Eduardo Barbosa Araújo

Scientific Collaboration Networks from Lattes Database: Topology, Dynamics and Gender Statistics

Fortaleza - Brazil

June 17, 2016

Eduardo Barbosa Araújo

Scientific Collaboration Networks from Lattes Database: Topology, Dynamics and Gender Statistics

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Física da Universidade Federal do Ceará como requisito parcial para obtenção do título de Doutor em Física.
Área de Concentração: Física da Matéria Condensada.
Orientador: Prof. Dr. José Soares de Andrade Jr.

Trabalho aprovado. Fortaleza - Brazil, 25 de maio de 2016.

BANCA EXAMINADORA

Prof. Dr. José Soares de Andrade Jr.
(Orientador)
Universidade Federal do Ceará (UFC)

Prof. Dr. André Auto Moreira
Universidade Federal do Ceará (UFC)

Prof. Dr. Humberto de Andrade Carmona
Universidade Federal do Ceará (UFC)

Prof. Dr. Tarcísio Haroldo Cavalcante Pequeno
Universidade de Fotaleza (UNIFOR)

Profa. Dra. Suani Tavares Rubim de Pinho
Universidade Federal da Bahia (UFBA)

Für Laura Barth

Acknowledgements

Many persons, and institutions, were undoubtedly essential to the development of this thesis, not only on the practical aspects of the research undertaken but, and equally important, on the support given during these years. I could not thank these individuals enough and hope that I can somehow retribute their contributions in the future. Some of them I feel obliged to name. They are listed in the following. Those who I might forget, I ask to be forgiven and, next time you meet me, ask me for a drink.

Professor Dr. José Soares de Andrade Jr., for the opportunity of initiating this journey and, in this long path, for the many discussions, advice, helps and, more important, for your patience.

Professor Dr. André Auto Moreira, for all discussions and the opportunity of collaboration.

Prof. Dr. Hans Jürgen Herrmann, for the opportunity of collaborating with your group at ETH-Zürich.

Luciano da Silva, Vasco Furtado, Tarcísio Pequeno and Nuno Araújo, for the works we did together.

Laura Barth, for the companionship, the motivation and the patience since the beginning of the Doctorate. I can not thank you enough.

The friends I met during these years: Saulo, Erneson, Hygor, Heitor, Rilder, Tatiana, Felipe, César, Rubens, Diego, Davi, Levi, Vagner, Vitor, Alexandre, Marcos (UFSC), Lineu (CBPF), Julian, Miller, Nicolas, Fahrang, Klara, Lucas, Gustavo, Amanda and Anabela.

My family.

The institutions: ETH-Zürich for the collaboration. FUNCAP, CAPES and CNPq for the financial support.

The man who comes back through the Door in the Wall will never be quite the same as the man who went out. He will be wiser but less sure, happier but less self-satisfied, humbler in acknowledging his ignorance yet better equipped to understand the relationship of words to things, of systematic reasoning to the unfathomable mystery which it tries, forever vainly, to comprehend.

–Aldous Huxley

Resumo

Compreender a dinâmica de produção e colaboração em pesquisa pode revelar melhores estratégias para carreiras científicas, instituições acadêmicas e agências de fomento. Neste trabalho nós propomos o uso de uma grande e multidisciplinar base de currículos científicos brasileira, a Plataforma Lattes, para o estudo de padrões em pesquisa científica e colaborações. Esta base de dados inclui informações detalhadas acerca de publicações e pesquisadores. Currículos individuais são enviados pelos próprios pesquisadores de forma que a identificação de coautoria não é ambígua. Pesquisadores podem ser classificados por produção científica, localização geográfica e áreas de pesquisa. Nossos resultados mostram que a rede de colaborações científicas tem crescido exponencialmente nas últimas três décadas, com a distribuição do número de colaboradores por pesquisador se aproximando de uma lei de potência à medida que a rede evolui. Além disso, ambas a distribuição do número de colaboradores e a produção por pesquisador seguem o comportamento de leis de potência, independentemente da região ou áreas, sugerindo que um mesmo mecanismo universal pode ser responsável pelo crescimento da rede e pela produtividade dos pesquisadores. Também mostramos que as redes de colaboração investigadas apresentam um típico comportamento assortativo, no qual pesquisadores de alto nível (com muitos colaboradores) tendem a colaborador com outros semelhantes. Em seguida, mostramos que homens preferem colaborar com outros homens enquanto mulheres são mais igualitárias ao estabelecer suas colaborações. Isso é consistentemente observado em todas as áreas e é essencialmente independente do número de colaborações do pesquisador. A única exceção sendo a área de Engenharia, na qual este viés é claramente menos pronunciado para pesquisadores com muitas colaborações. Também mostramos que o número de colaborações segue o comportamento de leis de potência, com um *cutoff* dependente do gênero. Isso se reflete no fato de que em média mulheres produzem menos artigos e têm menos colaborações que homens. Também mostramos que ambos os gêneros exibem a mesma tendência quanto a colaborações interdisciplinares, exceto em Ciências Exatas e da Terra, nas quais mulheres tendo mais colaboradores são mais propensas a pesquisas interdisciplinares.

Palavras-chave: redes de colaboração; Lattes; redes complexas; grafos; análise de redes sociais.

Abstract

Understanding the dynamics of research production and collaboration may reveal better strategies for scientific careers, academic institutions and funding agencies. Here we propose the use of a large and multidisciplinary database of scientific curricula in Brazil, namely, the Lattes Platform, to study patterns of scientific production and collaboration. Detailed information about publications and researchers is available in this database. Individual curricula are submitted by the researchers themselves so that co-authorship is unambiguous. Researchers can be evaluated by scientific productivity, geographical location and field of expertise. Our results show that the collaboration network is growing exponentially for the last three decades, with a distribution of number of collaborators per researcher that approaches a power-law as the network gets older. Moreover, both the distributions of number of collaborators and production per researcher obey power-law behaviors, regardless of the geographical location or field, suggesting that the same universal mechanism might be responsible for network growth and productivity. We also show that the collaboration network under investigation displays a typical assortative mixing behavior, where teeming researchers (*i.e.*, with high degree) tend to collaborate with others alike. Moreover, we discover that on average men prefer collaborating with other men than with women, while women are more egalitarian. This is consistently observed over all fields and essentially independent on the number of collaborators of the researcher. The solely exception is for engineering, where clearly this gender bias is less pronounced, when the number of collaborators increases. We also find that the distribution of number of collaborators follows a power-law, with a cut-off that is gender dependent. This reflects the fact that on average men produce more papers and have more collaborators than women. We also find that both genders display the same tendency towards interdisciplinary collaborations, except for Exact and Earth Sciences, where women having many collaborators are more open to interdisciplinary research.

Key-words: collaboration networks; Lattes; complex networks; graphs; social network analysis.

Contents

1	INTRODUCTION	23
2	NETWORKS	26
2.1	Origins	26
2.2	Elements and Types of Graphs	26
2.3	Graph connectivity	29
2.4	Representation of Graphs	31
2.5	Properties of graphs	33
2.6	Calculating paths	34
2.7	Classical Random Graphs	35
2.8	Barabási-Albert Model	39
2.9	Small-world phenomenon	39
2.10	Strogatz and Watts Model	40
2.11	Random graphs with specified degree sequence	41
3	LATTES COLLABORATION NETWORKS	44
3.1	Scientific collaborations	44
3.2	Co-authorship versus collaboration	45
3.3	The network approach to scientific collaboration	46
3.4	Lattes Platform	47
3.5	Data acquisition and parsing	48
3.6	Building a collaboration network	52
3.7	The Total Collaboration Network	53
3.8	Conclusions	66
4	GENDER AND COLLABORATION	68
4.1	Introduction	68
4.2	Methods	71
4.3	Results	71
4.4	Conclusions	78
5	CONCLUSIONS	80
	BIBLIOGRAPHY	82

List of Figures

Figure 1 – The problem of the seven bridges of Königsberg. (a) The bridges connect land masses. It is asked if it is possible for one person to make a walk through the city while crossing each bridge once and only once. (b) Euler solved the problem considering only the connectivity of the land masses. Considering the land masses points connected by lines representing the bridges, one can see that excluding the starting and ending points, for each bridges to be crossed only once, the number of lines must be even: half are used to arrive at that land mass and half to leave it. Since all land masses are connected by an odd number of bridges, such walk is impossible. 26

Figure 2 – (a): An undirected graph $G(V, E)$, with vertex set $V = \{1, 2, 3, 4, 5\}$ and edge set $E = \{(1, 2), (1, 5), (2, 3), (2, 4), (2, 5), (3, 4)\}$, which elements are unordered pairs. Vertices 1 and 5 are adjacent since there is an edge $(1, 5) \in E$. Vertices 3 and 4 are the endpoints of the edge $(3, 4)$, and incident with this edge. (b) A digraph $G(V, E)$, where $V = \{1, 2, 3, 4, 5\}$ and $E = \{\langle 1, 2\rangle, \langle 1, 5\rangle, \langle 2, 3\rangle, \langle 2, 4\rangle, \langle 4, 3\rangle, \langle 5, 2\rangle\}$, which elements are ordered pairs. Arrows point from the initial vertex of the edge to the final vertex. Vertex 1 is the initial vertex of the edge $\langle 1, 5\rangle$, while vertex 5 is the final vertex of such edge. Note that (a) is the undirected version of this graph. 27

Figure 3 – Examples of graphs. (a): A simple undirected graph with 11 vertices and 17 edges. (b): A digraph with 5 vertices and 6 directed edges, one of them being a self-loop. Arrows point from initial vertex to final vertex. (c): A multigraph, in which there is more than one edge incident with the same pair of vertices. . . . 28

Figure 4 – A bipartite graph. Vertices can be partitioned into two disjoint sets, black circles and red squares. Edges must join vertices of different partitions. For example, for a citation network we may take squares as papers and circles as authors. Authors who are adjacent to a same paper are co-authors. 28

Figure 5 – Connectivity relationships on a digraph $G(V, E)$. The correspondence (also called neighbors) of vertices are: $\Gamma(1) = \{2, 3\}$, $\Gamma(2) = \{3\}$, $\Gamma(3) = \emptyset$, $\Gamma(4) = \emptyset$ and $\Gamma(5) = \{3, 5\}$. The inverse correspondences are: $\Gamma^{-1} = \emptyset$, $\Gamma^{-1}(2) = \{1\}$, $\Gamma^{-1}(3) = \{1, 2, 5\}$, $\Gamma^{-1}(4) = \emptyset$ and $\Gamma^{-1}(5) = \{5\}$. The in-degree of vertices are: $\delta^{in}(1) = 0$, $\delta^{in}(2) = 1$, $\delta^{in}(3) = 3$, $\delta^{in}(4) = 0$ and $\delta^{in}(5) = 1$. The out-degree of vertices are: $\delta^{out}(1) = 2$, $\delta^{out}(2) = 1$, $\delta^{out}(3) = 0$, $\delta^{out}(4) = 0$ and $\delta^{out}(5) = 1$. The vertex 4 is isolated, since $\delta^{in}(4) = \delta^{out}(4) = 0$. 30

Figure 6 – An undirected graph with 5 vertices and 7 edges. 32

Figure 7 – Particular random graphs obtained using the $G_{N,p}$ model with $N = 20$ for different probabilities p: (a) $p = 0.00$, (b) $p = 0.15$, (c) $p = 0.30$ and (d) $p = 1.00$. 36

Figure 8 – Histogram of the degree distribution for a $G_{N,p}$ network with $N = 10000$ and $p = 0.004$. Dots represents the data points. The gray line is a fit using a Poisson distribution (Eq. 2.24). The parameter $z = 39.8898$ was obtained by maximum likelihood estimation, close to the expected value for the average degree, 39.996. 37

Figure 9 – $G_{1000,p}$ random graph with p varying from 0.00001 to 0.00500. Thus, the average degree of the networks varies from 0.001 to 5.00. For each value of p we have built 10 different networks. The size of the giant component for each of these was obtained and the average value was taken for each value of p. 38

Figure 10 – The Watts-Strogatz model [1]. We start with a regular lattice, here a ring lattice with $N = 20$ vertices. Each vertex has degree 4, being adjacent to its nearest neighbors and second-nearest neighbors. Initially, the graph displays high clustering and high average path length, l. For each edge there is a probability p of changing one endpoint of the edge with the condition that we do not allow self-loops and multiple edges. The rewiring of edges creates shortcuts, causing l to diminish. If $p = 1$ we obtain a ER random graph, but with low clustering. For intermediate values of p the graph displays both small average path length and high degree of clustering. 41

Figure 11 – Document tree of an xhtml document. <html> is the root element, with two children: <head> and <body>. <head> has a child <title>. 49

Figure 12 – (a) Bipartite network B containing node classes R and P representing researchers (circles) and papers (rectangles), respectively. (b) R-projection of B, where researchers are connected if the share a paper in B. The weight of the link is given by the number of shared papers. 52

Figure 13 – Sample network extracted from the collected Lattes Database. Links shown are between researchers (nodes) who were granted a scholarship and working in fields of Medicine in the state of São Paulo. Node size is proportional to the degree of the researcher in the whole database. Researchers were grouped according to the year of their first published paper. The first cohort (dark blue) comprises all researchers who published their first paper before 1975. Each subsequent one, in the counterclockwise direction, comprises researchers who published within 5 years from the previous one, up to 2000. The edges are directed, colored according to the most senior. 54

Figure 14 – Evolution of the fraction of the giant component of TCN since 1982. For every two years, the respective cumulative network was produced and the number of researchers in the largest component was divided by the number of researchers in collaboration in that year. 55

Figure 15 – Distribution of component sizes s in the TCN, excluding the giant component. Remaining component sizes are distributed as a power-law $p(s) = Ae^{-\lambda s}$, with $\lambda = 3.770 \pm 0.126$. The red line is the best power-law fit. 56

Figure 16 – Left: Number of researchers with published papers (black circles) and collaborations between them (red stars) present in the cumulative collaboration network. Dashed lines are exponential fits in the form $s = ae^{\alpha t}$ up to 2009, seen as straight lines in the linear-log plot. The coefficient α is shown in the picture for each curve. Deviations of the 2012 data points from the exponential fit are due to the early acquisition of the curricula, in June of 2012. Right: Superlinear scaling of the number of collaborations with the number of researchers. Dashed line is a power-law curve with exponent $\alpha_c/\alpha_r = 1.31$. 57

Figure 17 – Evolution of the largest component. Data points represent the fraction of researchers present in the largest component for a five year time window centered in the respective year. More than 80% of the researchers engaged in collaborations in the last 5 years are in the largest component. They represent 61% of the researchers in TCN. 57

Figure 18 – Scaling of $C(k)$ with k for the TCN (main graph) and the SCN (inset). The red lines are power-law fits of the respective data ($C(k) = Ak^{-\sigma}$). For the TCN, $\sigma = 0.71 \pm 0.009$. For the SCN $\sigma = 0.58 \pm 0.020$. 59

Figure 19 – Distribution of scientific production of researchers belonging to the TCN group. The solid red line is the best fit to the data points of a power-law with exponential cutoff, $P(n) = A_p n^{-\beta_p} e^{-n/l_p}$, where $\beta_p = 1.58$ and $l_p = 129$. The dashed black line is a power-law with exponent -1.58. 60

Figure 20 – Normalized distribution of the number of collaborators (k) of researchers with scholarship (blue stars), without (black circles) and for the TCN (red triangles). The distribution for researchers with scholarship decreases slowly up to one hundred collaborators, although most of them still have a small number of collaborators. The higher proportion of researchers with high k might reflect the CNPq policy of considering the proponent's participation in research groups, international immersion and human resources development to grant the scholarship. 61

Figure 21 – Variation of the average nearest-neighbor degree (k_{nn}) with k. Being an increasing function of k, the network displays assortative mixing. Researchers with high k are more likely to collaborate with other well connected researchers. This tendency, however, increases logarithmically with k, as indicated by the regression fit (dashed line). 61

Figure 22 – (a) Time evolution of the distribution of the number of collaborators in the TCN. (b) Rescaling the distribution in (a) by the relative number of collaborators for each year shows a collapse onto a single curve. We also show the respective cumulative distributions in (c) and (d). As the network ages, the fraction of researchers with high k increases (c), but the evolution of the network shows that the distribution is constrained to the average production (d). 62

Figure 23 – Interstate collaborations obtained from the Lattes Database. Vertices radii are proportional to the fraction of researchers in TCN. Edges are proportional to the total number of collaborations. Southeast states concentrate most of the collaborations, specially São Paulo. 63

Figure 24 – Top: Distribution of number of collaborators in the TCN for the 26 Brazilian states and the Federal District. The distributions display the same behavior as the TCN (Fig. 22). The dashed line is guide for the eye in the form of a power-law with exponent 1.63 ($P(k) = Ak^{-1.63}$). Bottom: the average number of collaborators versus the number of researchers in each state. The circles correspond to the results for 26 Brazilian states and the Federal District. The dashed line is the best fit obtained by linear regression of the data to a power-law $\langle k \rangle_s \sim N_s^{\delta}$ in logarithmic scale, with exponent $\delta = 0.12 \pm 0.01$. 64

Figure 25 – Cumulative distributions P_C of the number of papers published per researcher n (a) and number of collaborators (b) for each of the 8 major fields. The respective distributions for the rescaled data are shown in (c) and (d). Lines represent different fields, colored according to the symbol in the legend. Scientists working on social sciences and related fields (Lin, Soc and Hum) are less likely to have published more than one hundred papers than others. They also are less likely to have more than one hundred collaborators. Considering the average publication count $\langle n \rangle_f$ and average number of collaborations $\langle k \rangle_f$ in each field, all the curves collapse to a single universal behavior. The insets show the respective (non-cumulative) distributions. 65

Figure 26 – Study of interdisciplinary collaborations. Vertices represent different fields with sizes proportional to the fraction of collaborations with researchers of other fields. The directed edges are colored according to the source vertex and the width is proportional to the fraction of collaborations made with the target vertex. While some pairs are expected as Exa–Eng and Lin–Hum, the small fractions of collaborations between researchers of Bio with Eng could indicate that biotechnology is still a maturing field. 67

Figure 27 – Average of the number of papers for male (blue bars) and female (red bars) researchers for the TCN and for each of the eight major fields: Agricultural Sciences (Agr), Applied Social Sciences (Soc), Biological Sciences (Bio), Exact and Earth Sciences (Exa), Humanities (Hum), Health Sciences (Hea), Engineering (Eng) and Linguistics and Arts (Lin). 71

Figure 28 – Average of the number of collaborators for male (blue bars) and female (red bars) researchers for the TCN and for each of the eight major fields: Agricultural Sciences (Agr), Applied Social Sciences (Soc), Biological Sciences (Bio), Exact and Earth Sciences (Exa), Humanities (Hum), Health Sciences (Hea), Engineering (Eng) and Linguistics and Arts (Lin). 72

Figure 29 – Top: number of male (blue squares) and female (red circles) researchers who published their first paper per year. It is clear that since 2000 women are the majority joining the collaboration network. In the bottom, we show the same data with the number of researchers in logarithmic scale, for better visualization of the transition point. 72

Figure 30 – a) Distribution of the number of recurrent collaborations (weights) between researchers, divided into male and female researchers. b) Distribution of the number of collaborators, divided into male and female researchers. Women are less likely to display large values for both quantities. 74

Figure 31 – k_{nn} as a function of the number of collaborators (k) for male (blue squares) and female (red circles) researchers. 75

Figure 32 – Fraction of new collaborators acquired as a function of time since the first published paper. Both male and female researchers display an exponential decay with a change in behavior occurring after 30 years. The dashed lines are exponential fits for the first 30 years: $p(t) = Ae^{-\lambda t}$. The calculated exponents are $\lambda_w = 0.132 \pm 0.002$ for women and $\lambda_m = 0.0998 \pm 0.001$ for men. 75

Figure 33 – Fraction of new collaborators who joined the collaboration network after the researcher, for men (blue squares) and women (red circles). While in the beginning of a researcher career most of his or her collaborators are older (i.e., joined the network before them), after only 5 years they display a balance between older and younger collaborators. After 20 years of research, around 90% of new collaborators are younger and after 30 years, only a small fraction of new collaborators are older. The same behavior is observed for both genders. 76

Figure 34 – Mean values of *g-ratio* for each field of expertise: Agricultural Sciences (Agr), Applied Social Sciences (Soc), Biological Sciences (Bio), Exact and Earth Sciences (Exa), Humanities (Hum), Health Sciences (Hea), Engineering (Eng) and Linguistics and Arts (Lin). Red and blue bars represent values for female and male researchers, respectively. Yellow triangles show the proportion of female researchers in the respective field. 77

Figure 35 – Correlation between *g-ratio* and number of collaborators for Biological Sciences (a) and Engineering (b) for female (red circles) and male (blue circles) researchers. Lines represent the female proportion in the respective area. Female researchers are more likely to collaborate with other female researchers than their male peers, without regard to the expertise. For technology related fields, *g-ratio* is above the proportion of female researchers. The bars indicate the standard deviation. 78

Figure 36 – Correlation between *m-ratio* and number of collaborators in Biological Sciences (a) and Exact Sciences (b) for female (red circles) and male (blue circles) researchers. For all fields, except Exact and Earth Sciences, there is only a small difference regarding multidisciplinary collaborations. The bars indicate the standard deviation. 79

List of Tables

Table 1 – Some relevant elements in the Lattes CV XML files and some corresponding attributes. 50
Table 2 – Cost and example of operations in Damereau-Levenshtein Algorithm implementation used in this work. 53
Table 3 – Fraction of fields in the last 5 years. The network was constructed by projecting the bipartite network onto a network containing only researchers connected if they share a paper published in the last 5 years. Sum of fractions is not 100% because the field information is not available for all researchers. 56
Table 4 – Statistics for the networks studied in this work. 58
Table 5 – Statistics for researchers working on the 8 major fields associated with the TCN. 66
Table 6 – Number of male and female professors and researchers in Brazil. Source: Lattes Database (http://lattes.cnpq.br/), collected on 2015 January. 70
Table 7 – Number of researchers in the TCN, proportion of female researchers, average number of papers and collaborators for male and female researchers for each of the eight major fields: Agricultural Sciences (Agr), Applied Social Sciences (Soc), Biological Sciences (Bio), Exact and Earth Sciences (Exa), Humanities (Hum), Health Sciences (Hea), Engineering (Eng) and Linguistics and Arts (Lin). 73

1 Introduction

Nowadays, scientific collaboration is understood as extremely valuable, as it integrates skills, knowledge, apparatus and resources, allows division of labor and the study of more difficult problems, including interdisciplinary ones. It also brings recognition and visibility and increases the network of contacts of the researchers involved [2–4]. Scientific collaboration is strongly correlated with production measured by publication output and other indexes in Scientometrics [5–7], which has substantially contributed to raise the interest of the scientific community in studying itself over the last decades [3, 5, 8–11]. More recently, due to the fast growth and enormous development of the complex network science [1, 12–22] the subject of scientific collaboration has been extensively studied under the framework of rather powerful and universal paradigms [23–29].

The Internet and the fact that traveling became substantially less costly have facilitated international collaborations. Still, geographical constraints affect the dynamics of research [30–32]. Different countries have different funding policies and this fact impacts the publication outcome, which is correlated to collaboration. For a country to be above the world average number of citations, it must spend more than one hundred thousand US dollars per researcher per year [32]. At the same time, scientists with more investment in their research projects are more engaged in collaborations [33].

The social nature of collaboration [3, 34] might be the cause for the big disparity in production and number of collaborators [35]. Inequalities in income (Pareto distribution [36]) and movie co-appearance [37] are examples of social distributions, characterized by a power-law profile. For scientific collaborations, such distributions also appear, as demonstrated by Lotka in 1926 [38], from the analysis of two empirical sets of publications data in natural sciences.

Although in Lotka's analysis [38] only the senior authorship has been considered, the obtained power-law distributions was shown to be consistent with empirical bibliometric data taking all authors into account [39]. The so called Lotka's Law therefore seems to be valid even in different fields than those originally considered [39, 40]. It is also worth noting that highly prolific authors were excluded in Lotka's procedure due to the limited number of persons in the samples. These teeming researchers might lie outside the pure power-law distribution. Considering that engaging in collaboration is a time consuming activity, the number of collaborators can not be arbitrarily large, i.e., must be somehow limited. An exponential cutoff has then been suggested as a correction to fit the distribution of productivity [27]. Measuring the distributions of citations by city or country, a power-law distribution also arises [32], which indicates the presence of self-similarity in the science system [41].

Nonetheless, the definition of research collaboration is problematic due to the subjective understanding of its essential ingredients [3,4]. This can be avoided by considering as scientific collaboration a research which resulted in a co-authored scientific paper. This approach, although traditional, is not free of criticism as there are fruitful and relevant collaborations which do not necessarily involve a publication. Notwithstanding, there is evidence that division of labor of theoretical or ex-

perimental work is usually rewarded with a co-authorship [4]. Also, analysing co-authorship makes it feasible to study collaboration of a greater number of researchers as compared by interviewing each individual.

Despite the numerous studies about scientific production, citations and collaborations found in the literature, it is difficult to compare these variables as the databases used in these studies are usually unrelated [42]. Another problem is the small number of samples, due to a low number of respondents in questionnaires or data used only from a specific journal. To analyse the big picture is paramount to work with a dense information database. Here, we used data from Lattes Platform (http://lattes.cnpq.br), an online database maintained by CNPq (National Council of Technological and Scientific Development), a government agency that finances scientific research in Brazil. It contains the curricula of almost all researchers in Brazil and some of their collaborators abroad, as well as information concerning their research groups. The Lattes Curriculum became the standard national scientific curriculum in Brazil, and compulsory for those requiring financial support from the Brazilian government. The curricula present detailed information concerning the researcher, including, but not limited to, full name, professional address, academic titles, field of expertise and list of papers. Researchers are classified in 9 major fields: Agricultural Sciences (Agr), Applied Social Sciences (Soc), Biological Sciences (Bio), Exact and Earth Sciences (Exa), Humanities (Hum), Health Sciences (Hea), Engineeing (Eng), Linguistics and Arts (Lin), Technologies[1], and Others (Oth). Most information in the curriculum is provided by the researcher themselves, for example, their list of publications.

