set of recommender systems.

Precision and Recall: Precision and recall can be defined in the following manner:

$$recall = \frac{\mid test \cap top - N \mid}{\mid test \mid} \tag{5.1}$$

$$precision = \frac{\mid test \cap top - N \mid}{\mid N \mid} \tag{5.2}$$

where top-N is the top N recommendations and test is the test set. From

equations 5.1 and 5.2 we calculate the F1 parameter, which relates precision and recall with equal importance.

$$F1 = \frac{2 \cdot recall \cdot precision}{recall + precision} \tag{5.3}$$

The Somers D method follows the following procedure [98].

1. For each user we take the vector of similarities for each movie from the training set.

2. Take only the non-watched movies (not in the test set).

3. Rank the non-watched movies taking in consideration all movies.

4. Compute the degree of agreement: consider each pair (a, b) of movies from ranking, with a in the test set and b not. If a ahead of b: correct pair (agreement), b ahead of a: incorrect pair, 5.4.

$$d = \frac{\#agreements}{\#total - of - pairs} \tag{5.4}$$

5. Compute the global degree of agreement.

This variant of Somers D degree of agreement give us a measure of how well our retrieved set of recommendations is distributed in the first positions of our list of relevant items.

5.4 Results

We compare our results with the ones of Fouss et al [98]. Table 5.1 shows our results for the proximity and semi-metric approaches for item-based and user-based recommender systems.

	Prox-Item-based	SM-Item-based	Prox-User-based	SM-User-based
Agreement (in %)	89.53	90.16	88.20	88.16
F1	0.1827	0.1832	0.2130	0.2179

Table 5.1: Results for recommendation system. Somers'D degree of agreement [36] [98] and F1 measure.

The semi-metric thresholds were set on the below average ratio distribution around the cut-off point of the distribution, as shown in figure 5.1 for the item-based approach. In the item-based, we tested values below and above the threshold pointed in the figure 5.1, and the performance decreased in both directions, having its maximum at the point chosen ($b = 8$). For the user-based approach, we also set the threshold around the cut-off point, according to figure 5.2. In this case we observe at the cut-off point a small decrease on the Somers D measure and an increase on the F1 measure. We also search for thresholds values below and above the cut-off point. We observed that the Somers D measure decreased below the cut-off point and increased slightly for values above the cut-off point reaching 88.20% when we insert just a few semi-metric edges. The F1 measure remained almost the same 0.2179 for values around the cut-off. This indicates that the threshold should be chosen around the cut-off point of the distribution.

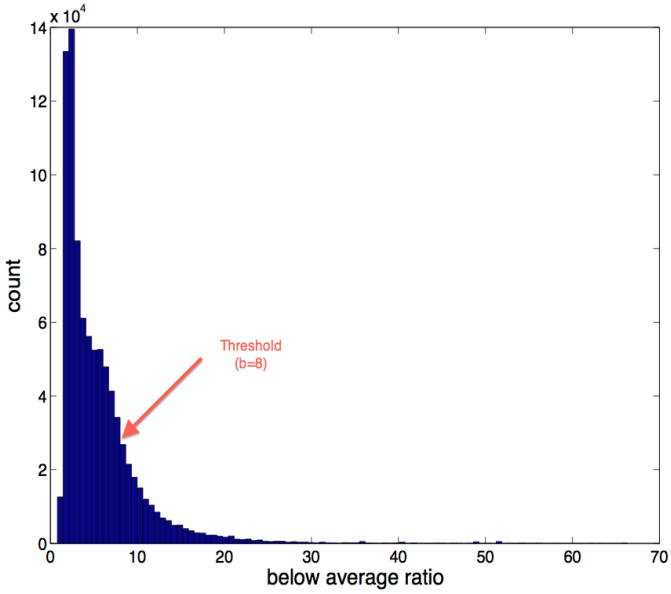

Figure 5.1: below average ratio distribution of semi-metric edges with direct edge $= \infty$ (item-based). The threshold was chosen at the indicated point.

Fouss et al in [98] obtained the following results presented in tables 5.2 and 5.3 for several item and user based algorithms described below. A more detailed description of these algorithms can be found in [98].

As we can see from tables 5.1, 5.2 and 5.3 the semi-metric approach improves the item-based proximity method and is as good as the best result of Fouss et al [98]. Our item-based algorithms are also among the best algorithms of user-based on table 5.3. Our semi-metric user-based approach is improved on F1 measure, but on the Somers D does not improve the results, they remain almost the same. This maybe can be explained with the fact that User-based approaches have to set the number of neighbors around a given user, and this has a strong impact on the results. We leave this further

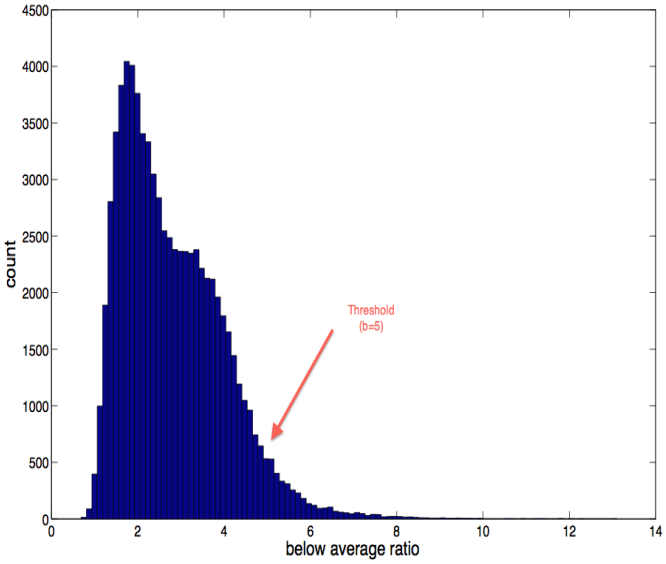

Figure 5.2: below average ratio distribution of semi-metric edges with direct edge $= \infty$ (user-based). The threshold was chosen at the indicated point.

analysis for future work, since the objective of this chapter is to show that semi-metric behavior in some way improves the predictions of a given user watch a movie in the future.

Next we describe according to Fouss [98] the algorithms involved in this comparison.

	MaxF	CT	PCA CT	One-way	Return	L^+	kNN	Cosine	Katz	Dijkstra
Agreement (in %)	85.69	85.66	87.08	85.64	80.65	90.99	−−	−−	87.90	49.11

Table 5.2: Results for item-based recommendation system from [98]. Somers'D degree of agreement [36] [98].

Maximum frequency algorithm (MaxF). This method ranks the movies by the number of users who watched them. The most watched movie

	MaxF	CT	PCA CT	One-way	Return	L^+	kNN	Cosine	Katz	Dijkstra
Agreement (in %)	−−	74.48	82.46	74.48	54.30	93.02	92.63	92.73	89.82	76.09
#Neighbors	−−	100	60	100	100	100	100	60	20	100

Table 5.3: Results for user-based recommendation system from [98]. Somers'D degree of agreement [36] [98].

(a blockbuster) is simply suggested first to each user. The ranking is thus the same for all the users. MaxF will be considered as a reference to which all the other methods will be compared; it may be viewed as an equivalent of basing the decision only on the a priori probabilities in supervised classification.

Average commute time (CT). We use the average com- mute time $n(i, j)$ to rank the elements of the considered set, where i is an element of the people set and j is an element of the set to which we compute the dissimilarity (the movie set). For instance, if we want to suggest movies to people for watching, we compute the average commute time between people elements and movie elements. The lower the value is, the more similar the two elements are. In the sequel, this quantity will simply be referred to as commute time.

Principal components analysis based on ECTD (PCA CT). Based on the eigenvector decomposition of L^+, the nodes can be mapped into a new Euclidean space that preserves the ECTD, or a m-dimensional subspace keeping as much variance as possible, in terms of ECTD. Thus, after performing a PCA and keeping a given number of principal components, we recompute the distances in this reduced subspace. These approximate ECTD between people and movies are then used to rank the movies for each user (the closest

first).

Average first-passage time (one-way). In a similar way, we use the average first-passage time, $m(i|j)$, to rank element i of the movie set with respect to element j of the people set. This provides a dissimilarity between person j and any element i of the movie set. This quantity will simply be referred to as one-way time.

Average first-passage time (return). As a dissimilarity between element j of the people set and element i of the movie set, this method uses $m(j|i)$ (the transpose of $m(i|j)$), that is, the average time used to reach j (from the people set) when starting from i (from the movie set). This quantity will simply be referred to as return time.

Pseudoinverse of the Laplacian matrix (L_+). L_+ provides a similarity measure since it is the matrix containing the inner products of the node vectors in the Euclidean space where the nodes are exactly separated by the ECTD. Once we have computed the similarity matrix, movies are ranked according to their similarity with the user, and the closest movie that has not been watched is proposed first.

Nearest neighbors (kNN). This method is one of the simplest and oldest methods for performing general classification tasks. It can be represented by the following rule: to classify a new item, choose the class of the nearest example in the training set as measured by a similarity metric. When choosing the k nearest examples to classify the unknown pattern, one speaks about *k-nearest* neighbors techniques.

Cosine coefficient. The cosine coefficient between persons i and j, which measures the strength and the direction of a linear relationship between two variables. The predicted value of person i for movie j is computed in a similar way as in the k-nearest neighbors method.

Katz. This similarity index has been proposed in the social sciences field and has been recently rediscovered in the context of collaborative recommendation and kernel methods where it is known as the von Neumann kernel. Katz proposed a method of computing similarities, taking into account not only the number of direct links between items but, also, the number of indirect links (going through intermediaries) between items.

Shortest path algorithm (Dijkstra). This algorithm solves a shortest path problem for a directed and connected graph with nonnegative edge weights. As a distance be- tween two elements of the database, we compute the shortest path between these two elements.

5.5 Discussion and Conclusions

From table 5.1 we can see that by introducing semi-metric edges into the proximity graph $I \times I$ and $U \times U$ we improve our predictions (recommendations), confirming the previous evidence in Rocha et al [76]. Moreover, the combination of our generalized Jaccard proximity graphs with addition of semi-metric edges are among the best recommender systems tested in previous works [98].

As suspected semi-metric edges show some evidence of predicability in recommender systems. In the next chapter we will see how semi-metric edges are related to the structure of graphs and how can we use this information to better characterize the structural properties of a complex networks.

Chapter 6

Weighted graphs and the small-world phenomenon

In this chapter we propose a new methodology to analyze weighted graphs, based on semi-metric behavior. We compare our methodology with the traditional approach by studying the affect on the average path length, clustering coefficient and semi-metricity and their implications to measure the small-world phenomenon. We analyzed six real-world networks: US Airports, Structural Human Cerebral Cortex, Functional Human Brain, Scientific Collaboration, Astrophysics Collaborations and High-Energy Theory Collaboration.

6.1 Introduction

A traditional approach to study weighted graphs, especially to characterize small-world behavior, is by to apply thresholds to the weights of the graph, and then study the properties of the resulting crisp graphs. We propose a new approach based on the *semi-metric threshold*. As detailed below this method removes an edge $e_{i,j}$ only if the strongest path between corresponding vertices v_i and v_j is below a certain threshold value. Let us describe the idea with an example: I know the president of the United States (met him in a formal dinner), but my direct tie to him is very weak. However, I have a strong indirect influence on him via a chain that includes other people (strongest path). In this case we do not remove the direct tie (link) between the president and me, because there is a strong indirect way in which I can influence the president. This may imply that in the future the direct tie can become stronger. However, when there is no indirect path (above a threshold), the ability to influence the president is too small, and so we can in effect remove the direct tie from the network. After we apply this reasoning by removing all direct ties that do not have an indirect or direct path above the (semi-metric) threshold we end up with a weighted sub-network of the original weighted network. This process is different from the traditional, where we apply a threshold to all edge weights of the network, thus removing the tail of the weights distribution. In our case, depending on the level of semi-metric behavior associated with the network, we end up

with a weighted sub-network with weights distribution more similar to the original network.

The shortest path length, as we have seen in chapter 4, depends on the specific graph closure used, which is defined by a t-norm/t-conorm pair (in proximity graphs) or a TD-norm/TD-conorm pair (in distance graphs). As shown in chapter 4, the isomorphism between these spaces, can be controlled by a t-norm generator function.

There is an unlimited number of t-norm generators, which follow in two classes: parametric and non-parametric [53]. In this chapter we apply a parametric t-norm, which allows us to sweep the range of graph closures. We choose the Dombi t-norm generator:

$$\varphi(x) = \left(\frac{1-x}{x}\right)^{\lambda} \tag{6.1}$$

where λ is the sweeping parameter. This t-norm generator takes values in $]0, +\infty[$: $\lambda \to 0$ lower bound (*drastic product*) and $\lambda \to \infty$ is the upper bound (*minimum*, ultra-metric). As in chapter 4, we choose this t-norm generator because it yields the more commonly used transformation from proximity to distance graphs, when $\lambda = 1$, [2] [89]. Moreover, as we have seen in chapter 4, this particular t-norm with $\lambda = 1$ leads to the closure with most desirable properties.

In this chapter we study the impact of the proposed methodology to better characterize the small-world phenomenon in weighted networks, which is typ-

ically based on two measures: *average path length* and *clustering coefficient*. This way, here we propose semi-metric behavior as one additional measure for this phenomenon. Note, that all measures which depend in some way on path lengths can profit from ours new methodologies. Examples of such measures are: Efficiency, Modularity, Betweenness, Clustering Coefficient, etc. It is easy to see, for instance that the weighted efficiency of a network with the Dombi t-norm generator corresponds to the average strongest path, when $\lambda = 1$. However, in this thesis we will restrict ourselves to studying average path length, clustering coefficient and semi-metric behavior to better characterize the small-world phenomenon in weighted networks.

The Small-world phenomenon is well understood in crisp graphs. When such graphs have a small *average shortest path* between any two vertices (that is, in the same order of magnitude of the logarithm of the size of the network) and a high *clustering coefficient* (high local transitivities), they are said to be organized as a small-world. An example is: I know a person who knows a person, who knows a person who is friend of the president of United States. Therefore, I am three steps away (degrees of separation) from the president of United States. Milgram [90] showed that, on average a person that lives in the United States is six degrees of separation from any other person in United States. In weighted graphs however, the situation is different: links between vertices (e.g. persons) are weighted according to some measure of proximity or distance, which is a more realistic situation; for example when we have a degree of friendship. Imagine the previous case where I am three

degrees of separation from the president of the United States. If the weights between each person are weak, which normally is the case, there is no way that I can influence the president of the United States. However, if all links on the path to the president of United States are strong, there is a greater possibility that I can influence the president. This simple example shows us that by moving from crisp to weighted graphs, the same data can lead to the presence or absence of the small-world phenomenon.

6.2 Semi-metric thresholding

As we have seen in chapter 4, when we produce distance graphs from proximity data, shortest paths depend on the isomorphism we apply, which can be constrained by the t-norm in the proximity space: see figure 6.1 for Dombi t-norm example. By choosing a parameterized t-norm we can sweep various by isomorphisms.

With the Dombi t-norm we can identify four regimes that depend on the λ parameter: I, II, III and IV, figure 6.2. Regime I has been well studied and it includes binary or crisp graphs, which do not observe any semi-metric behavior. However, regimes II, III and IV are not so well studied. In figure 6.2 we see these regimes according to the semi-metric percentage computed with the respective closure defined by λ.

