UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Multi-scale analysis of languages and knowledge through complex networks

Henrique Ferraz de Arruda

Tese de Doutorado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Henrique Ferraz de Arruda

Multi-scale analysis of languages and knowledge through complex networks

Doctoral dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP, in partial fulfillment of the requirements for the degree of the Doctorate Program in Computer Science and Computational Mathematics. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Luciano da Fontoura Costa

Co-advisor: Prof. Dr. Diego Raphael Amancio

USP – São Carlos
March 2019

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

F381m Ferraz de Arruda, Henrique
Multi-scale analysis of languages and knowledge
through complex networks / Henrique Ferraz de
Arruda; orientador Luciano da Fontoura Costa;
coorientador Diego Raphael Amancio. -- São Carlos,
2019.
220 p.

Tese (Doutorado - Programa de Pós-Graduação em
Ciências de Computação e Matemática Computacional) --
Instituto de Ciências Matemáticas e de Computação,
Universidade de São Paulo, 2019.

1. Mineração de Textos. 2. Redes Complexas. 3.
Dinâmicas Relacionadas ao Conhecimento. 4.
Classificação de Textos. I. da Fontoura Costa,
Luciano, orient. II. Amancio, Diego Raphael,
coorient. III. Título.

Henrique Ferraz de Arruda

**Análise multi-escala de línguas e conhecimento por meio de
redes complexas**

Tese apresentada ao Instituto de Ciências
Matemáticas e de Computação – ICMC-USP,
como parte dos requisitos para obtenção do título
de Doutor em Ciências – Ciências de Computação e
Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e
Matemática Computacional

Orientador: Prof. Dr. Luciano da Fontoura Costa
Coorientador: Prof. Dr. Diego Raphael Amancio

USP – São Carlos
Março de 2019

ACKNOWLEDGEMENTS

First of all, I would like to thank God for having the opportunity to attend excellent universities and to be part of a family that has always encouraged me to go ahead.

I would like to express my gratitude to my advisor, Prof. Luciano da Fontoura Costa, and my co-advisor Prof. Diego Raphael Amancio for their support, friendship, and patience. I also acknowledge Prof. Santo Fortunato, who supervised me in my period abroad, at Indiana University. These professors were extremely important to my growth as a scientist. I would like to thank my English teacher, Mara, that helped me directly and indirectly with this thesis and all the texts I wrote during my Ph.D. Additionally, I would like to express appreciation to all of my professors and teachers, which helped me to pursue this Ph.D. thesis.

In particular, I acknowledge my family, my mother Ana Sylvia B. de Arruda my father Juarez A. F. de Arruda, my brother Guilherme F. de Arruda, and my grandparents (in memoriam) for all their encouragement and support. I would like to thank my girlfriend, Caroline Ceribeli for her attention, comprehension, and companionship. Additionally, I would like to express my gratitude to my godfather Frank, my cousin Cristiani and her husband Cesar for their encouragement and support.

Furthermore, I express appreciation to all of my friends from USP, that supported me during this process, in special Alexandre Benatti, Bruno M. F. de Resende, César H. Comin, Cynthia M. V. Couto, Diego H. Negretto, Edilson A. Corrêa Júnior, Edmilson R. dos Santos, Eduardo C. Padovani, Felipe L. Gewers, Filipi N. Silva, Juliana V. Silva, Paulo C. V. da Silva, Paulo E. P. Burke, Paulo J. P. de Souza, Thales S. Lima, Thomas K. D. M. Peron, and Vanessa Q. Marinho. I would like to thank my friends Sandro M. Reia, Ariadne A. Costa, and Guoqiang Liang from Indiana University. I would like to acknowledge two special friends that have inspired me since I was in the college, André L. Pilastrri and Carlos Alex S. J. Gulo. Also, I would like to thank all USP staff, especially the concierges Ivan, Adriano, and Anésio. I would like to acknowledge all my friends from jiu-jitsu. Furthermore, I would like to thank all my friends, so as not to risk forgetting anyone, I do not list all the names here.

I also acknowledge the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil* (CAPES) - Finance Code 001, for financial support and the *Mobilidade Internacional Santander Universidades*, for a period abroad in Indiana University – Bloomington, Indiana, USA. This research was carried out using the computational resources of

the Center for Mathematical Sciences Applied to Industry (CeMEAI) funded by FAPESP (grant 2013/07375-0). This work has been supported also by FAPESP grants 11/50761-2 and 2015/22308-2.

“Persistence is the shortest path to success”.
(Charles Chaplin)

RESUMO

ARRUDA, H. F. **Análise multi-escala de línguas e conhecimento por meio de redes complexas**. 2019. 220 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2019.

Existem diversos aspectos das linguagens naturais e de dinâmicas relacionadas que estão sendo estudadas. No caso das línguas, algumas análises quantitativas foram feitas usando modelos estocásticos. Ademais, linguagens naturais podem ser entendidas como sistemas complexos. Para analisar linguagens naturais, existe a possibilidade de utilizar o conjunto de ferramentas que já foram desenvolvidas para analisar redes complexas, que são representadas computacionalmente por grafos. Além disso, tais ferramentas podem ser utilizadas para representar e analisar algumas dinâmicas relacionadas a redes complexas. Observe que o conhecimento está intrinsecamente relacionado à linguagem, pois a linguagem é o veículo usado pelos seres humanos para transmitir novas descobertas, sendo que a própria linguagem também é um tipo de conhecimento. Esta tese é dividida em dois tipos de análise: (i) textos e (ii) aspectos dinâmicos. Na primeira parte foram propostas representações de redes de texto em diferentes escalas de análise. A partir da análise do estilo de escrita, considerando redes de adjacência de palavras (co-ocorrência) para entender padrões locais de palavras, até uma representação mesoscópica, que é criada a partir de pedaços de texto e que representa informações do texto de acordo com o desenrolar da história. Na segunda parte, foram consideradas a estrutura e dinâmica relacionadas ao conhecimento e à linguagem. Neste caso, partiu-se da escala maior, com a qual estudamos a conectividade entre física aplicada e física teórica. A seguir, simulou-se a aquisição de conhecimento por pesquisadores em uma dinâmica multi-agente e uma máquina inteligente que resolve problemas, que é representada por uma rede. Como a menor escala considerada, foi simulada a transmissão de redes. Essa transmissão considera os dados como uma série de símbolos organizados que são obtidos a partir de uma dinâmica. Para melhorar a velocidade de transmissão, a série pode ser compactada. Para tanto, foi utilizada a teoria da informação e o código de Huffman. As propostas de abordagens baseadas em rede foram consideradas adequadas para lidar com a análise empregada, em todas as escalas testadas.

Palavras-chave: Mineração de Textos, Redes Complexas, Dinâmicas Relacionadas ao Conhecimento e Classificação de Textos.

ABSTRACT

ARRUDA, H. F. **Multi-scale analysis of languages and knowledge through complex networks**. 2019. 220 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2019.

There are many different aspects in natural languages and their related dynamics that have been studied. In the case of languages, some quantitative analyses have been done by using stochastic models. Furthermore, natural languages can be understood as complex systems. Thus, there is a possibility to use the set of tools developed to analyze complex networks, which are computationally represented by graphs, also to analyze natural languages. Furthermore, these tools can be used to represent and analyze some related dynamics taking place on the networks. Observe that knowledge is intrinsically related to language, because language is the vehicle used by humans beings to transmit discoveries, and the language itself is also a type of knowledge. This thesis is divided into two types of analyses: (i) texts and (ii) dynamical aspects. In the first part, we proposed network representations of text in different scales of analyses, starting from the analysis of writing style considering word adjacency networks (co-occurrence) to understand local patterns of words, to a mesoscopic representation, which is created from chunks of text and grasps information of the unfolding of the story. In the second part, we considered the structure and dynamics related to knowledge and language, in this case, starting from the larger scale, in which we studied the connectivity between applied and theoretical physics. In the following, we simulated the knowledge acquisition by researchers in a multi-agent dynamics and an intelligent machine that solves problems, which is represented by a network. At the smallest considered scale, we simulate the transmission of networks. This transmission considers the data as a series of organized symbols that is obtained from a dynamics. In order to improve the speed of transmission, the series can be compacted. For that, we considered the information theory and Huffman code. The proposed network-based approaches were found to be suitable to deal with the employed analysis for all of the tested scales.

Keywords: Text Mining, Complex Networks, Dynamics related to Knowledge, and Text Classification.

LIST OF FIGURES

Figure 1	– A network representation of the relationships among concepts studied in this thesis, in which the nodes and edges represent concepts and their relationships. Furthermore, the edge labels represent the part of the thesis where the connected concepts are described.	40
Figure 2	– Example of a network and its respective adjacency matrix. The matrix is symmetrical because the network is undirected.	48
Figure 3	– Example of degree considering the red node. The left-hand figure shows the measurement for undirected networks, and the right-hand figure depicts the two possibilities of both degrees for directed networks, in-degree (k^{in}) and out-degree (k^{out}).	49
Figure 4	– Examples of the strength of the node represented in red color. The left-hand example illustrates the measurement without direction, and the right-hand represents the directed version, in-strength, s^{in} , and out-strength, s^{out}	49
Figure 5	– Concentric levels of the red node, in which the nodes represented by the colors blue and orange take part in the concentric levels 1 and 2, respectively.	50
Figure 6	– Examples of the simplification of the network shown in item (a), by considering the symmetry of the red node and the access of all other nodes. These access probabilities are computed according to the self-avoiding random walk. Item (b) and (c) represent backbone and merged symmetries, respectively.	51
Figure 7	– Examples of clustering coefficient of node i , shown in red color. When all of the triples connected to i are also triangles, $cc_i = 1$ and when there are no triangles connected to i , $cc_i = 0$. Furthermore, in all other cases $0 \leq cc_i \leq 1$	52
Figure 8	– Example of the eccentricity measurement for all nodes of a network, in which the network diameter is 3 and the radius is 2.	54
Figure 9	– Example of closeness centrality for a tree network, in which the highest value consists of the root and the lowest of the leaves.	55
Figure 10	– Example of betweenness centrality for all nodes of a network. The nodes with higher values are part of the path connecting the communities.	55

Figure 11 – Example of ER network visualization (1000 nodes and $p = 0.01$). The color bar represents the network node degrees. The employed software of visualization was implemented by Silva (2015).	59
Figure 12 – Example of WS network visualization (1000 nodes, $n = 3$ and $p = 0.01$). The color bar represents the network node degrees. The employed software of visualization was implemented by Silva (2015).	60
Figure 13 – Example of BA network visualization (1000 nodes and $m = 3$). The color bar represents the network node degrees. The employed software of visualization was implemented by Silva (2015).	61
Figure 14 – Example of the visualization of the community network proposed by Lancichinetti, Fortunato and Radicchi (2008) (1000 nodes, $\mu = 0.01$, $s_{min} = 100$, $s_{max} = 600$, $k_{min} = 4$, $k_{max} = 50$, and $\mu = 0.01$). The color bar represents the network node degrees. The employed software of visualization was implemented by Silva (2015).	62
Figure 15 – Example of Wax network visualization (1000 nodes, $\alpha = 1$, and $\beta = 0.02$). The color bar represents the network node degrees. The employed software of visualization was implemented by Silva (2015).	63
Figure 16 – Example of Geo network visualization (1000 nodes, $d = 0.05$). The color bar represents the network node degrees. The employed software of visualization was implemented by Silva (2015).	64
Figure 17 – Example of KN network visualization (50 nodes).	65
Figure 18 – The tokens of the phrase "What is a scientist?" and the respective part-of-speech tags.	69
Figure 19 – Network representing a paragraph extract from the novel Emma. Each node is a lemma in the text.	75
Figure 20 – LSA visualization of informative and imaginative texts. The ten most frequent words of the texts are employed as features. It is possible to identify two groups. The contribution of the axis D1 and D2 are 0.45 and 0.24, respectively.	85
Figure 21 – PCA of the network with stopwords, LSS, where it is possible to see a clear separation between the groups.	87
Figure 22 – Density of the backbone symmetry ($Sb^{(h=3)}$) of the word "the". In the imaginative texts, $Sb^{(h=3)}$ tend to be higher.	88
Figure 23 – PCA of the features of stopwords count, by employing the bag-of-words approach. In this plot there is a clear separation between the two groups.	88
Figure 24 – Results obtained from the approach based on bag-of-characters. It is possible to see a clear separation between the groups.	89