By using this database, we may overcome some of the limitations found by other authors [23,24]. Due to the lack of individual information of the researcher, the problem of author name disambiguation [24,43] becomes relevant, when, for example, two or more authors share initials and surnames. This is not the case with the Lattes Platform, where co-authorship is unambiguous. Researchers themselves update their curricula with detailed information about their publications and professional activity. As a consequence, this type of data allows us to study scientific production and collaborations of individual researchers and correlations between fields of expertise.

The science of networks is built upon the mathematical concept of graph. In physics, its application can be traced back to the works of Kirchhoff on electric circuits. Networks can be used as a model to study the dynamics of non-linear phenomena, as disease spreading [44–46] and the physics of the brain [47, 48]. The technological advancement allowed scientists to study massive networks, a prohibitive task decades ago. This was facilitated not only by means of more powerful processors and greater data storages but also by the availability of high quality data and faster means of communication, in special the Internet.

The sequence of this work is as follows. In the second chapter we present the concept of graph and discuss different types of graphs. Following this, we summarize some concepts which are relevant for this work, such as degree, degree distribution, components of graphs, path length and

[1] This field was recently included and was not available when this research was conducted. Hence, it will not be considered in our results.

clustering. We discuss real network properties and models, starting with the Erdős-Renyi random graph [49, 50] and its inadequacy to describe real networks. We present the preferential attachment model [23], which reproduces the degree distribution observed in real networks. In the sequence, we discuss the small world phenomena [51] and a small-world model which incorporates clustering of vertices [1].

In the third chapter we give a review of works on scientific collaboration networks. We present the views on what consists a scientific collaboration and how collaboration networks are built. Afterwards, we present findings which characterize scientific collaboration networks. We present the Lattes Platform as an object of study, discussing the extensive available information. Afterwards, we discuss our method of data mining and data storage for further processing. The construction of the networks are described. We discuss results on the collaboration network of the Lattes Platform. We investigate the evolution of the network, the differences and similarities between fields of expertise.

In the fourth chapter we characterize the general aspects, performance and differences between male and female researchers for different fields of expertise. We introduce two original metrics to investigate divergent behaviour of male and female researchers and show the existence of a gender bias, which is a relevant factor contributing to the underrepresentation of women in academy.

In the last chapter we summarize our conclusions and present perspectives for future work.

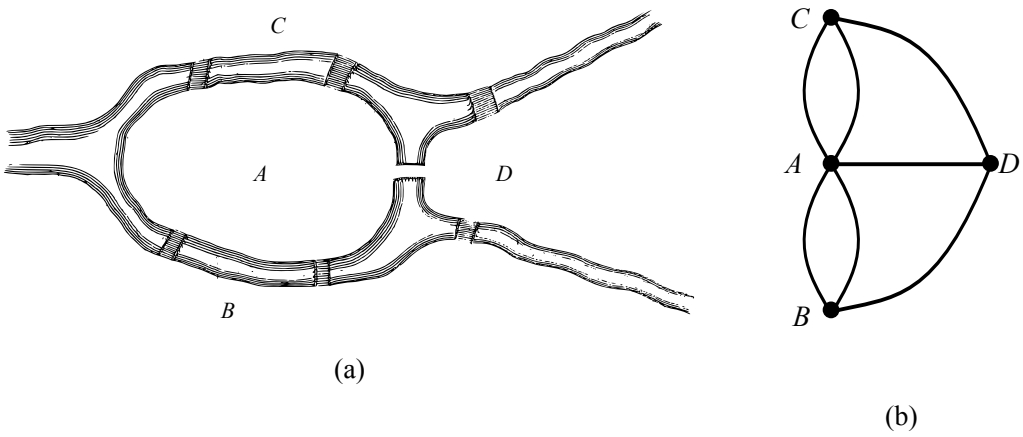
2 Networks

2.1 Origins

The study of networks is built upon the mathematical concept of graph, which had its inception in 1736 with a work of the Swiss mathematician Leonhard Euler (1707-1783). Euler presented the solution to the following problem: in Königsberg, Prussia (nowadays Kaliningrad, Russia), there was a river which surrounded an island and was divided into two branches, as shown in Fig. 1. There were seven bridges crossing the river and connecting the land masses. Concerning these bridges, would it be possible to find a route crossing each bridge once and only once? Euler proved that this was impossible and his ingenious solution resulted in the birth of a new branch of mathematics.

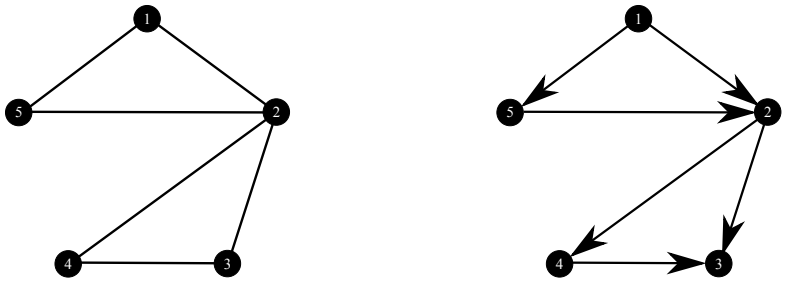
2.2 Elements and Types of Graphs

A *graph* consists of a collection of elements which we call vertices (*nodes*, *actors*, *site* and *points* are also used in the literature) whose relationships to one another are represented as *edges* (or *links*, *ties*). Thus, we characterize a graph G by its sets of vertices V and edges E as $G(V, E)$. This very broad concept allow us to use graphs to represent very different systems: the world



(a) (b)

Figure 1 – The problem of the seven bridges of Königsberg. (a) The bridges connect land masses. It is asked if it is possible for one person to make a walk through the city while crossing each bridge once and only once. (b) Euler solved the problem considering only the connectivity of the land masses. Considering the land masses points connected by lines representing the bridges, one can see that excluding the starting and ending points, for each bridges to be crossed only once, the number of lines must be even: half are used to arrive at that land mass and half to leave it. Since all land masses are connected by an odd number of bridges, such walk is impossible.



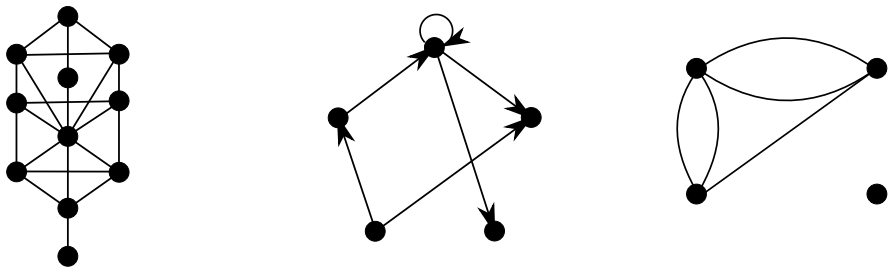
(a) (b)

Figure 2 – (a): An undirected graph $G(V, E)$, with vertex set $V = \{1, 2, 3, 4, 5\}$ and edge set $E = \{(1, 2), (1, 5), (2, 3), (2, 4), (2, 5), (3, 4)\}$, which elements are unordered pairs. Vertices 1 and 5 are adjacent since there is an edge $(1, 5) \in E$. Vertices 3 and 4 are the endpoints of the edge $(3, 4)$, and incident with this edge. (b) A digraph $G(V, E)$, where $V = \{1, 2, 3, 4, 5\}$ and $E = \{\langle 1, 2\rangle, \langle 1, 5\rangle, \langle 2, 3\rangle, \langle 2, 4\rangle, \langle 4, 3\rangle, \langle 5, 2\rangle\}$, which elements are ordered pairs. Arrows point from the initial vertex of the edge to the final vertex. Vertex 1 is the initial vertex of the edge $\langle 1, 5\rangle$, while vertex 5 is the final vertex of such edge. Note that (a) is the undirected version of this graph.

wide web (WWW) [52], the internet [53], actors networks [1, 37], network of sexual contacts [54], products networks for recommendation systems, mobility [55], food webs [56], genetic networks, the brain [48, 57], metabolic networks [58], evolution of diseases [59] and even networks of networks [60]. Graphs can be pictorially represented as dots, symbolizing the vertices, and bars joining dots, symbolizing the edges.

Each edge has two vertices as its *endpoints* and these are said to be *adjacent* to each other and *incident* with this edge. The relationship expressed by the edge may have a directional quality. For example, due to the subjective concept of friendship, it is possible that a person considers another as his or her friend but this might not be reciprocal. In this case, we call the edges of the graph *directed* and the graph itself a *directed graph* or *digraph*. In a directed edge, we have an *initial vertex* and a *final vertex*. If there is no such directional property, we have *undirected edges* and an *undirected graph*. This is the case of co-authorship networks, when we consider that there is a link between co-authors of a paper. We represent the edges of a digraph as an ordered pair of vertices $\langle v_1, v_2\rangle$, where v_1 is the initial vertex and v_2 the final vertex. For undirected graphs this pair is unordered and represented as (v_1, v_2).

The *directed version* of an undirected graph is obtained by replacing each edge (v_1, v_2) by a pair of directed edges $\langle v_1, v_2\rangle$ and $\langle v_2, v_1\rangle$. The *undirected version* of a digraph is obtained by replacing all directed edges by undirected edges. It is possible to have multiple edges between the vertices on a graph, depending on how we define the interaction in the system. Such graph is called a *multigraph*. This is the case of the Königsberg graph in Fig. 1b. Its is also possible for a vertex to be adjacent to itself: we call such an edge a *self-loop*. Graphs without multiple edges or self-loops are called *simple graphs*. Regarding the number of vertices and edges, a graph with no vertices



(a) (b) (c)

Figure 3 – Examples of graphs. (a): A simple undirected graph with 11 vertices and 17 edges. (b): A digraph with 5 vertices and 6 directed edges, one of them being a self-loop. Arrows point from initial vertex to final vertex. (c): A multigraph, in which there is more than one edge incident with the same pair of vertices.

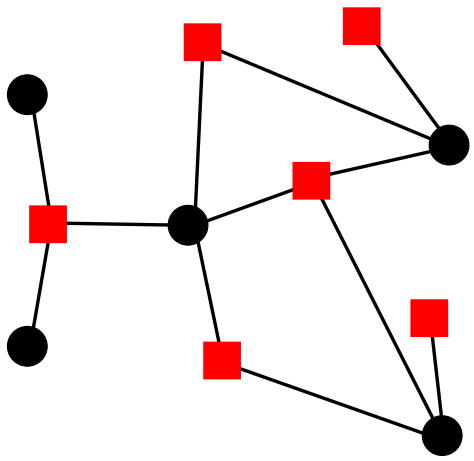


Figure 4 – A bipartite graph. Vertices can be partitioned into two disjoint sets, black circles and red squares. Edges must join vertices of different partitions. For example, for a citation network we may take squares as papers and circles as authors. Authors who are adjacent to a same paper are co-authors.

at all is an *empty graph* and a graph with all possible edges is a *complete graph*: every vertex is adjacent to one another.

For all graphs considered until now, there is no conceptual correlation between edges and vertices. This is not the case for all possible graphs. For example, consider as vertices of a graph the students enrolled in disciplines and the disciplines themselves. Consider also that there is an edge between a student and a discipline if the student is enrolled in that specific discipline. We can see that in this case the graph will not have any edge between students or between disciplines. This is what we call a *bipartite graph*. In bipartite graphs, we can partition the vertices into two disjoint sets such that there are no edges between elements belonging to the same partitioned set (see Fig. 4).

We may generalize the definition of a graph including another set W of elements called *weights*. Weights are numbers which can be associated to vertices or edges of a graph, now denoted by $G(V, E, W)$. A graph with weights associated to edges is called an *edge-weighted graph*. Numbers can also be associated to vertices, forming a *vertex-weighted graph*. In this work we refer to edge-weighted graphs simply as *weighted graphs*. We call the *strength* of a vertex $s(v)$ the sum of the weights w_i of the edges incident with v

$$s(v) = \sum w_i. \tag{2.1}$$

2.3 Graph connectivity

We define the *correspondence* $\Gamma(v)$ of a vertex v in a digraph $G(V, E)$ as the set

$$\Gamma(v) = \{v' \in V | \langle v, v' \rangle \in E\}. \tag{2.2}$$

For undirected graphs, the correspondence of a vertex is defined as the the correspondence of such vertex in the directed version of the graph. The correspondence of a set of vertices is defined as the union of the correspondence of each of those vertices,

$$\Gamma(\{v_1, v_2, \ldots, v_n\}) = \Gamma(v_1) \bigcup \Gamma(v_2) \bigcup \cdots \bigcup \Gamma(v_n). \tag{2.3}$$

The *inverse correspondence* $\Gamma^{-1}(v)$ of a vertex v in a digraph $G(V, E)$ is the set

$$\Gamma^{-1}(v) = \{v' \in V | \langle v', v \rangle \in E\}. \tag{2.4}$$

For a vertex in an undirected graph, the inverse correspondence is obtained by computing its value for the directed version of the graph. For a set of vertices, the inverse correspondence is analogous to the correspondence,

$$\Gamma^{-1}(\{v_1, v_2, \ldots, v_n\}) = \Gamma^{-1}(v_1) \bigcup \Gamma^{-1}(v_2) \bigcup \cdots \bigcup \Gamma^{-1}(v_n). \tag{2.5}$$

We may apply the correspondence function successively and define

$$\Gamma^p(v) = \Gamma(\Gamma^{p-1}(v)), \quad p \in \mathbb{N}. \tag{2.6}$$

The set $\Gamma(v)$ is called the *neighboring set* of v and all its elements are called *neighbors* of v. The cardinality of $\Gamma(v)$ is called the *degree* of vertex[1] v, $\delta(v)$,

$$\delta(v) = n(\Gamma v). \tag{2.7}$$

A vertex with $\delta(v) = 0$ is called *isolated* (node 4 in Fig. 5). For a digraph, a vertex v has two degrees associated. The *in-degree* $\delta^{in}(v)$ is the number of edges with v as the final vertex,

$$\delta^{in}(v) = n(\Gamma^{-1}(v)). \tag{2.8}$$

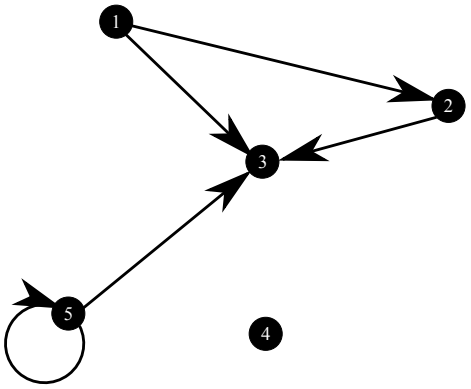


Figure 5 – Connectivity relationships on a digraph $G(V, E)$. The correspondence (also called neighbors) of vertices are: $\Gamma(1) = \{2, 3\}$, $\Gamma(2) = \{3\}$, $\Gamma(3) = \emptyset$, $\Gamma(4) = \emptyset$ and $\Gamma(5) = \{3, 5\}$. The inverse correspondences are: $\Gamma^{-1} = \emptyset$, $\Gamma^{-1}(2) = \{1\}$, $\Gamma^{-1}(3) = \{1, 2, 5\}$, $\Gamma^{-1}(4) = \emptyset$ and $\Gamma^{-1}(5) = \{5\}$. The in-degree of vertices are: $\delta^{in}(1) = 0$, $\delta^{in}(2) = 1$, $\delta^{in}(3) = 3$, $\delta^{in}(4) = 0$ and $\delta^{in}(5) = 1$. The out-degree of vertices are: $\delta^{out}(1) = 2$, $\delta^{out}(2) = 1$, $\delta^{out}(3) = 0$, $\delta^{out}(4) = 0$ and $\delta^{out}(5) = 1$. The vertex 4 is isolated, since $\delta^{in}(4) = \delta^{out}(4) = 0$.

The *out-degree* $\delta^{out}(v)$ is the number of edges with v as the initial vertex

$$\delta^{out}(v) = n(\Gamma(v)), \tag{2.9}$$

see Fig. 5.

A *walk* in an undirected graph is a sequence of alternating vertices and edges, starting with a *source* vertex and ending with a *target* vertex, with each edge having as endpoints the adjacent vertices in the sequence. If the source and target vertices are the same, we have a *closed walk*. A *trail* is a walk in which each edge appears only once. A *path* is a trail in which each vertex appears only once. A cycle is a closed trail in which all vertices but the source and target are distinct. A graph containing a cycle is a *cyclic graph*. A graph with no cycle is an *acyclic graph* or a *forest*. We say that two vertices are *connected* on an undirected graph if there is a path with one of them as source and the other as target. An undirected graph in which all pairs of distinct vertices are connected is called a *connected graph*. A connected forest is a *tree*. For digraphs, we define a *directed-walk* as a sequence of vertices and directed edges in which the edges have as ending points the adjacent vertices in the sequence. Likewise, we define *directed-trails*, *directed-paths* and *directed-cycles* when the edges in sequence are directed.

The set of vertices for which there is a directed-path starting from a vertex v is the reachable set of vertex v, $R(v)$. If we have the correspondences of each vertex of a graph $G(V, E)$, $R(v)$ can be found computing,

$$R(v) = \{v\} \bigcup \Gamma(v) \bigcup \Gamma^2(v) \cdots \bigcup \Gamma^N(v), \tag{2.10}$$

[1] The degree may also be called *connectivity*, but we shall avoid it since it also has another meaning in graph theory.

where $N = n(V)$ is the number of vertices in the graph. The set of vertices for which there is a directed-path ending in a vertex v is the reaching set of vertex v, $Q(v)$. If we have the correspondences of each vertex of a graph $G(V, E)$, $R(v)$ can be found computing

$$Q(v) = \{v\} \bigcup \Gamma^{-1}(v) \bigcup \Gamma^{-2}(v) \cdots \bigcup \Gamma^{-N}(v), \tag{2.11}$$

where $N = n(V)$ is the number of vertices in the graph.

We call a digraph *strongly-connected* if $R(v) = V$, $\forall v \in V$. If its underlying graph is connected, we say that a digraph is *weakly-connected*. For an undirected graph, if $R(v) \neq V$ for any $v \in V$ we call this a *disconnected graph*.

Let $G(V, E)$ be a graph. If we take $V' \subseteq V$ and $E' \subseteq E$ such that for every endpoint v of $e \in E'$, $v \in V'$, $G'(V', E')$ is a *subgraph* of $G(V, E)$. If $G' \neq G$, we say G' is a *proper subgraph* of G. For any disconnected graph $G(V, E)$, we call *components* of G, the disjoint set of connect subgraphs of G which are not contained in a connected subgraph with more vertices or edges. For any digraph G we may consider the maximum subgraph G' which is strongly connected. G' is then called the *strongly-connected component* of G. In the same sense, we call the *connected component* of a graph G, the maximum subgraph of G which is weakly-connected. In network science, this connected component may also be referred as the *giant component* or the *largest component*.

2.4 Representation of Graphs

Although a graph $G(V, E)$ is completely defined by the sets V and E, there are more convenient representations, specially when they are created and/or processed by computer programs. A suitable way of representing a graph $G(V, E)$ is by an *adjacency matrix* $A = [a_{ij}]$. For a graph with n vertices ($n(V) = n$), this is a square $n \times n$ binary matrix which rows and columns represent individual vertices and an element a_{ij} is 1 if vertex i is adjacent to vertex j and 0 otherwise. More precisely, its elements are defined as

$$a_{ij} = \begin{cases} 1 & \text{, if } j \in \Gamma(i) \\ 0 & \text{, if } j \notin \Gamma(i) \end{cases} \tag{2.12}$$

For undirected graphs, this is a symmetric matrix and the sum of the elements of a line or column equals the degree of the respective vertex. Notice that this representation is equivalent to consider $\Gamma(i)$ as an ordered tuple and list each set as lines in the same order they appear in $\Gamma(i)$.

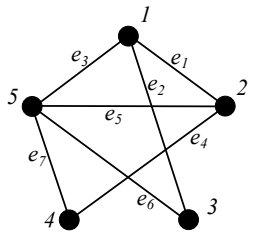


Figure 6 – An undirected graph with 5 vertices and 7 edges.

As an example, consider the graph shown on Fig. 6. Its adjacency matrix A is given by

$$A = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{array} \begin{array}{c} \begin{array}{ccccc} x_1 & x_2 & x_3 & x_4 & x_5 \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{pmatrix} \end{array}$$

Another useful representation is by an *incidence matrix*. Consider $n(V) = n$ and $n(E) = m$. The incidence matrix is a $n \times m$ matrix with rows representing the vertices and columns representing the edges. For undirected graphs this is a binary matrix and an element b_{ij} is 1 if vertex i is incident with edge j and 0 otherwise. For directed graphs b_{ij} can be -1, 0 or 1, depending on whether vertex i adjacent with edge j as a start vertex (1), end vertex (-1) or is not adjacent (0). The incident matrix for the graph shown in Fig. 6 is

$$B = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{array} \begin{array}{c} \begin{array}{ccccccc} e_1 & e_2 & e_3 & e_4 & e_5 & e_6 & e_7 \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 & 1 & 1 \end{pmatrix} \end{array}$$

These representations may become inefficient if the graph is *sparse*: the number of edges on the graph is comparable to the number of vertices. In this case, the matrix representations contain many zeros, being computationally costly. A more common way of representing a graph is using an *edges list*, which uses $2m$ data units, where m is the number of edges in the graph.

Other matrices of interest are the reachable matrix and reaching matrix. The former is defined as

$$R_{ij} = \begin{cases} 1 & \text{, if } j \in R(i) \\ 0 & \text{, if } j \notin R(i) \end{cases} \tag{2.13}$$

and the latter

$$Q_{ij} = \begin{cases} 1 & \text{, if } j \in Q(i) \\ 0 & \text{, if } j \notin Q(i). \end{cases} \tag{2.14}$$

They represent all vertices that can be reached (reach) by paths starting (ending) on vertex i. For undirected graphs $R = Q$. For a connected undirected graph all elements are 1. For disconnected undirected graph, the matrices can be put in the form of a diagonal block matrix, where each block represents a connected component of the graph.

2.5 Properties of graphs

The *degree distribution* $p(k)$ of a graph $G(V, E)$ is defined as the number of vertices of V which have degree equal to k,

$$p(k) = |\{v \in V | \delta(v) = k\}|. \tag{2.15}$$

We can normalize $p(k)$ and interpret it as the probability of randomly selecting a vertex with degree k. The degree distribution is usually represented as a histogram. In some cases this histogram is better visualized when displayed in a log-log scale, as for a power-law degree distribution.

The histogram of the degree distribution is a first step in network analysis, as the degree distribution is related to the mechanism of network evolution. For a completely random network, where the probability of an edge to be present in the network is the same for all edges, the degree distribution follows a Poisson distribution [49], as it shall be demonstrated latter in this chapter. But in general the degree distributions for real networks seldomly display this behavior [61]. This deviation from the expected behavior for a random network provides evidence for an underlying mathematical structure governing the formation of edges.

For an undirected graph $G(V, E)$, the *shortest path* between two vertices $v_1, v_2 \in V$ is the path from v_1 to v_2 with the least number of edges. The number of edges in the shortest path is the *shortest path distance*, $d(v_1, v_2)$. It is possible that there are several paths with the same distance d. If vertices v_1 and v_2 are in different components, $d(v_1, v_2) = \infty$. The *eccentricity* $e(v)$ of a vertex v if defined as the maximum shortest distance from v to any other vertex of the network,

$$e(v) = \max_{x \in V} \{d(v, x)\}. \tag{2.16}$$

The *radius* $r(G)$ of a graph $G(V, E)$ is the eccentricity of the vertex with lowest eccentricity in the graph,

$$r(G) = \min_{v \in V} e(v). \tag{2.17}$$

The *diameter* $diam(G)$ of a graph $G(V, E)$ is the eccentricity of the vertex with largest eccentricity in the graph,

$$diam(G) = \max_{v \in V} e(v). \tag{2.18}$$

The sum of the distances from a vertex to all other vertices in the network is called *distance sum*, $d_{sum}(v) = \sum_{u \in V} d(v, u)$. An important statistical measure for graph $G(V, E)$ is the average path length l,

$$l = \frac{\sum_{v \in V} d_{sum}(v)}{n(n+1)/2}, \tag{2.19}$$

where $n = n(V)$. Notice that in this definition we included the distance from a vertex to itself, which is zero. Excluding these distances from the mean is equivalent to multiply l by $(n-1)/(n+1)$. When the graph is disconnected, this definition leads to an infinite l. One solution is to consider only paths with finite distance when computing the mean. This can be avoided by changing the average path length definition to the harmonic mean of the distance sums,

$$l^{-1} = \frac{\sum_{v \in V} d_{sum}^{-1}(v)}{n(n+1)/2}. \tag{2.20}$$

Computing all-pair shortest paths for unweighted graphs can be done using a simple breadth-first search algorithm [62] (also known as 'burning algorithm' in physics), whereas for weighted graphs Dijkstra [63] or Bellman-Ford [64, 65] algorithms give the desired metric.

The *clustering coefficient* C introduced by Watts and Strogatz [1] measures the likelihood that an edge exists between two vertices connected to a given vertex. For example, it tell us the probability that two friends of a given person are also friends. It is defined as follows. Consider a vertex v of a graph $G(V, E)$ with degree k_v. Consider the subgraph V' formed by v and its k_v neighbors and all the n_e edges $e \in E$ between them. If this subgraph were complete, it would have $k_n(k_n - 1)/2$ edges. C_v defined as the ratio of the n_e existing edges and the total of the complete graph. The clustering coefficient is them the average of C_n over all vertices of the graph.

Another metric used in network analysis is the *clustering coefficient* $c(v)$ of a vertex, defined as

$$c(v) = \frac{\text{number of triangles connected to vertex } v}{\text{number of triples centered on vertex } v}. \tag{2.21}$$

Note that $c(v)$ is ill-defined if $\delta(v) = 0$ or 1. In this case, we set $c(v) = 0$. Hence, the clustering coefficient of a graph C is

$$C = \frac{1}{n} \sum_{v \in V} c(v), \tag{2.22}$$

where $n = n(V)$.

2.6 Calculating paths

There are several algorithms for calculating the distance between vertices on a graph. For unweighted simple graphs, the breadth first search algorithm [62] can be used to calculate the average path distance, diameter and radius of the graph. For weighted graphs, Dijkstra algorithm [63] can be used for the task, when weights are positive. For network with negative, weights the Bellman-Ford algorithm [66] can be used.

The breadth first search algorithm calculates the distance from a source vertex to any other reachable vertex of the graph. This algorithm can also be used to determine the number of connected components of an undirected graph. By applying the algorithm for every vertex in the graph, we can calculate the average path length l.