In regime I ($\lambda \to 0$; drastic product) the shortest path distribution has a characteristic average path length; because the graphs resulting from a

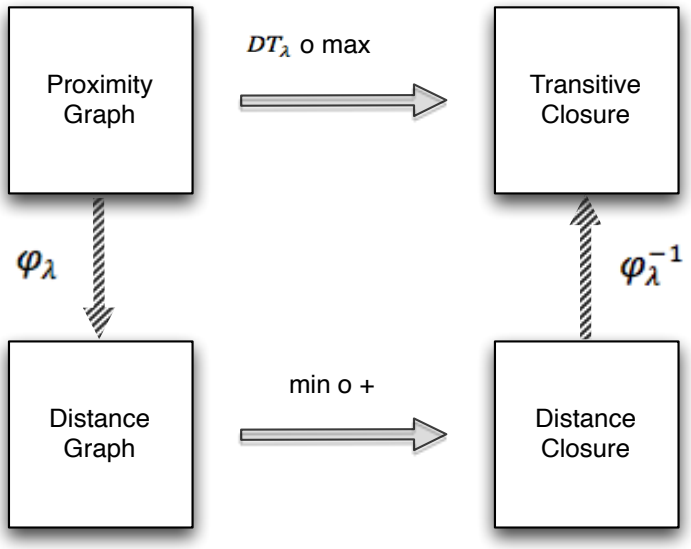

Figure 6.1: Closures with Dombi t-norm generator φ_λ

drastic product closure is essentially crisp, the semi-metric percentage of the graph (SM) is 0. In regime II ($\lambda > 0$ and inferior to the curve inflection point), the shortest path distribution still has a characteristic average path length with semi-metric percentage ≈ 0. In regime III the variance of the shortest path distribution starts increasing and diverges in regime IV (see below). The semi-metric percentage increases in regime III, and stabilizes in IV. *lambda* $= 1$ is in regime III (see case studies below), and we have seen that for the Dombi t-norm the best λ in this regime is $\lambda \approx 1$ (see chapter 4), therefore in our methodology we will fix $\lambda = 1$.

The counterpart of shortest path distribution (distance space) in our isomorphism space, is the strongest path distribution (proximity space).

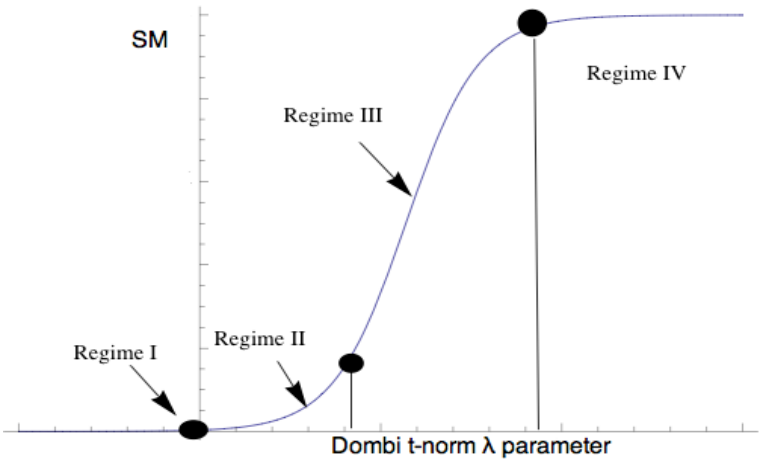

Figure 6.2: Regimes with dombi t-norm generator φ_λ

Our semi-metric threshold (SMT) method is based on the removal of all direct edges, which have both a direct and an indirect path below a threshold. This is exemplified in figures 6.3 and 6.4 in the proximity graph. In figure 6.3 we have a weak (proximity) direct edge between vertices 1 and 2, but a strong indirect path between these two vertices, via the path $1-3-4-2$. With a $0.5 \geq SMT > 0.1$, we do not remove the direct edge between vertices $1-2$. Otherwise, if the direct edge is weak and so is any indirect path (below the SMT threshold) between the two vertices, we remove the direct edge as shown in figure 6.4. The motivation for this kind of procedure can be exemplified in the context of a social network. Suppose a person has a direct weak link with the president of the united states – say, this person has been in a social event where the president was; the president knows this person exists but has a really weak social connection. However, this person has a strong indirect path through his social network to the president, via which

he/she can influence the president. In this case the weak direct connection is preserved because it denotes a link that can be indirectly realizable, and thus we expect that it may become stronger in the future. Otherwise, if there are no strong indirect paths in the network between the person and the president, the weak direct link is assumed useless, and it can be removed.

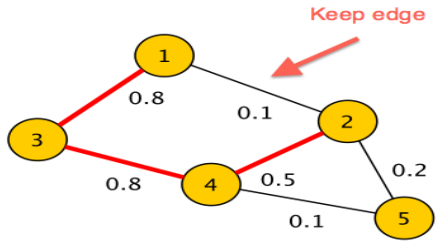

Figure 6.3: Indirect path length shorter than the direct – keep direct edge between vertices 1 and 2.

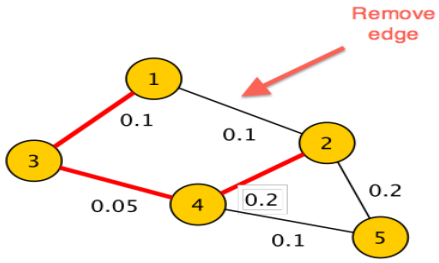

Figure 6.4: Indirect path length bigger than the direct – remove direct edge between vertices 1 and 2.

We show (case studies below) that this form of thresholding the network preserve better its original structure compared with traditional methodologies [81]. Moreover, by preserving the structure of the network, we gain

measures that can be characteristic of the network even in the weighted case such as: clustering coefficient, transitivity, semi-metric ratios, vertices degree, modularity, etc.

6.3 Semi-metric edges and the Network Metric backbone

In a network and its graph representation we have two types of edges: semi-metric and metric. Semi-metric edges are edges that have an indirect path that is shorter[1] than the direct edge. Metric edges are edges for which there is no indirect path shorter than the direct edge.

By definition, semi-metric edges do not affect the distribution of shortest path, since there is at least one alternative path shorter than the direct edge. The metric edges are the ones responsible for the shortest path distribution, since at least one shortest path crosses these edges.

The smallest sub-graph that preserves the distribution of shortest path of a given graph $G(V, E)$ is the graph $G_m(V_m, E_m)$, $V_m \subseteq V$ and $E_m \subseteq E$, where E_m is the set of metric edges. If we remove one edge from E_m the average shortest path increases, since any of these edges participate at least in one of the shortest paths. If the graph G has no isolated vertices it is easy to see that $V_m \equiv V$.

[1]We use shorter in distance space, which isomorphically stronger in the proximity space. Here, we center the discussion on the distance space. But, via the isomorphism, every concept has a proximity counterpart

Sub-graph G_m is what we call the *metric backbone* of a network. The graph G is robust to semi-metric edges removal, but not to metric edges. By removing a certain amount of metric edges from the network, the network starts breaking into modules. Semi-metric edges only fill-up the metric sub-graph G_m, increasing its connectivity. However, the evolution of a network in time can turn semi-metric edges into metric.

The metric backbone by its definition, has a main impact on all measures associated to the shortest path, such as: efficiency, betweenness, etc. Semi-metric edges by definition, have an impact only on measures that are related to connectivity structure such as: clustering coefficient, transitivity, vertices degree, modularity (or community structure), etc. They do not impact the shortest path measures, since they do not participate in shortest path. Modularity in particular, depends on both types of edges: semi-metric and metric.

In the next subsections we will study the impact of these two types of edges on real-world networks.

6.4 Measures and processing of weighted graphs

6.4.1 Normalization Procedures

In all networks we normalized the weighted graph using a linear normalization that maps an interval

$$[a, b] \to [\epsilon, 1 - \epsilon],$$

where ϵ is small or zero. For $\epsilon = 0$ the normalized edges are:

$$en_{ij} = \frac{e_{ij} - \underset{i,j}{MIN}(e_{ij})}{\underset{i,j}{MAX}(e_{ij}) - \underset{i,j}{MIN}(e_{ij})} \tag{6.2}$$

In the general case with $\epsilon > 0$ (avoids merging vertices):

$$en_{ij} = \frac{(1-2\epsilon)e_{ij} + (2\epsilon - 1)\underset{i,j}{MIN}(e_{ij})}{\underset{i,j}{MAX}(e_{ij}) - \underset{i,j}{MIN}(e_{ij})} + \epsilon \tag{6.3}$$

In our networks we used equation 6.3 with $\epsilon = 0.01$ in order to preserve the strongest and weakest edge in the network.

6.4.2 Average path length

The average path length $< l >$ in graph is defined according to the following equation:

$$< l >= \frac{1}{n \cdot (n-1)} \cdot \sum_{i,j} d(i,j) \tag{6.4}$$

where n is the number of vertices of the graph, $d(i,j)$ is the distance between vertex v_i and vertice v_j. If $i = j$ or $d(i,j) = \infty$ then some consider for the calculation $d(i,j) = 0$, which only happens when the graph is not connected. Here we only deal with networks that are connected.

6.4.3 Clustering coefficient

The clustering coefficients have been defined in chapter 2. Here, we recall the definitions. According to Watts the clustering coefficient $C(i)$ for a given vertex i in a crisp graph is defined by the following equation:

$$C(i) = \frac{e_i}{k_i(k_i - 1)/2} \tag{6.5}$$

where k_i is the degree of vertex i and ei the edges between vertex i and the neighbors of vertex i. The average clustering coefficient is defined by the equation:

$$< C >= \frac{1}{N} \sum_i C(i) \tag{6.6}$$

For direct and undirected weighted graphs the following clustering coefficients are proposed by Onnela et al. [67]. They renormalize the clustering coefficient of equation 6.6 of the equivalent crisp network. The renormalization factors, *intensity* and *coherence* are defined in the following way:

$$I(h) = \left(\prod_{(i,j) \in l_h} e_{i,j} \right)^{\frac{1}{|l_h|}} \tag{6.7}$$

$$Q(h) = \frac{I(h) \times |l_h|}{\sum_{(i,j) \in l_h} e_{i,j}} \tag{6.8}$$

where $I(h)$ is the Intensity for vertex h, $e_{i,j}$ the weight between vertex i and j, l_h is the number of edges for vertex h and $Q(h)$ is the coherence for vertex

h. The *average intensity* and the *average coherence* are defined as:

$$< I(h) >= \frac{1}{e_h} \sum_{i \in N(h)} I(i) \tag{6.9}$$

$$< Q(h) >= \frac{1}{e_h} \sum_{i \in N(h)} Q(i) \tag{6.10}$$

where $N(h)$ denotes the neighborhood of vertex h and e_h the edges among vertex h neighbors. The clustering coefficient for vertex h in a weighted network is now defined as:

$$C^w(h) =< I(h) > \times C(h) \tag{6.11}$$

$$C^w(h) =< Q(h) > \times C(h) \tag{6.12}$$

where $C(h)$ is the clustering coefficient defined in equation 6.6 for the binary contra-part of the weighted graph. The clustering coefficient for the network is now defined by the renormalization as:

$$CI \equiv< C^w >= \frac{\sum_i < I(i) > \times C(i)}{\sum_i < I(i) >} \tag{6.13}$$

$$CQ \equiv< C^w >= \frac{\sum_i < Q(i) > \times C(i)}{\sum_i < Q(i) >} \tag{6.14}$$

6.4.4 Coefficient of variability

The fluctuations to the average path length in the distance space are measured using the coefficient of variability:

$$CV_d = \frac{\sigma_l}{< l >} \tag{6.15}$$

where the l is the shortest path. The counterpart in the proximity space is:

$$CV_p = \frac{\sigma_{Stp}}{< Stp >} \tag{6.16}$$

where the Stp is the strongest path.

6.4.5 Traditional thresholds

We apply the traditional way to analyze weighted networks, which consists on applying to the edge weights of the network several thresholds. For each threshold we obtain a weighted sub-graph, which we can study either as crisp graph or as a weighted graph.

6.4.6 Semi-metric behavior

In chapter 2 we described three ways to measure the semi-metric behavior. Here we use the *semi-metric ratio* from an edge.

$$s(v_i, v_j) = \frac{d_{direct}(v_i, v_j)}{d_{shortest}(v_i, v_j)} \tag{6.17}$$

Where $d_{shortest}$ is the shortest path between any two vertices in the distance graph. $s \geq 1$ for semi-metric pairs, and $s = 1$ for metric pairs of a distance graph. Therefore we define the *semi-metric percentage* (SM) as:

$$SM = \frac{\sum_{i,j} (s(v_i, v_j) > 1)}{|E|} \tag{6.18}$$

where $|E|$ is the total number of direct edges.

For each network we study the *average shortest path fluctuations* by sweeping several values of λ for the Dombi t-norm generator. We study the properties of the metric backbone of the network, which determines measures such as average shortest path, efficiency, etc.

We apply, the semi-metric thresholding methodology, described above. In the limit of metric weighted graphs the traditional and semi-metric methodologies are equivalent, since with the semi-metric thresholding we obtain a sub-graph that contains the subgraph obtained by the traditional thresholding approach.

6.4.7 Null model

For each real network we randomize the topology of the corresponding weighted graph maintaining the same weight distribution. In this process we keep constant the number of edges, each vertex degree and the number of vertices. Our null model is an average of 11 randomizations of the original network.

6.4.8 Small-World phenomenon in weighted networks

The small-world phenomenon in weighted networks has been studied in the last years. Two main approaches have been proposed: the first one uses the traditional methodology of thresholding to produce crisp subgraphs, which are used to study, the average path length and clustering coefficient [66]; the second studies the small-world phenomenon of networks as weighted graphs analyzing the average path length and clustering coefficient compared to a null model based on randomization of the original graphs [56].

In the following sections, first we analyze how the average path length is affected by sweeping λ values, for the Dombi t-norm. This is equivalent to sweeping the various ways of computing distance closure. We first use this study to confirm that high distortion (see chapter 4) implies higher fluctuations shortest paths. Second, we determine the metric backbone for $\lambda = 1$, and study the average shortest path fluctuations. Third, we apply the traditional and semi-metric thresholding and compare the effects on the crisp sub-graph obtained, regarding the small-world phenomenon. Fourth, we apply the traditional and semi-metric thresholding, and analyze the resulting weighted sub-graphs by comparing it with our null model.

6.5 US Airport Network

6.5.1 Introduction

In the US Airport network [25] each vertex represents an airport and two vertices are joined by an edge (link) if there exists a direct airline connection between the corresponding airports. The edge weights are computed from the number of available seats on these direct connections. This undirected graph has 500 vertices. The edge weights represent the strength between the airports (proximity graph).

6.5.2 Results and Discussion

Average shortest path fluctuations

After normalization, we apply the Dombi t-norm generator for various values of its parameter and compute the metric closure of the respective distance graphs. Figure 6.5 shows the of semi-metric percentage (SM) for the US Airport network and for its randomized null model. In tables 6.1 and 6.2 we present various measures for the US Airport network and the randomized null model, respectively.