Figure 25 – Examples of EC, PB and APB networks generated from a text about cars. The colors of the nodes represent the respective communities obtained by using Fast Greedy method.	91
Figure 26 – Example of the employed subject detection method, in which three communities were found, and we aim at finding only two subjects. . . .	93
Figure 27 – Examples of EC, PB and APB networks generated from a text about cars by applying the method of subject detection. The node colors represent the respective communities and subjects, by considering $n_s = 2$ and Fast Greedy community detection.	94
Figure 28 – Example of the evolution of modularity measurement and the respective normalization for the methods EC, PB, and APB. The presented plots were obtained by employing the FGS database. There are no significant changes of modularity value with $\omega > 20$	96
Figure 29 – NMI computed between the found topics against the real topics and subtopics of CGF texts.	98
Figure 30 – NMI computed between the found topics against the real topics and subtopics of FGS texts.	99
Figure 31 – Example of the paragraph-based network for a text with five paragraphs, which is shown in item (a). Item (b) represents the tf-idf of the paragraphs, and (c) and (d) are the weighted and unweighted versions of the network. In the unweighted network the threshold T provides an average degree of two.	101
Figure 32 – Comparisons of the measurements that provided lowest values CV, by considering the average between CV computed for Books and Bible datasets. The error bars represent the confidence interval of a mean for the interval of 95% of confidence.	109
Figure 33 – Probability density function (pdf) of the first component of the linear discriminant analysis projection of the selected features which were employed in classifying the texts into two classes. In his analysis we compared two classes, which are <i>Real</i> against <i>Rhuffled</i> (SW and SS) texts.	110
Figure 34 – PCA projection computed from the <i>doc2vec</i> features. There are two groups, where one represents SW, and the other group represents RT and SS.	111

Figure 35 – Paragraphs-based network of Voynich manuscript. The labels consist in topics that were manually defined according to the manuscript pictures, as follows: (i) <i>text</i> , when there are no images; (ii) <i>plants</i> ; (iii) <i>bath</i> , with figures of women and bath-like shape; (iv) <i>astronomy</i> , with spatial-like figures (e.g., stars); and (v) <i>vase</i> . The employed software of visualization was implemented by Silva (2015).	113
Figure 36 – Distributions of Probability Density Function, PDF, by considering the modularity measurement of RT, SW, and SS of networks. The modularity of the two versions of the Voynich manuscript, RT and SW, are represented by blue and orange circles, respectively.	114
Figure 37 – Example of mesoscopic network for text with six paragraphs, shown in item (a). Item (b) represents the tf-idf of the paragraphs, and (c) and (d) are the weighted and unweighted versions of the network, respectively. When T is employed, the resultant unweighted network has an average degree equals to two.	115
Figure 38 – Scheme illustrating the mapping of texts into networks and the classification process. Initially, documents are mapped into networks. Then, two types of features are calculated: network features and attributes derived from network visualizations. The classification can then be performed using both network and image features.	117
Figure 39 – Visualization of the $\mathcal{G}_1$ network, which represents the book “Alice’s Adventures in Wonderland”, by Lewis Carroll. Such a visualization was projected by considering the threshold $T_1 = 0.31$, and the nodes represent sets of 20 paragraphs. The paragraphs order are shown in item (a), and item (b) represents the book chapters. The node is assigned to the paragraph that is in the middle of the set of paragraphs. In this example is 10. The employed software of visualization was implemented by Silva (2015).	119
Figure 40 – Visualization of the $\mathcal{G}_2$ network, which represents the book “Alice’s Adventures in Wonderland”, by Lewis Carroll. Such a visualization was projected by considering the threshold $T_1 = 0.31$, and the nodes represent sets of 20 paragraphs. The paragraphs order are shown in item (a), and item (b) represents the book chapters. The node is assigned to the paragraph that is in the middle of the set of paragraphs. In this example is 10. The employed software of visualization was implemented by Silva (2015).	121

Figure 41 – Visualization of the networks representing “Alice’s Adventures in Wonderland”. Item (a) represents the network $\mathcal{G}_1$ ($T_1 = 0.31$) and item (b) represents the network $\mathcal{G}_2$ ($T_2 = 0.18$). Clustering coefficient cc is indicated by the node colors, in which nodes with the highest values are represented in orange. There is a non-trivial pattern of clustering coefficient along the network nodes. The employed software of visualization was implemented by Silva (2015).	122
Figure 42 – Clustering coefficient, cc , measured from all network nodes of real and shuffled versions created from the book “Alice’s Adventures in Wonderland” (RT, SW, and SP). This example was measured for $T_1 = 0.31$, in order to select the strongest semantical links.	123
Figure 43 – PCA projections of the networks generated from Real Texts (RT), Shuffled Paragraphs (SP) texts, and Shuffled Words (SW) texts. The projections (a) and (b) represent the mesoscopic and the co-occurrence networks, respectively. The circles and stars represent English and Other languages datasets. In co-occurrence analysis, only the English texts were considered. These plots also show the densities of the samples according to the PCA components, in which the continuous and dashed lines represent English and Other languages datasets, respectively.	125
Figure 44 – Example of time difference between linked nodes against their respective edges weights, by considering “Alice’s Adventures in Wonderland” by Lewis Carroll. This measurement was computed for all network edges and the inset represents a region of long-range links. Comparing the different classes of texts, it is evident that strong long-range connections are more likely to appear in real networks.	127
Figure 45 – Examples of the visualization of networks created for texts of three different authors: Edgar Allan Poe, Thomas Hardy, and Mark Twain. The visualizations of each author seem to follow a specific type of pattern. The employed software of visualization was implemented by Silva (2015).	129
Figure 46 – LDA projection by considering all measurements obtained from the selected network measurements and image-based measurements. Each color represents a writer. The plots also show the densities of samples according to the respective classes (writers). Points that represent books of the same writer seem to be grouped on the scatter plot. . . .	131

Figure 47 – Cohesion (blue), defined as the reciprocal of the shortest path length, and betweenness centrality (green) relationships between Phys. Rev. D, Phys. Rev. Applied and Other APS journals for the directed network. The legend at the upper right corner indicates the width range of the links. Observe that all links have similar width, suggesting symmetric relationships between all groups.	136
Figure 48 – Cohesion (blue) and betweenness centrality (green) relationships between Phys. Rev. D, Phys. Rev. Applied and Other APS journals for the undirected network. The legend at the upper right corner indicates the width range of the links.	137
Figure 49 – Cohesion (blue) and betweenness centrality (green) relationships between Phys. Rev. D, Phys. Rev. B and Other APS journals for the directed network. The legend at the upper right corner indicates the width range of the links.	138
Figure 50 – Cohesion (blue) and betweenness centrality (green) relationships between Phys. Rev. D, Phys. Rev. Applied and Other APS journals for the directed network. The legend at the upper right corner indicates the width range of the links.	139
Figure 51 – Average of the total number of explored nodes ($\langle TE \rangle$) and the corresponding standard deviations, represented by the shadow regions, executed in LA.	142
Figure 52 – Average of the total number of explored nodes ($\langle TE \rangle$) and the corresponding standard deviations, represented by the shadow regions, computed by varying the jump probability (γ).	143
Figure 53 – Average of the total number of explored nodes ($\langle TE \rangle$) and the corresponding standard deviations of our dynamics, obtained by varying the values of γ when executed on WAX network.	144
Figure 54 – Average of the total number of explored nodes ($\langle TE \rangle$) and the corresponding standard deviations when executed on BA network. The influence of η and τ parameters were not found to be significative. However, higher values of jump probability, γ , results in a more efficient exploration.	145
Figure 55 – $\langle TE_{1000} \rangle$ performances and the respective standard deviations computed from the dynamics execution for each network region, defined by Chebyshev distance.	146
Figure 56 – Performances, $\langle TE_{1000} \rangle$, computed for network regions for distinct values of η . These regions are defined according to accessibility measurement with $h = 3$	147

Figure 57 – Performances, $\langle TE_{1000} \rangle$, computed for network regions for distinct values of γ . These regions are defined according to accessibility measurement with $h = 3$	149
Figure 58 – Scheme of how the dynamic works. The first part represents the set of nodelets, which are disconnected, and the second part represents the network. The connections are created only when there is a correct fit between a pair of nodelets. Additionally, the arrow represents RE	150
Figure 59 – Fraction of assembled N_P nodes as a function of normalized time. Here we plot all combinations between N_P and N_{RE} . Observe that for each type of N_P a different behavior was found, but N_{RE} did not change the results significantly.	153
Figure 60 – By considering $N_P = ER$, and N_{RE} networks as a variation of the WAX model, we plot the average and standard deviation of the time to mount an entire puzzle (T_F). Observe that the lowest values of T_F were found when the assigned priorities of N_P nodelets on N_{RE} were considered as the higher values of EC.	154
Figure 61 – Average and standard deviation of the fraction of completed puzzles against buffer size. In this case, we considered $N_{RE} = BA$ and $N_P = BA$. Due to the high standard deviations, here we represent only 20% of this measurement. For some values of buffer size, the assigned priority of k -lower is more effective than the random choice.	155
Figure 62 – Comparisons of times to mount at least 90% of the nodelets of one N_P among the assigned priorities. For all nodes, we considered buffer size of four positions. The colors of the plots represent the different versions of N_{RE}	157
Figure 63 – Comparisons of times to mount at least 90% of the nodelets of one N_P among the assigned priorities. For all nodes, we considered buffer size equal to the respective node degree. Consequently, the average buffer size is four. The colors of the plots represent the different versions of N_{RE}	158
Figure 64 – Normalized time computed between the time to mount at least 90% of the nodelets of N_P . Observe that the colors represent the different versions of N_{RE}	159
Figure 65 – Scheme representing the proposed framework. Items (a), (b), and (c) represent the original network, the obtained time series, and the reconstructed network, respectively.	161

Figure 66 – Results regarding all the employed network models and varying the random walk dynamics for undirected networks. The parallel coordinates present the average and standard deviation. Items (a), (b), (c), (d), (e), and (f) represent T_{90} , T_{90}^C , S_{90} , S_L , R_{90} , and R_L	164
Figure 67 – Results of all the employed network models and varying the random walk dynamics for directed networks. These results are presented as parallel coordinates and present the averages and standard deviations of the measurements. Items (a), (b), (c), (d), (e), and (f) represent T_{90} , T_{90}^C , S_{90} , S_L , R_{90} , and R_L	166
Figure 68 – Average compression rations (R_L) computed by employing word statistics of real books and the average R_L from network topology (degree) as Huffman dictionary. In the case of network topologies, the dashed lines represent the average value of R_L	168
Figure 69 – PCA projection computed from network measurements obtained from real and artificial languages (30 artificial samples for each real network). The circles and squares represent artificial and real languages, respectively. More specifically, item (a) represents the PCA of all considered measurements, items (b) and (c) represent the PCA only with the standard deviation of the measurements and with the remaining measurements (without the standard deviation), respectively.	217
Figure 70 – The heatmap represents the Euclidean distances between networks created from all (real and artificial) networks, by considering the two first dimensions of PCA. The character star represents the artificial languages. The dendrogram and the matrix order are computed from the heatmap matrix by considering Ward's method.	218
Figure 71 – The heatmap represents the cosine distances between networks created from all (real and artificial networks) by considering all employed network measurements. The character star represents the artificial languages. The dendrogram and the matrix order are computed from the heatmap matrix by considering Ward's method.	219
Figure 72 – For all the considered measurements, the module of z-score computed from the measurement obtained from the real and the group of artificial language networks. The measures are ordered in increasing order, according to the average z-scores of all languages.	220

LIST OF CHARTS

Chart 1	– Example of the pre-processing steps by the novel <i>Emma</i> by Jane Austen.	74
Chart 2	– Chapters of Alice’s Adventures in Wonderland, by Lewis Carroll.	118
Chart 3	– The adopted groups of articles and some of their properties. The used dataset comprises articles from 1893 to 2016. Control (A) and Control (B) represent the Control group of articles when, respectively, <i>Phys. Rev. Applied</i> and <i>Phys. Rev. B</i> is employed.	135
Chart 4	– Informations concerning the networks used to execute knowledge acquisition dynamics.	141
Chart 5	– Configurations of N_P and N_{RE} networks employed to Case 1.	152
Chart 6	– Configurations of N_P and N_{RE} networks employed to Case 2.	154