In order to explain the algorithm, we use the analogy of a forest fire, with vertices represeting trees. These trees can be in three states, namely, unburnt, burning or burnt, following the rules:

Listing 2.1 – Pseudocode describing the breadth-first search algorithm. G is a graph and $G.V$ is the vertex set of G. s is the source vertex. Status is a vertex property with string values unburnt, burning or burnt. Time is a vertex property given by an integer number. Parent is a vertex property indicating which adjacent vertex changed the status of the former. The shift operation remove the first element of a set.

```
for each vertex u ∈ G.V − {s}
 u.time = ∞
 u.status = 'unburnt'
␣u.parent␣=␣nil
s.status␣=␣'burning'
s.time = 0
s.parent = nil
burning = {s}
while burning ≠ {∅}
 u = burning.shift
 for each v in Γ(u)
  if v.status == 'unburnt'
␣␣␣v.status␣==␣'burning'
   v.time = u.time + 1
   v.parent = u
   burning << v
 u.status = 'burnt'
```

1. only unburnt trees can be ignited and become burning trees in the next time step;

2. at each time step, any burning tree ignites its adjacent unburnt trees and becomes burnt.

Hence, to calculate distances from any tree to a specific one (called the *source*), we consider that in the begining ($t = 0$) the source is burning and every other tree is unburnt. At each time step, we increase the time by 1 and apply the rules above. The algorithm stops when there are no burning trees in the begining of the time step. The time in which a tree becomes burnt is the desired distance. The pseudocode describing the algorithm is given in Listing 2.1.

2.7 Classical Random Graphs

The study of particular networks[2] may elucidate observed relatioships between topological and dynamical variables of the systems represented. For example, it may help us to identify the most critical transmission lines in a power grid [67] or to how to minimize the effect of genetic disorders [68]. Nonetheless, in many situations we do not have knowledge of the complete topology of the systems under consideration. Furthermore, it is also interesting to study general properties of similar systems, which evolve based on the same underlying principle; however, due to randomness, they develop into different networks. Concerning these subjects is the study of random graphs. A random graph is not a specific graph obtained by a random procedure but a statistical ensemble of such graphs, each with some realization probability. This is similar to the concept of ensemble in statistical physics, where we obtain thermodynamical properties by averaging that property over all

[2] From now on we will adopt the term network instead of graph when describing real systems, in order to emphasize that it is not only a mathematical construction.

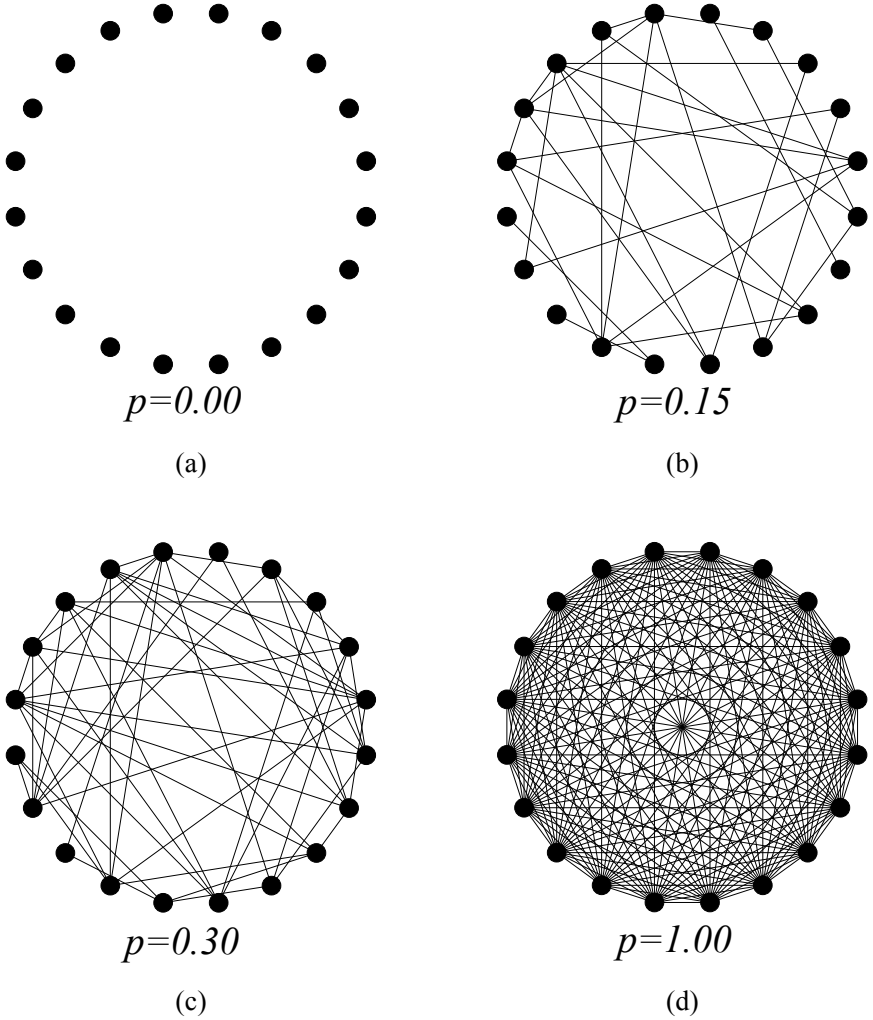


Figure 7 – Particular random graphs obtained using the $G_{N,p}$ model with $N = 20$ for different probabilities p: (a) $p = 0.00$, (b) $p = 0.15$, (c) $p = 0.30$ and (d) $p = 1.00$.

the systems comprising the ensemble. Once we define our random graph we may obtain properties which are shared by a large fraction of the ensemble elements.

During the 1950s, the study of random graphs was undertaken by several authors. The first authors to consider this subject were Solomonoff and Rapoport [69, 70]. The ensemble defined by them is known today as the Gilbert model $G_{N,p}$, due to Gilbert's latter description of the model in 1959 [71]. In this model we have a set of N vertices and a probability p that any two vertices are connected. To construct the ensemble elements, we start with N isolated vertices. For each pair of vertices, v_i and v_j, we add an edge (v_i, v_j) to $G_{N,p}$ with probability p. Thus, the existence of an edge is independent of any other. Using only these two parameters we form an ensemble of graphs. In Fig. 7 we show some random graphs obtained by different probabilities p.

By the end of the 1950's, Erdős and Rényi rediscovered the random graphs [49, 50], however, considering a different ensemble. Instead of considering a probability p, they studied all graphs with a fixed number of vertices and edges, which is labeled $G_{N,E}$. Note that their model include self-

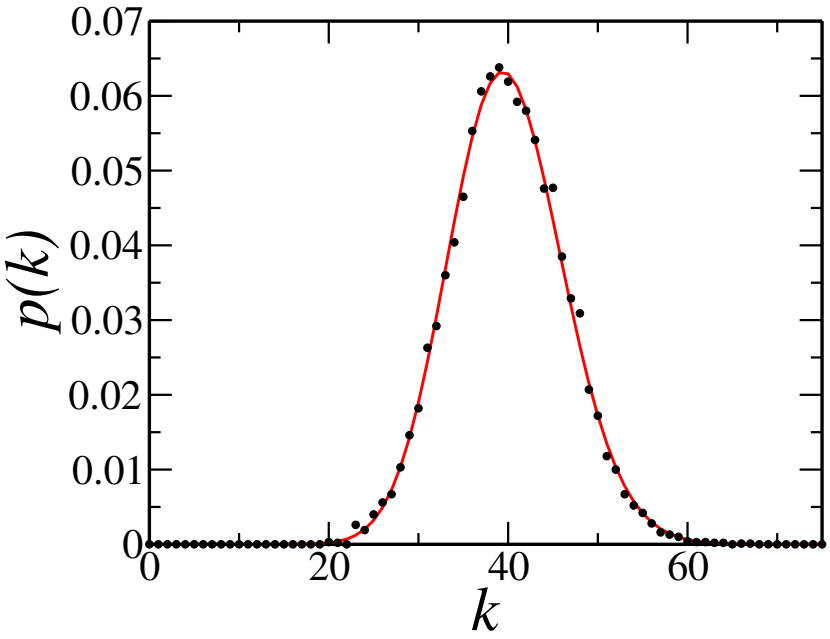


Figure 8 – Histogram of the degree distribution for a $G_{N,p}$ network with $N = 10000$ and $p = 0.004$. Dots represents the data points. The gray line is a fit using a Poisson distribution (Eq. 2.24). The parameter $z = 39.8898$ was obtained by maximum likelihood estimation, close to the expected value for the average degree, 39.996.

loops, which are not present in the $G_{N,p}$ model. Although these models are conceptually different, in the limit of large sparse graphs the difference can be neglected. Thus, in the following we shall discuss properties of the $G_{N,p}$ model.

For each vertex in $G_{N,p}$ there are $N - 1$ vertices to which it may be connected by an edge. Let us consider a vertex v with degree $\delta(v) = k$. There are $\binom{n-1}{k}$ ways of connecting v to the other vertices. Meanwhile the probability of v to be connected to exactly k other vertices is given by $p^k(1-p)^{n-1-k}$. Therefore, the probability distribution of degrees in the network is given by

$$P[\delta(v) = k] = \binom{n-1}{k} p^k (1-p)^{n-1-k}, \tag{2.23}$$

which is the binomial distribution. The average degree for such distribution is $z = p(n-1)$. In the limit of large N while keeping z constant, this distribution can be approximated by the Poisson distribution,

$$P(k) = \frac{e^{-z} z^k}{k!}. \tag{2.24}$$

Figure 8 shows the degree distribution for a network with 10000 vertices and $p = 0.004$.

Consider a vertex v in a $G_{N,p}$ random graph. Let k be the degree of v. The number of triangles containing v is $\binom{k}{2}$. The number of actual edges between neighbors of v is $p\binom{k}{2}$. Thus, $c(v) = p\binom{k}{2}/\binom{k}{2} = p$. The clustering coefficient of any vertex is equal to the probability p defining the

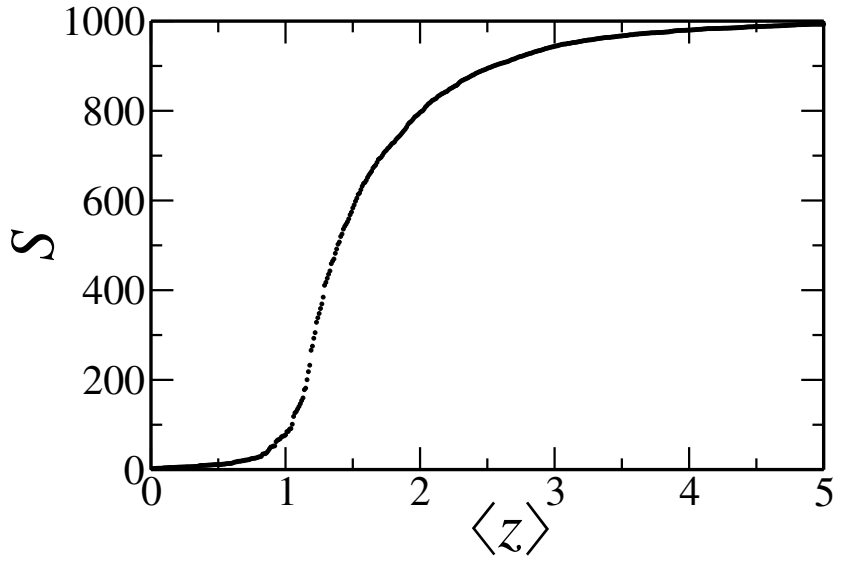


Figure 9 – $G_{1000,p}$ random graph with p varying from 0.00001 to 0.00500. Thus, the average degree of the networks varies from 0.001 to 5.00. For each value of p we have built 10 different networks. The size of the giant component for each of these was obtained and the average value was taken for each value of p.

graph. Hence, C is also equal to p. Hence, $G_{N,p}$ random graphs with a small linking probability also display a small degree of clustering.

The most striking result concerning $G_{N,p}$ random graphs is the emergence of a giant component. For small value of p the graph is disconnected, consisting of several small connected components. There is a critical value of p for which there is a non-zero probability for any given vertex to belong to this giant component. This critical value marks a phase transition from a low-density state, with many small components, to a high-density state with a very large component containing most of the vertices of the graph and the remaining vertices grouped in very small components. The critical value for this phase transition is $p = z/(N-1)$ [14]. For large N, this is equivalent to $z = 1$. Hence, we have the following behavior, depending on the value of z. For $z < 1$, there are only small components, the largest with size $s = O(\ln n)$. At $z = 1$, we have a critical point marking the emergence of a giant component with size $s = O(N^{2/3})$. At this point, the cluster size distribution follows a power-law with exponent $5/2$. Finally, for $z > 1$, there is a giant component with size $O(N)$. The second largest component has size $s = O(\ln n)$.

Figure 9 shows the emergence of the giant component. We have built $G_{1000,p}$ random graphs with p varying from 0.00001 to 0.00500. Thus, the average degree of the networks varies from 0.001 to 5.00. For each value of p we have built 10 different networks. The size of the giant component for each of these was obtained and the average value was taken for each value of p.

The average path length for ER random graphs is given by [72]

$$l = \frac{\ln N - \gamma}{\ln z} + \frac{1}{2}, \tag{2.25}$$

where γ is the Euler-Mascheroni constant ($\gamma \approx 0.5772$). We see that the average path length increases slowly, namely, with the logarithm of the size of the graph. As a result, even for large networks, vertices are at a small distance from each other.

2.8 Barabási-Albert Model

The degree distribution of real networks does not follow that observed in ER networks. Moreover, many real networks display a heavy-tailed degree distribution, which in some cases, agrees with a power-law distribution [61]. Such networks are known as *scale-free networks*.

In a seminal 1999 paper [52], Barabási and Albert studied the degree distribution of networks of actors, the world wide web (WWW) and the power-grid network. They found that these distributions were better represented by a power-law, instead of a Poisson distribution. In order to understand why different systems as these give rise to a similar behavior, they proposed a model today known as the Barabási-Albert (BA) Model, composed of two ingredients:

1. The networks evolve continuously by the addition of new vertices. This is not the case for the ER $G_{N,p}$ model in which the number of vertices is fixed from the beginning and unrelated to the formation of edges.

2. When new vertices are added, they are linked preferentially to vertices with a high degree.

Barabási and Albert [52] have shown through computational simulation that, by following these simple rules, a power-law degree distribution is obtained, with exponent 2.9. Real scale-free networks exhibit diverse exponents; nonetheless modifications of the BA model can produce different values. One distinctive feature of scale-free networks is the existence of hubs, namely, vertices with a very high degree. This is not observed in Poisson or Gaussian degree distributions due to their fast decrease for values distant from their mean. This feature is preserved by power-law distributions, which decay slowly compared to the former ones. Although successful in identifying a mathematical structure leading to observed degree distribution, the BA model fails to produce the clustering observed in social networks.

2.9 Small-world phenomenon

An ubiquitous characteristic of social networks is the small-world phenomenon. This corresponds to the surprising feeling we have when discovering that someone, far from our social group, knows one of our acquaintances. But the phenomenon is not limited to this; if we think of people in society as vertices in a social network and acquaintances linked by an edge, it consists in the notion

that the distance between any two persons in this network is awkwardly small. In the network of our society, everyone would be close to a movie start, to a politician or, even more surprisingly, to an undistinguished factory worker living in the suburb.

By the 1960's, there were two views on this matter. First it was believed that the small-world phenomenon was true for our society and everyone would be close, as stated in the last paragraph. The second view considered that it was also true but that not every pair of persons would have a path between them: people would be close to the ones in their community, but people from other communities could not be linked. In other words, in this view, the network of the society did not form a connected network: each connected component would be a community, isolated from each other.

In 1967, Stanley Milgram carried out an experiment to investigate the possibility of studying the small-world phenomenon [51]. The experiment consisted in asking subjects, sampled from men and women from all walks of life, to deliver a letter to a target person living somewhere in the United States. If the person did not know the target on a first-name basis, he or she was asked to send to one of his or her acquaintances which, by their judgment, were the most likely to know the target on a personal basis, with the same instructions. Along with the letter there was a roster on which everyone who received the letter should write his or her name down. This allowed tracking the number of persons the letter passed through until it was delivered to the target. As a side effect, it would also prevent the letter to loop around a group of acquaintances.

The result of the experiment was startling: the median of intermediate acquaintances for the letter to reach the target person was only five. Thus, between those persons there were six links, giving origin to the now widely known expression 'six degrees of separation'. Even the following criticism on the poor statistics and low response rate of the experiment did not reduce the importance of the work. Recently, the experiment was reproduced using e-mail instead of letters [73]. More than 60,000 e-mails users were asked to attempt to reach 18 target persons in 13 countries. The 348 completed tasks have an average of only 4.05 links.

2.10 Strogatz and Watts Model

The small-world phenomenon is not exclusive to the network of acquaintances. Many real networks exhibit a rather small average path length, which is linked to clustering of vertices: there is a high probability that neighbors of a vertex have an edge between them as compared to a random graph as the ER model. This is measured by the clustering coefficient, which for ER graphs is equal to the probability of two vertices sharing an edge, p. When modeling real networks, this value must be usually small, since it is related to the average degree of the network $z = p(N-1)$. Thus, ER graphs, although having a small average path length, do not reproduce the clustering observed in the real world. Nonetheless, clustering is not a sufficient condition to produce a small-world network. Regular lattices are highly clustered while exhibiting a high average path length.

Watts and Strogatz conceptualized a model (WS model) to reproduce these two features of real

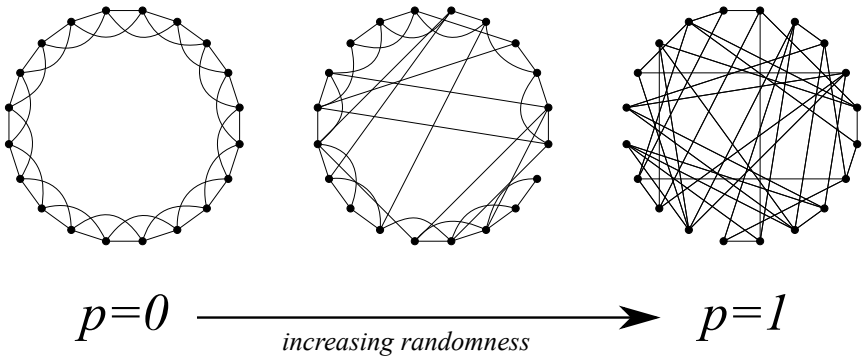


Figure 10 – The Watts-Strogatz model [1]. We start with a regular lattice, here a ring lattice with $N = 20$ vertices. Each vertex has degree 4, being adjacent to its nearest neighbors and second-nearest neighbors. Initially, the graph displays high clustering and high average path length, l. For each edge there is a probability p of changing one endpoint of the edge with the condition that we do not allow self-loops and multiple edges. The rewiring of edges creates shortcuts, causing l to diminish. If $p = 1$ we obtain a ER random graph, but with low clustering. For intermediate values of p the graph displays both small average path length and high degree of clustering.

networks [1]. Starting from a clustered regular lattice, the disorder observed in random networks can be obtained by rewiring some edges, changing one of the endpoints. This procedure, while keeping the clustering of the network, creates shortcuts between distant vertices on the original lattice, shrinking the average path length.

Consider a ring lattice, with n vertices and each vertex with k edges. Each edge has a probability p of being rewired at random. The p parameter acts as a controller for a continuous phase transition from an ordered state ($p = 0$) to an ER random graph ($p = 1$).

We show in Fig. 10 a ring lattice with $N = 20$ vertices. Each vertex has degree 4, being adjacent to its nearest neighbors and second-nearest neighbors. For $p = 0$, the graph displays high clustering and high average path length, l. With probability p, we change one endpoint of each edge (but not allowing self-loops and multiple edges). This rewiring procedure creates shortcuts between vertices in the graph, diminishing l. For $p = 1$ the WS model reproduces the ER model, which displays low clustering. The parameter p can be ajusted for the obtained graph to exhibit high clustering and low l.

2.11 Random graphs with specified degree sequence

The classical random graph models have a Poisson degree distribution which is not followed by real graphs. The BA preferential attachment model generates networks with power-law degree distributions. However, it is not clear if a pure power-law is the best representation for real empirical data [61, 74].

It is possible to generate graphs which do not follow the preceding degree distributions, but any desired p_k. This can be done by working with a specific degree sequence $\{k_i\}, i = 1, \ldots, n$, where n is the number of vertices in the network, chosen in such a way that the number of vertices with degree k tends to p_k in the limit of large n [75]. Such degree sequence can be obtained by numerically drawing random numbers from p_k.

The process to obtain a network can be pictured as giving sticks to the n vertices according to the degree sequence. Then, we chose two sticks uniformly at random and add an edge to the vertices owing them. Notice that this procedure allows self-loops. Also, the sum of the terms in the degree sequence generated must be even. This procedure defines an ensemble of graphs with degree distribution p_k.

Some properties of this model can be obtained in the limit of large n. In order to study the size of the largest component, we must focus our attention to the number of neighbors of a vertex at a specific distance. The average degree of the first neighbors of a vertex is $z = \langle k \rangle$. For second neighbors, we should in general subtract the average number of neighbors at distance 2 which are also at distance 1 from the vertex. In other words, we should take into account the clustering of vertices. However, we shall consider that the probability of two neighbors of a vertex to be joined by an edge goes as n^{-1}.

Now, to find out the average number of second neighbors, we must consider the degree distribution of a vertex reached by following an edge. This is not p_k since it is more likely that such edge belongs to a vertex with high degree. The actual distribution is kp_k. Furthermore, the edge we used to reach that vertex leads back to the former, hence not increasing its number of second neighbors. Thus, we are interested in the remaining degree distribution of a vertex reached by following an edge, q_k. The normalized distribution is

$$q_k = \frac{(k+1)p_{k+1}}{\sum_j j p_j}. \tag{2.26}$$

The average degree of this vertex is

$$\sum_{k=0}^{\infty} k q_k = \frac{\sum_{k=0}^{\infty} k(k+1)p_{k+1}}{\sum_j j p_j} = \frac{\sum_{k=0}^{\infty} k(k-1)p_k}{\sum_j j p_j} = \frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle}. \tag{2.27}$$

Multiplying this second neighbor average degree by the average number of first neighbors we obtain the average number of second neighbors, z_2,

$$z_2 = \langle k^2 \rangle - \langle k \rangle. \tag{2.28}$$

We can repeat this procedure for more distant neighbors, obtaining the following expression for the number of neighbors at distance m,

$$z_m = \frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle} z_{m-1} = \frac{z_2}{z_1} z_{m-1} = \left(\frac{z_2}{z_1} \right)^{m-1} z_1. \tag{2.29}$$

Thus, if z_2 if greater than z_1, z_m diverges exponentially as m increases, characterizing the existence of a giant component. Besides, if z_2 if less than z_1, z_m converges to a finite number and such

giant component do not exist (notice that we are considering the limit of infinite n). Thus, the network exhibits a phase transition at the point where $z_2 = z_1$. This condition can be rewritten as $\langle k^2 \rangle - 2\langle k \rangle = 0$ or

$$\sum_{k=0}^{\infty} k(k-2)p_k = 0, \tag{2.30}$$

known as Molloy-Reed criterion for the existence of a giant component [76].

Thus, the presence of a giant component is not unique to classical random graphs but expected for arbitrary degree distributions as long as the Molloy-Reed criterion is satisfied. Notice that if $z_2/z_1 \gg 1$, most of the vertices in the giant component will be away from a specific vertex v. Since they outnumber the immediate neighborhood of v, the average path length l to v is approximately the distance to these peripheral vertices. Setting $z_l \approx n$, we obtain

$$l \approx \frac{\ln(n/z_1)}{\ln(z_2/z_1)} + 1. \tag{2.31}$$

For example, for an ER graph, we have

$$z_1 = z, \tag{2.32}$$

$$z_2 = \langle k^2 \rangle - \langle k \rangle = z^2. \tag{2.33}$$

Then, in the limit of large n we have $l \approx \ln(n)/\ln(z)$, which agrees with the exact result (Eq. 2.25). This shows that the small-world effect is a common feature of random networks, since $l \sim \ln(n)$.

3 Lattes Collaboration Networks

3.1 Scientific collaborations

A traditional approach in studying scientific collaborations emerged in the second half of the twentieth century, with scientists from several fields of expertise focusing in papers with multiple authorship, also referred as co-authored papers.

The trend in multiple authorship and scientific collaboration have been studied since the beginning of the twentieth century. Nevertheless, it has not been considered a scientific fact, with criticism ranging from the universality of the results to the very concept of scientific collaboration. Smith [77] studied the trend of multiple authorship in psychology after WWII and showed that, while the fraction of single authored papers was diminishing, a larger fraction of papers were authored by two or three researchers. Although Smith initially considered the decrease in single authored papers could represent a shift of interest of the academic community to problems which require multiple researchers, the partitioning in research topics showed that the multiple authorship trend was consistent. Price [78], studying papers in the field of Chemistry ranging from 1910-1960, concluded that a steadily increase in multiple authorships can be observed since the beginning of the 20th century. Clarke [79] also studied research articles in the same period as Smith, though in biomedicine. This study showed evidence of two regimes separated by the war, with the trend of multiple authorships existing before the war. Clarke criticized the generalization made by Price, and speculated that multiple authorships saturate at different levels on different fields. Nonetheless, more recent studies confirm the tendency of multiple authorships [80].

This field dependence investigated by Frame and Carpenter [9] for international collaborations. They concluded that the more basic the field (in this sense, earth sciences, physics, mathematics, and biomedical research) more collaborations were made as compared to applied or clinical fields (psychology, clinical medicine, engineering, and biology). Also, basic fields received more government support, although more productive countries (as measured by paper output) displayed a smaller fraction of international collaborations.

Several reasons for collaborating are outlined by authors. Katz and Martin [3] considered that the competition for research founding, the lower costs of travel and communication, the institutional organization of science, the specialization and consequential division of labour and the advent of interdisciplinary fields are sociological reasons for collaborations. Beaver and Rosen argued that collaborations emerged as a response to the professionalization of science, rooted in the french scientific community of the early nineteenth century. “[. . .] professionalization defines the rules, rights and rites of access to the group, what holds the members of the group together, and what sets them apart from other individuals in the larger society.” [5, p. 66]. In their view, collaboration can be seem as a means of gaining and sustaining access to this professional community [5, p. 68]. Wray [81] argued that a strictly sociological explanation as Beaver and Rosen’s fails to uncover the differences

between fields. He adopted a functional explanation [82], in which collaborations contribute to the realization of the epistemic goals of the scientific community while it persists because its efficient in doing so. Bozeman and Corley [33] studied scientific collaboration strategies, focusing on the choices made by researchers involved in collaboration. Using data from 451 researchers, they found that those with larger grants have more collaborators and are more prone to work with researchers from different expertises.