As suspected the fluctuations increase with λ. We have seen in chapter 4, that $\lambda = 1$ is the value which best preserves the characteristic properties between proximity and distances with small fluctuations for path length. The shortest path length fluctuations for $\lambda = 1$ are high $CV_d \approx 0.66$ compared

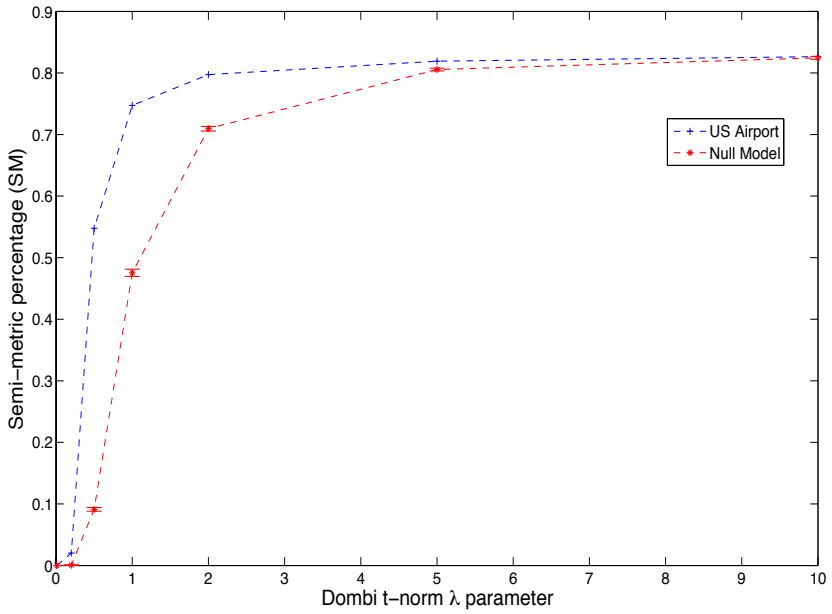

Figure 6.5: Semi-metric percentage (SM) for Dombi t-norm generator

λ	$< l >$	σ_l	CV_d	CI	CQ	C	SM
0.01	3.1	0.9	0.29	–	–	0.62	0
0.2	5.4	1.9	0.35	0.86	0.79	0.62	0.02
0.5	13.6	6.1	0.45	0.86	0.79	0.62	0.54
1	75.2	50.0	0.66	0.86	0.79	0.62	0.75
2	$3.7E3$	$3.6E3$	0.97	0.86	0.79	0.62	0.80
5	$1.0E9$	$1.9E9$	1.9	0.86	0.79	0.62	0.82
10	$2.9E18$	$1.2E19$	4.14	0.86	0.79	0.62	0.83

Table 6.1: Variation in the US Airport Network, for various values of the Dombi parameter λ. $< l >$ average path length, σ_l standard deviation of the shortest path, CV_d coefficient of variability, CI and CQ weighted clustering coefficients, C clustering coefficient (crisp), SM semi-metric percentage.

with the null model $CV_d \approx 0.30$ and as we have seen in chapter 4, $\lambda = 1$ minimizes these fluctuations in the both proximity and distance spaces given

λ	$< l >$	σ_l	CV_d	CI	CQ	C	SM
0.01	2.8	0.6	0.21	–	–	0.02	0
0.2	4.7	1.0	0.20	0.03	0.03	0.02	$6.04E-4$
0.5	9.6	2.2	0.23	0.03	0.03	0.02	0.09
1	25.6	7.6	0.30	0.03	0.03	0.02	0.48
2	154.8	91.7	0.59	0.03	0.03	0.02	0.71
5	$1.9E5$	$1.5E6$	7.24	0.03	0.03	0.02	0.81
10	$9.4E14$	$1.5E16$	12.79	0.03	0.03	0.02	0.82

Table 6.2: Variation in the null model of the US Airport Network, for various values of the Dombi parameter λ. $< l >$ average path length, σ_l standard deviation of the shortest path, CV_d coefficient of variability, CI and CQ weighted clustering coefficients, C clustering coefficient (crisp), SM semi-metric percentage.

the network real fluctuations CV_d, in this case $CV_d \approx 0.66$. Since, the real fluctuations $CV_d \approx 0.66$, in order to study the US network at $\lambda = 1$ we have to reduce the real shortest paths fluctuations, by studying sub-graphs of the original US network.

Metric backbone

As we can see from table 6.1 for $\lambda = 1$ the SM is 0.75, which means that 75% of the direct edges are semi-metric and the backbone has 25% of the network edges. Figure 6.6 and table 6.3 show us the metric graph representation and some of its properties.

From figure 6.6 we can qualitatively see that the backbone is mainly compose by hubs, bridges and peripheral vertices. All edges in this network participate on the shortest path between any two verices. We can see from this figure if we remove the hubs or bridges the graph becomes partitioned

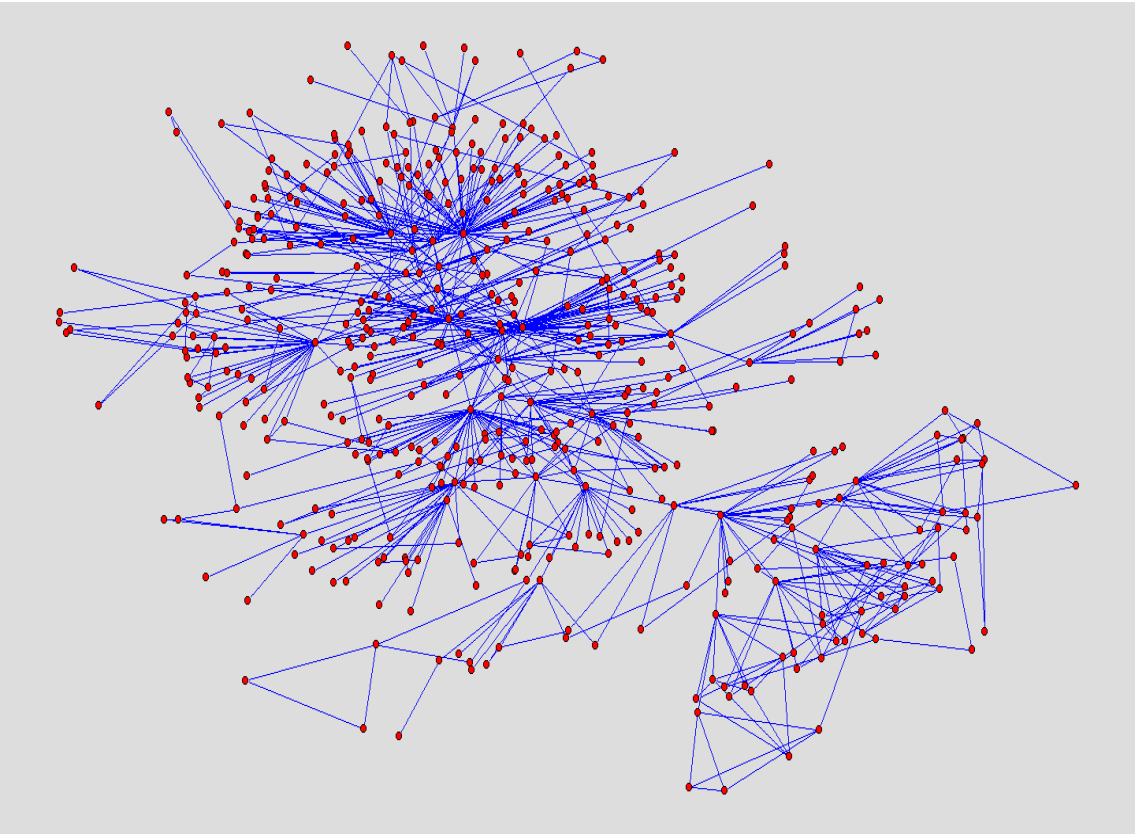

Figure 6.6: US Airport metric backbone

	λ	$< l >$	CV_d	$< Stp >$	CV_p	k_{avg}	CI	C	N	SM
USN	1	75.2	0.66	0.03	1.44	3.01	0.46	0.23	500	0
RNM	1	25.6	0.30	0.04	0.53	6.23	0.05	0.006	500	0

Table 6.3: US Airport metric backbone (USN) and Null Model (RNM), for $\lambda = 1$. $< l >$ average path length, CV_d coefficient of variability, $< Stp >$ average strongest path (Global Efficiency), CVp coefficient of variability, k_{avg} average degree, CI weighted clustering coefficient, C clustering coefficient (crisp), N number of vertices on the main component, SM semi-metric percentage.

in modules. From table 6.3 we can see that the average path length remains intact compared with the values obtain in tables 6.1 and 6.2 for $\lambda = 1$, as

expected. However, the clustering coefficients has been reduced considerably for the US Airport metric backbone, showing that semi-metric edges affect mainly the clustering coefficients.

Next, we compare the traditional and semi-metric (SMT) thresholding applied to weighted graphs.

Traditional versus Semi-metric thresholding

In figure 6.7 we can see the number of vertices removed with the traditional and the semi-metric thresholding. It is noteworthy that the semi-metric thresholding preserves more edges in the network. Even for small thresholds the traditional thresholding removes more than 90% of the edges. In figure 6.8 we can see that for threshold values approximately larger than 0.02 we start affecting the bridges of the backbone, and the graph starts being partitioned. Since, the removal of semi-metric edges do not affect bridges, with the semi-metric thresholding we obtain the same modularization as in the traditional thresholding, figure 6.7.

In tables 6.4 and 6.5 we show some properties of the crisp sub-graphs with traditional and semi-metric thresholding, respectively. We then study various properties of the main component obtained. We observe the main differences on the clustering coefficient and average degree as expected. In the case of semi-metric thresholding we observe an increase on the average degree. This is observed because the main component decreases in size. We also observe minor differences in the average shortest path because we are

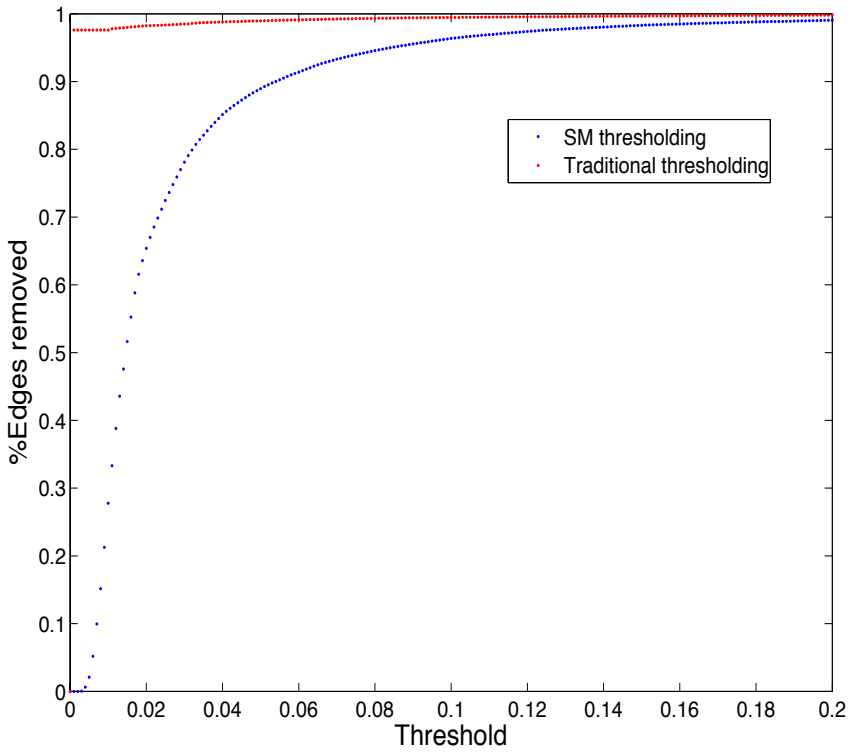

Figure 6.7: Traditional versus Semi-metric thresholding, which naturally coincide, for the US Airport network.

treating the sub-graphs as crisp and some semi-metric edges become metric.

The average shortest path is of the same order of magnitude of $log(N)$ and the clustering coefficients are high for both traditional and semi-metric thresholding, therefore the crisp sub-graphs obtained by thresholding are considered small-worlds. However, this is more pronounced on the semi-metric thresholding, since the clustering coefficients are considerably higher.

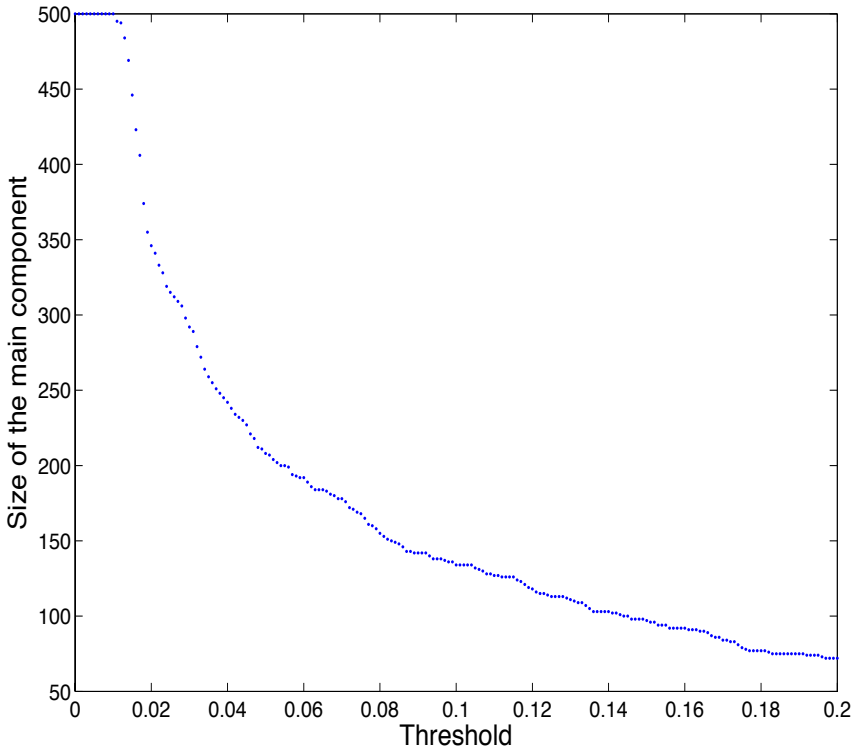

Figure 6.8: Main component size using traditional and semi-metric thresholding for the US Airport network.

Weighted network analysis

Here we study the network considering its weights, we do not turn the sub-graphs into crisp graphs. We apply the traditional and semi-metric thresholding. Table 6.6 show us the properties for both methods. Table 6.7 show the same properties for the Null Model.

We can see in these tables that for thresholds bigger than 0.04 the traditional and semi-metric sub-graphs are small-worlds, since they have average

th	$< l >$	CV_d	$< Stp >$	CV_p	C	k_{avg}	N	SM
0.02	2.6	0.28	0.10	1.05	0.53	12.8	346	0
0.04	2.5	0.27	0.13	0.86	0.57	12.5	242	0
0.06	2.5	0.28	0.16	0.73	0.50	11.5	192	0
0.1	2.4	0.31	0.21	0.58	0.45	10.1	134	0
0.15	2.4	0.34	0.27	0.48	0.50	8.6	97	0

Table 6.4: Results for the US Airport main component sub-networks (USN) with the traditional thresholding. SMT semi-metric threshold, $< l >$ average path length, CV_d coefficient of variability, $< Stp >$ average strongest path, CV_p coefficient of variability, C clustering coefficient (crisp), k_{avg} average degree, N number of vertices, SM semi-metric percentage.