LIST OF ALGORITHMS

Algorithm 1 – Algorithm of subject detection by using network communities. . .	93
Algorithm 2 – K-means algorithm.	207

LIST OF TABLES

Table 1	– Accuracy obtained from the three different approaches and the three classifiers. The highest accuracy was obtained by using LSS.	86
Table 2	– Measurements obtained for RT, SS, and SW networks by considering Holy Bible and Books dataset. To be possible to compare different measurements, all the presented data is standardized.	105
Table 3	– Percentage of documents in each dataset where the difference between a real text (English part) and the corresponding shuffled version was found to be significant. Apart from the modularity, the informativeness seems to depend on the type of dataset used.	107
Table 4	– Average distance among mesoscopic networks from the same class. When using mesoscopic networks, it is possible to discriminate real texts from those generated by both shuffled words and paragraph.	124
Table 5	– Average distance among co-occurrence networks from the same class. If co-occurrence networks are used, real texts and texts of shuffled paragraphs cannot be discriminated.	126
Table 6	– Comparison of different network approaches, by considering the K-means clustering performance. Two different measurements were applied: Adjusted Rand Index (ARI) and Accuracy. In both cases, 1 indicates that all instances are correctly classified, and 0 indicates the opposite.	126
Table 7	– Classification accuracy and number of used features considering different measurements.	130
Table 8	– Fractions of N_P mounted entirely, in which the dynamics was executed, by considering the buffer size of 4. Observe that rows and columns represent N_{RE} and N_P , respectively.	156
Table 9	– Fractions of N_P mounted entirely, in which the dynamics were executed, by considering the buffer size equal to the degree of the nodes in N_{RE} . The rows and columns represent N_{RE} and N_P , respectively. Observe that the average node degree is four, which means that the average buffer sizes are the same as the results presented in Table 8.	156

LIST OF ABBREVIATIONS AND ACRONYMS

KNN	<i>K</i> Nearest Neighbors
APB	Adapted Paragraph-Based
APS	American Physical Society
BA	Barabási–Albert
BOC	Bag Of Characters
BOW	Bag-Of-Words
CART	Classification And Regression Tree
CN	Community Network Lancichinetti Benchmark (LB)
DOI	Digital Object Identifier
EC	Extended Co-occurrence
EM	Expectation Maximization
ER	Erdős–Rényi
Geo	Random geometric graph
GS	Global Strategy without stopwords
IF	Image-based Features
IG	Information Gain
LA	Bidimensional Lattice
LB	Lancichinetti Benchmark
LDA	Linear Discriminant Analysis
LS	Local Strategy without stopwords
LSA	Latent Semantic Analysis
LSS	Local Strategy with stopwords
NF	Network Features
NLKT	Natural Language ToolKit
PB	Paragraph-Based
PDF	Probability Density Function
RT	Real Texts
RWD	Degree biased random walk
RWID	Inverse degree biased random walk
SAW	Self-Avoiding Walk
SMO	Sequential Minimal Optimization

SP	Shuffled Paragraphs
SS	Shuffled Sentences
SVD	Singular-Value Decomposition
SVM	Support Vector Machine
SW	Shuffled Words
TLA	Bidimensional Toroidal Lattice
TSAW	True Self-Avoiding Walk
Wax	Waxman
WIKI	Wikipedia network
WOS	Web of Science
WS	Watts-Strogatz

LIST OF SYMBOLS

G — Graph

V — Set of nodes (nodes)

E — Set of edges

N — Number of nodes

A — Adjacency matrix

W — Weighted matrix

w_{ij} — Weight of the edge between the nodes i and j

k — Degree

k^{in} — In-degree

k^{out} — Out-degree

s — Strength

s^{in} — In-strenght

s^{out} — Out-strenght

Nc — Neighborhood connectivity

h — Concentric level

α^h — Accessibility by considering the h th concentric level

$p_{ij}^{(h)}$ — Probability of the agent to reach to a given node j , departing from i , by walking through h edges

S^h — Symmetry by considering the h th concentric level

$\Gamma_h(i)$ — Set of neighbors of i that take part in the same hierarchical level h

Sb — Backbone symmetry

Sm — Merged symmetry

$N_{\Delta}(i)$ — Number of triangles connected to node i

cc — Clustering centrality

μ — Matching index

$d(i, j)$ — Shortest path length

E — Eccentricity

C — Closeness centrality

B — Betweenness centrality

EB — Edge Betweenness centrality

α^∞ — Generalized accessibility

Pr — Page rank

Q — Modularity

c_i — Community of the vertex i

δ — Kronecker delta

M — Number of edges

μ — Mixing parameter

k_{min} — Minimum degree

k_{max} — Maximum degree

s_{min} — Minimum community size

s_{max} — Maximum community size

$d_E(u, v)$ — Euclidean distance between u and v .

L — Maximum Euclidean distance between all pairs of nodes

β — Constant of Waxman model

α — Constant of Waxman model

d — Document

D — Set of documents

tf-idf — Term frequency-inverse document frequency

w — Word

$|D|$ — Number of documents

tf — Term frequency

idf — Inverse document frequency

F_w^d — Frequency of the word w in a given document d

F_w^D — Frequency of the word w in the a given documents D

f_w — Number of documents with the word w

idf^* — Variation of idf

tf^* — Variation of tf

θ — Angle

M — Latent semantic analysis matrix

n_w — Total number of words

U — Matrix of the coordinates of words in the LSA space

V^* — Matrix of the coordinates of texts in the LSA space

Σ — Matrix of singular values of LSA

f — Frequency

ω — Number of words

n_s — Desired number of subjects

$\mathcal{C}$ — Set of communities

Π — Set of paragraphs

n_c — Number of obtained communities

$\mathcal{L}$ — Set of labels

$\sigma(c_i)$ — Overlap of the community c_i

Q_N — Normalized modularity

Q_S — Modularity of shuffled networks

T — Threshold

$\langle \dots \rangle$ — average

σ — Standard deviation

$z(X)$ — z-score of a distribution X

CV — Coefficient of Variation

$\mathcal{O}$ — Organized text

Δ — Number of paragraphs

$\mathcal{G}_1$ — Network of “Alice’s Adventures in Wonderland”, by considering $T_1 = 0.31$

$\mathcal{G}_2$ — Network of “Alice’s Adventures in Wonderland”, by considering $T_2 = 0.18$

δ — window size

$\mathcal{X}$ — Subsequence of X

$\mathcal{C}_v^\delta(X)$ — Average values of $CV(X)$ by considering subsets of window size δ

$\mathcal{F}$ — Set of features

$\psi(i, j)$ — Function of True Self-Avoiding Walk

Ω — constant of True Self-Avoiding Walk

$n_v(j)$ — Number of times node j was already visited

$P_w(i, j)$ — Transition probabilities of True Self-Avoiding Walk

a — Random walk agent

γ — Jump probability

$E_a(i)$ — Function of the field of attraction

τ — Constant of function decay (field of attraction)

η — Vector of importance (field of attraction)

ε — Resultant field

A — Set of agents

TE — Total number of explored nodes

n_R — Window size

RE — Reasoning Engine

N_{RE} — Reasoning engine network

N_P — Puzzle network

T_F — Time to mount the first puzzle

T_{90} — Time series size that provides 90% of the network reconstruction

T_{90}^C — Size of the compressed time series that provides 90% of network reconstruction

R_L — Long term transmission ratio

T_L — Constant of large time

T_L^C — Compressed time series size

R_{90} — Transmission ratio that provides 90% of the network reconstruction

S — Steering coefficient

r — reciprocity

$Ne^{\leftrightarrow}$ — Number of edges that have both directions

Ne — Total number of edges

b — Book

CONTENTS

1	INTRODUCTION	37
1.1	Objectives	41
1.2	Contributions	42
1.3	Thesis organization	45
2	BASIC CONCEPTS	47
2.1	Complex network definitions	47
2.2	Characterization of complex networks	48
2.2.1	<i>Local measurements</i>	48
2.2.2	<i>Global measurements</i>	52
2.2.3	<i>Network communities</i>	57
2.3	Network models	58
2.3.1	<i>Erdős-Rényi model</i>	58
2.3.2	<i>Watts-Strogatz model</i>	58
2.3.3	<i>Barabási-Albert model</i>	59
2.3.4	<i>Network model of communities</i>	61
2.3.5	<i>Waxman model</i>	62
2.3.6	<i>Random geometric graph</i>	63
2.3.7	<i>Knitted</i>	64
2.4	Network visualization	65
2.5	Complex networks dynamics	66
2.6	Characteristics of natural languages	67
2.6.1	<i>Text pre-processing</i>	68
2.6.2	<i>Bag-of-words</i>	69
2.6.2.1	<i>Tf-idf</i>	70
2.6.2.2	<i>Cosine similarity</i>	71
2.6.2.3	<i>Latent Semantic Analysis</i>	71
2.6.3	<i>Bag-of-character</i>	72
2.7	Complex network-based models	72
2.7.1	<i>Word networks</i>	72
2.7.2	<i>Topic segmentation</i>	73
2.7.3	<i>Applications of complex networks to text mining problems</i>	75

3	NETWORK-BASED APPROACHES TO TEXT MINING	79
3.1	Databases	79
3.1.1	<i>Brown Database</i>	79
3.1.2	<i>Gutenberg</i>	80
3.1.3	<i>Holy Bible dataset</i>	81
3.1.4	<i>Voynich Manuscript</i>	81
3.1.5	<i>Artificial texts</i>	81
3.2	Classification between imaginative and informative texts	82
3.2.1	<i>From texts to co-occurrence networks</i>	83
3.2.2	<i>Pattern recognition and evaluation</i>	84
3.2.3	<i>Results and discussion</i>	84
3.2.4	<i>Conclusions</i>	89
3.3	Network-based subject detection	90
3.3.1	<i>Network approaches</i>	90
3.3.2	<i>Detection of subjects via community detection in networks</i>	92
3.3.3	<i>Modularity analysis</i>	93
3.3.4	<i>Results and discussion</i>	95
3.3.5	<i>Conclusions</i>	97
3.4	Paragraph-based networks	100
3.4.1	<i>Paragraph-based network approach</i>	100
3.4.2	<i>Network variations</i>	101
3.4.3	<i>Informativeness analysis</i>	102
3.4.4	<i>Dependency with language and semantics</i>	103
3.4.5	<i>Results and discussion</i>	103
3.4.5.1	<i>Informativeness analysis</i>	104
3.4.5.2	<i>Dependency with language and semantics</i>	108
3.4.5.3	<i>Classification tests</i>	108
3.4.5.3.1	<i>Discriminating between real and shuffled texts</i>	108
3.4.5.3.2	<i>Case Example: Voynich manuscript</i>	111
3.4.6	<i>Conclusions</i>	112
3.5	Mesoscopic analysis of texts	114
3.5.1	<i>Mesoscopic network approach</i>	114
3.5.2	<i>Classification between real and shuffled texts</i>	115
3.5.3	<i>Image-based features</i>	116
3.5.4	<i>Results and discussion</i>	117
3.5.4.1	<i>Case study: Mesoscopic analysis of "Alice's adventures in wonderland"</i>	118
3.5.4.2	<i>Discriminating real from shuffled texts</i>	120
3.5.4.3	<i>Use case: Authorship attribution</i>	127
3.5.5	<i>Conclusions</i>	130

4	ANALYSIS OF LANGUAGE AND KNOWLEDGE VIA NETWORK DYNAMICS	133
4.1	Datasets	133
4.1.1	<i>APS dataset</i>	134
4.1.2	<i>Knowledge representation</i>	134
4.2	Integration between theoretical and applied physics	134
4.2.1	<i>Methods</i>	134
4.2.2	<i>Results and discussion</i>	136
4.2.2.1	<i>Case Study #1 - Phys. Rev. Applied</i>	136
4.2.2.2	<i>Case Study #2 - Phys. Rev. B</i>	137
4.2.3	<i>Conclusions</i>	138
4.3	Simulating knowledge acquisition	139
4.3.1	<i>Problem definition</i>	140
4.3.2	<i>Results</i>	141
4.3.2.1	<i>Global analysis</i>	142
4.3.2.2	<i>Analysis by regions</i>	144
4.3.3	<i>Conclusions</i>	148
4.4	Intelligent complex networks	150
4.4.1	<i>Dynamics definitions</i>	150
4.4.2	<i>Results</i>	151
4.4.2.1	<i>Case 1: Infinite Buffer Size</i>	152
4.4.2.2	<i>Case 2: Limited Buffer Size</i>	154
4.4.3	<i>Conclusions</i>	156
4.5	Connecting network science and information theory	160
4.5.1	<i>Problem definition</i>	160
4.5.2	<i>Random walk dynamics</i>	162
4.5.3	<i>Analysis of network compression</i>	162
4.5.4	<i>Results and discussion</i>	163
4.5.5	<i>Use case: language compression</i>	165
4.5.6	<i>Conclusions</i>	168
5	CONCLUSIONS AND FUTURE WORKS	171
	BIBLIOGRAPHY	177
APPENDIX A	LIST OF PUBLICATIONS	199
APPENDIX B	LIST OF STOP WORDS	201
APPENDIX C	MACHINE LEARNING	203