The role of geographical distance between researchers was investigated by Katz [30], considering intra-national university-university collaborations within United Kingdom, Canada and Australia. From these data, it was found that the frequency of collaborations decreased exponentially with the distance between the universities. Pan, Kaski e Fortunato [32] studied collaborations on a city and country level. For cities, they found that the fraction of collaborations within the city increased with its size. They also showed that the probability of collaboration between two cities or two countries decayed as a power law of the distance between them. Ponds [31] found that collaborations were more localized when the involved organizations are of different kinds.

Finally, Heffner [10] studied research funding and found a positive correlation between the number of authors and the research support. Nonetheless, when partitioning the collaborations between technical support and theoretical support, he found that only technical support has an impact on funding.

3.2 Co-authorship versus collaboration

The identification of co-authorship as collaboration is not free of criticism. Katz and Martin argued that

> For co-authorship to be a truly accurate reflection of collaboration, it would require that, in all cases where the 'level' or intensity of joint work by collaborating researchers was above a certain minimum threshold, a jointly authored paper always resulted. […] Conversely, if the level of working together of a number of scientists was below this minimum threshold, they would never appear as co-authors of a publication. Having expressed it in this way, one can immediately appreciate how unrealistic such a criterion would be. Therefore, co-authorship can never be more than a rather imperfect or partial indicator of research collaboration between individuals. [3, p. 8]

Laudel noticed that, when considering collaborations, is imperative to be 'either working with implicit definitions or formulating an explicit (albeit incomplete) definition.' [4, p. 4]. Laudel chooses the latter, nevertheless opting for a broad definition of collaboration: "A research collaboration is defined as a system of research activities by several actors related in a functional way and coordinated to attain a research goal corresponding with these actors." [4, p. 32]. This definition does not include a shared research goal as necessary for a collaboration to take place. Considering this definition, an empirical investigation based on interviews of the members of 57 German research groups showed that many forms of collaboration are invisible when analysing co-authorship. Further, about

one third of the collaborations that resulted in a publication were rewarded only with acknowledgements. Nevertheless, the forms of collaboration labeled 'without reward' involved preparation of samples, access to equipment and laboratories and other technical work, which possibly did not involve a common research goal. Also, when a collaborator took part in the division of labour on the research activities, he was most likely rewarded as a co-author. In this study, from a total of 242 acknowledgements, only in one case the partner was involved in division of labour.

Melin and Persson [83] also discussed the relationship between co-authorship and collaboration. From a study at Umeå University, they concluded that only 5% of authors had collaborations that did not resulted in a co-authored paper, usually for having done minor contributions to the research. We thus conclude that, considering that collaboration involves a common research goal and division of labour, co-authorship and collaboration are safe to be used as interchangeable terms.

Another approach in studying affiliation networks like collaboration networks is through questionnaires and interviews of the subjects. For a high number of authors, this option turns counterproductive. Besides providing an objective way of measuring collaborations, the study of co-authorships allows the use of the extense bibliometrical data available in the Internet, avoiding contacting each subject and thus providing a larger number of collaborations and better statistical accuracy [24].

3.3 The network approach to scientific collaboration

While co-authorship networks have previously being studied in other fields, in 2000's physicists turned their attention to these networks. Frequently, such networks are constructed with authors being the nodes of the network and a link exists between two authors if there is a paper which both co-authored. Thus, self-loops are not present. A weight might be assigned to an edge according to a specific weighting scheme. These networks can be represented as simple sparse (un-)weighted graphs.

In a series of two papers [24, 25], Newman used publicly available databases of papers to construct scientific collaboration networks, comprising research in physics, biomedical science and computer science. Using computational techniques, he analysed and compared these databases. The construction of the networks was done by linking coauthors of scientific papers, a procedure which became standard in network science.

In the first paper [24], he found that the average number of papers per author varies with each discipline, where high-energy physics has the highest count. Also, analysing the distributions of the number of authors per number of papers, he found fat-tailed distributions, roughly resembling power-laws. High-energy physics also seems to be the most collaborative discipline, with the highest count of authors per paper. The distributions of the number of collaborator of the authors display some curvature, but not in agreement with the two power-law regimes suggested by the BA model [52]. Most disciplines presented a very large giant component, with 80%+ researchers belonging to it. The exception was the computer science community, although this may be an artifact

of the construction of the database. The clustering coefficients for such networks were shown to be relatively large, ranging from 0.3 to as large as 0.7. This indicated that collaborations of three individuals are quite common in scientific works.

In the second paper [25], non-local properties of collaboration networks were analysed, namely, shortest path distance, diameter of the network and betweeness centrality. Authors were found to be close to any other author, with the scientific community forming a "small world". It has shown that funneling occurs in these networks: of all shortest paths from a researcher to another, most pass through only one or two of the first neighbors.

Barrat *et al.* [16] studied the influence of recurring collaborations. From the preprints relative to condensed matter from arXiv (the same database used by Newman, as discussed previously), a weighted collaboration network was built, using a weighting scheme in which the contribution of a paper to the weight depends inversely to the number authors in that particular paper [25]. Both the distributions of the strength of the nodes and of the weights of the edges were found to display a heavy-tail.

The evolution of collaboration networks was a topic of interest of several authors. Pioneering these studies, Barabási *et al.* [23] presented a model to mimic the network evolution. The analytic solution of the model presents a power-law degree distribution. They argued that this preferential attachment rule together with evolution are the basic ingredients to explain the heavy tailed degree distributions found on real networks.

Several studies of evolution of regional and/or field specific collaboration networks have been conducted [84–91]. The role of geography was also a topic of interest for researchers of scientific collaboration. In a study with data from three countries, Katz [30] found that the frequency of collaborations between universities inside a country decayed exponentially with distance. Pan *el al.* [32] studied citation and collaborations between cities and countries. They found that these obeyed gravity laws, with a power-law dependency on distance. Even more interesting, they have found that the impact of the research of a country were linearly dependent on the national funding, although only countries which invested more than 100'000 UDS per researcher annually were above the world citations average.

3.4 Lattes Platform

The Lattes Curriculum (http://lattes.cnpq.br) is the Brazilian standard national scientific curriculum, being compulsory for those requiring financial support from the Brazilian government and related founding agencies for scientific and technological research. It was developed by CNPq (National Council for Scientific and Technological Development), a Brazilian organization under the Ministry of Science and Technology. Launched in Brazil in August 1999, it was translated to Spanish in 2002 and implemented in several South American countries (Colombia, Chile, Peru and Argentina). It is also present in Portugal and Mozambique.

It features a data-rich individual scientific curriculum containing detailed information concern-

ing the researchers. In addition to the publication list, which includes scientific papers, books, audiovisual production and patents, several other data are available, such as: professional address, academic records, scholarships, alma mater, research topics, research projects, idioms and academic advising. Until June 2012, gender information was also available. Most of these data are self-reported, thus depend on updates provided by the researchers themselves.

CNPq classifies research topics in terms of a hierarchical structure. On the top level there are presently 10 major areas: Agricultural Sciences (Agr), Applied Social Sciences (Soc), Biological Sciences (Bio), Exact and Earth Sciences (Exa), Humanities (HuM), Health Sciences (Hea), Engineering (Eng), Linguistics and Arts (Lin), Others (Oth) and Technologies (Tec)[1].

3.5 Data acquisition and parsing

The collaboration networks are built based on data of approximately 2.7 million curricula downloaded in June 2012 from the Lattes Platform website. Its advanced search functionality allows negative searches. A search for all researchers without numbers on their names was performed. From the page results HTML's, the links for the individual CV's were obtained. A robot written in Ruby programming language accessed and saved locally all the individual curricula. Currently, automated download is rather more difficult due to the implementation of a CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) when accessing any curriculum.

Curricula data are stored in XML (Extensible Markup Language) files, which are data containers and not intended for direct visualization. An application translates these files to XHTML, an application of XML, for browser visualization. The source code of these pages is stored locally. XML documents can be represented as digraphs without loops, i.e., they form tree structures. Vertices are called elements. If there is a direct edge from element a to element b, we call a the *parent* of element b and b a *child* of element a. As the graph must be a direct tree, every element may have several children, however only one parent. Elements with the same parent are called *siblings*. Also, the tree must be connected. Thus, there must be one and only one element without parent, which we call the *root* element. In XHTML this is the <html> element. Four elements are mandatory in XHTML documents: <html>, <head>, <title> and <body>, where <head> and <body> are children of <html> and <title> is a children of <head>. The structure of a HTML document is shown in Fig. 11.

In XML documents, elements may have attributes, which are included in the element tag. For example, we may have an <article> element with a title attribute. This is represented as <article title="“title of the article”>. The quoted string is the *value* of the attribute. Elements may also have text associated, which is written after the element tag. Any element <element> must be always closed, either by using the associated closing tag </element> after its text field or, if there is no text associated with such element, by including a slash character before the closing bracket: <element

[1] The latter major area was not present on June 2012 when we acquired the data.

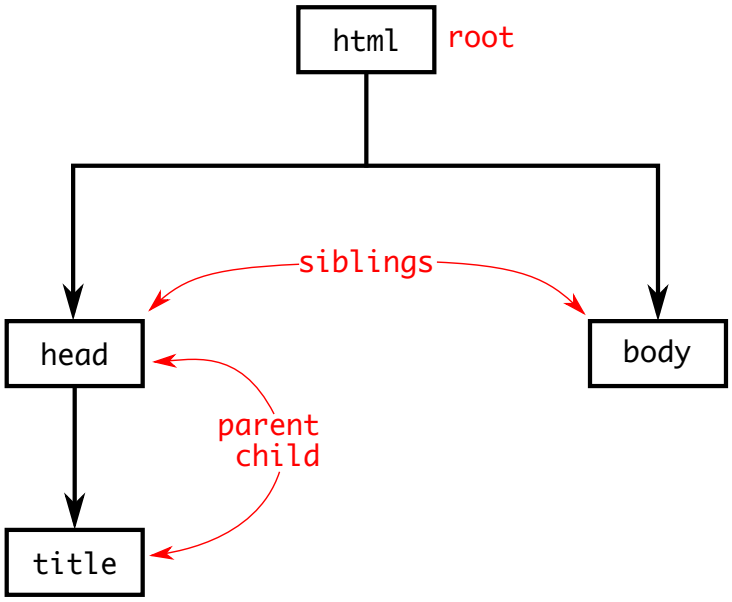


Figure 11 – Document tree of an xhtml document. <html> is the root element, with two children: <head> and <body>. <head> has a child <title>.

Listing 3.1 – A XHTML document example.

```
<html>
  <head>
    <title>Hello World!</title>
  </head>
  <body>
    content of the page.
  </body>
</html>
```

attribute="attribute"/>. Elements must be properly nested, i.e., they must be closed in the reverse order they were opened: <a>.

These properties make XHTML documents easier to parse, compared to standard HTML documents. An existing XML parser can be used for this task, for instance the REXML Library for Ruby language. In this thesis we developed our own parser.

In Lattes Platform implementation, elements containing information concerning articles published by the researcher can be identified by the specific value *informacao-artigo* of the *class* attribute from elements. The parent encodes the sought information across its children. Author information (name, gender, professional address and field of expertise) and paper information (list of authors, title, journal and year of publication) are extracted from either text elements or attributes of elements matched by regular expressions. The parser developed identifies these data. We notice that all paper information were stored in a single string. They were separated by splitting the string on the period mark. For each researcher, we saved the corresponding list of papers, giving each paper a unique identifier. Each entry has a unique identifier, the year of the publication, the number of authors and title. Even though there is a DOI (digital object identifier, http://www.doi.org) attribute which could be used to identify articles, its adoption is not universal

Table 1 – Some relevant elements in the Lattes CV XML files and some corresponding attributes.

Element	Attributes
CURRICULO-VITAE	DATA-ATUALIZACAO, NUMERO-IDENTIFICADOR
DADOS-GERAIS	PAIS-DE-NASCIMENTO, NOME-COMPLETO, PAIS-DE-NASCIMENTO, UF-NASCIMENTO, CIDADE-NASCIMENTO
ENDERECO-PROFISSIONAL	NOME-INSTITUICAO-EMPRESA, PAIS, UF, CIDADE
GRADUACAO	TITULO-DO-TRABALHO-DE-CONCLUSAO-DE-CURSO, NOME-DO-ORIENTADOR, NOME-INSTITUICAO, NOME-CURSO, STATUS-DO-CURSO, ANO-DE-INICIO, ANO-DE-CONCLUSAO, FLAG-BOLSA, CODIGO-AGENCIA-FINANCIADORA, NOME-AGENCIA
IDIOMA	IDIOMA, DESCRICAO-DO-IDIOMA, PROFICIENCIA-DE-LEITURA, PROFICIENCIA-DE-FALA, PROFICIENCIA-DE-ESCRITA, PROFICIENCIA-DE-COMPREENSAO
DADOS-BASICOS-DO-ARTIGO	NATUREZA, TITULO-DO-ARTIGO, ANO-DO-ARTIGO, IDIOMA
DETALHAMENTO-DO-ARTIGO	TITULO-DO-PERIODICO-OU-REVISTA, ISSN, VOLUME, PAGINA-INICIAL, PAGINA-FINAL

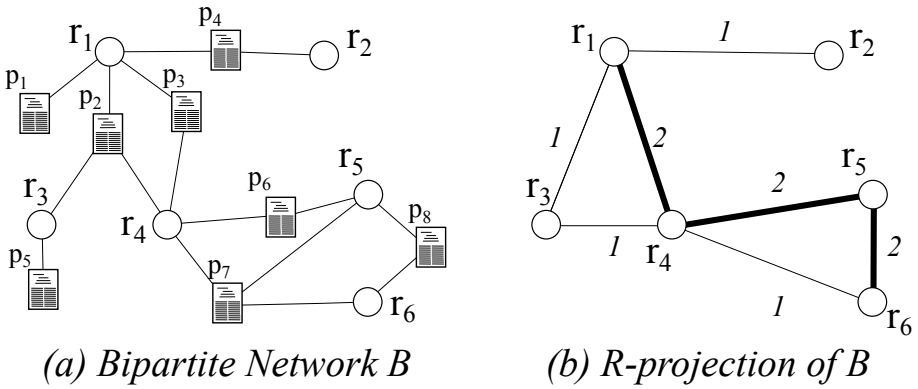
and were not reliable. Nowadays, raw data is available in XML format but, as already mentioned, access is restricted due to a CAPTCHA implementation.

The XML documents used to generate the XHTML CV webpage of a researcher are now available for download in Lattes Platform. They provide more reliable data, since every class of information has its own element or attribute. As stated before, when translated to XHTML, the paper title, authors, journal title and year of publication are merged into a single string, separated by periods, acting as the text of an element. Retrieving this information is not error free, as when the paper title contains one or more periods. The original XML files are easier to parse. Nonetheless, some information are not included in these files, e.g., if the researcher currently has a scholarship. A summary of the structure of relevant elements is given in Listing 3.2. A summary of relevant attributes of some of the elements is listed in Table 1.

Listing 3.2 – Lattes XML structure of some relevant elements. Attributes are not shown.

```
<CURRICULO–VITAE>
    <DADOS–GERAIS>
        <RESUMO–CV/>
        <ENDERECO>
            <ENDERECO–PROFISSIONAL/>
        </ENDERECO>
        <FORMACAO–ACADEMICA–TITULACAO>
            <GRADUACAO/>
            <ESPECIALIZACAO/>
        </FORMACAO–ACADEMICA–TITULACAO>
        <ATUACOES–PROFISSIONAIS>
            <ATUACAO–PROFISSIONAL>
                <VINCULOS/>
            </ATUACAO–PROFISSIONAL>
        </ATUACOES–PROFISSIONAIS>
        <AREAS–DE–ATUACAO>
            <AREA–DE–ATUACAO>
        </AREAS–DE–ATUACAO>
        <IDIOMAS>
            <IDIOMA>
        </IDIOMAS>
    </DADOS–GERAIS>
    <PRODUCAO–BIBLIOGRAFICA>
        <TRABALHOS–EM–EVENTOS>
            <TRABALHO–EM–EVENTO>
                <DADOS–BASICOS–DO–TRABALHO/>
                <DETALHAMENTO–DO–TRABALHO/>
                <AUTORES/>
            </TRABALHO–EM–EVENTO>
        </TRABALHOS–EM–EVENTOS>
        <ARTIGOS–PUBLICADOS>
            <ARTIGO–PUBLICADO>
                <DADOS–BASICOS–DO–ARTIGO/>
                <DETALHAMENTO–DO–ARTIGO/>
                <AUTORES/>
                <AREAS–DO–CONHECIMENTO>
                    <AREA–DO–CONHECIMENTO–1/>
                    <AREA–DO–CONHECIMENTO–2/>
                    <AREA–DO–CONHECIMENTO–3/>
                </AREAS–DO–CONHECIMENTO>
            </ARTIGO–PUBLICADO>
        </ARTIGOS–PUBLICADO>
    </PRODUCAO–BIBLIOGRAFICA>
    <PRODUCAO–TECNICA>

    </PRODUCAO–TECNICA>
    <DADOS–COMPLEMENTARES>
        <PARTICIPACAO–EM–EVENTOS–CONGRESSOS>
            <DADOS–BASICOS–DA–PARTICIPACAO–EM–CONGRESSO/>
            <DETALAHAMENTO–DA–PARTICIPACAO–EM–CONGRESSO/>
            <PARTICIPANTE–DE–EVENTOS–CONGRESSOS/>
        </PARTICIPACAO–EM–EVENTOS–CONGRESSOS>
    </DADOS–COMPLEMENTARES>
</CURRICULO–VITAE>
```



(a) Bipartite Network B (b) R-projection of B

Figure 12 – (a) Bipartite network B containing node classes R and P representing researchers (circles) and papers (rectangles), respectively. (b) R-projection of B, where researchers are connected if the share a paper in B. The weight of the link is given by the number of shared papers.

3.6 Building a collaboration network

The procedure to construct the collaboration networks is based on a previous application of collaboration networks to the Lattes Database [91] and is carried out as follows:

1. A list containing all papers from all authors is then constructed concatenating the list of the individual authors. The list contains many duplicates as papers are listed in each co-author curriculum. We may restrict our analysis to some years of publications to obtain a collaboration network for a specific time window;

2. In order to identify collaborations, we identify the duplicates comparing title, year of publication and number of authors. Due to possible typographical errors [92], an approximate string matching is used to compare paper titles. We use Damereau-Levenshtein distance [93] as the metric and compare papers of the same year and with the same number of authors starting with the same letter. By comparing paper titles, papers differing by 10% or less of the maximum distance are considered to be the same paper;

3. From the string matching results, we build a unweighted bipartite network B, with node classes R and P, representing researchers and papers, respectively (see Fig. 12a). A researcher r_i in R is connected to a paper p_i in P if r_i is identified as one of the authors of p_i in the former procedure. Nodes store the information parsed previously: r_i contains gender, fields of expertise, professional address and scholarships information, while p_i contains title, number of coauthors and year;

4. A collaboration network containing only researchers may be obtained by projecting the bipartite network on the vertex set. In this operation, a weight may be assigned according to the number of papers that two researchers appear as co-authors.

Table 2 – Cost and example of operations in Damereau-Levenshtein Algorithm implementation used in this work.

Operation	Initial String		Final String	Cost
Insertion	cat	→	cats	1
Deletion	cats	→	cat	1
Substitution	bat	→	cat	1
Swap	act	→	cat	1

The Damereau-Levenshtein algorithm [93] used is an approximated string matching algorithm, which assigns an edit distance between two strings s_a and s_b. A distance between strings s_1 and s_2 is defined as the minimum number of character insertion, deletion, substitution of a single character or swap of adjacent character to transform s_1 into s_2. In our implementation, all operations have the same cost, 1. Table 2 exemplifies each operation. The maximum distance between two strings is then the size of the largest one. The algorithm was implemented in C++ programming language, for better performance.

3.7 The Total Collaboration Network

We focus our study on a projection of the bipartite network onto R. There are many ways to accomplish this [94], the simplest being to project B onto an unweighted undirected network, with researchers r_i and r_j connected if both are connected to a paper p_k in B (see Fig. 12b). We used this method to construct a cumulative network containing collaborations of all researchers in the database, the Total Collaboration Network (TCN). One should note that, with this database, we are not limited to the simple projecting scheme, since information on researchers and papers can be used in the projection. In order to illustrate this procedure, we show in Fig. 13 a network constructed only with researchers working on fields of Medicine in the state of São Paulo and with a grant from the Brazilian government. We did the projection in such way that the edges are directed, pointing to the researcher with the earliest date of publication of a paper. Unless noted otherwise, all the network projections analysed in this work are unweighted and undirected.

TCN includes 275,061 researchers, with 90.4% belonging to the largest component. The total number of identified papers written in collaboration is 623,984, the number of collaborations is 1,095,871 and the network comprises all 8 major fields used by the Brazilian agency CNPq to classify researchers.

The extracted papers have publication date extending for several decades, the oldest paper in collaboration being from 1949. By analysing the growth of the network, we show in Fig. 16 (left) that the number of researchers (s_r) as well as collaborations (s_c) grew exponentially in the last three decades, $s_r \propto e^{0.139t}$ and $s_c \propto e^{0.181t}$, with t in years. We also show that the number of collaborations increases superlinearly with the number of researchers in the network. This accelerated growth

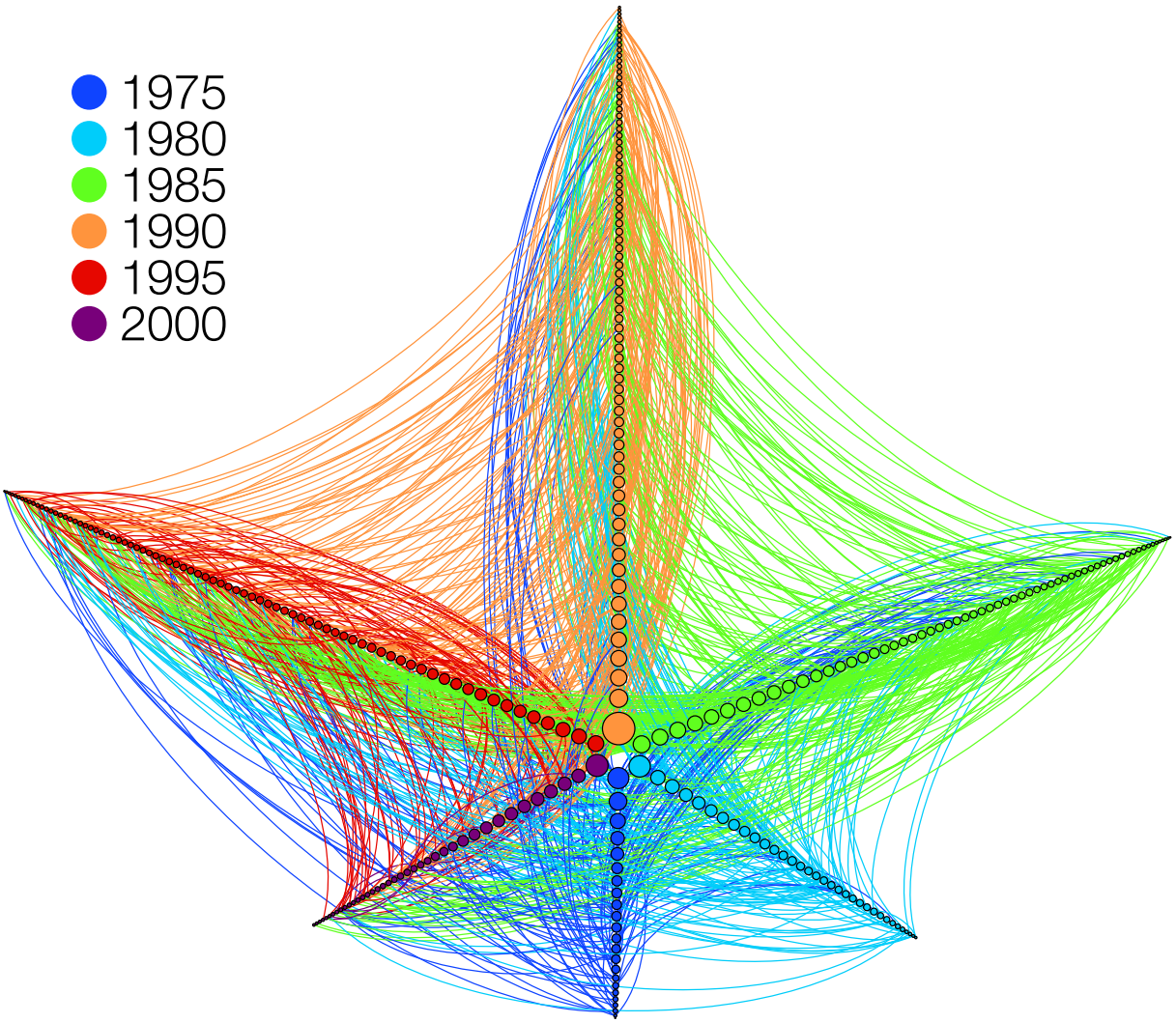


Figure 13 – Sample network extracted from the collected Lattes Database. Links shown are between researchers (nodes) who were granted a scholarship and working in fields of Medicine in the state of São Paulo. Node size is proportional to the degree of the researcher in the whole database. Researchers were grouped according to the year of their first published paper. The first cohort (dark blue) comprises all researchers who published their first paper before 1975. Each subsequent one, in the counterclockwise direction, comprises researchers who published within 5 years from the previous one, up to 2000. The edges are directed, colored according to the most senior.