SMT	$< l >$	CV_d	$< Stp >$	CV_p	C	k_{avg}	N	SM
0.02	2.6	0.28	0.09	1.16	0.62	15.1	346	0
0.04	2.3	0.27	0.09	1.11	0.73	19.4	242	0
0.06	2.1	0.29	0.10	1.08	0.76	22.5	192	0
0.1	1.9	0.30	0.11	1.01	0.76	26.3	134	0
0.15	1.9	0.34	0.14	0.91	0.70	24.7	97	0

Table 6.5: Results for the US Airport main component sub-networks (USN) with the SMT. SMT semi-metric threshold, $< l >$ average path length, CV_d coefficient of variability, $< Stp >$ average strongest path, CV_p coefficient of variability, C clustering coefficient (crisp), k_{avg} average degree, N number of vertices, SM semi-metric percentage.

shortest paths at the same order of magnitude of the Null Model and high clustering coefficient. From table 6.6 the semi-metric thresholding preserves better the degree , clustering coefficient and semi-metric percentage than the traditional thresholding, as expected.

th	$< l >$	CV_d	$< Stp >$	CV_p	CI_T	CI_{SM}	k_T	k_{SM}	N	SM_T	SM_{SM}
0.02	37.5	0.60	0.04	1.12	0.89	0.92	12.8	15.1	346	0.78	0.82
0.04	20.3	0.51	0.07	0.88	0.91	0.95	12.5	19.4	242	0.76	0.85
0.06	14.5	0.50	0.09	0.77	0.89	0.94	11.5	22.5	192	0.73	0.86
0.1	8.9	0.46	0.13	0.62	0.76	0.96	10.1	26.3	134	0.65	0.87
0.15	6.2	0.47	0.17	0.53	0.78	0.93	8.6	24.7	97	0.56	0.85

Table 6.6: Results for the US Airport sub-network (USN) with metric distance closure ($\lambda = 1$). SMT semi-metric threshold, $< l >$ average path length, CV_d coefficient of variability, $< Stp >$ average strongest path, CV_p coefficient of variability, CI_T and CI_{SM} the clustering coefficient for traditional and semi-metric thresholding, k_T and k_{SM} are the average degree for the traditional and semi-metric thresholding, N number of vertices, SM_T and SM_{SM} semi-metric percentage for traditional and semi-metric thresholding.

th	$< l >$	CV_d	$< Stp >$	CV_p	CI_T	CI_{SM}	k_T	k_{SM}	N	SM_T	SM_{SM}
0.02	17.6	0.30	0.06	0.49	0.03	0.04	12.8	15.1	346	0.46	0.54
0.04	12.7	0.30	0.08	0.47	0.06	0.07	12.5	19.4	242	0.39	0.61
0.06	10.2	0.29	0.10	0.45	0.05	0.11	11.5	22.5	192	0.32	0.65
0.1	7.5	0.30	0.13	0.43	0.07	0.19	10.1	26.3	134	0.19	0.68
0.15	5.7	0.32	0.17	0.41	0.08	0.25	8.6	24.7	97	0.13	0.66

Table 6.7: Results for the null model (RNM) with metric distance closure ($\lambda = 1$). SMT semi-metric threshold, $< l >$ average path length, CV_d coefficient of variability, $< Stp >$ average strongest path, CV_p coefficient of variability, CI_T and CI_{SM} the clustering coefficient for traditional and semi-metric thresholding, k_T and k_{SM} are the average degree for the traditional and semi-metric thresholding, N number of vertices, SM_T and SM_{SM} semi-metric percentage for traditional and semi-metric thresholding.

6.6 Structural Human Cerebral Cortex Network

6.6.1 Introduction

The Human Cerebral Cortex Network is a weighted network with 66 vertices. The vertices represent anatomical regions and the edges the interaction

(strength) between these regional anatomical areas in the Human Cerebral Cortex. The strength of these connections are obtain using high-resolution T1 weighted and diffusion spectrum MRI (DSI), which estimates the axonal trajectories across regions of interest (ROI's), e.g. anatomical regions. This gives an approximation of the large scale human cortex structural network (connectome). A more deep description in how this network was build is in [34]. We study the network obtained by averaging five subjects networks.

6.6.2 Results and Discussion

Average shortest path fluctuations

After normalization, we apply the Dombi t-norm generator for several parameters and apply the APSP Dijkstra algorithm to calculate the shortest path distribution. Figure 6.9 shows the semi-metric percentage (SM) for the Human Cerebral Cortex Network and for randomized weighted random network.

In tables 6.8 and 6.9 we present the results for the Human Cerebral Cortex network and the randomized network, respectively.

In tables 6.8 and 6.9 we can confirm when we increase λ the fluctuations increase. The high values of $CV_d \approx 0.91$ for $\lambda = 1$ conduct us to study sub-networks of the Human Cortex network.

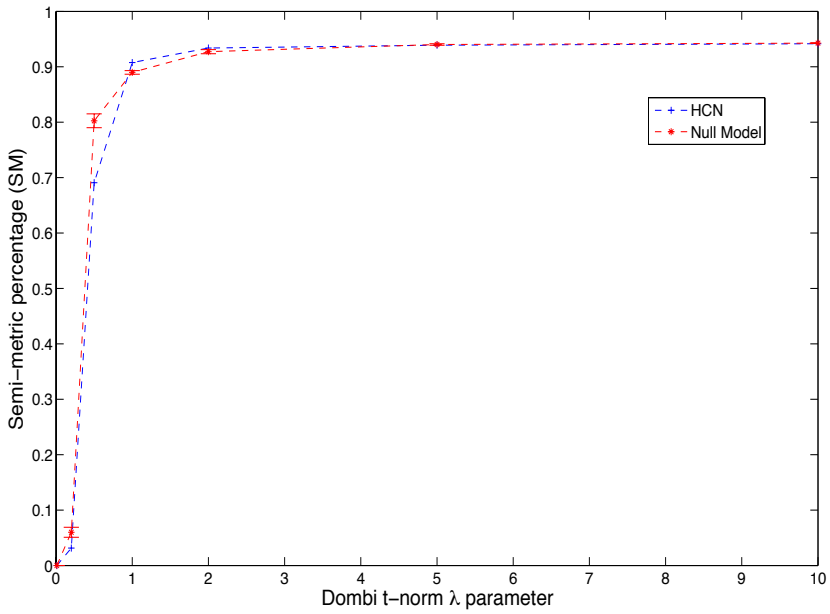

Figure 6.9: Semi-metric percentage (SM) for Dombi t-norm generator

λ	$< l >$	σ_l	CV	CI	CQ	C	SM
0.01	1.5	0.5	0.33	–	–	0.72	0
0.2	3.0	0.9	0.30	0.73	0.73	0.72	0.03
0.5	8.0	3.5	0.44	0.73	0.73	0.72	0.69
1	21.5	19.6	0.91	0.73	0.73	0.72	0.91
2	288.7	699.2	2.42	0.73	0.73	0.72	0.93
5	$1.9E7$	$9.7E7$	5.11	0.73	0.73	0.72	0.94
10	$9.5E15$	$5.4E16$	5.68	0.73	0.73	0.72	0.94

Table 6.8: Variation in the Human Cortex Network, for various values of the Dombi parameter λ. $< l >$ average path length, σ_l standard deviation of shortest path, CV coefficient of variability, CI and CQ weighted clustering coefficients, C clustering coefficient (crisp), SM semi-metric percentage.

Metric backbone

As we can see from table 6.8 for $\lambda = 1$ the SM is 0.91, which means that 91% of the direct edges are semi-metric and the backbone has 9% of the network

λ	$< l >$	σ_l	CV	CI	CQ	C	SM
0.01	1.5	0.5	0.33	–	–	0.54	0
0.2	2.7	0.6	0.24	0.53	0.53	0.53	0.06
0.5	5.5	1.8	0.33	0.53	0.53	0.53	0.80
1	11.9	6.2	0.52	0.53	0.53	0.53	0.89
2	65.0	112.0	1.57	0.53	0.53	0.53	0.93
5	$1.5E5$	$7.8E5$	3.43	0.53	0.53	0.53	0.94
10	$2.6E11$	$8.8E11$	4.50	0.53	0.53	0.53	0.94

Table 6.9: Variation in the null model of the Human Cortex Network, for various values of the Dombi parameter λ. $< l >$ average path length, σ_l standard deviation of shortest path, CV coefficient of variability, CI and CQ weighted clustering coefficients, C clustering coefficient (crisp), SM semi-metric percentage.

edges. Figure 6.10 and table 6.10 show us the metric graph representation and some of its properties.

	λ	$< l >$	CV_d	$< Stp >$	CV_p	k_{avg}	CI	C	N	SM
HCN	1	21.5	0.91	0.08	1.10	3.21	0.31	0.13	66	0
RNM	1	11.9	0.46	0.10	0.78	3.86	0.11	0.02	66	0

Table 6.10: Human Cortex metric backbone (HCN) and Null Model (RNM), for $\lambda = 1$. $< l >$ average path length, CV_d coefficient of variability, $< Stp >$ average strongest path (Global Efficiency), CVp coefficient of variability, k_{avg} average degree, CI weighted clustering coefficient, C clustering coefficient (crisp), N number of vertices on the main component, SM semi-metric percentage.

From figure 6.10 we can see that the backbone has some modular structure and contain mainly bridges between these modules. All edges in this network participate on the shortest path between any two vertices. We can see from this figure if we remove some of these bridges the graph becomes partitioned in modules. From table 6.10 we can see that the average path length remains

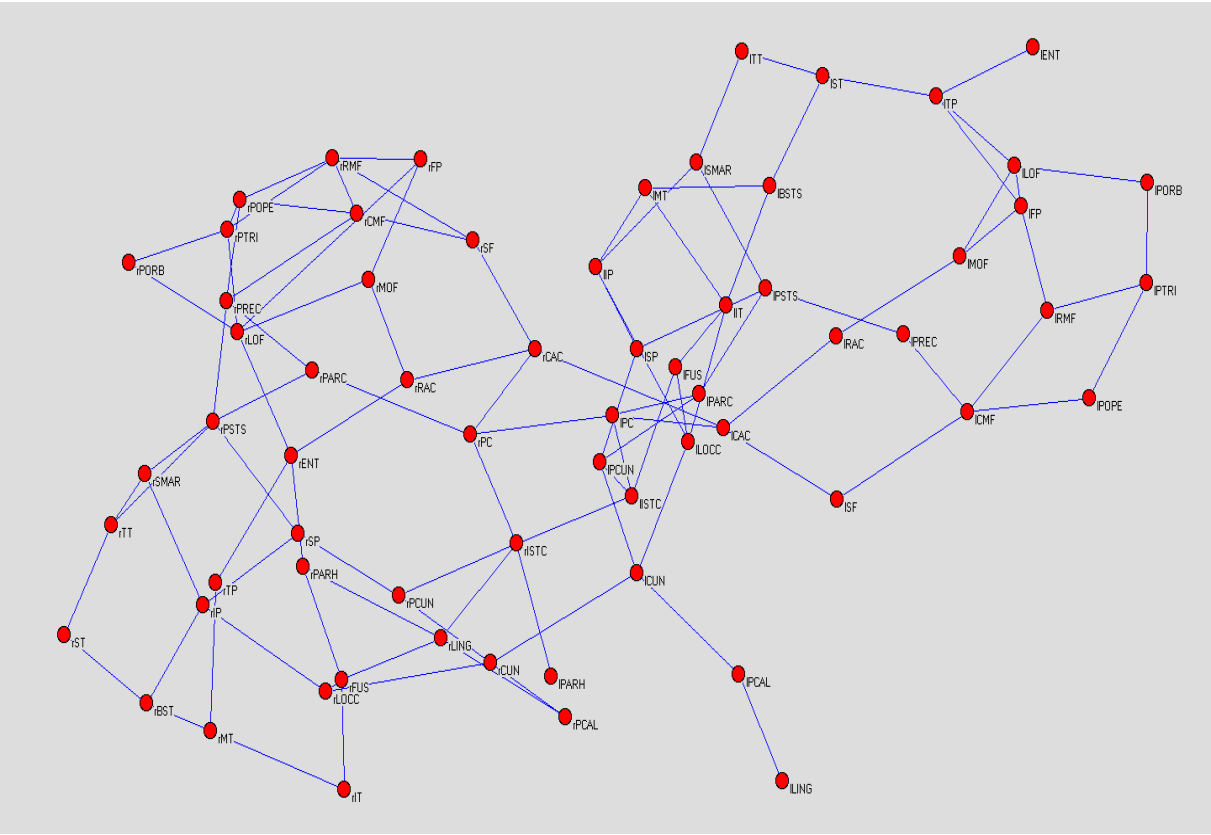

Figure 6.10: Human Cortex metric backbone

intact compared with the values obtain in table 6.8 for $\lambda = 1$, as expected. However, the clustering coefficients reduced considerably.

Next, we compare the traditional and semi-metric (SMT) thresholding applied to weighted graphs.

Traditional versus Semi-metric thresholding

In figure 6.11 we can see the number of edges removed with the traditional and the semi-metric thresholding. It is noteworthy that the semi-metric

thresholding preserves more edges in the network. Even for small thresholds the traditional thresholding removes more than 80% of the edges. In figure 6.12 we can see how the threshold partitioned the network. The partitioning comes in blocks with a big jump around $th = 0.15$. This block partitioning can be explained by the modular structure observed in figure 6.10 (backbone).

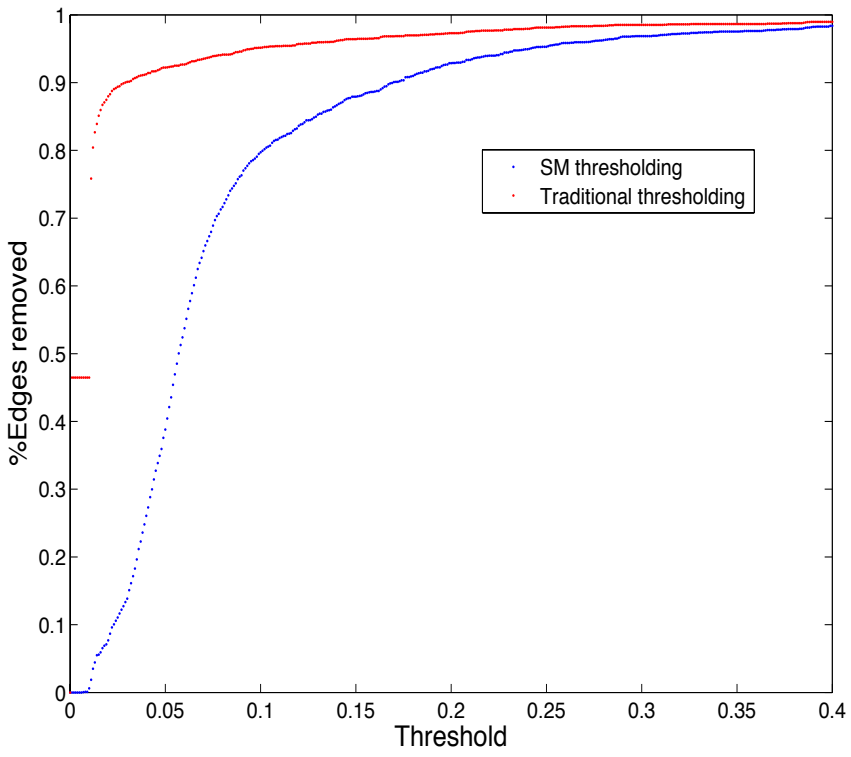

Figure 6.11: Traditional versus Semi-metric thresholding, which naturally coincide, for the Human Cortex network.