C.1	Standardization	203
C.2	Attribute selection	203
C.2.1	<i>Information gain</i>	204
C.2.2	<i>SVM attribute selection</i>	204
C.3	Classification	204
C.3.1	<i>Naive Bayes</i>	204
C.3.2	<i>KNN</i>	205
C.3.3	<i>CART</i>	205
C.3.4	<i>SVM</i>	206
C.4	Data clustering	206
C.4.1	<i>K-means</i>	206
C.4.2	<i>Expectation maximization</i>	206
C.5	Quantification of relevance of the characteristics	207
APPENDIX D	IMAGE ANALYSIS	209
APPENDIX E	INFORMATION THEORY	211
E.1	Entropy	211
E.2	Information gain	212
E.3	Huffman	212
APPENDIX F	LANGUAGE COMPARISON: SYLLABLES-BASED AP- PROACH	213
F.1	Syllable dataset	213
F.2	Language model	214
F.2.1	<i>Network model</i>	215
F.2.2	<i>Results</i>	215

INTRODUCTION

Natural languages vary widely among different groups of people. There are evident differences between distinct languages, or the same language when they are spoken by people that live in distinct regions (MEECHAM; REES-MILLER, 2001). These variations can be found through characteristics of grammar or in spoken language and can be influenced by the demography of the regions (AGUIRRE; NYERGES, 2014). Furthermore, natural languages go through a complex and natural biological adaptation (PINKER; BLOOM, 1990). However, apart from these differences, it is arguable that natural languages follow a universal grammar law that determines a “single human language” (CHOMSKY *et al.*, 2000). In other words, natural languages share many of their characteristics. Furthermore, its acquisition is intrinsically related to the characteristic structure of the humans brain (CHOMSKY *et al.*, 2000).

Methodologies regarding statistics are amply applied to analyze characteristics of language, such as in the comparison between statistics of different languages (PIANTADOSI, 2014). One approach widely used consists in considering only the frequency of words, and is known as *bag of words* (LEBANON; MAO; DILLON, 2007). By doing so, characteristics that describe texts and language were observed, such as the scale-free distribution of word frequencies (PIANTADOSI, 2014). Texts can be converted into diverse network representations, such as the co-occurrence of words (BREDE; NEWTH, 2008) and semantical dependency networks (FERRER I CANCHO; SOLÉ; KÖHLER, 2004). However, there are few quantitative analyses and, in general, these studies tend to be constrained (GHANBARNEJAD *et al.*, 2014).

Language can be understood as a complex system (STEELS, 2000), and a proper type of approach that models this type of system is the use of complex networks (FERRER I CANCHO; SOLÉ, 2001). Networks, or graphs, are vastly used to describe many real systems, in which there is a high number of connected components (STROGATZ, 2001). Nodes and edges, respectively represent the elements of these systems and the

interactions between them. Some examples are power grids (MOTTER *et al.*, 2013), airlines (LORDANA *et al.*, 2014), World-Wide Web (ALBERT; BARABÁSI A, 2002), word networks (AMANCIO; OLIVEIRA JR; COSTA, 2012a), among others (NEWMAN, 2003). By considering a network representation, it is possible to use a vast number of tools that have been developed in the last decades to study its properties. Moreover, our society can be understood as a self-organizing complex system. Some examples of studied dynamics are simulations of game dynamics (PONCELA *et al.*, 2009), voting modeling (BARRAT; BARTHLEMY; VESPIGNANI, 2008), and epidemic propagation (NEWMAN, 2003; BARRAT; BARTHLEMY; VESPIGNANI, 2008). Many researchers have been representing texts as networks. Some applications include summarization (TOHALINO; AMANCIO, 2018), word sense disambiguation (JR; LOPES; AMANCIO, 2018; MIHALCEA; TARAU; FIGA, 2004), authorship attribution (AKIMUSHKIN; AMANCIO; JR, 2018; MARINHO; HIRST; AMANCIO, 2016), among others. The majority of these studies represent the texts as co-occurrence networks, which mainly grasp information regarding the writing style. Some studies simulate the evolution of language and knowledge, such as in (GHANBARNEJAD *et al.*, 2014), which used the concepts of epidemic dynamics. Furthermore, language and knowledge share many elements and can be understood as complementary (DICKMANN, 2011). Language can be understood as a type of knowledge and is used by human beings as a way to communicate all types of knowledge. Many researchers study knowledge representations by considering the structure of citation networks (LANCICHINETTI; FORTUNATO, 2012; MARTIN *et al.*, 2013; PAN; KASKI; FORTUNATO, 2012; SIDIROPOULOS; KATSAROS; MANOLOPOULOS, 2007), such as the analysis of the relationship between areas of study. For example, the integration between science and technology (TUSSEN; BUTER; LEEUWEN, 2000; SHIBATA; KAJIKAWA; SAKATA, 2010) that can be divided into the studies devoted to understand how science influences technology (MEYER; DEBACKERE; GLÄNZEL, 2010; SHIBATA; KAJIKAWA; SAKATA, 2010) and how technology influence science (GAZIS, 1979).

The objective of this study is to analyze texts and databases through complex networks. From this investigation, we aim at finding characteristics that can classify the documents according to a given characteristic. Furthermore, here we propose different network representations that vary according to the scale of analysis. More specifically, we start analyzing specific nodes of co-occurrence networks, in which the nodes represent words of the text and edges are created when two words co-occur. In the following, we consider a network whose connections also consider words that are not co-occurring. Finally, we propose methods that connect chunks of texts, according to the semantical similarity between them. In this case, the scale also varies according to the size of these chunks.

The second part of this thesis consists on analyzing the knowledge representation and acquisition. We also investigate knowledge representations in terms of many different

scales. However, in this part, we start from the higher scale analysis, in which we examine the structure of a citation network affecting the connection between two areas of study. We also propose two dynamics, the first consists of a model of knowledge acquisition that considers the influence between researchers and the second is a reasoning engine that solves problems. In the last part, we propose a framework, based on random walks and information science, to transmit the knowledge through a time series. As an example, we consider the application of text transmission. In summary, we studied from the knowledge structure to the transmission, where we also considered the knowledge acquisition and the intelligence to solve problems.

The relationships between all the concepts of both parts of this thesis are illustrated in Figure 1, which are presented as a network. The two central aspects are *language* and *knowledge*, which have a correspondence concerning their dynamics. In the first part, texts are modeled as *network representations*, which contain characteristics from the language and the text subject. By considering these networks, many network-based *measurements* are computed. These measurements are employed as features for classification tasks. As a crucial point of this thesis, for each of the classification tasks we considered a *different scales of analysis*. More specifically, we propose the use of some *network representations* that extract characteristics of the documents from the relationship between words to chunks of texts. Furthermore, we also considered some network *measurements* that are also related to a multi-scale analysis (with respect to the node *different scales of analysis*). In the first text analysis, we extracted information of local patterns of the network in order to grasp stylistic information of the texts. These characteristics were employed to classify texts between fact or fiction. In the following, we considered a word-based network that could represent texts in the scale of the semantics. So, we employ this representation to detect subjects in texts. As a higher scale, we represent the networks as the semantical relationships between paragraphs. From this networks, we showed many different features and selected the best set that better represent texts as being real or shuffled. Moreover, we compared networks obtained from translations of the same text and characterized their features. This network model was also employed to compare the unknown manuscript, Voynich, with real and shuffled topologies. As the last network model, we represent nodes as chunks of adjacent paragraphs, which we called mesoscopic networks. This representation was found to be able to grasp information regarding the unfolding of the story.

In the second part, knowledge is also represented by networks (represented by the node *network representation*). First of all, we analyzed the knowledge organization in terms of groups of nodes. So, by considering the network topology, it was possible to describe the relationship between two areas of study (theoretical and applied physics). Additionally, *Simulations of dynamics* were executed on these networks. In order to describe the behavior of the dynamics, *measurements* are also employed to extract the respective

The diagram illustrates the organization of the book, centered around two main themes: **Language** and **Knowledge** (orange circles). Surrounding these are six other topics: **Classification**, **Different scales of analysis**, **Information theory**, **Data transmission**, **Simulation of dynamics**, and **Network representation** (blue circles). **Measurements** is also a blue circle, positioned between **Classification** and **Network representation**.

The diagram is divided into sections by curved lines, with various section numbers (Sections 3.2 - 3.5, Sections 3.2 - 3.5 and 4.2 - 4.5, Section 4.5, Sections 4.2 - 4.5, Sections 4.3 - 4.5, Sections 4.3 - 4.5) placed along the curved lines separating the topics. A dashed arrow points to **Chapter 3** (between **Classification** and **Different scales of analysis**), and a solid line points to **Chapter 4** (between **Language** and **Knowledge**).

information. For the first dynamics, knowledge acquisition, a multi-agent dynamics was employed to simulate the relationships between researchers learning new concepts. In this case, apart from measuring global information on the dynamics, we considered regions of the network, which were defined from network measurements. As the next dynamics, we considered reasoning engines that can solve problems. This machine consists of a network that assembles another network from their disconnected parts. Topological measurements were employed to assign the initial positions of the disconnected parts on the reasoning engine network. This dynamics solves the problem by sending their parts to their neighbors. Finally, we use some characteristics of this dynamics to simulate the *data transmission*. In this case, the dynamics return a temporal series that can be converted into a knowledge representation. The obtained series can be compressed by considering techniques of *infor-*

mation theory. Information theory methods can be employed for sequences of organized symbols. Here we obtain these symbols by using a random walk dynamics and used information theory to simulate the transmission of the symbols. Because the network topology is already known, this knowledge was employed as a piece of information to compress the time series, and consequently the network transmission. By considering these concepts, we proposed a connection between network science and information theory. As a case example, we compare the differences in transmission efficiency by considering different language representations.

1.1 Objectives

The general objectives consist of comparing the differences among patterns in language, from different perspectives, by considering representations of various scales. Language and knowledge share many elements, and consequently, can be understood as complementary (DICKMANN, 2011). Language itself is also a piece of knowledge. So, we extend this study to the analysis of diverse aspects regarding knowledge representations.

The specific objectives are:

- **Analyze the text-based networks through previously available complex network measurements:** By considering that many different network-based measurements were already employed to text classification, we aim at distinguishing which measurement is more suitable to be employed for a given application. Furthermore, we focus on extending the already used methods which employ network-based measurements. For that purpose, well-defined machine learning approaches are also considered. This objective is necessary to develop the following items (see the nodes language, network representation, measurements, and classification of Figure 1);
- **Analysis of text-based networks by considering different scales:** The present study comprises comparisons among different representations of texts as complex networks. According to problems, we propose distinct representations, which are related to some levels of information on the documents. More specifically, these scales vary from the network representations as connected words to models of connected chunks of text (see the nodes language, network representation, measurements, classification, and different scales of analysis of Figure 1);
- **Characterization of texts written in different languages:** Certain characteristics of co-occurrence networks can be observed regardless of the language of the text, such as Zipf's law of the degree distribution. However, according to Gao *et al.* (2014) through the analysis with basis on complex networks, it is possible to characterize languages. Some languages were found to have similar features when

they have near origins. In this thesis, we aim to identify which network-based features are more related to the text subject or the language (see the nodes language, network representation, measurements, and classification of Figure 1);

- **Identification of subjects in a given text:** A text can comprise many different subjects. Here we propose a network model that is able to represent texts in the same scale of the text subjects. So, we aim to detect chunks of texts by using information extracted from network communities. We believe that this approach can be useful for other studies (see the nodes language, network representation, measurements, and classification of Figure 1);
- **Simulation of network dynamics:** Many different simulations were proposed by using complex networks. One dynamics that is intrinsically related to the language is knowledge. Here, we aim to better understand the structure of knowledge representations and the relationship between areas of study. As a case study, we study the relationship between theoretical and applied physics. Here we also propose a novel dynamics that simulates the knowledge acquisition, by considering a multi-agent dynamics, in which some researchers can influence the others. Another simulation consists in proposing a reasoning engine that can solve problems. In this case, the agents are not researchers, they represent parts of the engine that communicate to solve the problem (see the nodes language, knowledge, simulation of dynamics, network representation, and different scales of analysis of Figure 1);
- **Simulation of data transmission:** By considering that knowledge can be represented as a network, we compare some stochastic strategies to transmit this knowledge, which are based on variations of random walks. The transmission consists of a temporal series obtained by the random walk agent. This transmission can also be compressed by considering information theory. So, it is possible to find which strategy is most efficient for both time series (with or without compression). This dynamics can be compared to language because the transmission of knowledge through speaking and writing consists of a linear sequence of words. Additionally, the natural languages also comprise characteristics of compression, e.g., the most frequent words tend to be smaller (see the nodes language, knowledge, simulation of dynamics, network representation, data transmission, and information theory of Figure 1).