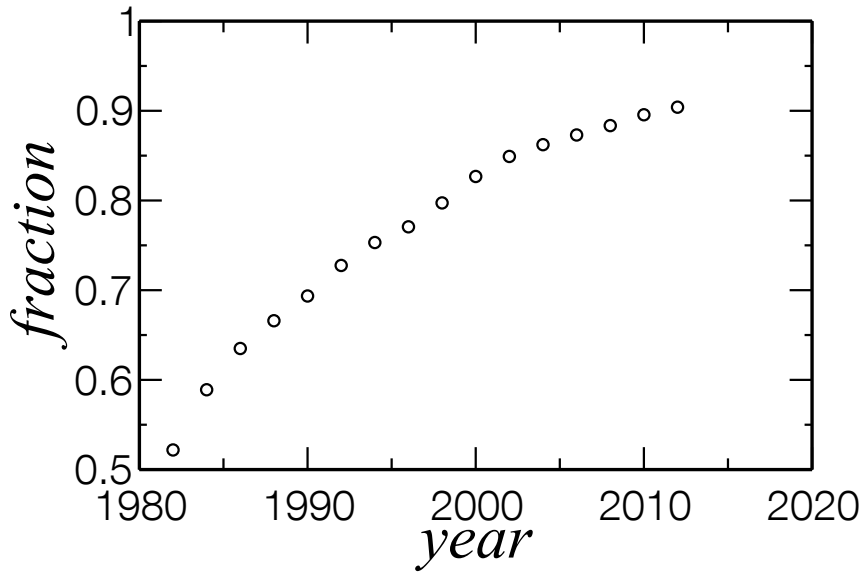


Figure 14 – Evolution of the fraction of the giant component of TCN since 1982. For every two years, the respective cumulative network was produced and the number of researchers in the largest component was divided by the number of researchers in collaboration in that year.

has been observed in collaboration networks [23, 95] and other types of empirical networks [96]. More recently, it was shown that the number of social contacts and total communication also scales superlinearly with city population size [97].

To analyse the evolution of the largest component of the network, we constructed networks comprising papers published until 1980 and for every 2 years onward. As shown in Fig. 14, the fraction of researchers belonging to the largest component increases from slightly over 50% to 90% in 2012. The remaining components are much smaller, the second largest being formed by 45 researchers. The remaining component sizes are distributed as a power-law, $p(s) = As^{-\lambda s}$ (see in Fig. 15), with $\lambda = 3.770 \pm 0.126$.

It is possible that many researchers in the cumulative networks retired from scientific publishing in recent years. This suggests that the high connectivity of TCN, with the largest component comprising 90% of the researchers, might be misleading. To investigate this possibility, we constructed networks with a limited time window spanning five years centered in 1990, 1995, 2000, 2005 and 2010. This was accomplished projecting the bipartite network linking researchers connected to papers published only within the respective time window. Figure 17 shows an increase in the largest component fraction over years, with a fraction 84.9% of researchers in the last data point. For this time window, we obtained the fraction of each field, shown in Table 3, indicating that fields are mixed in the largest component in the same proportion as in the complete network.

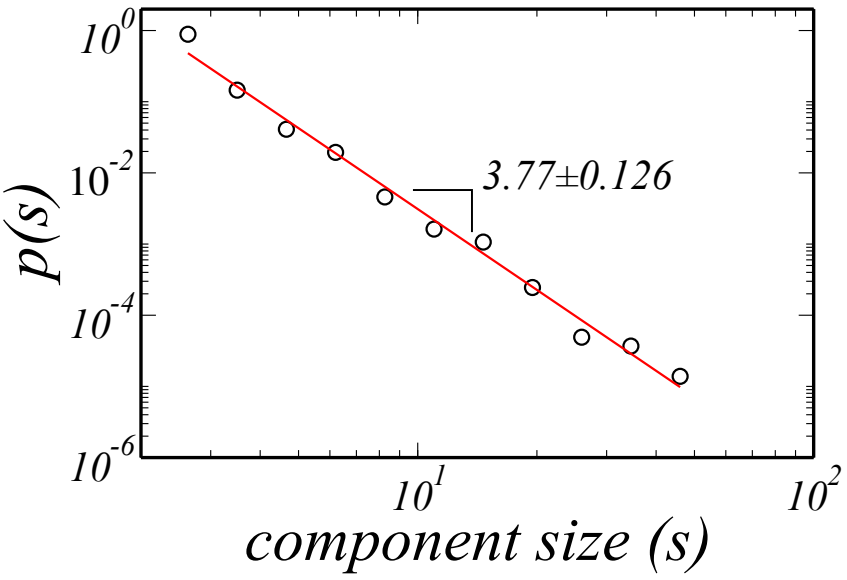


Figure 15 – Distribution of component sizes s in the TCN, excluding the giant component. Remaining component sizes are distributed as a power-law $p(s) = Ae^{-\lambda s}$, with $\lambda = 3.770 \pm 0.126$. The red line is the best power-law fit.

Table 3 – Fraction of fields in the last 5 years. The network was constructed by projecting the bipartite network onto a network containing only researchers connected if they share a paper published in the last 5 years. Sum of fractions is not 100% because the field information is not available for all researchers.

Field	fraction in largest component	fraction in the network
Agr	13.9%	12.2%
Bio	18.0%	15.8%
Hea	26.3%	24.1%
Exa	13.0%	12.3%
Hum	5.9%	8.9%
Soc	5.1%	7.3%
Eng	6.5%	6.5%
Lin	0.5%	1.8%

The fact that more than 80% of the network is connected together with the field distribution is an interesting sign, which indicates that discoveries from a field can spread in the communities through interdisciplinary collaborations. As this last network is a subgraph of TCN, most of the links in the latter were active in the last 5 years.

A commendable initiative of the Brazilian government is to award scholarships to distinguished researchers among their peers. Researchers with Ph.D. may apply for several levels of scholarship.

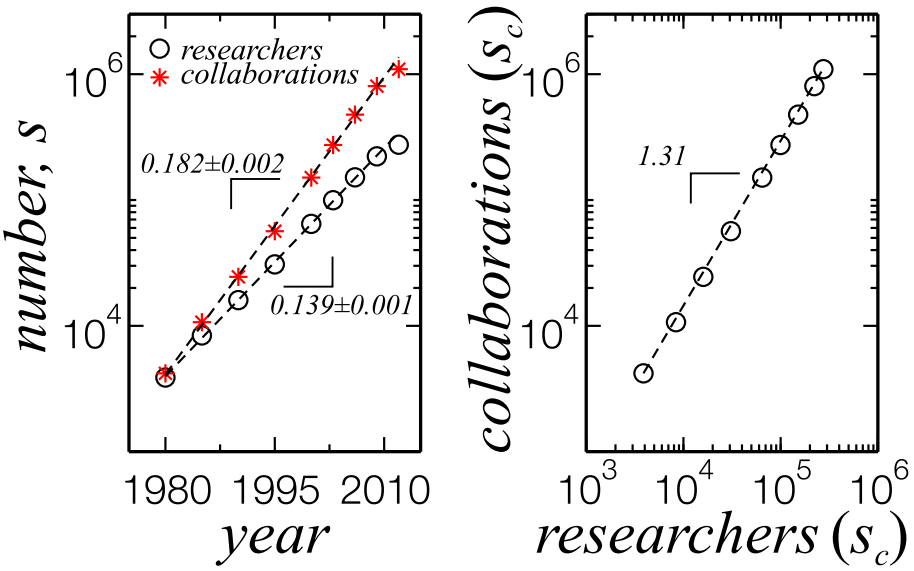


Figure 16 – Left: Number of researchers with published papers (black circles) and collaborations between them (red stars) present in the cumulative collaboration network. Dashed lines are exponential fits in the form $s = ae^{\alpha t}$ up to 2009, seen as straight lines in the linear-log plot. The coefficient α is shown in the picture for each curve. Deviations of the 2012 data points from the exponential fit are due to the early acquisition of the curricula, in June of 2012. Right: Superlinear scaling of the number of collaborations with the number of researchers. Dashed line is a power-law curve with exponent $\alpha_c/\alpha_r = 1.31$.

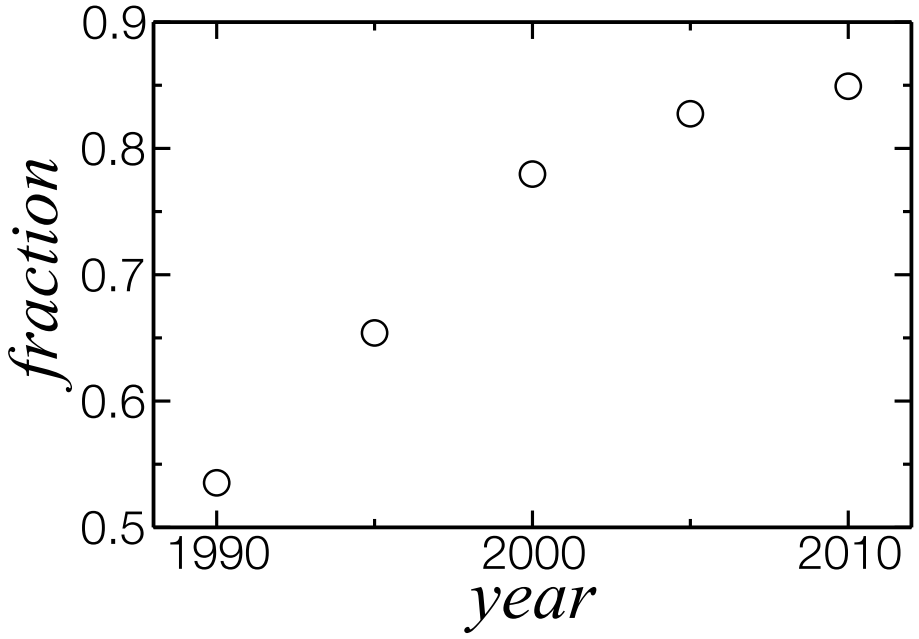


Figure 17 – Evolution of the largest component. Data points represent the fraction of researchers present in the largest component for a five year time window centered in the respective year. More than 80% of the researchers engaged in collaborations in the last 5 years are in the largest component. They represent 61% of the researchers in TCN.

Table 4 – Statistics for the networks studied in this work.

Property	TCN	SCN
Number of researchers (s_r)	275,061	12,302
Number of edges (s_c)	1,095,871	67,093
Total number of papers	623,984	129,699
Average researchers per paper	4.51	5.26
Average papers per author ($\langle n \rangle$)	11.1	61.4
Average number of collaborators ($\langle k \rangle$)	8.0	38.1
Largest component fraction	90.4%	94.6%
Clustering coefficient (C)	0.465	0.266
Assortativity coefficient (r)	0.094	0.230
Network radius R (largest component)	17	12
Network Diameter D	33	24
Average path length	6.46	5.44

Applications are judged by a committee based on the requestor's project, scientific contributions, participation as a journal editor, among other criteria. These scholarships correspond to a bonus payment in addition to their base salary. The scholarship information is included in the CV by CNPq, not by the researcher, and we obtain the list of researchers awarded when parsing their curricula. For comparison with the TCN, we built a collaboration network with only these researchers, projecting the bipartite network B onto R connecting only awarded researchers with shared papers on B, which we call the Scholarship Collaboration Network (SCN). SCN is therefore a subgraph of TCN. In Table 4 we show the basic statistical properties of TCN and SCN.

As already explained in second chapter, the clustering coefficient [1], C, measures the probability that two collaborators of a given researcher have papers in common (forming a triangle in the graph). Social networks are known to have high degree of clustering [14], which can be explained in terms of a hierarchical structure [15]. Here both networks display a high clustering coefficient but the average value for SCN is about half of TCN. This difference reflects the higher position in the research groups of the researchers with scholarship. They are more likely to have contacts in other research groups, which means being less clustered. We probed the scaling of the clustering coefficient with the researcher degree for both the TCN and the SCN, as shown in Fig. 18. Both decay as power-laws ($C(k) = Ak^{-\sigma}$) with exponents $\sigma = 0.71 \pm 0.009$ and $\sigma = 0.57 \pm 0.020$, respectively.

It is inviting to verify if the production of researchers on Lattes Platform obeys Lotka's Law. As shown in Fig. 19, the distribution of scientific production (in number of papers, n) obeys a power-law with exponential cutoff, $P(n) = A_p n^{-\beta_p} e^{-n/l_p}$, with exponent $\beta_p \approx 1.58$ and characteristic cutoff length $l_p \approx 129$. We estimate the parameters of the distributions using the Levenberg-Marquardt Algorithm (LMA). The value for onset of the power-law behavior is found by minimizing the the Kolmogorov-Smirnoff (KS) distance, defined as

$$D_{data} = |S(x) - P(x)|_{max},$$

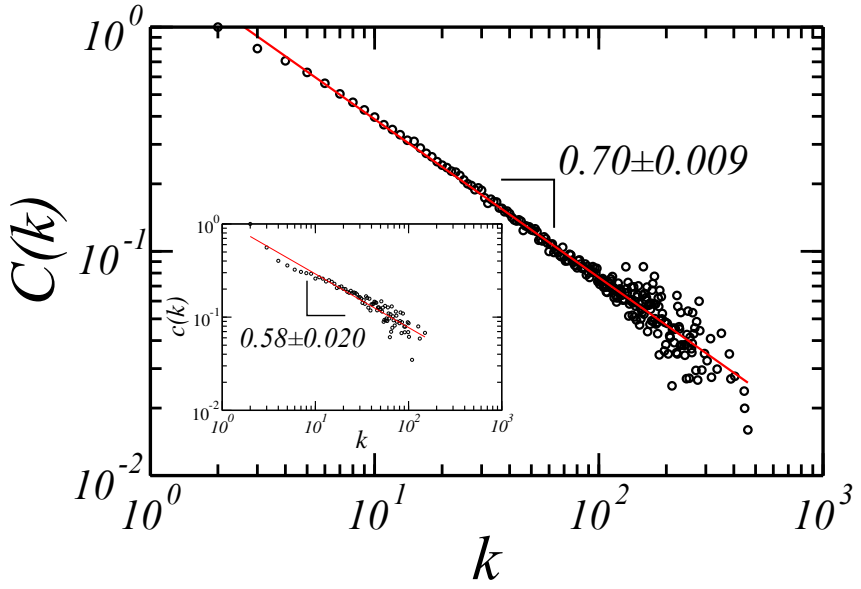


Figure 18 – Scaling of $C(k)$ with k for the TCN (main graph) and the SCN (inset). The red lines are power-law fits of the respective data ($C(k) = Ak^{-\sigma}$). For the TCN, $\sigma = 0.71 \pm 0.009$. For the SCN $\sigma = 0.58 \pm 0.020$.

where $S(x)$ is the data cumulative distribution and $P(x)$ is the power-law with exponential cutoff cumulative distribution, given by (for $\alpha > 1$)

$$P(x) = -C\left(\frac{\lambda^{\alpha}\Gamma(\lceil\alpha\rceil - \alpha, \lambda x)}{\prod_{j=1}^{\lfloor\alpha\rfloor}(j-\alpha)} - e^{-\lambda x}x^{1-\alpha}\sum_{j=1}^{\lfloor\alpha\rfloor}\frac{(\lambda x)^{j-1}}{\prod_{k=1}^{j}(k-\alpha)}\right),$$

where

$$C = \left(\frac{\lambda^{\alpha}\Gamma(\lceil\alpha\rceil - \alpha, \lambda x_{min})}{\prod_{j=1}^{\lfloor\alpha\rfloor}(j-\alpha)} - e^{-\lambda x_{min}}x_{min}^{1-\alpha}\sum_{j=1}^{\lfloor\alpha\rfloor}\frac{(\lambda x_{min})^{j-1}}{\prod_{k=1}^{j}(k-\alpha)}\right)^{-1}.$$

Once we have determined the onset of the power-law behavior, we test the hypothesized distribution by sampling from $P(x)$ with parameters given by the LMA fit and calculate the KS distance between the sample and its own fit (D_{sample}). For each fit we have sampled 1000 times. The fraction $(D_{sample} > D_{data})/1000$ is the p-value of our fit. The p-value is the probability of obtaining a distance as extreme as D given $P(x)$. Low values of the p-value ($p < 0.01$ or $p < 0.05$) constitute evidence against the hypothesized model, and, in such case, we should reject it.

A relevant question which naturally arises is how the scientific productivity and collaboration statistics of researchers awarded with scholarships differ from regular researchers. Studying our database, we find that researchers in the SCN represent less than 5% of the researchers in the TCN but contribute with 20% of the production. They are in average more than five times more productive, as measured by publication output. Also, SCN is more dense than TCN, as measured by the

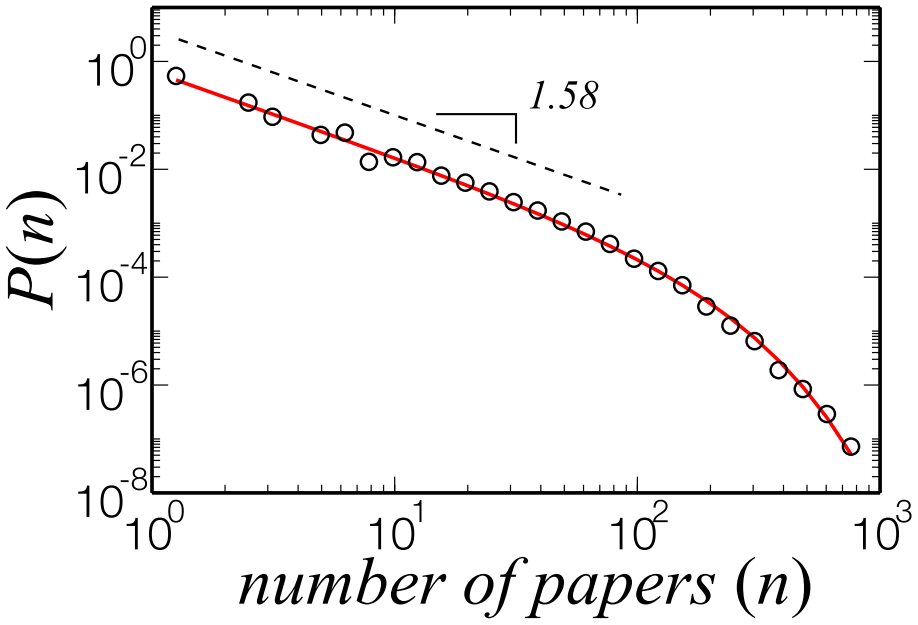


Figure 19 – Distribution of scientific production of researchers belonging to the TCN group. The solid red line is the best fit to the data points of a power-law with exponential cutoff, $P(n) = A_p n^{-\beta_p} e^{-n/l_p}$, where $\beta_p = 1.58$ and $l_p = 129$. The dashed black line is a power-law with exponent -1.58.

size of the giant component. To determine whether these characteristics are cause or consequence of their scholarship is not our aim, but previous research on collaborations strategies indicate that those with higher grants are more likely to have more collaborators [33]. The degree distributions shown in Fig. 20 clearly corroborate this difference between groups.

The assortativity coefficient [13], r, measures the correlation between degrees of nodes at either ends of an edge. Networks with $r < 0$ are said to display disassortative mixing, while $r > 0$ means assortative mixing. Social networks, including collaborations networks, are known to display assortative mixing [13, 16]. Another way of looking at the assortative properties of a network is through the average nearest-neighbor degree, $k_{nn}(k)$ [17], where k is the number of collaborators of a researcher. This measures how well connected the collaborators of a researcher are. If $k_{nn}(k)$ is an increasing function, then researchers with high k collaborate with other well-connected researchers, and the network displays assortative mixing. We show in Fig. 21 that this occurs in TCN, and that k_{nn} increases logarithmically with k. Assuming that researchers with a high number of collaborators are positioned in the top of the academic hierarchy, we can infer from Fig. 21 that prominent researchers and group leaders collaborate more among themselves. Nonetheless, k_{nn} does not grow fast but logarithmically, as researchers growing in importance absorb the influx of new actors in the network.

With this database, we can study the time evolution of the cumulative collaboration network by analysing different groups of papers that have been published within a specific range of years. We show in Fig. 22 (a) the evolution of the distribution of the number of collaborators in TCN, from 1980 to 2012. We show in Fig. 22 (b) a rescaling of these curves by the relative number of

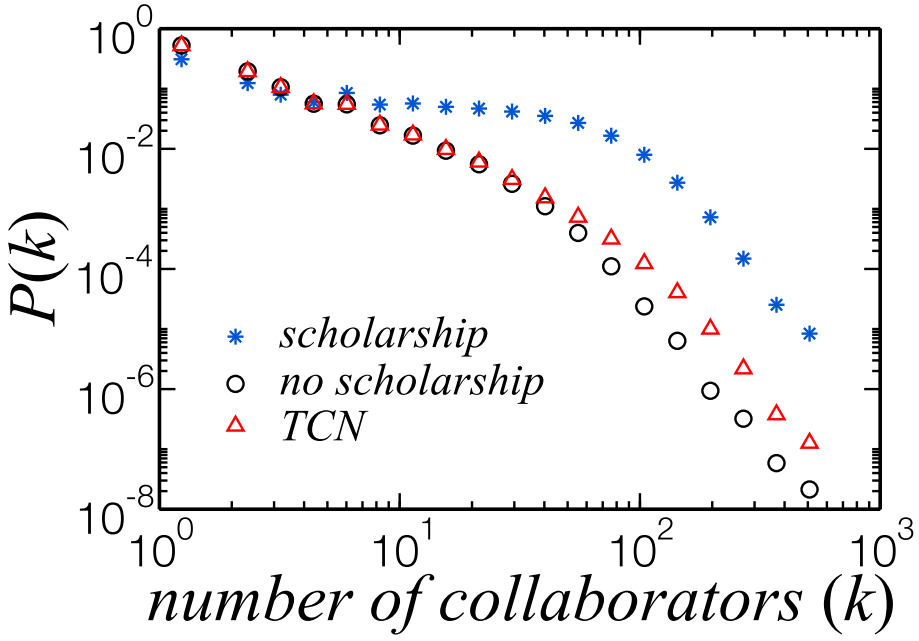


Figure 20 – Normalized distribution of the number of collaborators (k) of researchers with scholarship (blue stars), without (black circles) and for the TCN (red triangles). The distribution for researchers with scholarship decreases slowly up to one hundred collaborators, although most of them still have a small number of collaborators. The higher proportion of researchers with high k might reflect the CNPq policy of considering the proponent's participation in research groups, international immersion and human resources development to grant the scholarship.

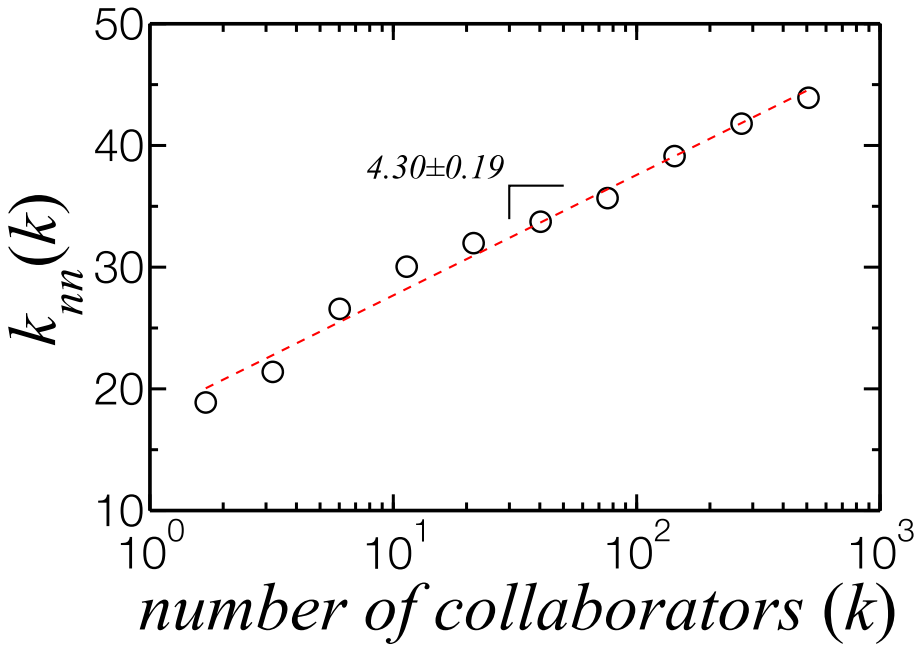


Figure 21 – Variation of the average nearest-neighbor degree (k_{nn}) with k. Being an increasing function of k, the network displays assortative mixing. Researchers with high k are more likely to collaborate with other well connected researchers. This tendency, however, increases logarithmically with k, as indicated by the regression fit (dashed line).

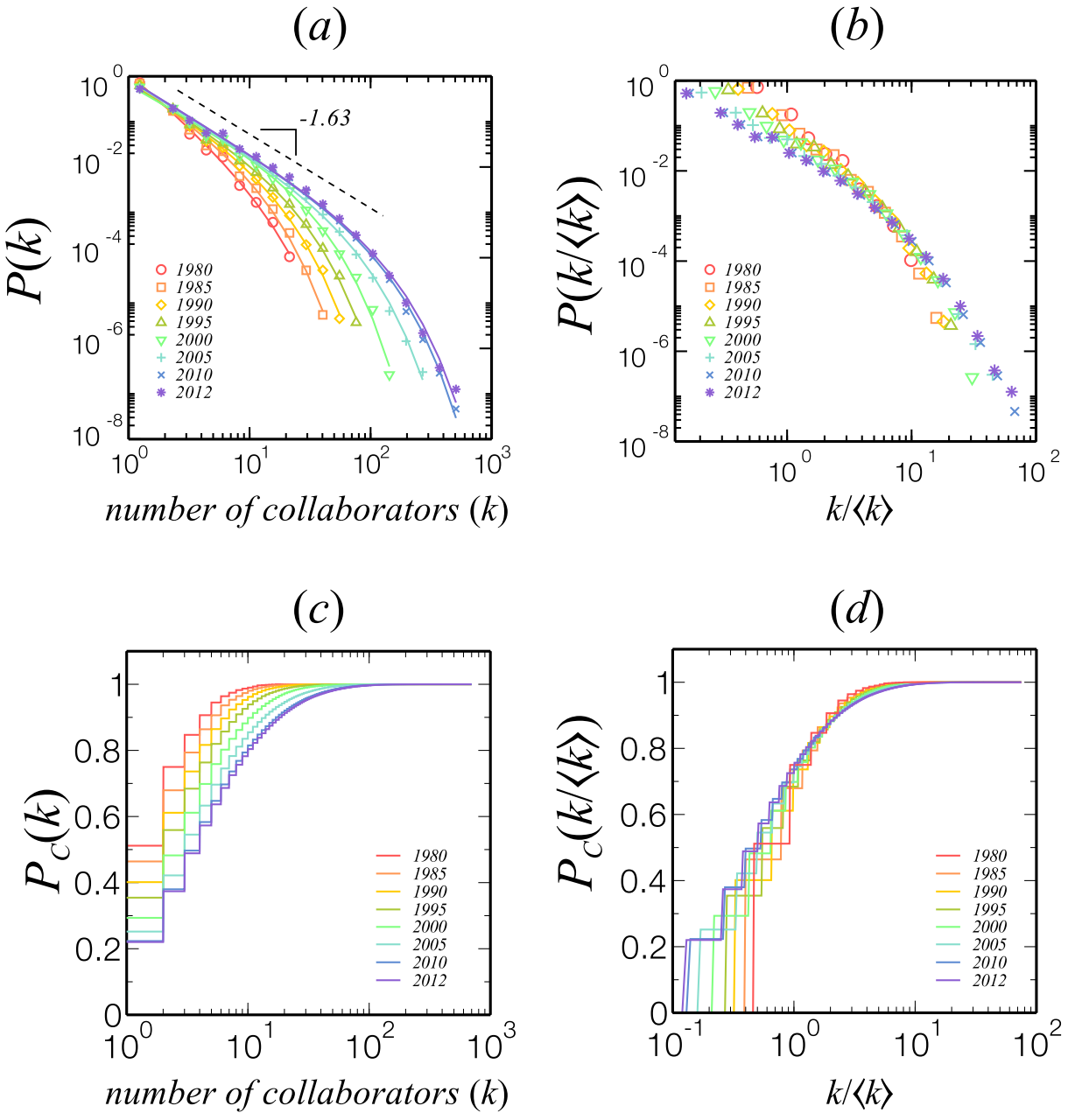


Figure 22 – (a) Time evolution of the distribution of the number of collaborators in the TCN. (b) Rescaling the distribution in (a) by the relative number of collaborators for each year shows a collapse onto a single curve. We also show the respective cumulative distributions in (c) and (d). As the network ages, the fraction of researchers with high k increases (c), but the evolution of the network shows that the distribution is constrained to the average production (d).

collaborators for each year, collapsing onto a single curve. Figs. 22 (c) and (d) show the respective cumulative distributions. Although the cumulative distribution varies with year, with the increase of highly connect researchers, this distribution is constrained to the average number of collaborators of TCN (d).