In tables 6.11 and 6.12 we show some properties of the crisp sub-graphs after thresholding with traditional and semi-metric, respectively.

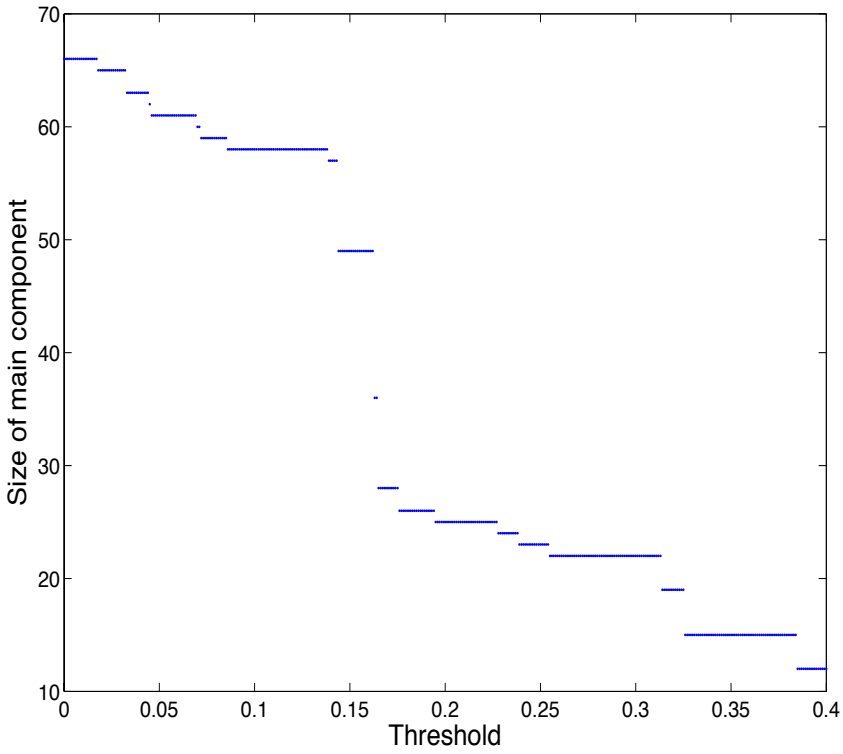

Figure 6.12: Main component size using traditional and semi-metric thresholding for the Human Cortex network.

We observe the main differences on the clustering coefficient and average degree as expected. The differences in the average shortest path are observed because we are treating the sub-graphs as crisp. As we will see below when we treat the sub-graphs as weighted there are no differences on average shortest path.

The average shortest path is of the same order of magnitude of $log(N)$ and the clustering coefficients are high for traditional thresholding for $th <$

th	$< l >$	CV_d	$< Stp >$	CV_p	C	k_{avg}	N	SM
0.025	3.1	0.43	0.02	4.41	0.46	7.02	65	0
0.05	3.5	0.42	0.02	4.39	0.43	5.48	61	0
0.1	4.9	0.46	0.02	4.77	0.32	3.59	58	0
0.15	6.3	0.56	0.02	4.70	0.19	2.86	49	0
0.02	3.8	0.52	0.05	3.16	0.03	2.72	25	0

Table 6.11: Results for the Human Cortex sub-network (HCN) with the traditional thresholding. SMT semi-metric threshold, $< l >$ average path length, CV_d coefficient of variability, $< Stp >$ average strongest path, CV_p coefficient of variability, C clustering coefficient (crisp), k_{avg} average degree, N number of vertices, SM semi-metric percentage.

SMT	$< l >$	CV_d	$< Stp >$	CV_p	C	k_{avg}	N	SM
0.025	1.5	0.39	0.02	3.43	0.70	33.78	65	0
0.05	1.5	0.34	0.02	3.42	0.69	28.39	61	0
0.1	2.5	0.46	0.02	3.92	0.72	13.07	58	0
0.15	3.9	0.62	0.03	3.91	0.56	8.90	49	0
0.02	1.9	0.42	0.06	2.59	0.71	8.72	25	0

Table 6.12: Results for the Human Cortex sub-network (HCN) with the semi-metric thresholding. SMT semi-metric threshold, $< l >$ average path length, CV_d coefficient of variability, $< Stp >$ average strongest path, CV_p coefficient of variability, C clustering coefficient (crisp), k_{avg} average degree, N number of vertices, SM semi-metric percentage.

0.05 and low for $th > 0.05$. The crisp sub-graphs obtained by traditional thresholding are considered small-worlds for $th < 0.05$ and not for $th > 0.05$. However, for semi-metric thresholding we can can consider the small-world for all range of thresholds.

Weighted network analysis

Here we study the network considering its weights, we do not turn the sub-graphs into crisp graphs. We apply the traditional and semi-metric thresholding. Table 6.13 show us the properties for both methods. Table 6.14 show the same properties for the Null Model.

th	$< l >$	CV_d	$< Stp >$	CV_p	CI_T	CI_{SM}	k_T	k_{SM}	N	SM_T	SM_{SM}
0.025	19.3	0.78	0.09	1.07	0.49	0.72	7.0	33.8	65	0.56	0.91
0.05	15.0	0.51	0.09	1.00	0.41	0.70	5.5	28.4	61	0.44	0.89
0.1	15.1	0.55	0.10	1.01	0.35	0.68	3.6	13.1	58	0.24	0.79
0.15	15.5	0.67	0.11	1.04	0.33	0.46	2.9	8.9	49	0.16	0.73
0.2	4.7	0.53	0.23	0.65	0.21	0.77	2.7	8.7	25	0.15	0.73

Table 6.13: Results for the Human Cortex sub-network (HCN) with metric distance closure ($\lambda = 1$). SMT semi-metric threshold, $< l >$ average path length, CV_d coefficient of variability, $< Stp >$ average strongest path, CV_p coefficient of variability, CI_T and CI_{SM} the clustering coefficient for traditional and semi-metric thresholding, k_T and k_{SM} are the average degree for the traditional and semi-metric thresholding, N number of vertices, SM_T and SM_{SM} semi-metric percentage for traditional and semi-metric thresholding.

We can see in these tables the weighted sub-graphs obtained by traditional thresholding are considered small-worlds for $th < 0.05$ and not for $th > 0.05$. However, for semi-metric thresholding we can consider the small-world for all range of thresholds.

th	$< l >$	CV_d	$< Stp >$	CV_p	CI_T	CI_{SM}	k_T	k_{SM}	N	SM_T	SM_{SM}
0.025	11.4	0.48	0.11	0.79	0.11	0.52	7.0	33.8	65	0.46	0.89
0.05	10.7	0.52	0.11	0.76	0.11	0.48	5.5	28.4	61	0.32	0.87
0.1	10.3	0.53	0.12	0.76	0.17	0.21	3.6	13.1	58	0.09	0.72
0.15	9.0	0.70	0.14	0.74	0.22	0.17	2.9	8.9	49	0.06	0.65
0.2	4.0	0.67	0.22	0.75	0.20	0.33	2.7	8.7	25	0.09	0.70

Table 6.14: Results for the null model (RNM) with metric distance closure ($\lambda = 1$). SMT semi-metric threshold, $< l >$ average path length, CV_d coefficient of variability, $< Stp >$ average strongest path, CV_p coefficient of variability, CI_T and CI_{SM} the clustering coefficient for traditional and semi-metric thresholding, k_T and k_{SM} are the average degree for the traditional and semi-metric thresholding, N number of vertices, SM_T and SM_{SM} semi-metric percentage for traditional and semi-metric thresholding.

6.7 Functional Human Brain Network

6.7.1 Introduction

The Functional Human Brain Network is a weighted network with 116 vertices. It was acquired in resting-state the fMRI images for 40 healthy subjects during 420000 ms at a TR=2000 ms. All participants were scanned using the same Siemens 3T Tim Trio Scanner at the Medical Research Council Cognition and Brain Sciences Unit, Cambridge, UK. Functional images were acquired with a gradient echo planar imaging sequence with the following parameters: repetition time TR= 2000 ms, echo time TE= 30 ms, voxel size $= 3 \times 3 \times 3$ mm, for 32 brain slices. A more detailed description of the acquisition of fMRI data can be found in [85]. The vertices represent anatomical regions (ROI's) and the edges the Maximal Overlap Discrete Wavelet Transform (MODWT) correlation between ROI's time series. We study the

network obtained by averaging the 40 subjects networks at wavelet scale 1. Scale 1 frequencies range is $[0.125, 0.25]$ Hz.

6.7.2 Results and Discussion

Average shortest path fluctuations

After normalization, we apply the Dombi t-norm generator for several parameters and apply the APSP Dijkstra algorithm to calculate the shortest path distribution. Figure 6.13 shows the semi-metric percentage (SM) for the Functional Human Brain Network and for randomized weighted random network.

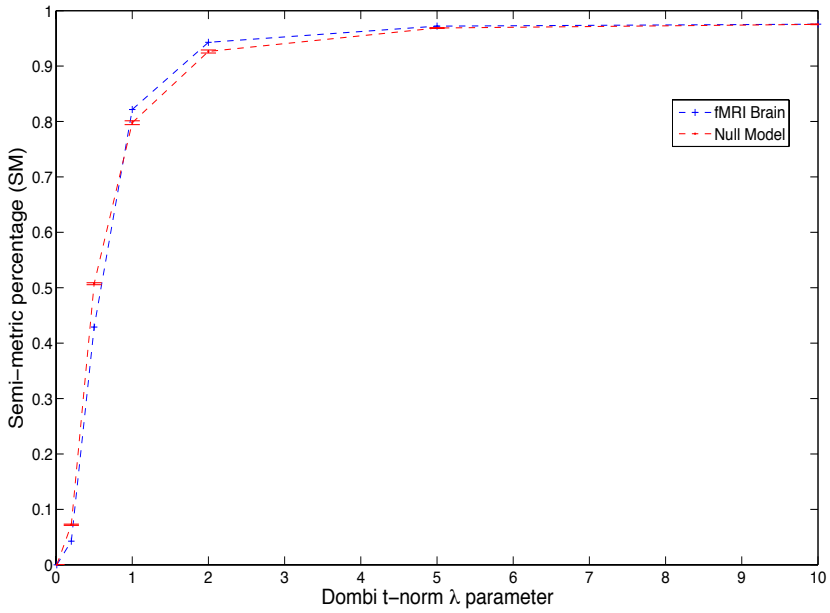

Figure 6.13: Semi-metric percentage (SM) for Dombi t-norm generator

In tables 6.15 and 6.16 we present the results for the Functional Human Brain network and the randomized network, respectively.

λ	$< l >$	σ_l	CV	CI	CQ	C	SM
0.01	1.3	0.4	0.34	–	–	0.85	0
0.2	1.6	0.5	0.33	0.84	0.83	0.85	0.04
0.5	2.3	0.8	0.37	0.84	0.83	0.85	0.39
1	2.9	1.3	0.46	0.84	0.83	0.85	0.81
2	3.2	2.2	0.69	0.84	0.83	0.85	0.94
5	7.5	20.1	2.77	0.84	0.83	0.85	0.97
10	422.9	$3.0E3$	7.14	0.84	0.83	0.85	0.98

Table 6.15: Variation in the Functional Human Brain Network, for various values of the Dombi parameter λ. $< l >$ average path length, σ standard deviation of shortest path, CV coefficient of variability, CI and CQ weighted clustering coefficients, C clustering coefficient (crisp), SM semi-metric percentage.

λ	$< l >$	σ_l	CV	CI	CQ	C	SM
0.01	1.3	0.4	0.34	–	–	0.76	0
0.2	1.5	0.4	0.24	0.76	0.76	0.76	0.07
0.5	1.8	0.4	0.22	0.76	0.76	0.76	0.52
1	1.7	0.4	0.26	0.76	0.76	0.76	0.81
2	1.2	0.4	0.35	0.76	0.76	0.76	0.93
5	0.3	0.2	0.76	0.76	0.76	0.76	0.97
10	0.04	0.07	1.83	0.76	0.76	0.76	0.97

Table 6.16: Variation in the null model of the Functional Human Brain Network, for various values of the Dombi parameter λ. $< l >$ average path length, σ standard deviation of shortest path, CV coefficient of variability, CI and CQ weighted clustering coefficients, C clustering coefficient (crisp), SM semi-metric percentage.

In tables 6.15 and 6.16 we can confirm when we increase λ the fluctuations increase.

In table 6.16 for $\lambda > 1$ the null model changes the tendency of increasing the average path length with λ. The reason for this type of behavior can be explained by its semi-metric values and weights distribution. We have seen that the average path length only depends on the metric backbone. For $\lambda > 1$ the semi-metricity is bigger than 94%, which means only 6% of the edges contribute for average shortest path, the ones in metric sub-graph. We have $5,071$ edges in the Functional Human Brain network, which means we have $0.06 \times 5,071 \approx 304$ edges that contribute to the average shortest path. From the weights distribution, we have 302 edges out of $5,071$ with weights bigger than 0.5, which means in the distance graph for this weights we have distances between 0 and 1, and for $\lambda > 1$ are considerable smaller than for $\lambda \leq 1$. With the randomization of the network we get a high probability of having some of these edges on the backbone, decreasing the average shortest path. Moreover, for $\lambda > 1$ the metric closure behaves as the ultra-metric closure (max,min), and the shortest path will be equal to max of the minimum distance weight which is very small for weights bigger than 0.5. This explains how the null model gets small average shortest path values for $\lambda > 1$.

Metric backbone

As we can see from table 6.15 for $\lambda = 1$ the SM is 0.81, which means that 81% of the direct edges are semi-metric and the backbone has 19% of the network edges. Figure 6.14 and table 6.10 show us the metric graph representation

and some of its properties.