1.2 Contributions

In this section, we present the contributions of this thesis and the respective publications. The entire list of papers is shown in Appendix A. The appendix list also includes publications that have subjects related to the thesis, but do not belong to the text.

- **Using complex networks for text classification: Discriminating informative and imaginative documents:** In this analysis, we employ co-occurrence networks to represent texts. From these networks, we computed many measurements and classified the texts between fact (informative texts) fiction (imaginative texts). As the main result, the connection pattern of stop words (words with low semantical content) was found to be more discriminative between the two classes. Published in: ARRUDA, H. F.; COSTA, L. F.; AMANCIO, D. R. Using complex networks for text classification: Discriminating informative and imaginative documents. **EPL(Europhysics Letters)**, IOP Publishing, v.113, n.2, p.28007, 2016;
- **Topic segmentation via community detection in complex networks:** By considering a higher scale than the previous test, we proposed three novel representations of texts. These approaches consider connections with words that are not adjacent. We found that the communities of these networks reflect the subjects of the texts. Published in: ARRUDA, H. F.; COSTA, L. F.; AMANCIO, D. R. Topic segmentation via community detection in complex networks. **Chaos: An Interdisciplinary Journal of Nonlinear Science**, AIP Publishing, v.26, n.6, p.063120, 2016;
- **Paragraph-based representation of texts: A complex networks approach:** We propose a new paragraph-based representation of texts as networks. In this case, the nodes and edge weights represent paragraphs and semantical similarity between paragraphs, respectively. We analyzed many complex network measurements and identified the dependency with language and semantics. Additionally, we applied the identified measurements to two classification tasks: (i) Discriminating between real and shuffled texts and (ii) A case study considering the unknown manuscript called Voynich. Because the order of paragraphs of Voynich manuscript is unknown, this approach can be useful to analyze the text. We have found that this manuscript is compatible with real texts in terms of the organization of its paragraphs. Published in: ARRUDA, H. F.; MARINHO, V. Q.; COSTA, L. F.; AMANCIO, D. R. Paragraph-based representation of texts: A complex networks approach. **Information Processing & Management**, Elsevier, v. 56, p. 479–494, 2019;
- **Representation of texts as complex networks: a mesoscopic approach:** By considering a higher scale, we proposed a network model in which the nodes represent chunks of paragraphs, and the connections also consider semantical similarity. This method was proposed to grasp information concerning the unfolding of the story along time. In order to evaluate this characteristic, we compare networks created from real and shuffled texts. As a result, we have found that the topology of networks created from real and shuffled texts are different. So, networks grasp information regarding the story. Published in: ARRUDA, H. F de; SILVA, F. N.;

MARINHO, V. Q.; AMANCIO, D. R.; COSTA, L. da F. Representation of texts as complex networks: a mesoscopic approach. **Journal of Complex Networks**, v. 6, n. 1, p. 125–144, 2018;

- **On the “Calligraphy” of Books:** In this study, we employ the mesoscopic networks to the task of authorship attribution. First of all, we found that the visualization of the networks seems to express a kind of “calligraphy” of the authors. In other words, the network visualizations of books of the same author seem to have particular characteristics. From these networks, we measured some network features and employed supervised classifiers. These features are able to distinguish between some authors. Published in: MARINHO, V. Q.; ARRUDA, H. F.; LIMA, T. S.; COSTA, L. F.; AMANCIO, D. R. On the “calligraphy” of books. In: **TextGraphs**. [S.l.]: Association for Computational Linguistics, 2017;
- **An image analysis approach to text analytics based on complex networks:** Because in the last study we found that the network visualizations represent a kind of visual “calligraphy”, here we analyze the networks in terms of image analysis features. We compare the features obtained from network measurements and image-based features, and the best results were found when both approaches were considered together. Published in: ARRUDA, H. F.; MARINHO, V. Q.; LIMA, T. S.; AMANCIO, D. R.; COSTA, L. da F. An image analysis approach to text analytics based on complex networks. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 510, p. 110–120, 2018;
- **How Integrated are Theoretical and Applied Physics?:** We employed the American Physical Society (APS) network of citations to analyze the connectivity between papers of theoretical and applied physics. As a result, we have found that there is no significant difference between the connection between theoretical and applied papers, and other papers of the network. Published in: ARRUDA, H. F.; COMIN, C. H.; COSTA, L. F. How integrated are theoretical and applied physics? **Scientometrics**, Springer, v. 116, p. 1113–1121, 2018;
- **Knowledge Acquisition: A Complex Networks Approach:** In this study, we proposed a dynamics that simulates the knowledge acquisition in a knowledge network. Our dynamics is based on a random walk dynamics in which the agent (walker) tends to visit unknown nodes. As the main result, we found that there are no significant differences in discovery performance when the researcher follows or not other researchers of the knowledge network. Published in: ARRUDA, H. F.; SILVA, F. N.; COSTA, L. F.; AMANCIO, D. R. Knowledge acquisition: a complex networks approach. **Information Sciences**, Elsevier, v. 421, p. 154–166, 2017;

- **Intelligent Complex networks:** We proposed a system that solves problems, which is henceforth called “reasoning engine”. As a case example, we employ a machine that mounts the puzzle. This engine consists of a network in which each node changes puzzle parts among the neighbors. Furthermore, the puzzle consists of another network that shall be mounted according to the original connections. As the main result, we found that the initial organization of the puzzle parts is intrinsically related to the performance of the reasoning engine. Published in (preprint): ARRUDA, H. F.; COMIN, C. H.; COSTA, L. F. Intelligent complex networks. **arXiv preprint arXiv:1809.09201**, 2018;
- **Connecting Network Science and Information Theory:** This study proposes a vectorial representation, which is a temporal series. This series consists of the nodes visited by a random walker. Two versions of the series are considered: the original and compressed series. We also studied a case example, in which we regard language networks and the topology information to compress texts. Published in: ARRUDA, H. F.; SILVA, F. N.; COMIN, C. H.; AMANCIO, D. R.; COSTA, L. F. Connecting network science and information theory. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 515, p. 641–648, 2019;
- **Text Networks (CDT-4):** This didactic manuscript is a tutorial of how text can be mapped as networks, in which we presented three types of networks: co-occurrence, paragraph-based, and mesoscopic. Furthermore, we show one application for each model. Published in (preprint): ARRUDA, H. F.; COSTA, L. F. Text networks (cdt-4). **Preprint on ResearchGate**, p. 1–7, 2018.

1.3 Thesis organization

The remaining of this thesis is organized as follows. Chapter 2 describes the basic concepts on complex networks (measurements, models, and dynamics) and text mining (characteristics of texts and methods of analysis). In Chapter 3, we proposed different models of network representations of texts, according to distinct scales of analysis. For each proposed model, the respective results are also shown. The methodology and analysis of knowledge and its respective dynamics are shown in Chapter 4. In Chapter 5 we present the conclusions and future works.

BASIC CONCEPTS

In this chapter, we present the statistical measurements we employ to characterize complex networks. We relate these measurements to the known properties of texts. We also describe random network models and their importance to better understand real systems. Regarding network dynamics, we depict some variations of random walks. This chapter also comprises the principal concepts concerning the synthesis of the main bibliography that is employed in the text analysis of this thesis. Furthermore, we present some of the traditional methods of text processing and complex networks techniques.

2.1 Complex network definitions

Complex networks can be understood as discrete representations of systems in which some elements are connected. A network is defined as a graph $G(V, E)$, where V is the set of nodes (also called nodes) and E is the set of edges (also called links). The nodes represent the discrete units of a given system and the edges their interactions. By considering a graph with N nodes, its structure can be stored as an $N \times N$ adjacency matrix, A (BOCCALETTI *et al.*, 2006). For all pair of nodes, when there is a connection between i and j , $A_{ij} = 1$, otherwise $A_{ij} = 0$. Figure 2 shows an example of such a structure. Another characteristic that can be added to this structure is the edge directionality. If $A_{ij} = 1$ and $A_{ji} = 0$, there is an edge that connects from i to j . In this case, the network is called directed networks. For the undirected networks, A is a symmetric matrix because all edges are connected in both directions. Furthermore, a similar structure, known as weighted matrix W , can be employed to represent the weighted networks. In this case, the edges are represented by w_{ij} , which is the weight of the edge from the nodes i to j .

Figure 2 – Example of a network and its respective adjacency matrix. The matrix is symmetrical because the network is undirected.

Source: elaborated by the author.

2.2 Characterization of complex networks

There are two possibilities to characterize complex networks nodes: (i) Local characterization, in which each node is described regarding its nearest neighborhood and (ii) Global characterization, in which each node is described regarding the larger proportion of the network. Such types of measurements have been used in many applications that take into consideration networks created from texts and knowledge-related dynamics. The measurements can be differently interpreted according to different purposes. Here, we present some measurements and also particular interpretations regarding text-based networks.

2.2.1 Local measurements

The local measurements are considered here because for some text-based networks they are associated with the syntactical relationship between the words of the text (NAVIGLI, 2009). The more straightforward measurement is the node degree (COSTA *et al.*, 2007a), k_i , which is quantified for each node i as the number of connections. For directed networks, two possibilities can be considered, in-degree (k^{in}) and out-degree (k^{out}), which accounts for the number of *in* and *out* connections, respectively. Examples of this measurement and its variations for directed networks are shown in Figure 3.

In case of weighted networks, the information regarding edges weights can be included to compute the weighted node degree. This measurement is also called strength (s_i) and is computed for each node i as the sum of all edges weights, w_{ij} , which are connected to the node i (BARRAT *et al.*, 2004). Similarly to the degree, this measurement is also defined for directed networks, in which there are two possibilities: in-strength, s^{in} , and out-strength, s^{out} . Figure 4 illustrates one example of each strength version. In text-based networks, it has been observed that the distribution of degree and strength reflect universal characteristics of particular languages (MASUCCI; RODGERS, 2006; SHENG;

Figure 3 – Example of degree considering the red node. The left-hand figure shows the measurement for undirected networks, and the right-hand figure depicts the two possibilities of both degrees for directed networks, in-degree (k^{in}) and out-degree (k^{out}).

Figure 4 – Examples of the strength of the node represented in red color. The left-hand example illustrates the measurement without direction, and the right-hand represents the directed version, in-strength, s^{in} , and out-strength, s^{out} .

LI, 2009). These characteristics are better described in section 2.6.

From the concept of degree, other measurements were derived. The more straightforward information could be obtained from the neighborhood of the nodes, which is defined as concentric levels (COSTA; ANDRADE, 2007). These levels are established as the set of nodes that can be reached from a given node i passing by h edges. Such edges should be part of the shortest paths between i and the reached nodes. Figure 5 illustrates an example of concentric levels. Among the measurements that consider this concept, there is the neighborhood connectivity, Nc , which is defined by the number of nodes that can be reached from a given node i that take part in the h th concentric level. An example is shown in Figure 5, where for the first concentric level $Nc^{(h=1)} = 6$ and for the second $Nc^{(h=2)} = 10$.

Another measurement that also extrapolates the concept of degree is accessibility, α^h , which is defined on the basis of random walks or other dynamics in networks (TRA-
VENÇOLO; COSTA, 2008). The random walk dynamics consists of a walker, also called agent, that randomly walks on the network, in which the agent goes from a node to oth-

Figure 5 – Concentric levels of the red node, in which the nodes represented by the colors blue and orange take part in the concentric levels 1 and 2, respectively.