We can use the professional address information included in the curricula to study the differences

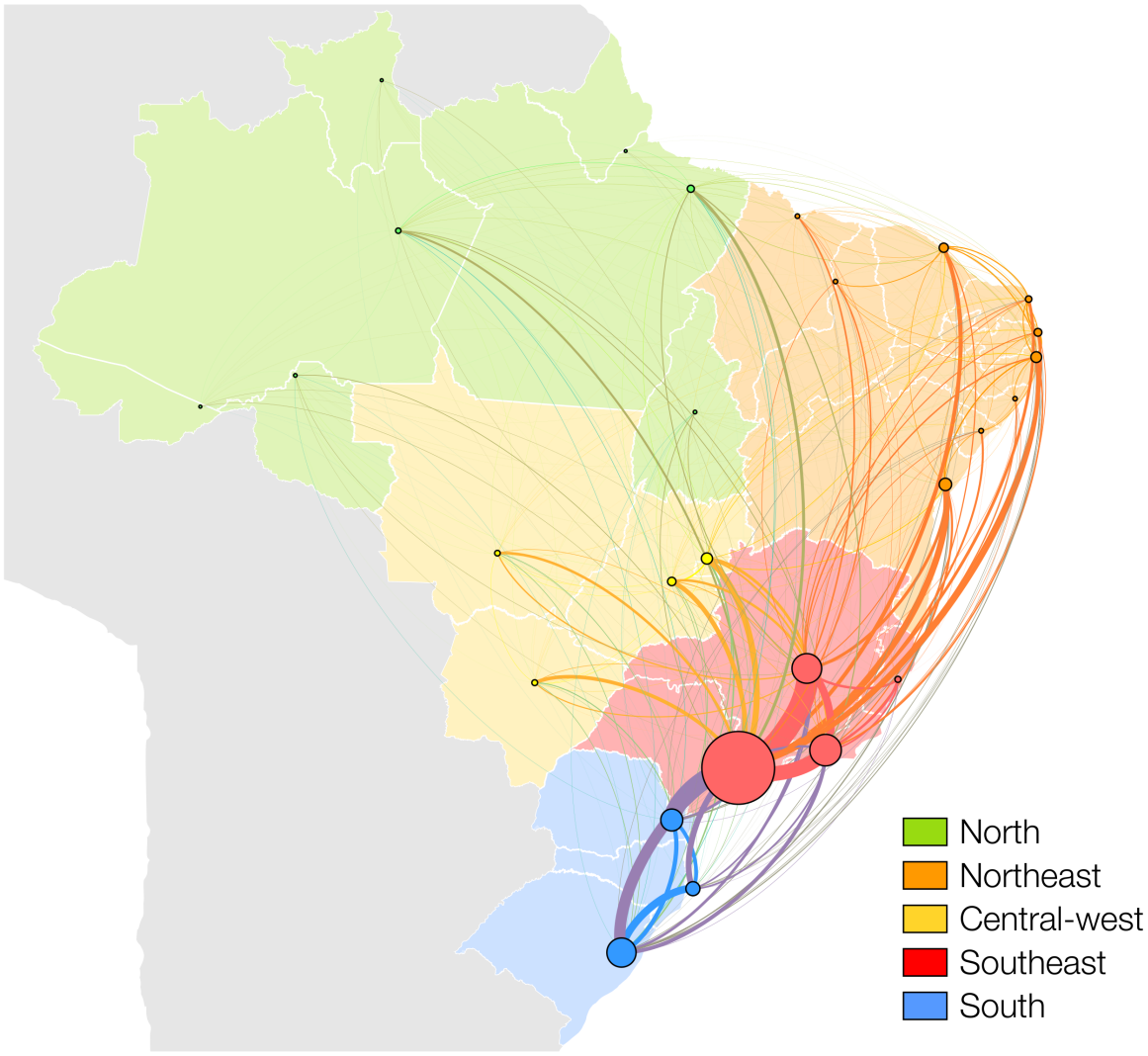


Figure 23 – Interstate collaborations obtained from the Lattes Database. Vertices radii are proportional to the fraction of researchers in TCN. Edges are proportional to the total number of collaborations. Southeast states concentrate most of the collaborations, specially São Paulo.

of collaboration profile due to geographical location. In Fig. 23 we use CV information concerning the professional address to construct a network of the Brazilian states and the Federal District. São Paulo concentrates most of the researchers (vertices radii are proportional to the number of researchers in the state), and for all states most of its collaborations involve peers in São Paulo.

We now focus on the degree distributions of researchers working on each state. As shown in Fig. 24 (top), the overlap of the degree distributions for the TCN at each of the 26 states of Brazil and Brasília, the Federal District, suggests universality in the collaboration mechanism. The geographical location of the researcher, while not changing the shape of the distribution, is correlated with the spectrum of the number of collaborators. Recent allometric studies show that a large number of urban indicators (e.g., R&D employment, total wages, GDP, gasoline sales, length of electrical cables) scale as a power-law of population of the city [98]. In Fig. 24 (bottom) we show that the

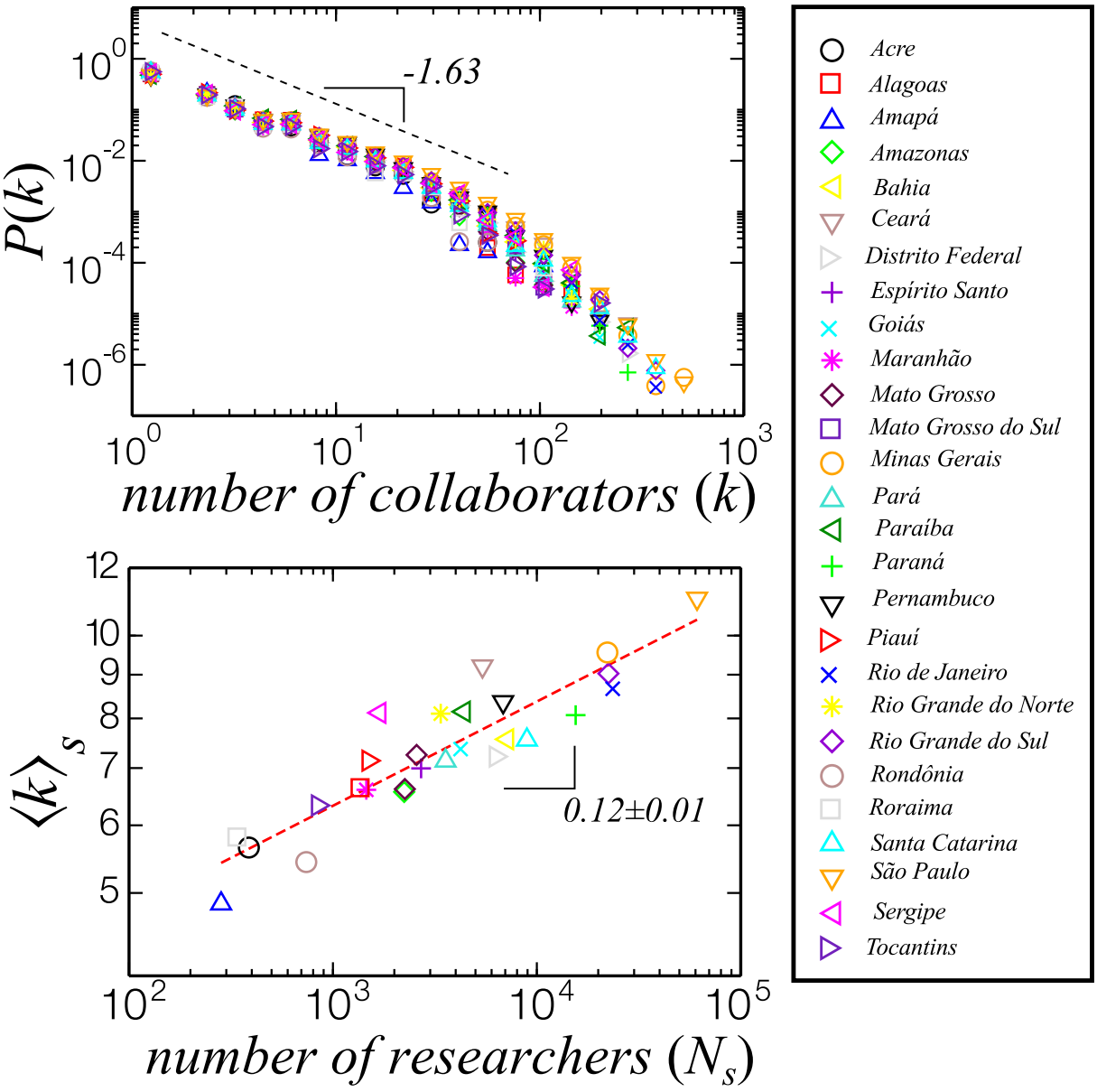


Figure 24 – Top: Distribution of number of collaborators in the TCN for the 26 Brazilian states and the Federal District. The distributions display the same behavior as the TCN (Fig. 22). The dashed line is guide for the eye in the form of a power-law with exponent 1.63 ($P(k) = Ak^{-1.63}$). Bottom: the average number of collaborators versus the number of researchers in each state. The circles correspond to the results for 26 Brazilian states and the Federal District. The dashed line is the best fit obtained by linear regression of the data to a power-law $\langle k \rangle_s \sim N_s^{\delta}$ in logarithmic scale, with exponent $\delta = 0.12 \pm 0.01$.

average number of collaborators per researcher in the Brazilian states $\langle k \rangle_s$ generally increases with their number of researchers as a power-law, $\langle k \rangle_s \sim N_s^{\delta}$ with an exponent $\delta = 0.12 \pm 0.01$.

Finally, the way researchers from different fields collaborate can also be investigated with the data downloaded from the Lattes platform. Fig. 25(a) and (b) show that the cumulative distributions of researcher productivity $P_C(n)$ as well as their corresponding degree distributions $P_C(k)$, respectively, can be rather different for distinct fields. However, since different fields are known to have

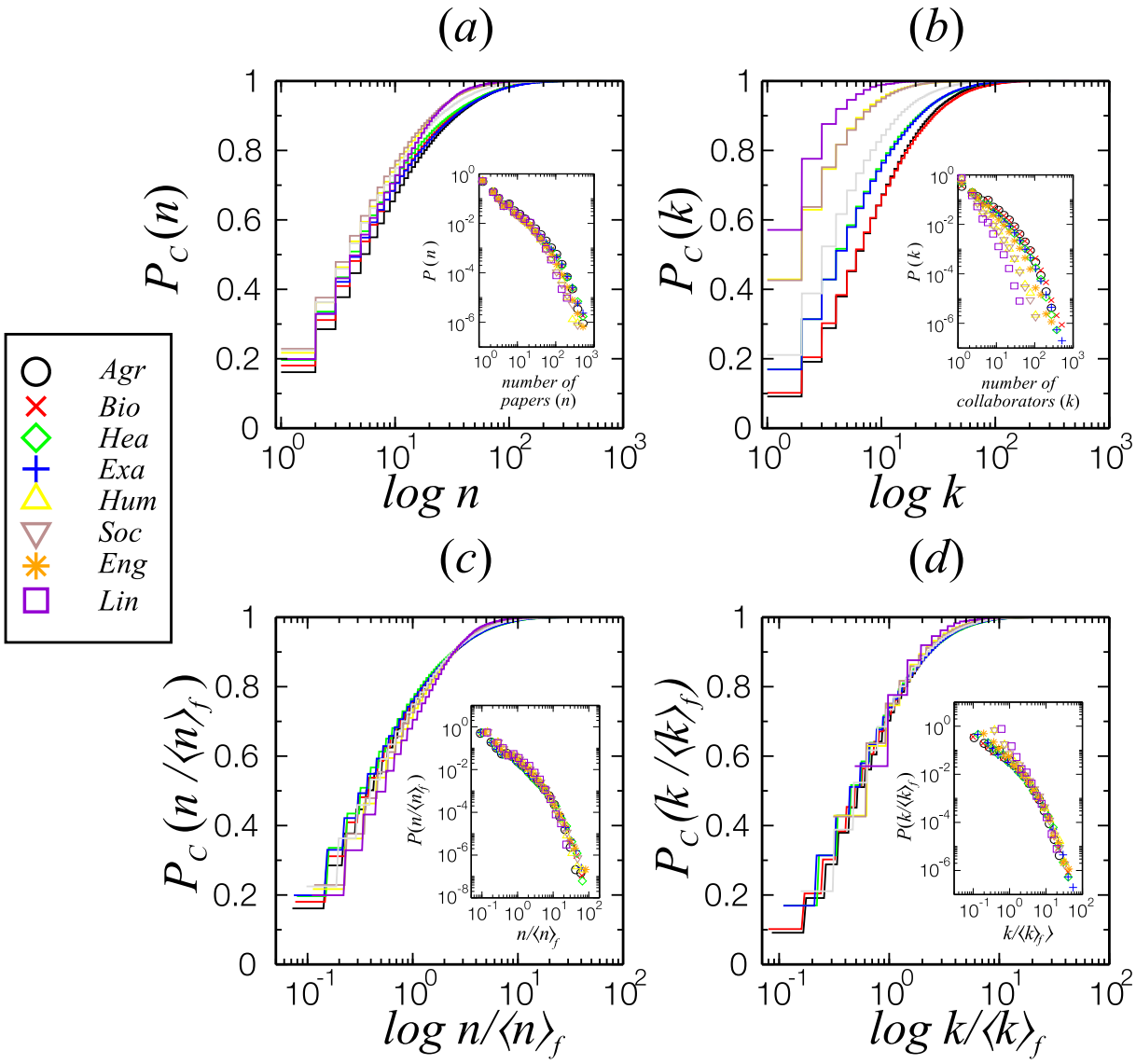


Figure 25 – Cumulative distributions P_C of the number of papers published per researcher n (a) and number of collaborators (b) for each of the 8 major fields. The respective distributions for the rescaled data are shown in (c) and (d). Lines represent different fields, colored according to the symbol in the legend. Scientists working on social sciences and related fields (Lin, Soc and Hum) are less likely to have published more than one hundred papers than others. They also are less likely to have more than one hundred collaborators. Considering the average publication count $\langle n\rangle_f$ and average number of collaborations $\langle k\rangle_f$ in each field, all the curves collapse to a single universal behavior. The insets show the respective (non-cumulative) distributions.

different levels of productivity [99], by rescaling k and n to the corresponding average values of the field (see Table 5), $\langle k\rangle_f$ and $\langle n\rangle_f$, both $P_C(n)$ and $P_C(k)$ distributions collapse to single universal curves, as depicted in Figs. 25(c) and (d), respectively.

In the last decades, there has been an increasing number of interdisciplinary research. Since the giant component of TCN comprises most of the researchers, it is interesting to verify the relationship between different fields. In Fig. 26 we show for each field the fraction of collaborations to other

Table 5 – Statistics for researchers working on the 8 major fields associated with the TCN.

	Number of researchers (N_f)	Researchers with scholarship (S_f)	Average number papers per researcher ($\langle n \rangle_f$)	Average number of collaborators ($\langle k \rangle_f$)
Agr	31812	1692	13.9	11.7
Bio	39767	2605	13.1	12.5
Hea	67561	1511	12.6	9.08
Exa	33310	3273	13.5	9.16
Hum	26263	1324	8.90	3.21
Soc	20806	742	8.66	3.23
Eng	18365	1841	10.2	6.37
Lin	5202	300	9.09	2.06

fields. As expected, most of the collaborations of engineers are made with researchers working on exact and earth sciences, and most of collaborations of researchers of arts and linguistics are made with researchers from humanities. But its clear that researchers of biological sciences are not strongly linked to engineers, while social scientists seek collaborators evenly across fields.

3.8 Conclusions

In this chapter, we have used the Lattes Platform, which contains detailed and unambiguous information of approximately 2.7 million curricula of researchers, as a database for analysing research collaboration in Brazil. It has the advantage of displaying individual curricula, allowing us to study collaborations in a mix of a paper-based approach and questionnaire data.

We therefore built collaboration networks including all researchers data from Lattes Platform as June 2012, and found that the network has grown exponentially for the last three decades. The calculated values of the assortativity coefficient and the average nearest-neighbor degree indicate that the networks display assortative mixing, where researchers having high k collaborate with others alike. Our results show that these teeming researchers are more likely to have a scholarship and to produce more papers than researchers with low k. The distribution $P(k)$ is also approaching a power-law as the network gets older.

We confirmed the validity of Lotka's Law for researchers working on different states of Brazil and found substantial correlations between $\langle k \rangle_f$ and N_f. Lotka's Law is shown to be valid for different fields: indeed, $P(n)$ and $P(k)$ follow a universal behavior.

We have shown evidence that the patterns in publication and collaboration are universal across different fields and geographical locations, with only a scale factor needed to account for differences.

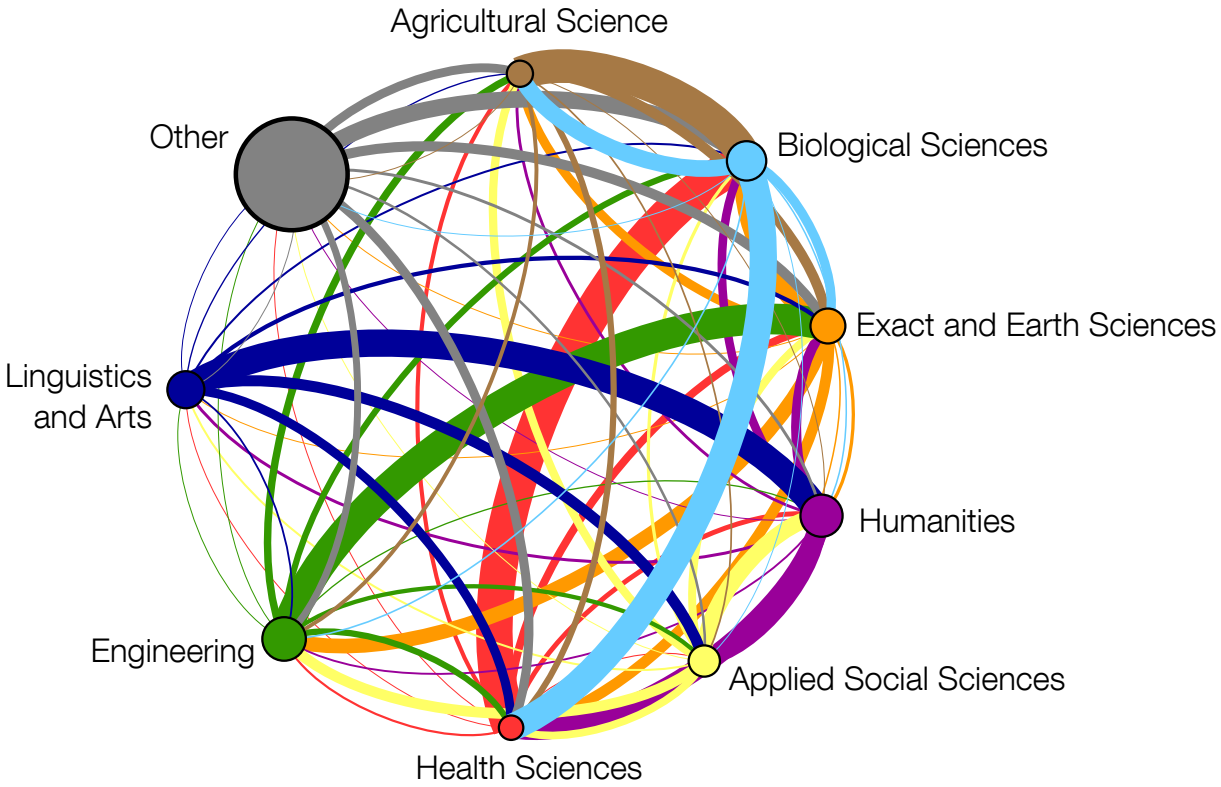


Figure 26 – Study of interdisciplinary collaborations. Vertices represent different fields with sizes proportional to the fraction of collaborations with researchers of other fields. The directed edges are colored according to the source vertex and the width is proportional to the fraction of collaborations made with the target vertex. While some pairs are expected as Exa–Eng and Lin–Hum, the small fractions of collaborations between researchers of Bio with Eng could indicate that biotechnology is still a maturing field.

4 Gender and Collaboration

4.1 Introduction

The underrepresentation of women in science is an enduring issue [100–102], notwithstanding government policies employed by several countries. Further, there is a clear disparity between the increase of women who earn Ph.D.'s and the increase in women occupying faculty positions [101].

An early studied disparity index between male and female researchers is the productivity, as measured by paper output. In the 1980's, Cole and Zuckerman [103] compiled and presented the results of several studies showing that female researchers publish consistently less than male researchers. This disparity persists nowadays. In a multidisciplinary global bibliometrical analyis of more than five million papers indexed in Thomson Reuters Web of Science, Larivière *et al.* [104] confirmed these paradigm and found that women are the majority of authors in only nine countries. Four of these had less than 1'000 articles. Futhermore, papers authored by female researchers are less likely to receive citations, impacting the visibility of these scientists.

Kyvik and Teigen [105] provided evidence that the productivity of female researchers is significantly influenced by child care. In their study, women with children under six were 60% less productive than their male peers. For researchers with children between 6 and 10 years old, female researchers were around 30% less productive. The impact of children over 10 years old is less significant. Further, when considering the age of the researcher, while women below 40 years publish considerably less than men in the same age range, the same is not observed for women older than 40 years. Thus, child birth, maternity leave and child care are important factors which must be taken into account when considering productivity.

Prpić [106] argued that gender differences in productivity are intrinsic to the social organization of science and "Explanations for the systematic differences in scientific productivity should be sought in the social organization of science where, just as society in general, hidden or (more rarely) open mechanisms of discrimination may exist." [106, p. 49]. The subjects of this study were a group of 840 young researchers. Prpić found that the most powerful predictors for productivity of female researchers are attendance at international conferences. At these meetings, researchers are able to establish new collaborations, which is strongly correlated with productivity.

Duch *et al.* [102] hypothesized that, since the presence of female researchers varies across fields, the observed differences in publishing could be a consequence of lower support (understood as resources) received by females [107, 108]. In fields which require lesser resources, the publishing rates should be similar. They found that the lower publication rates of female researchers can be explained by higher research expenditures, with a negative correlation existing between the two quantities.

Biases against women within academia might be one reason for this scarcity. Steinpreis *et al.* [109] sent to 238 male and female academic psychologists four versions of a real person curriculum,

only changing the name to traditional male and female names, for a job and tenure application. Their results show that both men and women were more likely to vote to hire the male applicant, even though the female applicant had an identical curriculum. In a more recent but similar double-blind study conducted by Moss-Racusin *et al.* [110], identical applications for a laboratory manager position were sent to science faculty members, with a randomly assigned male or female name. Those having male names were rated more competent, and offered higher starting salary. Again, the gender of the faculty member did not influence the results.

Bozeman and Gaughan [111] studied the differences between men and women regarding collaboration. They found that, when taking into account factors as tenure, family status and discipline, females have more collaborators on average than male scientists. This is in conflict with previous studies [7, 33, 103] which suggest that male scientists are more collaborative. It should be noted, however, that their concept of collaboration is not based on co-authorship, but self-reported.

West *et al.* [80] investigated the presence of female authors and the relationship between author position in the list of authors of multiple authored papers and gender. Over eight million papers from JSTOR (http://www.jstor.org/), a digital library of academic journals and books, were analysed. They found a steady increase in the proportion of female authorships since mid 1960's. Additionally, men predominate as first and last authors, and women seem to publish less single-authored papers than would be expected by their fraction in a field.

Worldwide, women became the majority of students enrolled in undergraduate and graduate courses [112]. In Brazil specifically, in the last ten years women are the majority to enroll in higher education, with 56.7% of the 6 million enrollments. They are also the majority to complete higher education, representing 60% of the total in presential courses. Nonetheless, data from 2015.01 (see Table 6) shows that women are still underrepresented when considering Ph.D.'s, comprising 47% of Ph.D.'s registered in Lattes Platform. The recent shift to a more egalitarian climate in academia can be seen by partitioning the number of Ph.D.'s by age: in the 19-34 cohort there are 6959 female Ph.D.'s, comparable to the 7103 males.