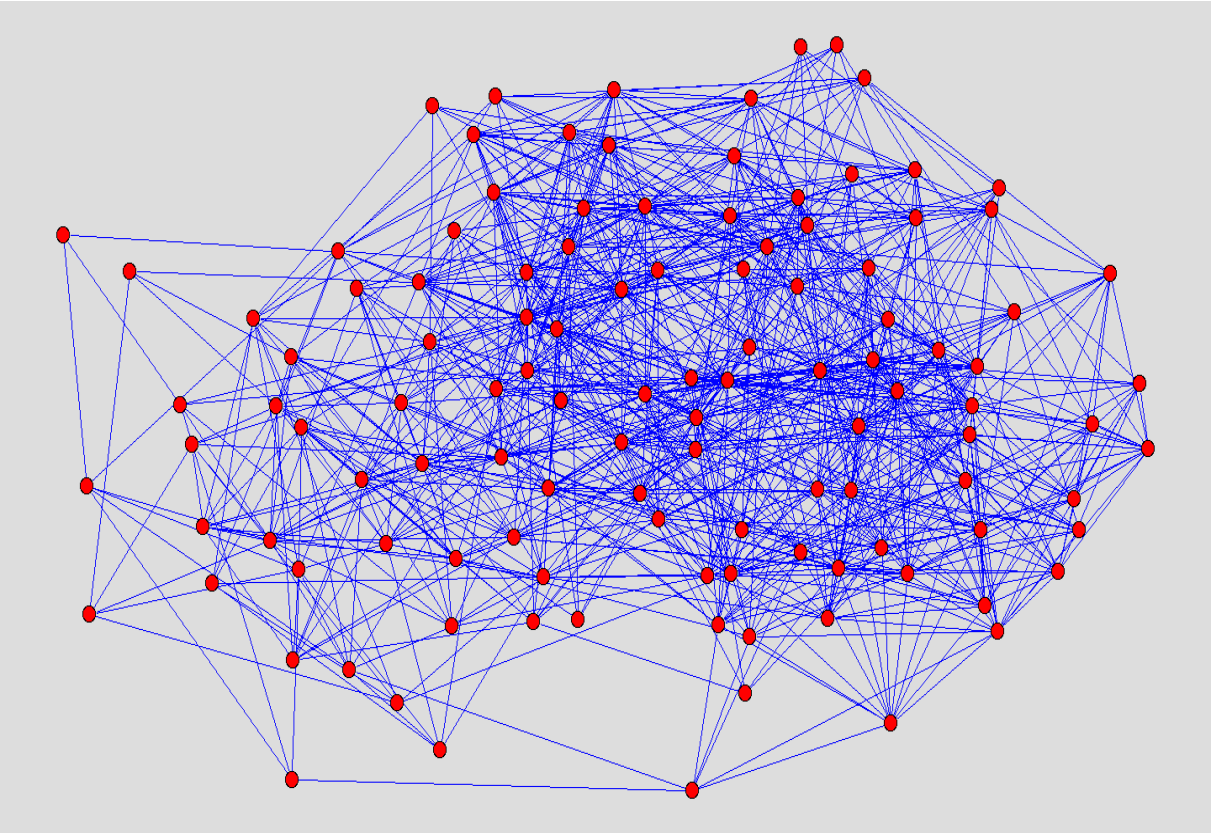

Figure 6.14: Functional Human Brain metric backbone

	λ	$< l >$	CV_d	$< Stp >$	CV_p	k_{avg}	CI	C	N	SM
HCN	1	2.9	0.46	0.29	0.38	16.8	0.34	0.38	116	0
RNM	1	1.7	0.26	0.41	0.19	17.9	0.13	0.14	116	0

Table 6.17: Functional Human Brain metric backbone (HCN) and Null Model (RNM), for $\lambda = 1$. $< l >$ average path length, CV_d coefficient of variability, $< Stp >$ average strongest path (Global Efficiency), CVp coefficient of variability, k_{avg} average degree, CI weighted clustering coefficient, C clustering coefficient (crisp), N number of vertices on the main component, SM semi-metric percentage.

From figure 6.14 we can see that the metric backbone has some modular structure and contains hubs and bridges between these modules. All edges in this network participate on the shortest path between any two vertices. We can see from this figure if we remove some of these bridges the graph becomes partitioned in modules. From table 6.17 we can see that the average path length remains intact compared with the values obtain in table 6.15 for $\lambda = 1$, as expected. However, the clustering coefficients reduced considerably.

Next, we compare the traditional and semi-metric (SMT) thresholding applied to weighted graphs.

Traditional versus Semi-metric methodology

In figure 6.15 we can see the number of edges removed with the traditional and the semi-metric thresholding. In figure 6.16 shows how the threshold partition the network. The partitioning comes in blocks with a big jump between $0.4 < th < 0.7$. This block partitioning can be explained by the modular structure observed in figure 6.14 (backbone).

This network can be considered as small-world for $\lambda = 1$, since the values of $CV_d \approx 0.46$ and $CV_p \approx 0.38$ are small. For the propose of characterizing the network as small-world we do not need to further inspect sub-graphs (thresholding). If we intend to characterize other measures we can apply the semi-metric thresholding, however in this thesis we do not proceed with that analysis.

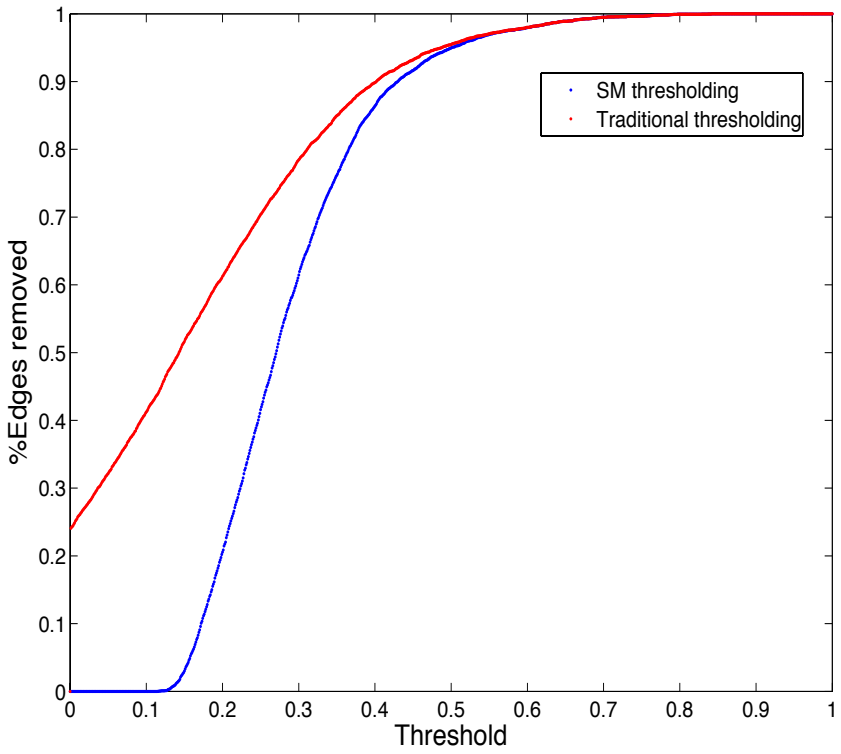

Figure 6.15: Traditional versus Semi-metric thresholding for the Functional Human Brain network.

6.8 Scientific Collaboration Network

6.8.1 Introduction

The structure and connectivity of the Scientific Collaborative Network has been deduced from co-authorship of scientists in a single paper, which a link between two scientists is established by their coauthorship of one or more scientific papers [66]. Thus the groups to which scientists belong in

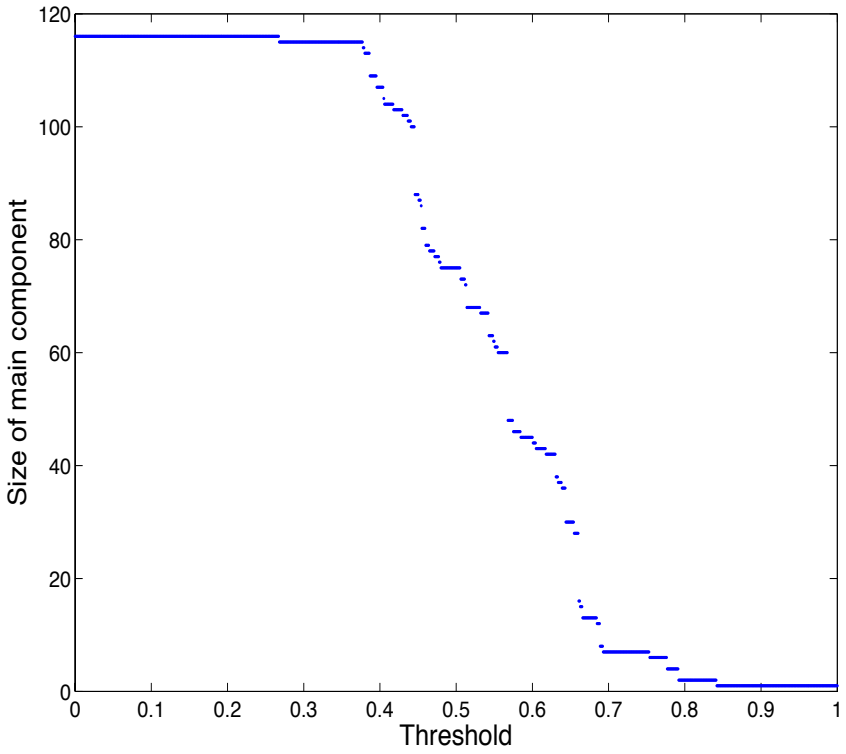

Figure 6.16: Main component size using traditional and semi-metric thresholding, which naturally coincide, for the Functional Human Brain network.

this network are the groups of coauthors of a single paper. This network is undirected and has a total of $15,179$ vertices and $86,022$ undirected edges.

6.8.2 Results and Discussion

After normalization, we apply the Dombi t-norm generator for several parameters and apply the metric closure to calculate the main components. We have one strong component of $12,722$ vertices and 735 small components of

2 vertices. We analyze only the strong component. Figure 6.17 shows the semi-metric percentage (SM) for the Scientific Collaboration Network and for randomized weighted random network (null model).

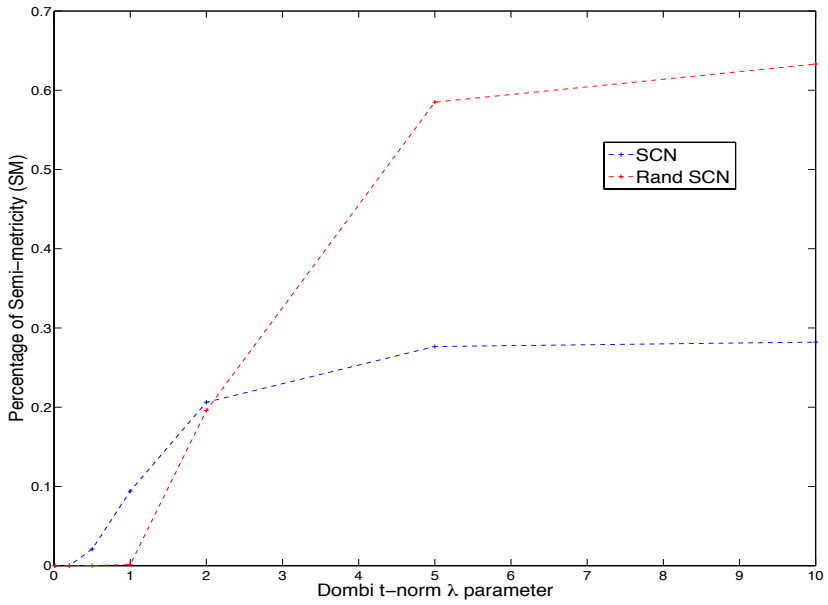

Figure 6.17: Percentage of semi-metricity for Dombi t-norm generator

In tables 6.18 and 6.19 we present the results for the Scientific Collaboration network and the randomized network, respectively.

The network is 91% metric, which means the metric backbone is the network itself. Moreover, the average shortest path is at the same order of magnitude of the null model and the $CV = 0.32$ and $CV_p = 0.44$ for the actual network with $\lambda = 1$, we can conclude that the network can be considered a small-world. Therefore, this network has no need for further

λ	$< l >$	σ_l	CV	CI	CQ	C	SM
0.01	7.1	1.7	0.24	–	–	0.65	0
0.2	13.2	3.1	0.23	0.69	0.70	0.65	$2.5E-5$
0.5	35.6	8.9	0.25	0.69	0.70	0.65	0.02
1	183.3	58.7	0.32	0.69	0.70	0.65	0.09
2	$5.5E3$	$3.3E3$	0.60	0.69	0.70	0.65	0.21
5	$7.6E8$	$1.2E9$	1.58	0.69	0.70	0.65	0.28
10	$1.6E18$	$5.5E18$	3.44	0.69	0.70	0.65	0.28

Table 6.18: Variation in the Scientific Collaboration Network, for various values of the Dombi parameter λ. $< l >$ average path length, σ_l standard deviation of shortest path, CV coefficient of variability, CI and CQ weighted clustering coefficients, C clustering coefficient (crisp), SM semi-metric percentage.

λ	$< l >$	σ_l	CV	CI	CQ	C	SM
0.01	5.5	0.9	0.16	–	–	$5E-4$	0
0.2	10.8	1.7	0.16	0.04	0.04	0.07	0
0.5	31.2	5.0	0.16	0.04	0.04	0.07	$1.0E-4$
1	174.4	33.1	0.19	0.04	0.04	0.07	$1.4E-3$
2	$4.8E3$	$1.5E3$	0.31	0.04	0.04	0.07	0.20
5	$1.4E8$	$3.2E8$	2.29	0.04	0.04	0.07	0.58
10	$1.0E17$	$1.2E18$	12.00	0.04	0.04	0.07	0.63

Table 6.19: Variation in the null model of the Scientific Collaboration Network, for various values of the Dombi parameter λ. $< l >$ average path length, σ_l standard deviation of shortest path, CV coefficient of variability, CI and CQ weighted clustering coefficients, C clustering coefficient (crisp), SM semi-metric percentage.

investigation since for $\lambda = 1$ the average path length is characteristic and we would take advantage with the semi-metric thresholding.

6.9 Astrophysics Collaborations Network

6.9.1 Introduction

The structure and connectivity of the Astrophysics Collaborations Network has been deduced from co-authorship of scientists in a single paper, which a link between two scientists is established by their coauthorship of one or more scientific papers [66]. Thus the groups to which scientists belong in this network are the groups of coauthors of a single paper. This network is undirected and has a total of $16,706$ vertices and $121,251$ undirected edges.

6.9.2 Results and Discussion

After normalization, we apply the Dombi t-norm generator for several parameters and apply the metric closure to calculate the main components. We have one strong component of $14,845$ vertices and many small components of 2 vertices. We analyze only the strong component. Figure 6.18 shows the semi-metric percentage (SM) for the Astrophysics Collaborations Network and for randomized weighted random network (null model).

In tables 6.20 and 6.21 we present the results for the Astrophysics Collaborations network and the randomized network, respectively.

The network is 80% metric, which means the metric backbone is the network itself. Moreover, the average shortest path is at the same order of magnitude of the null model and the $CV = 0.35$ and $CV_p = 0.43$ for the actual network with $\lambda = 1$, we can conclude that the network can be

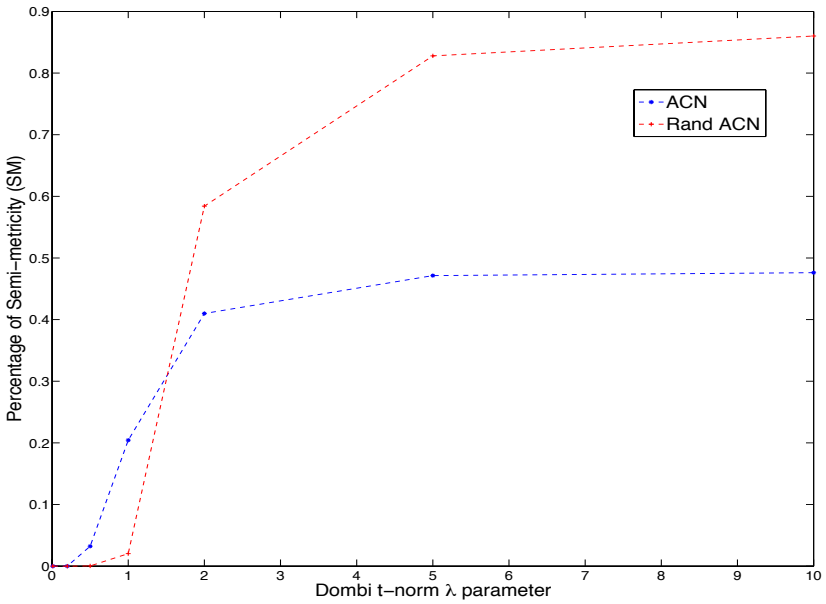

Figure 6.18: Percentage of semi-metricity for Dombi t-norm generator

λ	$< l >$	σ_l	CV	CI	CQ	C	SM
0.01	5.0	0.4	0.08	–	–	0.70	0
0.2	9.5	2.3	0.24	0.71	0.73	0.70	$2.5E-5$
0.5	26.1	6.4	0.25	0.71	0.73	0.70	0.03
1	132.4	45.8	0.35	0.71	0.73	0.70	0.20
2	$4.2E3$	$3.3E3$	0.79	0.71	0.73	0.70	0.41
5	$1.1E9$	$2.2E9$	2.00	0.71	0.73	0.70	0.47
10	$5.2E18$	$1.5E19$	2.88	0.71	0.73	0.70	0.48

Table 6.20: Variation in the Astrophysics Collaborations Network, for various values of the Dombi parameter λ. $< l >$ average path length, σ_l standard deviation of shortest path, CV coefficient of variability, CI and CQ weighted clustering coefficients, C clustering coefficient (crisp), SM semi-metric percentage.