Source: Elaborated by the author.

ers through the edges. Some variations of random walk dynamics are better described in Section 2.5. In order to define accessibility, the authors considered the random walk variation called self-avoiding walk, in which the agent cannot go through the same node more than once. On the bases of this concept, the accessibility measurement is defined for the concentric level h as

$$\alpha_i^{(h)} = \exp \left(- \sum_j p_{ij}^{(h)} \log p_{ij}^{(h)} \right), \quad (2.1)$$

where $p_{ij}^{(h)}$ is the probability of the agent to reach to a given node j , departing from i , by walking through h edges. This measurement is employed in unweighted networks. Furthermore, accessibility is called the effective degree of a node when computed for $h = 1$ (VIANA; BATISTA; COSTA, 2012; RODRIGUES; COSTA, 2009b). By considering other applications, accessibility was employed to describe the accessibility of the streets of cities (TRAVENTOLO; COSTA, 2008), and in a text mining application to find keywords in texts, as part of the task of summarization (TOHALINO; AMANCIO, 2018; AMANCIO *et al.*, 2012a).

In order to simplify this measurement, it is possible to compute the transition probabilities by considering the transition matrix, which is defined as

$$P_{ij} = A_{ij}/k_j, \quad (2.2)$$

where k_j is the degree of the node j . The probabilities can also be defined by the matrix power P^h , where the position ij of P^h contains the probability of a walker starting from node i to reach node j by walking through h edges. This simplification does not consider the self-avoiding walk.

Figure 6 – Examples of the simplification of the network shown in item (a), by considering the symmetry of the red node and the access of all other nodes. These access probabilities are computed according to the self-avoiding random walk. Item (b) and (c) represent backbone and merged symmetries, respectively.

Source: Elaborated by the author.

With basis on the accessibility and the concentric levels, another measurement was proposed, the concentric symmetry, S^h (SILVA *et al.*, 2016b). This measurement is computed for each node i , by considering the concentric level h and is defined as

$$S_i^{(h)} = \frac{\exp \left\{ - \sum_{j \in \Gamma_h(i)} p_{ij}^{(h)} \log p_{ij}^{(h)} \right\}}{|\Gamma_h(i)| + \sum_{r=0}^{h-1} \eta_r}, \quad (2.3)$$

where η_r is the number of nodes that do not have connections with the next hierarchical level, $\Gamma_h(i)$ is the set of neighbors of i that take part in the same hierarchical level h , and $p_{ij}^{(h)}$ is the probability of access, by considering the self-avoiding walk. These random walk dynamics is the same that was employed to the accessibility. In order to compute these probabilities, the measurement was defined in two different forms: backbone symmetry, Sb^h , and merged symmetry, Sm^h (see Figure 6). The first version (Sb^h) consists in removing all the edges that connect the nodes of the same hierarchical level. See an example in Figure 6b. In the case of Sm^h , the nodes that are connected to others in the same hierarchical level are merged into a new node, as can be seen in Figure 6c.

Symmetry measurement is considered in this study because it takes part in many real-world systems (LONGUET-HIGGINS, 1963; COLEMAN; WEINBERG, 1973). Furthermore, symmetrical patterns can also be found in text-based networks, such as in the task of detecting different writing styles (AMANCIO; SILVA; COSTA, 2015). Among the proposed possibilities to compute symmetry in networks (SILVA *et al.*, 2016b; HOLME, 2006; ROSSI *et al.*, 2013), we consider and present the approach proposed by Silva *et al.* (2016b). As one advantage, through this approach, it is possible to determine the symmetry of each node in a multi-scale fashion.

In addition to the degree-based measurements, other measurements were proposed

Figure 7 – Examples of clustering coefficient of node i , shown in red color. When all of the triples connected to i are also triangles, $cc_i = 1$ and when there are no triangles connected to i , $cc_i = 0$. Furthermore, in all other cases $0 \leq cc_i \leq 1$.

Source: Elaborated by the author.

with the basis of other characteristics provided by real systems. One of these measurements is the clustering coefficient (WATTS; STROGATZ, 1998), which is based on the cycles of size three, henceforth called triangles. This characteristic is observed in many real systems, such as the internet (ROSATO *et al.*, 2008), networks of metabolic interactions (RAVASZ *et al.*, 2002), and for quantifying the specificity of a word in text-based networks (AMANCIO *et al.*, 2011). The clustering coefficient is computed as

$$cc_i = \frac{N_{\Delta}(i)}{N_3(i)}, \quad (2.4)$$

where $N_{\Delta}(i)$ is the number of triangles connected to node i and $N_3(i)$ is the number of triples connected to i , which are the pairs of nodes. The maximum and minimum values of cc are 0 and 1, respectively. Examples of this measurement can be seen in Figure 7.

Apart from the local centralities of nodes, there are also measurements that concern characteristics of edges. In this thesis, we considered Matching Index ($\mu_{i,j}$), which is calculated for all network edges. This measure consists in the similarity between the two nodes connected to the edge and is computed according to the number of neighbors both node have in common (SPORNS, 2003; KAISER; HILGETAG, 2004), and is defined as

$$\mu_{i,j} = \frac{\sum_{k \neq i,j} a_{ik} a_{jk}}{\sum_{k \neq j} a_{ik} + \sum_{k \neq i} a_{jk}}, \quad (2.5)$$

where a_{ij} is an element of the adjacency matrix.

2.2.2 Global measurements

In this study, the centrality measurements which consider the network globally were also employed. This type of analysis consists of characterizing each network node by examining the connection pattern of the entire network. These measurements are different to the local, that consider only the immediate neighborhood of a node. For example

the Pearson correlation between the degree of connected nodes, which is called assortativity (NEWMAN, 2002).

One characteristic that can affect this kind of centrality measurements is the presence of more than one connected component in the network. Connected components are defined as a set of nodes for which there is at least one path that connects each vertex to all others (HARARY, 1969a). Paths are defined as a set of nodes and edges connected sequentially, from the initial node i to j (THULASIRAMAN; SWAMY, 2011). In a network in which there are no edges, the number of connected components is the same as the number of nodes. Furthermore, for directed networks, there are two types of connected components, strongly and weakly connected. The first type refers to a component for which there is a path connecting all pair of nodes ij and ji , by considering the direction of the edges. The weakly connected component consists in the set of nodes with at least one path between all pairs of nodes ij without considering the edges directions.

Among the centrality measurements that consider global characteristics of networks, many of them are based on concepts related to paths. One of the most straightforward is the path length between nodes i and j , also called distance, which is defined as the number of nodes that take part in the path. Based on this definition, the shortest path, or geodesical path (HARARY, 1969a), between a given pair of nodes i and j is the set of nodes that give rise to the lowest value of all possible distances between i and j . It is possible to have more than one shortest path for the same pair of nodes. Some of the possibilities to find the shortest paths are Floyd-Warshall (LARSON; ODONI, 1981) and the Dijkstra's algorithm (DIJKSTRA, 1959).

The concept of the shortest path length ($d(i, j)$) is used to define eccentricity measurement (HARARY, 1969b; ESTRADA, 2012), E_i , which is determined for each node i as the maximum value of $d(i, j)$ between i and all other nodes j of the network. E_i is computed as

$$E_i = \max_j(d(i, j)). \quad (2.6)$$

In a network with a single connected component, the eccentricity can be computed, but when there are more than one connected component, the eccentricity of all nodes is infinite (WEST, 2001). Additionally, from this measurement other information can be calculated. The maximum value of E corresponds to the network diameter, which is defined as the maximum value of all the network shortest path lengths. The lowest value defines the network radius. Figure 8 illustrates one example of this measurement.

Another measurement that is also based on network distances is called closeness centrality, C , which considers the shortest paths of the network. Through this measurement, it is possible to quantify the proximity between a vertex and all other nodes of the

Figure 8 – Example of the eccentricity measurement for all nodes of a network, in which the network diameter is 3 and the radius is 2.

Source: Elaborated by the author.

network. Such a measurement is defined as

$$C_i = \frac{1}{\sum_{j \in V} d(i, j)}. \quad (2.7)$$

In the cases in which more than one connected component takes part of the network, the distance between some pair of nodes cannot be computed. Thus, these distances are considered infinite. Therefore, when there is more than one connected component, $C_i = 0$ for all i nodes of the network because $\frac{1}{\infty} = 0$. Moreover, when this measurement is computed for a network without cycles, also called tree, the root is the node with the highest value of closeness centrality (see Figure 9). This measurement was employed to extract information related to the relevancy of words of text-based networks (AMANCIO *et al.*, 2011).

Instead of considering the network distances, some measurements are computed from the number of shortest paths a node takes part. One of them is the betweenness centrality (FREEMAN, 1977), B , which is computed for each node u as

$$B_u = \sum_{ij} \frac{\delta(i, u, j)}{\delta(i, j)}, \quad (2.8)$$

where $\delta(i, u, j)$ is the number of shortest paths that connect the nodes i and j , in which the node u take parte in the path, and $\delta(i, j)$ is the number of shortest paths that connects i and j . Many different tasks employed this measurement, such as to compute the flow in a network (FREEMAN; BORGATTI; WHITE, 1991; NEWMAN, 2005), the robustness of a power grid (MOTTER; LAI, 2002; MIRZASOLEIMAN *et al.*, 2011), flow of vehicles (HOLME, 2003), discovering of paths interconnecting communities in geographical networks (ARRUDA; COMIN; COSTA, 2016) (the concept of communities is described in Section 2.2.3), among others (MOTTER; LAI, 2002). In text-based networks betweenness

Figure 9 – Example of closeness centrality for a tree network, in which the highest value consists of the root and the lowest of the leaves.

Source: Elaborated by the author.

Figure 10 – Example of betweenness centrality for all nodes of a network. The nodes with higher values are part of the path connecting the communities.

Source: Elaborated by the author.

centrality was employed to determine the specificity and generality of words ([AMANCIO; OLIVEIRA JR; COSTA, 2012c](#)). Figure 10 shows an example of betweenness centrality, where the higher values take part in the path that connects the two communities.

Betweenness centrality was also defined to be computed in the network edges and is called edge betweenness centrality ([GIRVAN; NEWMAN, 2002](#)), *EB*. This measurement is calculated similarly to the original version but considers the network edges instead of the nodes. Furthermore, it was proposed to be employed in the task of community detection in networks ([GIRVAN; NEWMAN, 2002](#)).

A variation of the accessibility measurement is generalized accessibility α^∞ ([AR-RUDA *et al.*, 2014](#)) that is calculated by considering the entire network. As an essential characteristic of this measurement, it is not necessary to define the size of a concentric neighborhood of the node. So, the generalized accessibility is not related to a specific path

length but to all of the paths of the network that cross a given node i . In this way, the probabilities are computed based on a modified random walk that penalizes the longer paths. This measurement is computed as

$$\alpha_i^\infty = \exp \left(- \sum_j p_{ij}^{(\infty)} \log p_{ij}^{(\infty)} \right), \quad (2.9)$$

where p_{ij}^∞ are the elements of the matrix P^∞ , which is the matrix of transition probability between all pairs of nodes. Such transitions represent the probability of a given walker departing from the node i , to reach the node j and is computed as

$$P^{(\infty)} = \frac{1}{e} \sum_{k=0}^{\infty} \frac{1}{k!} P^k, \quad (2.10)$$

where $P^{(\infty)}$ is the matrix that stores the elements $p_{ij}^{(\infty)}$. In matrix form, this equation is described by

$$P^{(\infty)} = \frac{\exp(P)}{e}. \quad (2.11)$$

By considering the equation 2.10 the weights are higher when the considered pair of nodes are nearer because there is a factorial decay according to the distance. This measurement was found to be correlated with the phenomenon of epidemic spreading in networks (AR-RUDA *et al.*, 2014).

Apart from the measurements based on paths and their properties, there are the spectral measurements, which are the measurements based on the eigenvectors and eigenvalues of the adjacency matrix. For example, the eigenvector centrality (BONACICH, 1987), Ec , which is an extension of the degree and is computed as the eigenvector associated with the bigger eigenvalue of the adjacency matrix, A . In text-based approaches, this measurement was applied in the summarization task, to find the more relevant expressions of the text (PATIL; BRAZDIL, 2007).