Still, there are differences when considering distinct fields. Women are still the minority in Exact and Earth Sciences and Engineering, while being the majority in Biological Sciences and Humanities. Considering the discipline of Physics, women are greatly under-represented, specially when considering those in the higher end of academic career [113].

Initiatives to raise awareness of the problem of women's under-representation in academia have been made in last decades. In physics, the International Union of Pure and Applied Physics (IUPAP) created a working group on the subject in 1999, named Women in Physics (WiP), which through global and local initiatives aims to build an egalitarian climate for women in physics internationally [114]. In Brazil, the Brazilian Physical Society created in 2003 the Commission for Relations and Gender (CRG-BPS).

As highlighted by Cole [115], the development of science might be constrained by the social network formed by scientists. Under this light, it is desirable to study the presence of female scientists in this academic network. The results of such investigation may unfold some of the mechanisms

Table 6 – Number of male and female professors and researchers in Brazil. Source: Lattes Database (http://lattes.cnpq.br/), collected on 2015 January.

Field	Female Ph.D.	Male Ph.D.	Female Master	Male Master
Agricultural Sciences	4732	7248	2256	2379
Biological Sciences	8832	6477	4248	2559
Health Sciences	10282	7739	8393	3985
Exact and Earth Sciences	5578	11991	3395	6315
Humanities	10976	8562	9176	5528
Social and Applied Sciences	5164	6795	6706	7343
Engineering	2772	8431	1323	4246
Linguistics and Arts	4682	2638	3688	1805

behind the observed gender disparities in publishing and collaborating [106].

As already described, the Brazilian government requires every researcher in the country to display, on a public website, called the Lattes Platform (http://lattes.cnpq.br), the complete information about his or her scientific productivity and many other data, including (only until 2012) the gender. To the best of our knowledge, this is the largest available dataset about scientific collaboration containing information about gender, allowing making statistically significant studies on the role of gender in different fields. Here we analyse in detail the fraction of female co-authors as a function of the total number of collaborators and find that on average men prefer collaborating with men, while women are more egalitarian. Discriminating by fields, we discover that this effect is most pronounced in health sciences, while in engineering the effect is less pronounced for researchers with a high number of collaborators.

The challenges encountered by women in academia are reflected in their ubiquitous underrepresentation [100–102, 116]. Biases against women may contribute to this scarcity, presenting themselves in several academic related activities such as hiring [110], grant funding [108], collaborating strategies of researchers [111] and prestigious author position in papers [80]. A large portion of the previous work on this subject is devoted to study the differences in productivity between male and female researchers, called the ‘productivity puzzle’ [103, 105, 106]. More recent contributions are more focused in the social aspect of science and investigate the strategies followed by women in their career, in publishing and in collaborating [102, 111, 117, 118]. Nevertheless, these works only superficially regard researchers as actors in a large interconnected network environment. An application of network science to a social community have shown that women and men organize themselves differently [119]. The impact of the network structure on the development of science on large scale [115] and individual level deserves more attention. In the present work, we contribute to this topic analysing global and local metrics partitioned by gender in a large collaboration network [90].

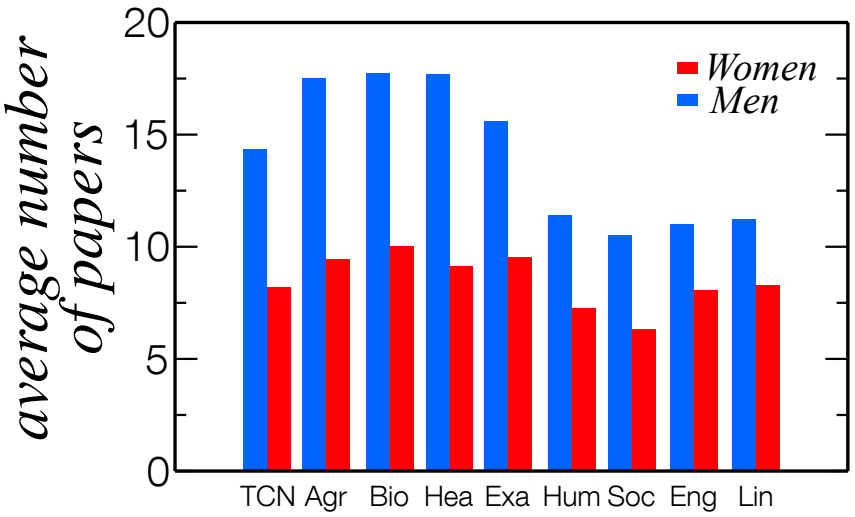


Figure 27 – Average of the number of papers for male (blue bars) and female (red bars) researchers for the TCN and for each of the eight major fields: Agricultural Sciences (Agr), Applied Social Sciences (Soc), Biological Sciences (Bio), Exact and Earth Sciences (Exa), Humanities (Hum), Health Sciences (Hea), Engineering (Eng) and Linguistics and Arts (Lin).

4.2 Methods

As previously presented, we build a cumulative collaboration network (the TCN) from the CV's available in Lattes Platform. In this chapter, we focus on the differences between male and female researchers in this network. The proportion of female researchers varies across fields [80] and the values for the TCN are shown in Table 7. The averages of the number of papers and collaborators are shown also in Figs. 27 and 28 for clarity.

Previous works on gender and collaboration [33, 105, 111, 117, 118] used information from a limited number of authors, usually much less than 10'000. We have information concerning the productivity (as measured by article output) of 275'061 researchers with published papers on periodicals, 130'525 men (47.4%) and 144'440 women (52.5%). Only 96 researchers do not display the gender information on the curriculum. 90.4% belong to the giant component.

4.3 Results

Women became the majority in higher education in Brazil in the last decade. Whether this higher proportion is reflected in research is not clear. In order to elucidate this question, we extracted the number of researchers who published his or her first paper for each year in Lattes Platform. Results are shown in Fig. 29. Clearly, women became also the majority of researchers joining the collaboration network since 2000.

Once these researchers are incorporated in the collaboration network, we can investigate how they are organized. The most simple characteristic is their degree. As shown in Fig. 30, both male

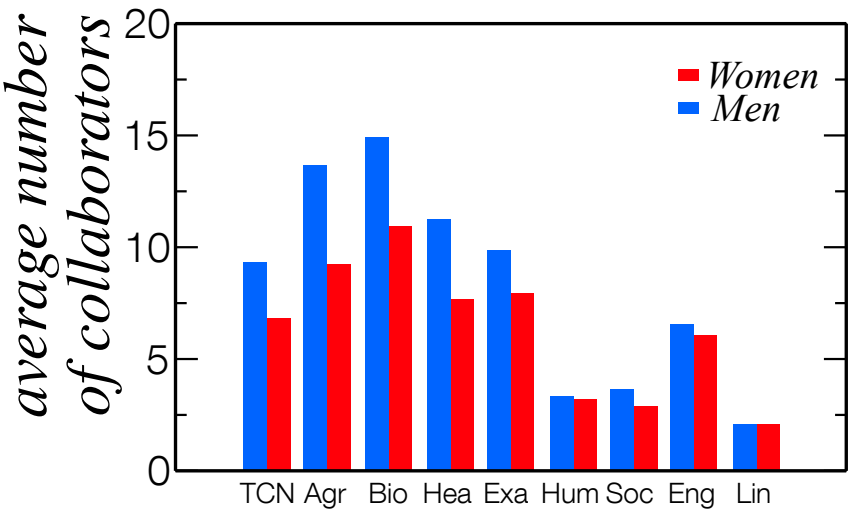


Figure 28 – Average of the number of collaborators for male (blue bars) and female (red bars) researchers for the TCN and for each of the eight major fields: Agricultural Sciences (Agr), Applied Social Sciences (Soc), Biological Sciences (Bio), Exact and Earth Sciences (Exa), Humanities (Hum), Health Sciences (Hea), Engineering (Eng) and Linguistics and Arts (Lin).

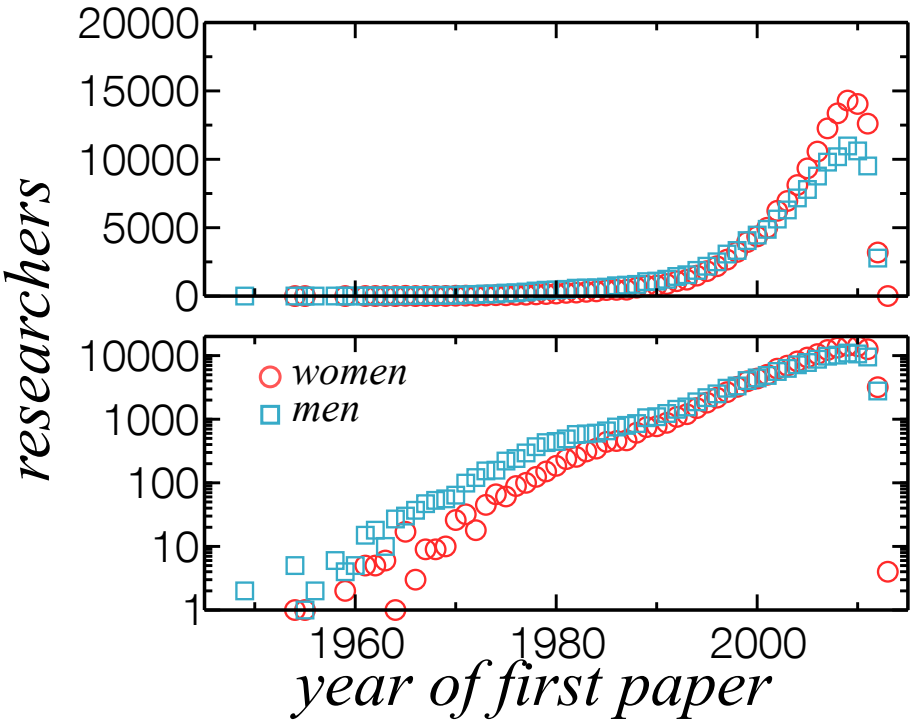


Figure 29 – Top: number of male (blue squares) and female (red circles) researchers who published their first paper per year. It is clear that since 2000 women are the majority joining the collaboration network. In the bottom, we show the same data with the number of researchers in logarithmic scale, for better visualization of the transition point.

Table 7 – Number of researchers in the TCN, proportion of female researchers, average number of papers and collaborators for male and female researchers for each of the eight major fields: Agricultural Sciences (Agr), Applied Social Sciences (Soc), Biological Sciences (Bio), Exact and Earth Sciences (Exa), Humanities (Hum), Health Sciences (Hea), Engineering (Eng) and Linguistics and Arts (Lin).

	Total	Agr	Bio	Hea	Exa	Hum	Soc	Eng	Lin
Researchers	275'061	31'812	39'767	67'561	33'310	26'263	20'806	18'365	5'202
Fraction of female researchers	51.9%	44.4%	60.1%	59.8%	34.7%	65.1%	47.3%	27.2%	71.6%
Average number of papers (men)	14.3	17.5	17.7	17.7	15.6	11.4	10.5	11.0	11.2
Average number of papers (women)	8.15	9.45	10.0	9.18	9.49	7.54	6.61	8.08	8.25
Average number of collaborators (men)	9.27	13.6	14.9	11.2	9.84	3.31	3.57	6.50	2.04
Average number of collaborators (women)	6.80	9.20	10.9	7.65	7.90	3.16	2.85	6.02	2.06

and female researchers display a similar degree distribution. Moreover, once a researcher establishes some collaborations, he or she may continue to work with these colleagues or look for newer collaborators. In the former case, when a pair of researchers produce more than one paper together, we call these collaborations 'recurrent'. We define as the weight of the collaboration between two researchers the total number of papers co-authored by both. Figure 30 shows the distribution of these weights, for male and female researchers. Again, the two curves are very similar, with women having slightly less recurrent collaborations. This might be a consequence of women displaying a lower average number of published paper than men. We analysed the Lattes Database and found the same behavior, as shown in Table 7, where we compare the mean productivity as measured by paper output and the mean number of collaborators, for male and female researchers. We notice this effect even in fields where female researchers are the majority, namely Bio, Hea, Hum and Lin. The average number of papers produced by men is higher for every field (for TCN, 10.6 for men and 6.26 for women), while the average number of collaborators is of male researchers is also higher

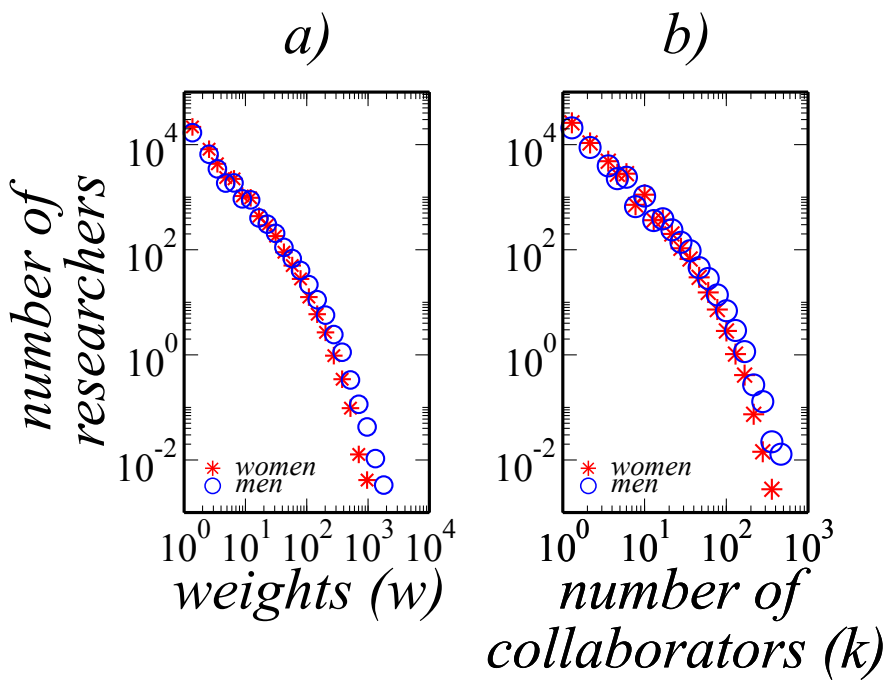


Figure 30 – a) Distribution of the number of recurrent collaborations (weights) between researchers, divided into male and female researchers. b) Distribution of the number of collaborators, divided into male and female researchers. Women are less likely to display large values for both quantities.

for every field except Lin, (for TCN, 9.27 for men and 6.80 for women).

We also compare researchers with their neighbors in the network. For both male and female researchers, we calculated their average nearest neighbor degree. Results are shown in Fig. 31. Both display the same logarithmic increase but, for intermediate values of k, male researchers are connected to more connected peers on average. This can be interpreted as a difference in the collaborating strategies, with men actively looking for highly connected researchers.

We now focus on the acquisition of new collaborators. Since researchers in the database display a varied career span, we adopted the following procedure. For each researcher, we counted the number of new collaborators for each year since he or she became part of the network. Hence, we sum the number of new collaborators for each year for all male and all female researchers. Then we divide the results for each year by the total number of new collaborators. Results are shown in Fig. 32. Both male and female researchers display a similar exponential decay of the number of new collaborators with career span,

$$p(t) = Ae^{-\lambda t}, \tag{4.1}$$

where $p(t)$ is the fraction of new collaborators after time t in years. Both exhibit two regimes separated by a threshold of 30 years from the beginning of collaborative work. The frequency of new collaborators acquired by female researchers decays faster in this first regime ($\lambda_w = 0.132 \pm 0.002$) than male's ($\lambda_m = 0.0998 \pm 0.001$).

This decay in acquiring new collaborators can be interpreted as a shift from an initial struggle

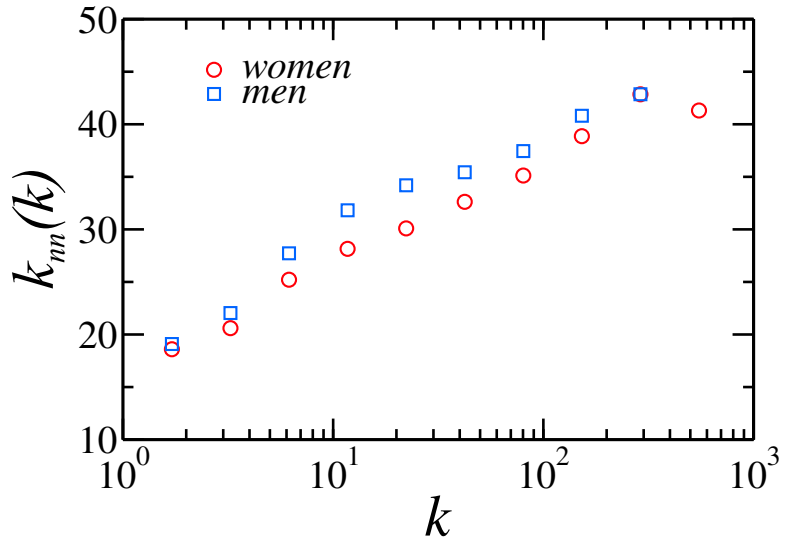


Figure 31 – k_{nn} as a function of the number of collaborators (k) for male (blue squares) and female (red circles) researchers.

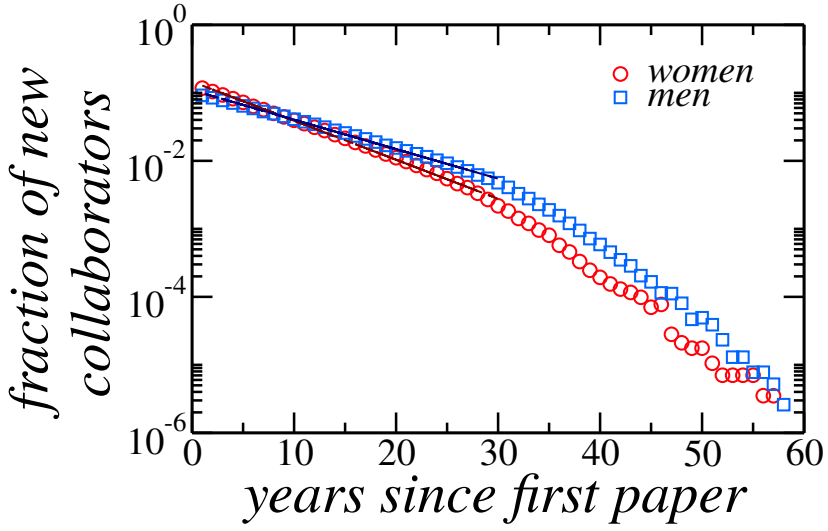


Figure 32 – Fraction of new collaborators acquired as a function of time since the first published paper. Both male and female researchers display an exponential decay with a change in behavior occurring after 30 years. The dashed lines are exponential fits for the first 30 years: $p(t) = Ae^{-\lambda t}$. The calculated exponents are $\lambda_w = 0.132 \pm 0.002$ for women and $\lambda_m = 0.0998 \pm 0.001$ for men.

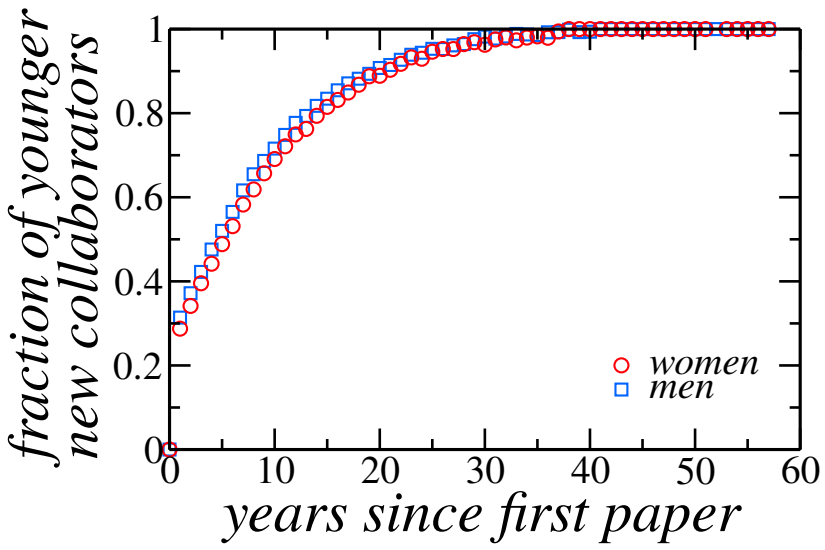


Figure 33 – Fraction of new collaborators who joined the collaboration network after the researcher, for men (blue squares) and women (red circles). While in the beginning of a researcher career most of his or her collaborators are older (i.e., joined the network before them), after only 5 years they display a balance between older and younger collaborators. After 20 years of research, around 90% of new collaborators are younger and after 30 years, only a small fraction of new collaborators are older. The same behavior is observed for both genders.

to advance in career ('publish or perish') to a mentoring position, where the researcher achieved career independence and can focus on the younger researchers who seek him or her. In order to quantify this behavior, we adopted the following procedure: for each researcher, we calculated the fraction of the new collaborators who joined the network after the researcher in question, this being performed for each year. Hence, we average these fractions for both male and female researchers. Results are shown in Fig. 33. During the evolution of a researcher career, one begins with most of his or her collaborators being older (i.e., joined the network before them). A balance between older and younger new collaborators is observed after 5 years. After 30 years, the researcher made contact with most of the teeming researchers in his or her field and acquires mostly younger collaborators. This behavior is gender independent. Note that this threshold of 30 years characterizes a change in the behavior of the fraction of new collaborators shown on Fig. 32.

The quantities calculated show only a slight distinction in their values for male and female researchers. Some of these quantities, as productivity, have long been discussed in the literature. However, since we have information concerning the gender of researchers in Lattes Platform, we can proceed further. We now focus on the study of homophily in the collaboration network. To address influence of gender in choosing a collaborator, we define the *g-ratio* as the number of

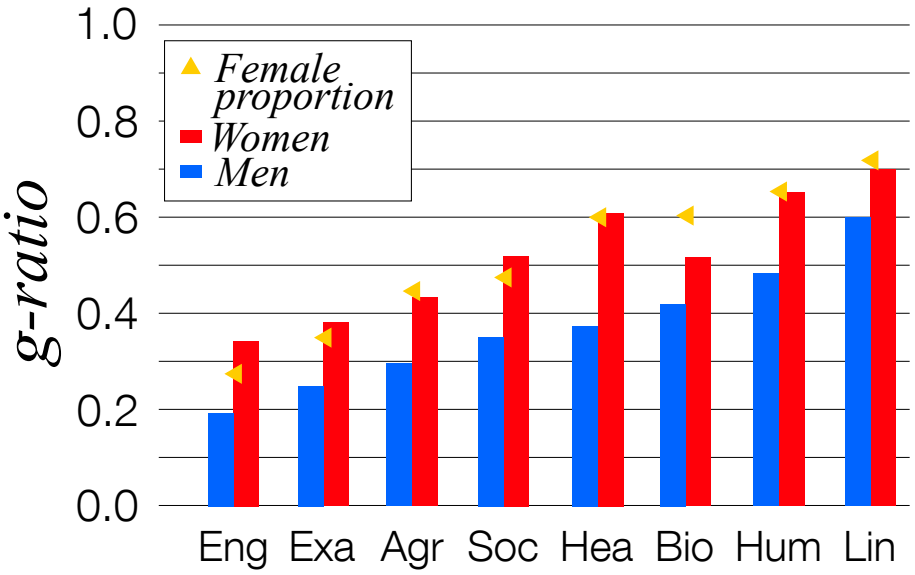


Figure 34 – Mean values of *g-ratio* for each field of expertise: Agricultural Sciences (Agr), Applied Social Sciences (Soc), Biological Sciences (Bio), Exact and Earth Sciences (Exa), Humanities (Hum), Health Sciences (Hea), Engineering (Eng) and Linguistics and Arts (Lin). Red and blue bars represent values for female and male researchers, respectively. Yellow triangles show the proportion of female researchers in the respective field.

female collaborators of researcher i weighted by the number of co-authored papers,

$$g - ratio_i = \frac{\sum'_{fem,j} w_{ij}}{\sum_j w_{ij}}, \tag{4.2}$$

where w_{ij} is the number of papers co-authored by i and j, and the top sum is over the female collaborators.

The *g-ratio* thus defined unveils a dissimilar behavior for male and female researchers regarding collaborations. For each of the eight major fields, we calculate the average *g-ratio* for both genders, see Fig. 34. Female researchers display a higher average *g-ratio*, regardless of the field, and close to the fraction of females in the respective field. Male researchers have relatively more collaborations with other men, indicating a homophilic behavior. Previous results based on a rather small number of researchers suggest that women collaborate more with other women [33, 117], but here we show that this is not the case.

The evidence of a bias against gender for male researchers in choosing a collaborator invites the question of how it is related to the number of collaborators, k, of a researcher. The correlation between *g-ratio* and this quantity is also investigated for each field. In Fig. 35, we show the results for Bio and Eng. Except for Eng, *g-ratio* varies only slightly with k. The values for women are close to the female proportion in the respective field and values for men consistently lower. The number of collaborators extenuate on a small scale the male bias. Eng seems to be a particular field, with women being preferred for collaborating, for both male and female researchers.

Women have been reported to be more involved in interdisciplinary research than their male peers [118]. In order to investigate this, we define a ratio related to interdisciplinary work, which

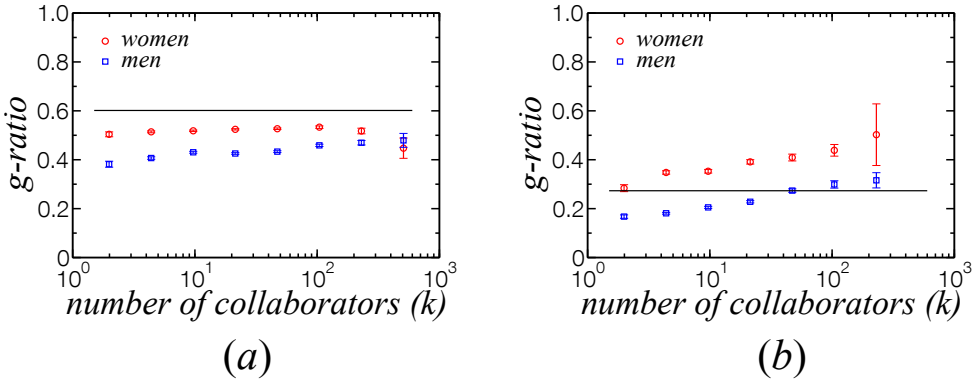


Figure 35 – Correlation between *g-ratio* and number of collaborators for Biological Sciences (a) and Engineering (b) for female (red circles) and male (blue circles) researchers. Lines represent the female proportion in the respective area. Female researchers are more likely to collaborate with other female researchers than their male peers, without regard to the expertise. For technology related fields, *g-ratio* is above the proportion of female researchers. The bars indicate the standard deviation.

we call *m-ratio*. More precisely, we define the *m-ratio* as the probability of randomly choosing a collaborator with expertise on a distinct major field,

$$\text{m-ratio}_i = \frac{\sum'_{field,j} w_{ij}}{\sum_j w_{ij}}, \tag{4.3}$$

where the sum in the numerator includes only collaborators of a researcher i working on different fields and the denominator is the number of papers co-authored by i.