λ	$< l >$	σ_l	CV	CI	CQ	C	SM
0.01	3.9	0.5	0.13	–	–	$1E-3$	0
0.2	7.8	1.0	0.13	$7E-3$	$7E-3$	$1E-3$	0
0.5	23.3	3.3	0.14	$7E-3$	$7E-3$	$1E-3$	$3.7E-4$
1	125.2	21.1	0.17	$7E-3$	$7E-3$	$1E-3$	0.02
2	$2.7E3$	$7.3E2$	0.27	$7E-3$	$7E-3$	$1E-3$	0.58
5	$2.9E7$	$8.2E7$	2.83	$7E-3$	$7E-3$	$1E-3$	0.83
10	$6.7E15$	$2.5E17$	36.7	$7E-3$	$7E-3$	$1E-3$	0.86

Table 6.21: Variation in the null model of the Astrophysics Collaborations Network, for various values of the Dombi parameter λ. $< l >$ average path length, σ_l standard deviation of shortest path, CV coefficient of variability, CI and CQ weighted clustering coefficients, C clustering coefficient (crisp), SM semi-metric percentage.

considered a small-world. Therefore, this network has no need for further investigation since for $\lambda = 1$ the average path length is characteristic and we would take advantage with the semi-metric thresholding.

6.10 High-Energy Theory Collaborations Network

6.10.1 Introduction

The structure and connectivity of the High-Energy Theory Collaborations Network has been deduced from co-authorship of scientists in a single paper, which a link between two scientists is established by their coauthorship of one or more scientific papers [66]. Thus the groups to which scientists belong in this network are the groups of coauthors of a single paper. This network

is undirected and has a total of $8,361$ vertices and $15,751$ undirected edges.

6.10.2 Results and Discussion

After normalization, we apply the Dombi t-norm generator for several parameters and apply the metric closure to calculate the main components. We have one strong component of $5,835$ vertices and many small components of 2 vertices. We analyze only the strong component. Figure 6.19 shows the semi-metric percentage (SM) for the High-Energy Theory Collaborations Network and for randomized weighted random network (null model).

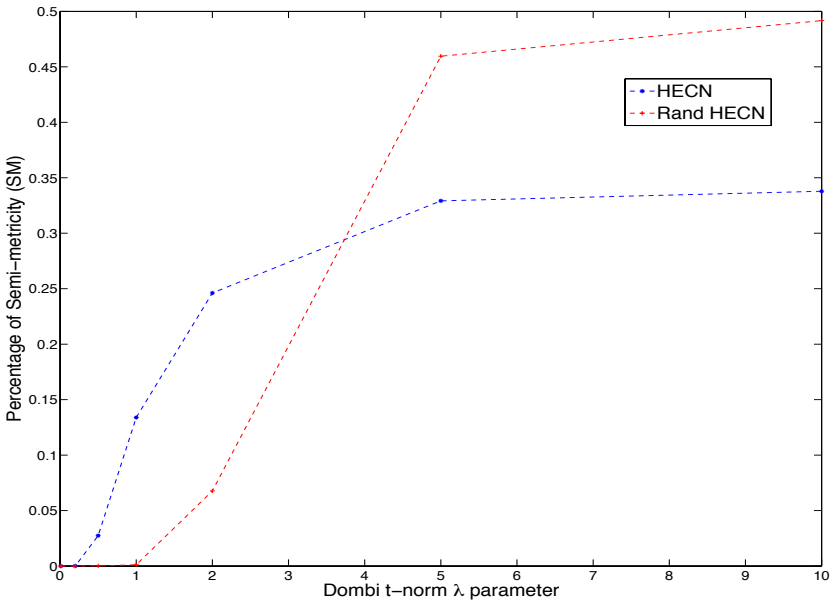

Figure 6.19: Percentage of semi-metricity for Dombi t-norm generator

In tables 6.22 and 6.23 we present the results for the C. Elegans network

and the randomized network, respectively.

λ	$< l >$	σ_l	CV	CI	CQ	C	SM
0.01	7.3	2.0	0.27	–	–	0.51	0
0.2	13.5	3.7	0.27	0.58	0.60	0.51	$7.2E-5$
0.5	36.5	10.3	0.28	0.58	0.60	0.51	0.03
1	186.4	61.4	0.33	0.58	0.60	0.51	0.13
2	$5.1E3$	$2.6E3$	0.51	0.58	0.60	0.51	0.25
5	$3.7E8$	$7.5E8$	2.03	0.58	0.60	0.51	0.33
10	$6.0E17$	$4.9E18$	8.17	0.58	0.60	0.51	0.34

Table 6.22: Variation in the High-Energy Theory Collaborations Network , for various values of the Dombi parameter λ. $< l >$ average path length, σ_l standard deviation of shortest path, CV coefficient of variability, CI and CQ weighted clustering coefficients, C clustering coefficient (crisp), SM semi-metric percentage.

λ	$< l >$	σ_l	CV	CI	CQ	C	SM
0.01	5.8	1.	0.17	–	–	$1E-3$	0
0.2	11.1	2.0	0.18	0.06	0.06	$1E-3$	0
0.5	30.9	5.8	0.19	0.06	0.06	$1E-3$	$1.0E-4$
1	167.4	35.6	0.21	0.06	0.06	$1E-3$	$1.2E-3$
2	$4.7E3$	$1.5E3$	0.32	0.06	0.06	$1E-3$	0.07
5	$1.4E8$	$3.9E8$	2.79	0.06	0.06	$1E-3$	0.46
10	$1.5E17$	$2.7E18$	18.00	0.06	0.06	$1E-3$	0.49

Table 6.23: Variation in the null model of the High-Energy Theory Collaborations Network, for various values of the Dombi parameter λ. $< l >$ average path length, σ_l standard deviation of shortest path, CV coefficient of variability, CI and CQ weighted clustering coefficients, C clustering coefficient (crisp), SM semi-metric percentage.

The network is 87% metric, which means the metric backbone is the network itself. Moreover, the average shortest path is at the same order of magnitude of the null model and the $CV = 0.33$ and $CV_p = 0.47$ for

the actual network with $\lambda = 1$, we can conclude that the network can be considered a small-world. Therefore, this network has no need for further investigation since for $\lambda = 1$ the average path length is characteristic and we would take advantage with the semi-metric thresholding.

6.11 Conclusion

We have seen the the previous chapter 5 how the semi-metric behavior is related to some latent association between vertices. In this chapter we explored the structural properties of semi-metric edges in a network and how they can differentiate two types of edges in a graph: metric and semi-metric. We have seen in this chapter that semi-metric edges work on connectivity structure properties of the network, such as vertex degree, clustering coefficient, etc. and metric edges work on network shortest path distribution. Since, the semi-metric edges work only on the graph structure connectivity, we proposed a new methodology of thresholding (semi-metric thresholding), which better preserves the connectivity of the networks, allowing us to better characterized the small-world phenomenon either than the traditional thresholding.

Chapter 7

Concluding Remarks

7.1 Summary of Contributions

In this thesis we have addressed three main problems: (1) the scale-free degree distribution with cut offs and (2) generalized transitive closures on complex networks (3) a new methodology to analyze complex networks, which allow us to study the small-world phenomena in weighted complex networks. The main contributions were:

(a) Produced an analytical solution and integrative model of cut-offs in the power-law degree distribution , which gives us the ability to better predict the organization of complex networks.

(b) Produced a relation between mathematical treatment of transitive closure in fuzzy graphs and the Dijkstra algorithm [29] in weighted graphs. This result bridges the gap between complex networks and fuzzy graphs and

gives an insight about how we measure the shortest paths between any two vertices in a weighted graph, since there are no unique way to perform this measurement.

(c) Proposed a new methodology to analyze complex networks and study properties such as: the average path length, semi-metric behavior and clustering coefficient in weighted graphs. This helps us to characterize more effectively the small-world phenomena in weighted networks.

In chapter 3 we addressed the first contribution of this thesis. We introduced a stochastic theoretical model as a mathematical explanation of the Amaral's et al. PAVA model [6]. We started by presenting our stochastic model for the Amaral et al. model of preferential attachment with vertex aging. We explained the exponential decay for degree distributions and the network stop growing estimation. We tested the predicting simulations done by Amaral et al. with our STM model.

We believe this work can provide a simple explanation for the dynamics of some scale-free networks and through this knowledge, obtain a better understanding of how these scale-free networks can emerge. As we have seen in the introduction, the field of complex networks is an interdisciplinary field. Therefore, a better understanding of the mechanisms behind complex networks can improve the understanding behind certain problems in areas like the Internet, World Wide Web, Neural Networks, Chemical Networks, Social Networks and so on.

In chapter 4 we addressed the second main contribution, the relations

between transitive closure in fuzzy graphs and the APSP Dijkstra algorithm. We proved corollary 4.1, which states there always exists a t-norm and t-conorm for (min,+) operators such that the two generated closures are isomorphic and the isomorphism between the fuzzy graphs and the distance graphs is in fact the generator for the t-norm. This theorem has a strong impact when we convert a proximity graph into a distance graph and then calculate the respective distance closure. The isomorphism we use defines a t-norm, that is, a metric in fuzzy or distance graphs, which influences the way we measure distances in our graphs.

It is implicit in the results that the transitive closure is a generalization of the APSP Dijkstra algorithm, and consequently these closures are not unique as already known in the theory of fuzzy graphs [53].

We also estimate the best t-norm in the family of Dombi t-norms, which preserves the characteristics in the proximity and distance spaces.

In chapter 5 we study proximity and semi-metric networks, and empirically enforce that the ultra-metric closure destroys several properties from the original networks. Moreover, we verify the result from chapter 4 that $\lambda = 1$ for Dombi t-norm, gives good experimental results in recommendation systems and also preserves properties such strength distribution from the original graphs.

In Chapter 6 we introduced a new methodology, based on semi-metric behavior, to analyze complex networks. This methodology has as main contribution find the core sub-network of a real-world network, which preserves

the structure of the network. Following from results on chapter 4 and this new methodology we discuss the small-world phenomena in six real-world weighted networks.

7.2 Future Work

For future work we foresee the following developments in the following areas.

7.2.1 Study other pairs of t-norms and t-conorms between Proximity and Distance spaces

We have studied in this thesis the Dombi t-norm, when we map a proximity into a distance. However, there are many other possible mappings, such as $\varphi = -log(x)$, which belongs to the family of Schweizer-Sklar t-norms and many others, [51]. We would like to pursue an in deep study of all these t-norms possibilities and it consequences to real problems such as how it affect the topological and dynamics of complex networks.

7.2.2 Depth study of the semi-metricity of Human Cortex Network

We intend pursue a study more in-depth of the Human Cerebral Cortex network, to detect the biological implications derived from this work. Where we found that the Human Cortex network is highly semi-metric. This future

work can provide us a better understanding of the functionalities the Human Cortex.

7.2.3 Community detection in weighted networks

Community detection in crisp networks (edges with values 1 or 0) studies the topology of connections to separate groups of vertices. Through utility functions we can optimize the communities in crisp networks. This works in crisp networks because they do not violate the triangle inequality, that is, all direct paths are shorter than the indirect paths. However, when dealing with weighted networks this is not the case. Real-world weighted networks generally do not obey the triangle inequality, they are semi-metric, that is, many indirect paths between vertices in the network are shorter than the direct paths. The proximity between vertices must be studied not only from the topology of crisp edges but also by the general topology with weighted edges. If we apply the general algorithms from crisp networks to detect communities in weighted networks we are not taking in consideration the semi-metric effect. For future work we intend to explore a framework where we calculate first the shortest distance between all vertices, through the Dijkstra algorithm. After using the matrix with all shortest paths calculate the communities by adapting the concept of distance with any clustering algorithm from the data-mining field; in general k-means or c-means. With this process we expect to be able to detect groups of vertices that have in consideration the semi-metric topology in weighted graphs.

7.2.4 Dynamics in weighted networks

In general real-world networks are heterogeneous in the degree distribution, weight distribution. We have proposed in this thesis a new methodology of analysis of complex networks. This methodology allow us to study subnetworks that have properties such as small-world and represent a more homogeneous subgraph of the network. We would like to explore the dynamics of epidemic spreading in vertices of this subnetworks versus vertices outside of these subnetworks.

7.2.5 Churning in telecommunication networks

Churning is a problem for telecommunication companies, since the market is saturated and is more expensive to try to acquire a new customer than to maintain the customer. To maintain profitability, telecommunication companies must control churn. Nowadays, the pre-paid services to around 80% of the marked. In the pre-paid the telecommunication companies have almost no information about the customer, turning the churn prediction very difficult by the use of traditional data mining tools.

From the Call Detail Record (CDR), we are able to build a weighted graph, representing the social connections of customers. We intend to explore the subnetwork using our methodology with epidemic spreading algorithms to predict churn.

Bibliography

[1] *Classification of protein-protein interaction documents using text and citation network features*. Proceedings of the BioCreative II.5 Workshop 2009: Special Session on Digital Annotations, 2009.

[2] *On Fuzzy vs. Metric Similarity Search in Complex Databases*, 2009.

[3] S.Kamal Abdali and B.David Saunders. Transitive closure and related semiring properties via eliminants. *Theoretical Computer Science*, 40(0):257 – 274, 1985. ¡ce:title¿Eleventh International Colloquium on Automata, Languages and Programming¡/ce:title¿.

[4] Alaa Abi-Haidar, Jasleen Kaur1, Ana Maguitman, Predrag Radivojac, Andreas Retchsteiner, Karin Verspoor, Zhiping Wang, and Luis M. Rocha. Uncovering protein interaction in abstracts and text using a novel linear model and word proximity networks. *Genome Biology*, page In Press, 2008.

[5] Reka Albert and Albert-Laszlo Barabasi. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47, 2002.

[6] L. A. N. Amaral, A. Scala, M. Barthelemy, and H. E. Stanley. Classes of small-world networks. *PNAS*, 97(21):11149–11152, 2000.

[7] T Antal and P L Krapivsky. Weight-driven growing networks. *Physical Review E*, 71:026103, 2005.

[8] R. C. BACKHOUSE and B. A. CARRÉ. Regular algebra applied to path-finding problems. *IMA Journal of Applied Mathematics*, 15(2):161–186, 1975.