Based on the eigenvector centrality, another measurement was proposed, page rank (LANGVILLE; MEYER, 2011), which is computed by the equation

$$Pr = \alpha AD^{-1} Pr + \beta l, \quad (2.12)$$

where α and β are positive constants (usually $\beta = 1$) l is a vector $(1; 1; 1; \dots)^T$ and D is a matrix represented by

$$D_{ij} = \begin{cases} \max(k_i^{out}, 1) & \text{if } c_i \neq c_j \\ 0 & \text{otherwise.} \end{cases} \quad (2.13)$$

This method is known to be used to rank the web pages of Internet according to its importance, and was applied as a part of the engine search of Google (MIHALCEA; RADEV, 2011). In another text-based task, page rank was employed in word disambiguation (MIHALCEA; TARAU; FIGA, 2004).

2.2.3 Network communities

The network communities are defined by subgraphs with a large number of nodes that are interconnected (internal connections) and a small number of nodes connected to the rest of the network (external connections). In the literature, this concept has been applied for many distinct purposes (ARENAS *et al.*, 2004; GUIMERA *et al.*, 2005; PALLA *et al.*, 2005), which includes text analysis (LONDHE *et al.*, 2014). By considering a complex network as a macro-scale representation and the centrality of the nodes as micro-scale, since the communities are represented by sets of nodes, it is possible to assume that communities are in meso-scale. In order to quantify how well defined the communities are, modularity measurement (NEWMAN, 2006) could be used, which is defined as

$$Q = \frac{1}{2M} \sum_i \sum_j \left(a_{ij} - \frac{k_i k_j}{2M} \right) \delta(c_i, c_j), \quad (2.14)$$

where c_i e c_j are the communities in which the nodes i and j take part, a_{ij} is an element of the adjacency matrix A , δ is the Kronecker delta, M is the number of the edges of the network, computed by

$$M = \frac{1}{2} \sum_{ij} a_{ij}, \quad (2.15)$$

the Kronecker delta is defined by

$$\delta(c_i, c_j) = \begin{cases} 1 & \text{if } c_i = c_j \\ 0 & \text{otherwise.} \end{cases} \quad (2.16)$$

By considering the modularity measurement, it is possible to find the optimal combination of the corresponding communities of the nodes. However, this procedure can imply a high computing cost, and cannot be computed for networks larger than a hundred nodes. However, there are many possibilities to find the network communities by optimizing the modularity measurement (BOCCALETTI *et al.*, 2006). One example is the Fast Greedy algorithm (CLAUSET; NEWMAN; MOORE, 2004), which consists of a greedy approach. This method was proposed to be executed on networks with a large number of nodes.

Modularity has been employed in many different problems to find network communities (BLONDEL *et al.*, 2008; SILVA *et al.*, 2016a; NORTON *et al.*, 2018), but this measurement can be imprecise due to the resolution limit (FORTUNATO; BARTHELEMY, 2007). This effect happens because in order to find the maximum value of Q , the modules with less than or equal to $\sqrt{2M}$ edges are neglected (FORTUNATO; BARTHELEMY, 2007). By considering that the optimization methods aim at finding the higher possible value of Q , in many cases, the found results may not be precise.

Despite this problem, many different studies employ modularity in their analysis. In this thesis, we consider modularity in the analysis of topics in a text, in which the

community represent a topic in a text, the smallest topics could be considered as the same. One example of the use of modularity in text analysis is (SILVA *et al.*, 2016a), where the authors proposed a network-based technique of topic modeling and considered the modularity-based algorithm called *multilevel* (BLONDEL *et al.*, 2008). Another reason to employ modularity is to use it as a feature in classification problems. However, in many tasks, modularity is avoided. For example, in text-based networks, the concept of community was also applied to find topics (MARTINEZ-ROMO *et al.*, 2011). In this case, the authors employed the approach called Walktrap (PONS; LATAPY, 2006), which is based on a random walk dynamics.

2.3 Network models

Many different systems have been studied through complex networks, and their topological characteristics can vary. For instance, by comparing networks that have a geographical dependency (e.g., networks of cities) and non-geographical representations (e.g., the web network). In order to study specific characteristics of these types of networks, a diverse range of models was proposed. Each network model plays a different characteristic of the real systems. In this section, we present the models employed in this study, which are unweighted and undirected networks.

2.3.1 Erdős-Rényi model

One of the simplest network models was proposed by Erdős and Rényi (1959), called ER model, in which the nodes are uniformly randomly connected. Starting with a given set of N nodes without connections, for each pair of nodes an edge can be created at random choice, with a given probability p . The degree distribution of this model follows a binomial, and for a large number of nodes tends to a Poisson. One characteristic of this network model that is part of real systems is the small world effect, which consists in small topological distance among the network nodes (COSTA *et al.*, 2007a). See an example of the ER network in Figure 11.

2.3.2 Watts-Strogatz model

In many real networks, the nodes have small distances among themselves, which is called the small world effect. For example, social networks (WASSERMAN; FAUST, 1994), food web (DUNNE; WILLIAMS; MARTINEZ, 2002), cerebral networks (SPORNS; ZWI, 2004; BASSETT; BULLMORE, 2006), word-based networks (FERRER I CANCHO; SOLÉ, 2001), among others (AMARAL *et al.*, 2000). Another characteristic that is also present in real networks is the high level of clustering coefficient (cc) (STROGATZ, 2001), such as in brain networks (BASSETT; BULLMORE, 2017). In order to represent these

Figure 11 – Example of ER network visualization (1000 nodes and $p = 0.01$). The color bar represents the network node degrees. The employed software of visualization was implemented by [Silva \(2015\)](#).

Source: Elaborated by the author.

networks characteristics, [Watts and Strogatz \(1998\)](#) proposed a model known as Watts-Strogatz (WS) (see Figure 12). This approach starts with a regular lattice, with N nodes, in which the nodes are connected to n nearest neighbors. A parameter defines a probability of rewiring, $0 < p \leq 1$. For each edge, a random number is generated and when a value less than or equal to p is reached, the edge between a pair of nodes (v_i, v_j) is replaced by (v_i, v_l) , where v_l is randomly chosen among all nodes $l \neq i$ with uniform probability. In this model the initial average degree is maintained.

2.3.3 Barabási–Albert model

Another common pattern found in real systems is the power-law degree distribution, which is described by

$$P(k) = k^{-\gamma}, \quad (2.17)$$

Figure 12 – Example of WS network visualization (1000 nodes, $n = 3$ and $p = 0.01$). The color bar represents the network node degrees. The employed software of visualization was implemented by [Silva \(2015\)](#).

Source: Elaborated by the author.

where k is the degree. This pattern is found in many real networks, such as Web ([REKA; JEONG; BARABASI, 1999](#)), internet ([SIGANOS *et al.*, 2003](#)), and social networks ([MIS-LOVE *et al.*, 2007](#)). In order to represent this type of system, [Barabási and Albert \(1999\)](#) proposed a network model, known as BA, that incorporates this characteristic by considering connections with the preferential attachment. This model starts by defining the network size N , and the number of edges that is previously connected to each node, m . The network construction starts with a small graph fully connected having n_0 initial nodes. The remaining $N - n_0$ nodes and their respective edges are sequentially added to the network with the probability of connection

$$p_i = \frac{k_i}{\sum_j k_j}, \quad (2.18)$$

where i is the new network node, and j means the already existing nodes. This process finishes when all the N nodes and their respective edges are added. During this process, the node degree defines the probability of creating a new edge connecting to it. In other

words, the nodes that already have the highest degree, also have the highest probability of creating new connections. By considering this methodology, the created network has $\gamma = 3$. Figure 13 shows an example of BA network.

Figure 13 – Example of BA network visualization (1000 nodes and $m = 3$). The color bar represents the network node degrees. The employed software of visualization was implemented by [Silva \(2015\)](#).

Source: Elaborated by the author.

2.3.4 Network model of communities

Apart from the characteristics of the previous network models, there are also many real systems that are organized into communities ([GIRVAN; NEWMAN, 2002](#)). In order to represent this type of network, in this study, we employ the community network proposed by [Lancichinetti, Fortunato and Radicchi \(2008\)](#) (see an example in Figure 14). As a further characteristic, this model generates degree and community sizes distributions that are scale-free. This network model is based on some parameters. The first is the mixing parameter, μ , which represents the proportion of edges that connect nodes from a given community to another. So, in a complementary manner, $1 - \mu$ is the proportion of edges that connects only nodes that take part in the same community. Other parameters control

the minimum k_{min} and maximum k_{max} network degree, respectively. Furthermore, the sizes of the communities are controlled by the parameters s_{min} and s_{max} , which define the minimum and maximum community sizes, respectively. To be possible to create this network, it is necessary to consider $s_{min} > k_{min}$ and $s_{max} > k_{max}$.

Figure 14 – Example of the visualization of the community network proposed by [Lancichinetti, Fortunato and Radicchi \(2008\)](#) (1000 nodes, $\mu = 0.01$, $s_{min} = 100$, $s_{max} = 600$, $k_{min} = 4$, $k_{max} = 50$, and $\mu = 0.01$). The color bar represents the network node degrees. The employed software of visualization was implemented by [Silva \(2015\)](#).

Source: Elaborated by the author.

2.3.5 Waxman model

Among the complex network models, there is a subset that represents geographical networks ([PENROSE, 2003](#); [ROZENFELD et al., 2002](#)). These models generate networks with nodes and edges positions having spatial dependency ([BARTHÉLEMY, 2011](#)). In this study, we employ the approach proposed by [Waxman \(1988\)](#), henceforth called Wax. The network starts with nodes randomly organized in the space (a rectangular region), and the edges are created with a probability that decays according to the Euclidean

distance between nodes. The probabilities are calculated by considering an exponential decay, as follows

$$p_{u,v} = \alpha \exp\left(\frac{-d_E(u,v)}{\beta L}\right), \quad (2.19)$$

where u and v is a pair of nodes, $d_E(u,v)$ is the Euclidean distance between the nodes, L is the maximum Euclidean distance between all pairs of nodes, β and α are the parameters that control the characteristics of the curve and are defined in the interval $(0,1]$. Higher values of β generate a network with a higher number of connections, and lower values of α generate networks with a higher density of local connections. See an example of the Wax network in Figure 15.

Figure 15 – Example of Wax network visualization (1000 nodes, $\alpha = 1$, and $\beta = 0.02$). The color bar represents the network node degrees. The employed software of visualization was implemented by [Silva \(2015\)](#).

Source: Elaborated by the author.

2.3.6 Random geometric graph

Another possibility of a geographical-based network is the model called Random Geometric Graph (Geo) ([PENROSE, 2003](#)) (see Figure 16). To create this network, the

number of nodes, N , is defined, and these nodes are randomly organized in a pre-defined region of a two-dimensional space. In order to create the network edges, a parameter d is defined, and the pairs of nodes that are at most with distance d became connected. The Euclidean distance is employed here.

Figure 16 – Example of Geo network visualization (1000 nodes, $d = 0.05$). The color bar represents the network node degrees. The employed software of visualization was implemented by [Silva \(2015\)](#).

Source: Elaborated by the author.

2.3.7 Knitted

Knitted (KN) ([COSTA, 2007](#)) model is created starting from a set of N unconnected nodes. A sequence of nodes is randomly defined, and the connections are established sequentially until a stop condition is reached. Observe that this network model is intrinsically directed, which is set by order of the connected nodes. A possible stop condition is to achieve a given average degree. See an example for an average degree equals two in Figure 17. This network model simulates the co-occurrence networks ([BREDE; NEWTH, 2008](#)), in which the nodes represent words and the connections are created

for the adjacent words. More information concerning co-occurrence networks is shown in Section 2.7.1.

Figure 17 – Example of KN network visualization (50 nodes).

Source: Elaborated by the author.

2.4 Network visualization

Many algorithms have been proposed to visualize networks ([TAMASSIA, 2013](#)). In this work, we employ the approach proposed by [Fruchterman and Reingold \(1991\)](#), which consists in a physics-inspired model based on forces of attraction and repulsion, known as force-directed. This model considers forces between all pairs of network nodes. The nodes are attracted by their neighbors, which is compared to springs connecting them. In a complementary manner, all the nodes are repelled by the others, which is similar to electrical forces. The sum of these forces defines the resultant direction of the nodes. The algorithm converges when the energy of the system is minimized, and the movements stop, reaching the equilibrium.