Our findings for *m-ratio* indicate that both male and female researchers display the same tendency to be involved in interdisciplinary collaborations in all fields, except for Exact and Earth Sciences in which females are more likely to be involved in such activities. For all fields, the correlation between *m-ratio* and k for both male and female researchers follow the same tendency. Researchers with a higher number of collaborators are more prone to display larger *m-ratio* values.

4.4 Conclusions

In studying a large scientific collaboration network, comprising more than 270'000 researchers, we have found gender differences regarding collaboration. From the Lattes Platform dataset, we observe that female researchers are the majority in Biological and Health Sciences, Humanities and Linguistics and Arts. Despite this greater presence, they have less published papers and collaborators.

Two metrics were introduced to investigate gender differences. The *g-ratio* measures the fraction of female collaborators of a given researcher and the *m-ratio* the fraction of collaborators working on a field distinct from the one of a given researcher. Male and female researchers display a heavy tailed and similar distribution of recurrent collaborations, which does not unveil any patent gender differences.

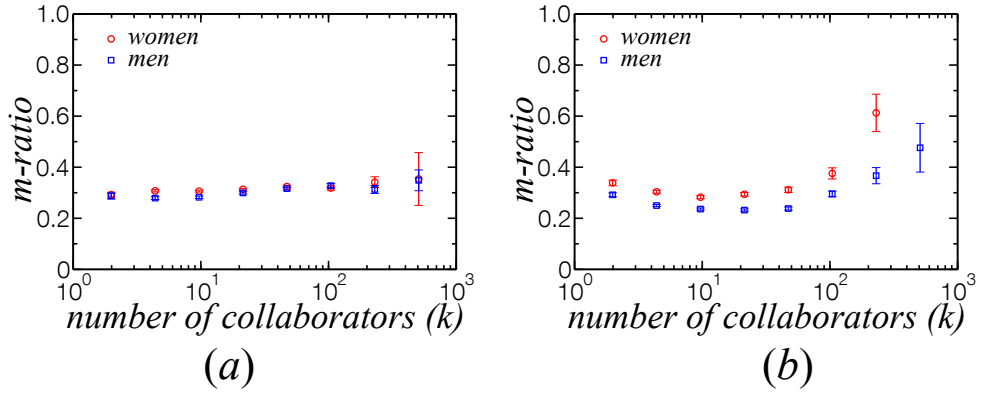


Figure 36 – Correlation between *m-ratio* and number of collaborators in Biological Sciences (a) and Exact Sciences (b) for female (red circles) and male (blue circles) researchers. For all fields, except Exact and Earth Sciences, there is only a small difference regarding multidisciplinary collaborations. The bars indicate the standard deviation.

From the results using our proposed metric, the *g-ratio*, we show that while female researchers display values consonant with their proportion on a field, male researchers display a homophilic bias, regardless of the field. Also, this tendency is the same, regardless of the number of collaborators of a given researcher, except for Engineering, where researchers with more collaborators are more likely to work with women.

The results obtained for the *m-ratio*, on the other hand, show that male and female researchers have the same tendency to participate in interdisciplinary research, except for Exact and Earth Sciences, where females are more interdisciplinary. Researchers with a higher number of collaborators are more likely to do their research on several fields.

The path to gender equity in academia must involve not only government and institutional support but also consciousness of the biases against women and proactive support from the researchers themselves in order to extinguish them. The causes for the homophilic behavior male researchers display must be investigated critically.

5 Conclusions

In summary, we have used the Lattes Platform, which contains detailed and unambiguous data of approximately 2.7 million curricula of researchers, as a database for analysing research collaboration in Brazil. It has the advantage of displaying individual curricula, allowing us to study collaborations in a mix of a paper-based approach and questionnaire data.

We developed a software to download all CV's as June 2012 and a parser to extract relevant information concerning research activities. We built collaboration networks including all researchers data from Lattes Platform, and found that the historical network (aggregating longitudinal data) has been growing exponentially for the last three decades. Even comprising several different disciplines, this collaboration network displays a giant component comprising 90% of the researchers. The calculated values of the assortativity coefficient and the average nearest-neighbor degree indicate that the networks display assortative mixing, where researchers having high k collaborate with others alike. Our results show that these teeming researchers are more likely to have a productivity grant and to produce more papers than researchers with low k. The distribution $P(k)$ is also approaching a power-law as the network gets older. We confirmed the validity of Lotka's Law for researchers working on different states of Brazil and found substantial correlations between $\langle k \rangle_f$ and N_f. Lotka's Law is shown to be valid for different fields: indeed, $P(n)$ and $P(k)$ follow an universal behavior. We have shown strong evidence that the patterns in publication and collaboration are universal across different fields and the same for both genders, with only a scale factor needed to account for differences.

We have investigated the participation of female researchers in the collaboration network. Collected data is in conformity with other studies, with women displaying a lower productivity and lower number of collaborators on average. To investigate endogenous factors contributing to this productivity puzzle, we have developed a metric, *g-ratio*, related to the fraction of collaborations with female researchers. We have shown that male researchers display a bias against gender in choosing their collaborations. The divergent results for productivity regarding gender might be a consequence of this dissociation. We have developed a metric, *m-ratio*, to study the behavior of male and female researchers regarding interdisciplinary research. There is no substantial gender differences, contrary to previous studies.

Due to the extensive data avaiable in Lattes database, there are many possible ramifications for this work. We restricted ourselves to research papers published in periodicals. A similar analysis can be used to include books, presentations, patents and software. For researchers working on more than one major field, we considered only the first cited. We can further analyse subfields for more detailed statistics. The correlations between these fields can be also investigated. From the mentoring activity data, we can study how the advisor's position in the collaboration network can influence that of his or her students and his or her productivity. With the information concerning the university or college where researchers undertook undergraduate and graduate studies, we may

investigate their mobility. All of these branchs may consider the gender of the researchers. The homophilic behavior of different vertices on a network can be modelled in order to investigate its implications to topological and dynamical properties of the system. Comparing results of computational simulations of such model to the collaborations networks studied in this thesis can give insights concerning strategies to prevent undesirable biases. The now avaiable XML CV files allow a better parsing of the desired information.

The underrepresention of women in research calls attention to the waste of human intellectual capital. If these women are hindered in collaborations due to a male bias, its imperative to find mechanisms to prevent it. We believe that the identification of such bias can guide academic policies devoted to bring equity in the relations between administration, funding agencies and male researchers and women, as the progress of science requires us to maximize intellectual capital, which can not be done while excluding talented women.

Bibliography

1 WATTS, D. J.; STROGATZ, S. H. Collective dynamics of ‘small-world’ networks. *Nature*, v. 393, p. 440, 1998. 16, 23, 25, 27, 34, 41, 58

2 FOX, M. F.; FAVER, C. A. Independence and cooperation in research: The motivations and costs of collaboration. *J. Higher Educ.*, v. 55, p. 347, 1984. 23

3 KATZ, J. S.; MARTIN, B. R. What is research collaboration? *Res. Policy*, v. 26, p. 1, 1997. 23, 44, 45

4 LAUDEL, G. What do we measure by co-authorships. *Res. Eval.*, v. 11, p. 3, 2002. 23, 24, 45

5 BEAVER, D. deB.; ROSEN, R. Studies in scientific collaboration part i. the professional origins of scientific co-authorship. *Scientometrics*, v. 1, p. 65–84, 1978. 23, 44

6 LAWANI, S. M. Some bibliometric correlates of quality in scientific research. *Scientometrics*, v. 9, p. 13, 1986. 23

7 LEE, S.; BOZEMAN, B. The impact of research collaboration on scientific productivity. *Soc. Stud. Sci.*, v. 35, p. 673, 2005. 23, 69

8 PRICE, D. J. de S.; BEAVER, D. Collaboration in an invisible college. *Am. Psychol.*, v. 21, p. 1011, 1966. 23

9 FRAME, J. D.; CARPENTER, M. P. International research collaboration. *Soc. Stud. Sci.*, v. 9, p. 481, 1979. 23, 44

10 HEFFNER, A. G. Funded research, multiple authorship, and subauthorship collaboration in four disciplines. *Scientometrics*, v. 3, p. 5, 1981. 23, 45

11 KRAUT, R.; EGIDO, C. *Patterns of Contact and Communication in Scientific Research Collaboration*. [S.l.]: ACM Press, 1988. 1–12 p. 23

12 ALBERT, R.; BARABÁSI, A.-L. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, v. 74, p. 47, 2002. 23

13 NEWMAN, M. E. J. Assortative mixing in networks. *Physical Review Letters*, v. 89, p. 208701, 2002. 23, 60

14 NEWMAN, M. E. J. The structure and function of complex networks. *SIAM Rev. Soc. Ind. Appl. Math.*, v. 45, p. 167, 2003. 23, 38, 58

15 RAVASZ, E.; BARABÁSI, A. L. Hierarchical organization in complex networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, v. 67, p. 026112, 2003. 23, 58

16 BARRAT, A. et al. The architecture of complex weighted networks. *Proc. Natl. Acad. Sci. U.S.A.*, v. 101, p. 3747, 2004. 23, 47, 60

17 BARRAT, A.; BARTHÉLEMY, M.; VESPIGNANI, A. Modeling the evolution of weighted networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, v. 70, p. 066149, 2004. 23, 60

18 MOREIRA, A. A. et al. Competitive cluster growth in complex networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, v. 73, p. 065101, 2006. 23

19 LIND, P. G. et al. Spreading gossip in social networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, v. 76, p. 036117, 2007. 23

20 MOREIRA, A. A. et al. How to make a fragile network robust and vice versa. *Physical Review Letters*, v. 102, p. 018701, 2009. 23

21 GALVÃO, V. et al. Modularity map of the network of human cell differentiation. *Proc. Natl. Acad. Sci. U.S.A.*, v. 107, p. 5750, 2010. 23

22 SCHNEIDER, C. M. et al. Mitigation of malicious attacks on networks. *Proc. Natl. Acad. Sci. U.S.A.*, v. 108, p. 3838, 2011. 23

23 BARABÁSI, A. L. et al. Evolution of the social network of scientific collaboration. *Physica A*, v. 311, p. 590, 2002. 23, 24, 25, 47, 55

24 NEWMAN, M. E. J. Scientific collaboration networks. i. network construction and fundamental results. *Phys Rev E Stat Nonlin Soft Matter Phys*, v. 64, p. 016131, 2002. 23, 24, 46

25 NEWMAN, M. E. J. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Phys Rev E Stat Nonlin Soft Matter Phys*, v. 64, p. 016132, 2002. 23, 46, 47

26 GOH, K. et al. Classification of scale-free networks. *Proc. Natl. Acad. Sci. U.S.A.*, v. 99, p. 12583, 2002. 23

27 NEWMAN, M. E. J. Coauthorship networks and patterns of scientific collaboration. *Proc. Natl. Acad. Sci. U.S.A.*, v. 101, p. 5200, 2004. 23

28 RAMASCO, J. J.; DOROGOVTSEV, S. N.; PASTOR-SATORRAS, R. Self-organization of collaboration networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, v. 70, p. 036106, 2004. 23

29 LI, M. et al. Weighted networks of scientific communication: the measurement and topological role of weight. *Physica A*, v. 350, p. 643, 2005. 23

30 KATZ, J. S. Geographical proximity and scientific collaboration. *Scientometrics*, v. 31, p. 31, 1994. 23, 45, 47

31 PONDS, R.; OORT, F. V.; FRENKEN, K. The geographical and institutional proximity of research collaboration. *Pap. Reg. Sci.*, v. 86, p. 423, 2007. 23, 45

32 PAN, R. K.; KASKI, K.; FORTUNATO, S. World citation and collaboration networks: uncovering the role of geography in science. *Scientific Reports*, v. 2, p. 902, 2012. 23, 45, 47

33 BOZEMAN, B.; CORLEY, E. Scientists' collaboration strategies: implications for scientific and technical human capital. *Research Policy*, Elsevier, v. 33, n. 4, p. 599–616, 2004. 23, 45, 60, 69, 71, 77

34 HAGSTROM, W. O. *The Scientific Community*. [S.l.]: Basic Books, 1964. 297 p. 23

35 MUCHNIK, L. et al. Origins of power-law degree distribution in the heterogeneity of human activity in social networks. *Scientific Reports*, v. 3, p. 01783, 2013. 23

36 PARETO, V. *Cours d'économie politique*. Lausanne: F. Rouge, 1897. 426 p. 23

37 GALLOS, L. K. et al. Imdb network revisited: unveiling fractal and modular properties from a typical small-world network. *PLOS ONE*, v. 8, p. e66443, 2013. 23, 27

38 LOTKA, A. J. The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, v. 16, p. 317–323, 1926. 23

39 NICHOLLS, P. T. Empirical validation of lotka's law. *Inf. Process. Manag.*, v. 22, p. 417, 1986. 23

40 PAO, M. L. An empirical examination of lotka's law. *J. Am. Soc. Inf. Sci.*, v. 37, p. 26, 1986. 23

41 KATZ, J. S. The self-similar science system. *Res. Policy*, v. 28, p. 501, 1999. 23

42 MARTIN, T. et al. Coauthorship and citation patterns in the physical review. *Physical Review E*, v. 88, 2013. 24

43 TANG, L.; WASH, J. P. Bibliometric fingerprints: name disambiguation based on approximate structure equivalence of cognitive maps. *Scientometrics*, v. 84, p. 763, 2010. 24

44 PASTOR-SATORRAS, R.; VESPIGNANI, A. Epidemic spreading in scale-free networks. *Physical Review Letters*, v. 86, p. 3200–3203, 2001. 24

45 KITSAK, M. et al. Identification of influential spreaders in complex networks. *Nature Physics*, v. 6, p. 888–893, 2010. 24

46 GOLTSEV, A. V. et al. Localization and spreading of diseases in complex networks. *Physical Review Letters*, v. 109, 2012. 24

47 BULLMORE, E.; SPORNS, O. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, v. 10, p. 186–198, 2009. 24

48 REIS, S. D. S. et al. Avoiding catastrophic failure in correlated networks of networks. *Nature Physics*, v. 10, p. 762–767, 2014. 24, 27

49 ERDŐS, P.; RÉNYI, A. On random graphs i. *Publicationes Mathematicae Debrecen*, v. 6, p. 290–297, 1959. 25, 33, 36

50 ERDŐS, P.; RÉNYI, A. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Science*, v. 6, p. 17–61, 1960. 25, 36

51 MILGRAM, S. The small-world problem. *Psychology Today*, v. 1, p. 61–67, 1967. 25, 40

52 BARABÁSI, A. L.; ALBERT, R. Emergence of scaling in random networks. *Science*, v. 286, p. 509–512, 1999. 27, 39, 46

53 FALOUTSOS, M.; FALOUTSOS, P.; FALOUTSOS, C. On power-law relationships of the internet topology. *Computer Communication Review*, v. 29, p. 251–262, 1999. 27

54 LILJEROS, F. et al. The web of human sexual contacts. *Nature*, v. 411, p. 907–908, 2001. 27

55 LOUAIL, T. et al. Uncovering the spatial structure of mobility networks. *Nature Communications*, v. 6, p. 1–8, 2015. 27

56 DUNNE, J. A.; WILLIAMS, R. J.; MARTINEZ, N. D. Food-web structure and network theory: The role of connectance and size. *Proceedings of the National Academy of Sciences*, v. 99, p. 12917–12922, 2002. 27

57 ACHARD, S. et al. A resilient, low-frequency, small-world human brain functional network with high highly connected association cortical hubs. *The Journal of Neuroscience*, v. 26, p. 63–72, 2006. 27

58 JEONG, H. et al. The large-scale organization of metabolic networks. *Nature*, v. 407, p. 651–654, 2000. 27

59 LEVENTHAL, G. E. et al. Evolution and emergence of infectious diseases in theoretical and real-world networks. *Nature Communications*, v. 6, p. 1–11, 2015. 27

60 GAO, J. et al. Networks formed from interdependent networks. *Nature Physics*, v. 8, p. 40–48, 2012. 27

61 CLAUSET, A.; SHALIZI, C. R.; NEWMAN, M. E. J. Power-law distributions in empirical data. *SIAM Rev.*, v. 51, p. 661, 2009. 33, 39, 41

62 MOORE, E. F. The shortest path through a maze. *Proceedings of the International Symposium on the Theory of Switching*, p. 285–292, 1959. 34

63 DIJKSTRA, E. W. A note on two problems in connexion with graphs. *Numerische Mathematik*, v. 1, p. 269–271, 1959. 34

64 BELLMAN, R. 34

65 FORD, L. R. J.; FULKERSON, D. R. *Flows in networks*. [S.l.]: Princeton University Press, 1962. 34

66 CORMEN, T. H. et al. *Introduction to Algorithms*. 3rd. ed. Cambridge: MIT Press, 2009. 34

67 ALBERT, R.; ALBERT, I.; NAKARADO, G. L. Structural vulnerability of the north american power grid. *Physical Review E*, v. 69, p. 025103–1,025103–4, 2004. 35

68 SARDIELLO, M. et al. A gene network regulating lysosomal biogenesis and function. *Science*, v. 325, p. 473–477, 2009. 35

69 SOLOMONOFF, R.; RAPOPORT, A. Connectivity of random nets. *Bulletin of Mathematical Biophysics*, v. 13, p. 107–117, 1951. 36

70 SOLOMONOFF, R. An exact method for the computation of the connectivity of random nets. *Bulletin of Mathematical Biophysics*, v. 14, p. 153–157, 1952. 36

71 GILBERT, E. N. Random graphs. *The annals of Mathematical Statistics*, v. 30, p. 1141–1144, 1959. 36

72 FRONCZAK, A.; FRONCZAK, P.; HOŁYST, J. A. Average path length in random graphs. *Physical Review E*, v. 70, p. 056110, 2004. 39

73 DODDS, P. S.; MUHAMAD, R.; WATTS, J. D. An experimental study of search in global social networks. *Science*, v. 301, p. 827–829, 2003. 40

74 VIRKAR, Y.; CLAUSET, A. Power-law distributions in binned empirical data. *The Annals of Applied Statistics*, v. 8, 2014. 41

75 BENDER, E. A.; CANFIELD, E. R. The asymptotic number of labeled graphs with given degree sequences. *Journal of combinatorial theory*, v. 24, p. 296–307, 1978. 42

76 MOLLOY, M.; REED, B. A critical point for random graphs with a given degree sequence. *Random Structures & Algorithms*, v. 6, p. 161–179, 1995. 43

77 SMITH, M. The trend toward multiple authorship in psychology. *American Psychologist*, v. 13, p. 596–599, 1958. 44

78 PRICE, D. J. de S. *Little Science, Big Science*. New York: Columbia University Press, 1963. 44

79 CLARKE, B. L. Multiple authorship trends in scientific papers. *Science*, v. 143, p. 822–824, 1964. 44

80 WEST, J. et al. The role of gender in scholarly authorship. *PLOS One*, v. 8, p. e66121, 2013. 44, 69, 70, 71

81 WRAY, K. B. The epistemic significance of collaborative research. *Philosophy of Science*, v. 69, p. 150–168, 2002. 44

82 KINCAID, H. *Philosophical Foundations of the Social Sciences: Analyzing Controversies in Social Research*. Cambridge: Cambridge University Press, 1996. 45

83 MELIN, G.; PERSSON, O. Studying research collaboration using co-authorship. *Scientometrics*, v. 36, p. 363–377, 1996. 46

84 TOMASSINI, M.; LUTHI, L. Empirical analysis of the evolution of a scientific collaboration network. *Physica A*, v. 385, p. 750–764, 2007. 47

85 PERC, M. Growth and structure of slovenia's scientific collaboration network. *Journal of Informetrics*, v. 4, p. 475–482, 2010. 47

86 ÇAVUŞOĞLU, A.; TÜRKER, İ. Scientific collaboration network of turkey. *Chaos, Solitons and Fractals*, v. 57, p. 9–18, 2013. 47

87 CHEN, Y.; BÖRNER, K.; FANG, S. Evolving collaboration networks in scientometrics in 1978-2010: a micro-macro analysis. *Scientometrics*, v. 95, p. 1051–1070, 2013. 47

88 HÂNCEAN, M.-G.; PERC, M.; VLASCEANU, L. Fragmented romanian sociology: Growth and structure of the collaboration network. *PLOS One*, v. 9, p. e113271, 2014. 47

89 SAVIĆ, M. et al. The structure and evolution of scientific collaboration in serbian mathematical journals. *Scientometrics*, v. 101, p. 1805–1830, 2014. 47

90 ARAÚJO, E. B. et al. Collaboration networks from a large cv database: Dynamics, topology and bonus impact. *PLOS One*, v. 9, p. e90537, 2014. 47, 70

91 MENA-CHALCO, J. P. et al. Brazilian bibliometric coauthorship networks. *Journal of the Association for Information Science and Technology*, v. 65, p. 1424–1445, 2014. 47, 52

92 O'NEILL, E. T.; ROGERS, S. A.; OSKINS, W. M. Characteristics of duplicate records in oclc's online union catalog. *Libr. Resour. Tech. Serv.*, v. 37, p. 59, 1993. 52

93 WAGNER, R. A.; LOWRANCE, R. An extension of the string-to-string correction problem. *J. Assoc. Comput. Mach.*, v. 22, p. 177, 1975. 52, 53

94 ZHOU, T. et al. Bipartite network projection and personal recommendation. *Phys Rev E Stat Nonlin Soft Matter Phys*, v. 76, p. 046115, 2007. 53

95 ZHANG, J. Growing random geometric graph models of super-linear scaling law. *arXiv e-print*, p. arXiv:1212.4914 [physics.soc–ph], 2012. 55

96 DOROGOVTSEV, S. N.; MENDES, J. F. F. Accelerated growth of networks. In: BORNHOLDT, S.; SCHUSTER, H. G. (Ed.). *Handbook of Graphs and Networks: From the Genome to the Internet*. [S.l.]: Wiley-VCH, 2003. cap. 14, p. 318–341. 55

97 SCHLÄPFER, M. et al. The scaling of human interactions with city size. *Journal of the Royal Society Interface*, v. 11, p. 20130789, 2014. 55

98 BETTENCOURT, L. M. A. et al. Growth, innovation, scaling, and the pace of life in cities. *P. Natl. Acad. Sci. U.S.A.*, v. 104, p. 7301, 2007. 63

99 ALLISON, P. D. Inequality and scientific productivity. *Soc. Stud. Sci.*, v. 10, p. 163, 1980. 65

100 LESLIE, L. L.; MCCLURE, G. T.; OAXACA, R. L. Women and minorities in science and engineering: A life sequence analysis. *The Journal of Higher Education*, v. 69, p. 239–276, 1996. 68, 70

101 HANDELSMAN, J. et al. More women in science. *Science*, American Association for the Advancement of Science, v. 309, n. 5738, p. 1190–1191, 2005. 68, 70

102 DUCH, J. et al. The possible role of resource requirements and academic career-choice risk on gender differences in publication rate and impact. *PLOS One*, v. 7, p. e51332, 2012. 68, 70

103 COLE, J. R.; ZUCKERMAN, H. The productivity puzzle: Persistence and change in patterns of publication of men and women scientists. *Advances in motivation and achievement*, v. 2, 1984. 68, 69, 70

104 LARIVIÈRE, V. et al. Global gender disparities in science. *Nature*, v. 504, p. 211–213, 2013. 68

105 KYVIK, S.; TEIGEN, M. Child care, research collaboration, and gender differences in scientific productivity. *Science, Technology & Human Values*, Sage Publications, v. 21, n. 1, p. 54–71, 1996. 68, 70, 71

106 PRPIĆ , K. Gender and productivity differerentials in science. *Scientometrics*, v. 55, p. 27–58, 2002. 68, 70

107 LEY, T. J.; HAMILTON, B. H. The gender gap in nih grant applications. *Science*, v. 322, p. 1472–1474, 2008. 68

108 BOYLE, P. J. et al. Gender balance: Women are funded more fairly in social science. *Nature*, v. 525, p. 181–183, 2015. 68, 70

109 STEINPREIS, R. E.; ANDERS, K. A.; RITZKE, D. The impact of gender on the review of the curricula vitae of job applicants and tenure candidates: A national empirical study. *Sex Roles*, v. 41, p. 509–528, 1999. 68

110 MOSS-RACUSIN, C. A. et al. Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, v. 109, p. 16474–16479, 2012. 69, 70

111 BOZEMAN, B.; GAUGHAN, M. How do men and women differ in research collaborations? an analysis of the collaborative motives and strategies of academic researchers. *Research Policy*, Elsevier, v. 40, n. 10, p. 1393–1402, 2011. 69, 70, 71

112 EDUCATION at Glance 2012. [S.l.]: Organisation for Economic Co-operation and Development, 2012. 69

113 SAITOVITCH, E. M. B. et al. Gender equity in the brazilian physics community at the present time. *AIP Conference Proceedings*, v. 1697, 2015. 69

114 BARBOSA, M. C. Equity for women in physics. *Physics World*, v. 7, p. 14–15, 2003. 69

115 COLE, S. *Making Science: Between Nature and Scociety*. Cambridge: Harvard University Press, 1992. 69, 70

116 SCHIEBINGER, L. Getting more women into science: knowledge issues. *Harvard Journal of Law & Gender*, v. 30, p. 350, 2007. 70

117 FOX, M. F. Women, science, and academia: graduate education and careers. *Gender and Society*, v. 15, p. 654–666, 2001. 70, 71, 77

118 RHOTEN, D.; PFIRMAN, S. Women in interdisciplinary science: Exploring preferences and consequences. *Research policy*, Elsevier, v. 36, n. 1, p. 56–75, 2007. 70, 71, 77

119 SZELL, M.; THURNER, S. How women organize social networks different from men. *Scientific Reports*, v. 3, p. 1–6, 2013. 70