[9] Ricardo Baeza-Yates, Berthier Ribiero-Neto, and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Pearson Education, 1999.

[10] Amir Baniamerian and Mohammad Menhaj. Fuzzy shortest paths in fuzzy graphs. In Bernd Reusch, editor, *Computational Intelligence, Theory and Applications*, pages 757–764. Springer Berlin Heidelberg, 2006.

[11] Albert-Laszlo Barabasi and Reka Albert. Emergence of scaling in random networks. *Science*, 286:509, 1999.

[12] Alain Barrat, Marc Barthelemy, Romualdo Pastor-Satorras, and Alessandro Vespignani. The architecture of complex weighted networks. *PROC.NATL.ACAD.SCI.USA*, 101:3747, 2004.

[13] Alain Barrat, Marc Barthelemy, and Alessandro Vespignani. Modeling the evolution of weighted networks. *Physical Review E*, 70:066149, 2004.

[14] Alain Barrat, Marc Barthélemy, and Alessandro Vespignani. Weighted evolving networks: coupling topology and weight dynamics. *Phys Rev Lett*, 92:228701, 2004.

[15] Alain Barrat, Marc Barthelemy, and Alessandro Vespignani. *Dynamical Processes on Complex Networks*. Cambridge, 2008.

[16] V. Batagelj and A. Mrvar. Pajek - program for large network analysis. *Connections*, 21(2):47–57, 1998.

[17] P. Behzadnia, S.M.M. Zarandi, R. Berangi, and A. Baniamerian. On updating the shortest path in fuzzy graphs. In *Computational Intelligence for Modelling Control Automation, 2008 International Conference on*, pages 1188 –1193, dec. 2008.

[18] Luis M. A. Bettencourt, Ariel Cintron-Arias, David I Kaiser, and Carlos Castillo Chavez. The power of a good idea: Quantitative modeling of the spread of ideas from epidemiological models. *Physica D*, 364:513–536, 2006.

[19] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. U Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424:175–308, 2006.

[20] M. Boguna, R. Pastor-Satorras, and A. Vespignani. Cut-offs and finite size effects in scale-free networks. *The European Physical Journal B*, 38:205–209, 2004.

[21] Bela Bollobas. *Random Graphs*. Cambridge University Press, 2004.

[22] Stefan Bornholdt and Hans George Schuster. *Handbook of Graphs and Networks*. Wiley-VCH, 2003.

[23] Ulrik Brandes and Thomas Erlebach. *Network Analysis Methodological Foundations*. Springer, 2005.

[24] Guido Caldarelli. *Scale-Free Networks*. Oxford University Press, 2007.

[25] Guido Caldarelli and Alessandro Vespignani. *Large Scale Structure and Dynamics of Complex Networks*. World Scientific, 2007.

[26] Vittoria Colizza, Alessandro Flammini, Amos Maritan, and Alessandro Vespignani. Characterization and modeling of proteinprotein interaction networks. *PHYSICA A: Statistical Mechanics and its Applications*, 352(1):1–27, 2005.

[27] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms*. MIT Press, 2nd edition edition, 2001.

[28] Chris Cornelis, Peter De Kesel, and Etienne E. Kerre. Shortest paths in fuzzy weighted graphs. *International Journal of Intelligent Systems*, 19(11):1051–1068, 2004.

[29] E. W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1:269–271, 1959.

[30] Chris H. Q. Ding, Xiaofeng He, and Stephen R. Holbrook. Transitive closure and metric inequality of weighted graphs-detecting protein interaction modules using cliques. *Int. J. Data Mining and Bioinformatics*, 2006.

[31] J. Dombi. A general class of fuzzy operators, the demorgan class of fuzzy operators and fuzziness measures induced by fuzzy operators. *Fuzzy Sets and Systems*, 8:149–163, 1982.

[32] S. N. Dorogovtsev and J.F.F. Mendes. *Evolution of Networks*. Oxford University Press, 2003.

[33] D. Dubois and H. Prade. *Fuzzy Sets and Systems*. Academic Press, New York, 1980.

[34] Patric Hagmann et al. Mapping the structural core of human cerebral cortex. *PLOS Biology*, 6(7), July 2008.

[35] Santo Fortunato. Community detection in graphs. arXiv:0906.0612, 2009.

[36] François Fouss, Stephane Faulkner, Manuel Kolp, Alain Pirotte, and Marco Saerens. Web recommendation system based on a markov-chainmodel. *ICEIS*, 4:56–63, 2005.

[37] F. Galvin and S.D. Shore. Distance functions and topologies. *American Mathematical Monthly*, 98:620–623, 1991.

[38] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA*, 99:7821–7826, 2002.

[39] K-I. Goh, B. Kahng, and D. Kim. Nonlocal evolution of weighted scale-free networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 72:017103, 2005.

[40] G. Grefenstette. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic, 1994.

[41] H. Hamacher. Uber logische verknupfungen unscharfer aussagen un deren zugehorige bewertungsfunktionen. In R. Trappl, G.J. Klir, and L. Ricciardi, editors, *Progress in Cybernetics and Systems Research*, volume 3, pages 276–288. Hemisphere, 1978.

[42] Carl Herrmann, Marc Barthelemy, and Paolo Provero. Connectivity distribution of spatial networks. *Physical Review E*, 68:026128, 2003.

[43] Lynette Hirschman, Alexander Yeh, Christian Blaschke, and Alfonso Valencia. Overview of biocreative: critical assessment of information extraction for biology. *BMC Bioinformatics*, 6 Suppl 1(S1), 2005.

[44] Daniel J. and Lehmann. Algebraic structures for transitive closure. *Theoretical Computer Science*, 4(1):59 – 76, 1977.

[45] 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.

[46] Peter Jones, Johan Bollen, Ronald Coifman, Andrew McCallum, and Karin Verspoor. Workshop iii: Social data mining and knowledge. 2007.

[47] A. Kaufmann. *Introduction to the theory of Fuzzy Subsets*. Academic Press, 1975.

[48] E.P. Klement, R. Mesiar, and E. Pap. Problems on triangular norms and related operators. *Fuzzy Sets and Systems*, 145:471–479, 2004.

[49] E.P. Klement, R. Mesiar, and E. Pap. Triangular norms. position paper ii: general constructions and parameterized families. *Fuzzy Sets and Systems*, 145:411–438, 2004.

[50] E.P. Klement, R. Mesiar, and E. Pap. Triangular norms. position paper iii: continuous t-norms. *Fuzzy Sets and Systems*, 145:439–454, 2004.

[51] Erich Peter Klement, Radko Mesiar, and Endre Pap. *Triangular Norms*. Kluwer Academic Publishers, 2000.

[52] George J. Klir. *Facets of System Science*, volume 15. ISFR International Series on Systems Science and Engineering, 2001.

[53] G.J. Klir and B. Yuan. *Fuzzy sets and fuzzy logic, theory and applications*. Prentice Hall PTR, 1995.

[54] Artemy Kolchinsky, Alaa Abi-Haidar, Jasleen Kaur, Ahmed Abdeen Hamed, and Luis M. Rocha. Classification of protein-protein inter-

action full-text documents using text and citation network features. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7:400–411, 2010.

[55] M. Kulenovic and O. Merino. *Discrete Dynamical Systems and difference equations*. Chapman and HallCRC, 2002.

[56] Wenyuan Li, Yongjing, and Ying Liu. The structure of weighted small-world networks. *PHYSICA A*, 376:708–718, 2007.

[57] Cerry M. and Klein. Fuzzy shortest paths. *Fuzzy Sets and Systems*, 39(1):27 – 41, 1991.

[58] K. Maes and B. De Baets. Rotation-invariant t-norm solutions of a system of functional equations. *Fuzzy Sets and Systems*, 157(373-397), 2006.

[59] B. Markines, L. Stoilova, and F. Menczer. Social bookmarks for collaborative search and recommendation. In *Proc. AAAI*, 2006.

[60] J. Mordeson and P. Nair. *Fuzzy Graphs and Fuzzy Hypergraphs*. Physica-Verlag, 2000.

[61] Stefano Mossa, Marc Barthelemy, H Eugene Stanley, and Luis A Nunes Amaral. Truncation of power law behavior in "scale-free" network models due to information filtering. *PHYS.REV.LETT*, 88:138701, 2002.

[62] Kiyohiko Nakamura, Sosuke Iwai, and Tetsuo Sawaragi. Decision support using causation knowledge base. *Systems, Man and Cybernetics, IEEE Transactions on*, 12(6):765 –777, nov. 1982.

[63] M. E. Newman. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Phys Rev E Stat Nonlin Soft Matter Phys*, 64:016132, 2001.

[64] M E J Newman. The structure and function of complex networks. *SIAM Review*, 45:167, 2003.

[65] M. E. J. Newman. Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46(5):323–352, 2005.

[66] M.E.J. Newman. Scientific collaboration networks.i. network construction and fundamental results. *Physical Review E*, 64(016131), 2001.

[67] Jukka-Pekka Onnela, Jari Saramaki, Janos Kertesz, and Kimmo Kaski. Intensity and coherence of motifs in complex networks. *Physical Review E*, 71, 2005.

[68] Romualdo Pastor-Satorras and Alessandro Vespignani. *Evolution and Structure of the Internet*. Cambridge University Press, 2004.

[69] M. Popescu, J.M. Keller, and J.A. Mitchell. Fuzzy measures on the gene ontology for gene product similarity. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 3:263–274, 2006.

[70] Proceedings of the Second BioCreative Challenge Evaluation Workshop. *Uncovering Protein-Protein Interactions in the Bibliome*, 2007.

[71] Albert-Laszlo Barabasi Reka Albert, Hawoong Jeong. Diameter of the world-wide web. *Nature*, 401(9):130–131, September 1999.

[72] L. M. Rocha. Semi-metric behavior in document networks and its application to recommendation systems. In V. Loia (Ed.), editor, *Soft Computing Agents: A New Perspective for Dynamic Information Systems*, International Series Frontiers in Artificial Intelligence and Applications, pages 137–163. IOS Press, 2002.

[73] L.M. Rocha. Evidence sets: Modeling subjective categories. *International Journal of General Systems*, 27:457–494, 1999.

[74] L.M. Rocha. Proximity and semi-metric analysis of social networks. Technical report, Los Alamos National Laboratory: LAUR 02-6557, 2002.

[75] L.M. Rocha and J. Bollen. *Biologically Motivated Distributed Designs for Adaptive Knowledge Management*. Oxford University Press, 2001.

[76] L.M. Rocha, T. Simas, A. Rechtsteiner, M. DiGiacomo, and R. Luce. Mylibrary@lanl: Proximity and semi-metric networks for a collaborative and recommender web service. In IEEE Press, editor, *Proc. 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*, 2005.

[77] Luis M. Rocha. Combination of evidence in recommendation systems characterized by distance functions. In *2002 IEEE International Conference on Fuzzy Systems: FUZZ-IEEE'02; May 12-17 2002; Honolulu, HI, United States*, 2001.

[78] Luis M. Rocha. Automatic conversation driven by uncertainty reduction and combination of evidence for recommendation agents. In *NATO Advanced Research Workshop on Systematic Organisation of Information in Fuzzy Systems; October 24-26, 2001; Vila Real, PORTUGAL*. I O S PRESS, 2003.

[79] Badrul M. Sarwar, George Karypis, Joseph A. Konstan, and John T. Riedl. Application of dimensionality reduction in recommender systems–a case study. WebKDD-2000 Workshop, 2000.

[80] B. Schweizer and A. Sklar. *Probabilistic Metric Spaces*. North-Holland, 1983.

[81] M. Angeles Serrano, Marian Boguna, and Alessandro Vespignani. Extracting the multiscale backbone of complex weighted networks. *PNAS*, 106(16):6483–6488, April 2009.

[82] S. Siegel and J. Castellan. *Nonparametric Statistics for the behavioral Sciences*. McGraw-Hill, 2nd edition edition, 1988.

[83] Jeremy Siek, Lie-Quan Lee, and Andrew Lumsdaine. *The Boost Graph Library*. Addison-Wesley, 2002.

[84] Tiago Simas and Luis M. Rocha. Generalized transitive closures on complex networks. *Fuzzy Sets and Systems*, submmited, 2011.

[85] MD Spencer, RJ Holt, LR Churra, J Suckling, AJ Calder, ET Bulmore, and S Baron-Cohen. A novel functional brain imaging endophenotype of autism: the neural response to facial expression of emotion. *Transl Psychiatry*, 1(e19), 2011.

[86] Jari Sramaki, Mikko Kivela, Jukka-Pekka Onnela, Kimmo Kaski, and Janos Kertesz. Generalizations of clustering coefficient to weighted complex networks. *Physical Review E*, 75(027105), 2007.

[87] Hrvoje Stefancic and Vinko Zlatic. Preferential attachment with information filtering–node degree distribution properties. *PHYSICA A*, (350):657–670, 2005.

[88] Lubomira Stoilova, Todd Holloway, Ben Markines, Ana G. Maguitman, and Filippo Menczer. Givealink: mining a semantic network of bookmarks for web search and recommendation. In *LinkKDD '05: Proceedings of the 3rd international workshop on Link discovery*, pages 66–73, New York, NY, USA, 2005. ACM.

[89] Alexander Strehl. *Relationship-based Clustering and Cluster Ensembles for High-dimensional Data Mining*. PhD thesis, Austin University of Texas, 2002.

[90] Jeffrey Travers and Stanley Milgram. An experiment study of the small world problem. *Sociometry*, 32(4):425–443, 1969.

[91] Peter D. Turney. Mining the Web for synonyms: PMI–IR versus LSA on TOEFL. *LNCS*, 2167:491–502, 2001.

[92] Stephen B. Vardeman and J. Marcus Jobe. *Data Collection and Analysis*. Duxbury, 2001.

[93] K. Verspoor, J. Cohn, C. Joslyn, S. Mniszewski, A. Rechtsteiner, L.M. Rocha, and T Simas. Protein annotation as term categorization in the gene ontology using word proximity networks. *BMC Bioinformatics*, pages 6(Suppl 1):S20. doi:10.1186/1471–2105–6–S1–S20, 2005.

[94] A. Wagner and D. A. Fell. The small world inside large metabolic networks. *Proc Biol Sci*, 268(1478):1803–1810, September 2001.

[95] Wen-Xu Wang, Bing-Hong Wang, Bo Hu, Gang Yan, and Qing Ou. General dynamics of topology and traffic on weighted technological networks. *Phys Rev Lett*, 94:188702, 2005.

[96] Stanley Wasserman and Katherine Faust. *Social Networks Analysis*. Cambridge, 1994.

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

[98] WI '05 Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence. *A Novel Way of Computing Similarities between Nodes of a Graph, with Application to Collaborative Recommendation*, 2005.

[99] L.A. Zadeh. Fuzzy sets and systems. In ed Fox, J., editor, *System Theory*, pages 29–37. Polytechnic Press, Brooklyn, NY, 1965.

[100] Uri Zwick. All pairs shortest paths using bridging sets retangular matrix multiplication. *Journal of the ACM*, 49(3):289–317, May 2002.

VITA

Tiago Manuel Louro Machado De Simas was born in Lisbon, Portugal. His licenciate (5 years degree) in Physics from Instituto Superior Tecnico,Technical University of Lisbon, and Master of Science in Artificial Intelligence from Universidade de Evora.