The force-directed method was implemented in the NetworkX ([SCHULT; SWART, 2008](#)) library and can be used in the Python language. This library allows the user to compute the node positions and plot the network in two dimensions. One example of visualization considering a directed network is shown in Figure 17. Another possibility is the 3D interactive visualizer implemented by [Silva et al. \(2016a\)](#), which also considers the

gravitational force that acts between all nodes. In this work, the small networks were visualized by using the NetworkX library, and the others were visualized by using software created by [Silva et al. \(2016a\)](#). This software allows us to visualize networks with a significant number of nodes (some examples are shown in Figures [11](#), [12](#), [13](#), [14](#), [15](#), and [16](#)).

2.5 Complex networks dynamics

There are many different types and objectives to study dynamics executed on complex networks. One of them is to extract properties of a network, such as to compute symmetry ([SILVA et al., 2016b](#)), accessibility ([TRAVENTÇOLO; COSTA, 2008](#); [ARRUDA et al., 2014](#)) and community detection ([PONS; LATAPY, 2006](#)). These dynamics can also be employed to improve the understanding of real systems, such as cascade failure ([MOTTER; LAI, 2002](#)), epidemics ([BARRAT; BARTHLEMY; VESPIGNANI, 2008](#)), rumors ([MORENO; NEKOVEE; PACHECO, 2004](#)), and the synapsis of neurons ([ROXIN; RIECKE; SOLLA, 2004](#); [GUARDIOLA et al., 2000](#)).

A type of dynamics that has been massively studied is random walk. This subject of the study came up before the studies about complex networks ([PEARSON, 1905](#); [FLORY, 1953](#)). The problem of random walks was proposed by [Pearson \(1905\)](#), in which he would like to know the most probable position to find a walker that walks on a two-dimensional space. That author modeled this problem as a walker, henceforth called agent, that represents a drunk man who starts his walk from the initial position 0 . In the proposed dynamics, for each step of time (discrete time), the agent turns around its axis with a randomly chosen angle and walks l yards straight ahead. This process is repeated n iterations. The answer was presented by [Rayleigh \(1905\)](#), which showed that with a enough number of iterations n , the most probable places to find the walker is near the initial position 0 .

Recently, many variations of the random walk dynamics have been proposed in the literature, incorporating the simulation of many different problems. For example, the analysis of movements of insects ([KAREIVA; SHIGESADA, 1983](#)), models of neuronal activity ([GERSTEIN; MANDELBROT, 1964](#)) and other biological processes ([CODLING; PLANK; BENHAMOU, 2008](#)). Additionally, many variations of this dynamics were simulated by considering the complex networks ([NOH; RIEGER, 2004](#); [COSTA; TRAVIESO, 2007](#); [COSTA et al., 2011](#)), which can represent real systems. In this case, the agent walks from a node to another through the edges. Some examples of applications of random walks in networks are: identification of rumor propagation ([ARRUDA et al., 2014](#)), characterization of proteins ([RODRIGUES; COSTA, 2009a](#)), text mining ([MIHALCEA, 2006](#)), and representation of knowledge ([LAO; MITCHELL; COHEN, 2011](#)).

The standard random walk dynamics considers that an agent has the same prob-

ability of walking from a node to all of its neighbors. However, some other variations were proposed, such as the biased random walks (ZLATIĆ; GABRIELLI; CALDARELLI, 2010; NES; LATORA, 2008), in which the bias modifies these probabilities. For example, by considering complex networks, the degree biased random walk (ZLATIĆ; GABRIELLI; CALDARELLI, 2010; NES; LATORA, 2008), where the probabilities are computed taking into account the degree of the target node. In other words, the probability of transition between a node i and a given neighbor j is proportional to the fraction between its degree (k_j) and the sum of the degree of all neighbors of node i . Interestingly, these local bias causes alteration in global characteristics of the dynamics.

Another possibility of random walk is the Self-Avoiding Walk (SAW) (HAVLIN; BEN-AVRAHAM, 1982), which was employed on complex network studies (MILLÁN *et al.*, 2012). In this walk variation, an already visited node cannot be revisited. So, the agent can become trapped. An alternative to this model is called True Self-Avoiding Walk (TSAW) (AMIT; PARISI; PELITI, 1983; LAM, 1984; STELLA *et al.*, 1984), which was also executed on complex networks (KIM; PARK; YOOK, 2016). In this case, the nodes already visited can be visited again, but for each time the node is visited, the probability of visit decreases according to a negative exponential of the number of visitations. One advantage of using this approach is that the agent cannot be trapped because there is always the probability of revisiting the nodes.

Apart from these variations, a characteristic that could be incorporated into the random walk models is the concept of flight. For example, the Lévy flight (IBE, 2013), in which the agent jumps from a given position to a target position. More specifically, by considering a two-dimensional space, the probability of the target position is computed in line with a power-law decay according to the distance. By considering the complex networks, the concept of flight consists of a possibility of the agent to go from a node to another without walking through the edges. Variations of this probability function can be employed to the random walk dynamics.

2.6 Characteristics of natural languages

The emergence of natural language represents one of the most critical parts of the human evolution (CHRISTIANSEN; KIRBY, 2003). Natural languages came from the evolution of the human species and are defined by the syntax, but the evolutive origins are unknown (ARNOLD; ZUBERBÜHLER, 2006). In order to understand how these phenomena occur, studies employed non-human primates. By considering the combination of different vocalizations that occurred during specific events, such as the presence of a predator and the movement of the group, some patterns were observed (ARNOLD; ZUBERBÜHLER, 2006). Due to the diverse characteristics, it is crucial to study language

from distinct perspectives.

The science that studies this type of phenomena is called linguistics (SAUSSURE, 2011), which includes: phonetics, phonology, morphology, lexicon, syntax, semantics, and pragmatics (BRISCOE, 2011). The area of phonetics studies properties of the sounds that can be produced by humans, which are related to acoustic and articulatory features of the language. Phonology is about the sounds, phonemes, employed in a specific language. Morphology regards the structures, classification, and meaning of the words in terms of the word forms. Lexicon studies the properties of words (e.g., the part-of-speech of a given word). Syntax takes into consideration the study of the disposition and combination of words to form phrases. Semantics studies the meanings of words, phrases, and expressions. Pragmatics investigates the language regarding the context, by including the context of a situation and the linguistics.

One characteristic present in the natural languages is Zipf's law, which means that the frequency of words in a given language decays according to a power law distribution (FERRER I CANCHO; SOLÉ, 2003). With the basis of this law, many statistical models were proposed (CONG; LIU, 2014). Apart from these type of models, it is also interesting to understand why this law happens because the natural language follows complex rules. Lots of regulations take part in these languages (e.g., syntax and lexical regularity and semantics), and, even so, the word frequency distribution in a text is complex (PIANTADOSI, 2014). Furthermore, it is known that for the texts that have word frequency following the Zipf's law, the degree and the strength distributions of the co-occurrence networks also follow the same law (SHENG; LI, 2009; COSTA *et al.*, 2007b; MASUCCI; RODGERS, 2006). In these networks, the nodes represent words, and the edges are created for each pair of adjacency words (a better description of co-occurrence networks is described in Section 2.7.1).

2.6.1 Text pre-processing

Before analysis and text mining, it is common to apply techniques of pre-processing. As a first step, there is a possibility to employ the Part-of-speech tagging, also called POS-Tagging, which consists of the identification of the grammatical class of the words. See an example in Figure 18. One possibility of POS-Tagging is the statistical method based on entropy maximization (RATNAPARKHI, 1996). The computation of these tags is critical because it could be employed in the following steps. This method was implemented as a part of the NLKT (Natural Language ToolKit) (MALECHA; SMITH, 2010), which can be used in Python language.

As the first step, punctuation, numbers, and stopwords (words with low semantic value) can be removed. Some examples of stopwords are “in”, “the”, and “its”. The entire list of the stopwords is shown in Appendix B. Another step is to convert the capital letters

Figure 18 – The tokens of the phrase "What is a scientist?" and the respective part-of-speech tags.

Tokens	What	is	a	scientist?
Part-Of-Speech tagging	pronoun	verb	determiner	noun

Source: Elaborated by the author.

into lowercase. In the following, other methodologies could be employed ([MANNING; SCHÜTZE, 1999](#)), such as to reduce words by using the stemming algorithms, which removes word suffixes, e.g., the stem of "studies" is "studi". There is also the possibility to employ the lemmatization, in which the words can be reduced to a common base form ([MATOS, 2009](#)). For example, the lemma of the word "studies" is "study". In this thesis, because we opted to use only lemmatization, stemming was not employed.

A possible method of lemmatization, developed for the English language, is the WordNet lemmatization ([MILLER *et al.*, 1990](#); [SIGMAN; CECCHI, 2002](#)), which provides better results when the information of POS-Tags is used. The WordNet approach consists of a search in a lexical database organized as a network, where for each word there is a respective lemma. When the word is not part of the database (e.g., proper names), the method returns the original string of characters. In the WordNet network, the words are the nodes, which are connected to their respective synonyms. Additionally, words that are part of the same cognitive synonyms are grouped in sets, also called synsets. Another type of connection links the synsets according to their concepts. WordNet also includes a list of exceptional form words. For example irregular verbs and plurals (e.g., "man" and its respective plural "men"). More information regarding WordNet and its respective characteristics are described in ([BECKWITH; MILLER, 1990](#); [MILLER, 1990](#); [GROSS; MILLER, 1990](#); [BECKWITH; MILLER; TENGI, 2006](#); [MILLER, 1995](#)). This method was implemented as a part of the NLKT ([BIRD; KLEIN; LOPER, 2009](#)).

2.6.2 Bag-of-words

Many of the approaches of text analysis employ models called Bag-Of-Words (BOW), which consists of the frequencies of words in the text ([BUNT; BOS; PULMAN, 2014](#)). These frequencies can be employed as features in a classification task. For that, a vector is created, in which the number of elements is equal to the number of considered words, and for each position of this vector, there is a value that represents the repetitions of the word in the text. This simplification does not consider the order of words in the text.

One alternative to the word count that considers the order of the words is the

count of n -grams, which is defined as the set of n adjacency words (FELDMAN; SANGER, 2007). For example, by considering the text “Alice sees the White Rabbit”, the respective 2-grams are: “Alice sees”, “sees the”, “the White”, and “White Rabbit”. For $n = 1$, the method provides the same result as when considering the words separately. Moreover, as in the word count method, vectors of n -grams counts can be employed for similar analysis and tasks. Another type of method is based on vectorial representations of embeddings, in which the representation of words and documents are called word2vec (MIKOLOV *et al.*, 2013a) and doc2vec (DAI; OLAH; LE, 2015), respectively. In this thesis, we compared one of the proposed methods with a doc2vec representation, and the details are shown in Section 3.4.5.3.

2.6.2.1 Tf-idf

Term frequency-inverse document frequency, *tf-idf*, quantifies the importance of a given word in a document, d , that takes part in a set of documents, D (MANNING; SCHÜTZE, 1999). This measurement can be defined as a map $\text{tf-idf}(w, d, D)$ that quantifies the importance of each word w of in document d of the set $|D|$ and is computed as

$$\text{tf-idf}(w, d, D) = \text{tf}(w, d, D) \times \text{idf}(w, D), \quad (2.20)$$

where $\text{tf}(w, d, D)$ is the term-frequency, which quantifies the relevance of $w \in d$ according to the set of documents D , and $\text{idf}(w, D)$ is the inverse document frequency, which quantifies the frequency of w in all documents $d \in D$.

Variations of the $\text{tf}(w, d)$ measurement were proposed in literature (DOMENICONI *et al.*, 2015). One of the most common manners to define term-frequency is according to the normalized frequency of the word in the text (LESKOVEC; RAJARAMAN; ULLMAN, 2014) as follows

$$\text{tf}(w, d, D) = \frac{F_w^d}{F_w^D}, \quad (2.21)$$

where F_w^d is the frequency of the word w in a given document d and F_w^D is the frequency of the word w in the given set of documents D . A variation of its measurement considers only F_w^d , as follows

$$\text{tf}^*(w, d) = F_w^d. \quad (2.22)$$

There are also many different manners to compute the $\text{idf}(w, D)$ term. One of the most usual manners is given by the equation

$$\text{idf}(w, D) = \ln \left(\frac{|D|}{f_w} \right), \quad (2.23)$$

where f_w is the number of documents with the word w . By considering the equation 2.23, when a word w takes part in all the texts, $\text{tf-idf}(w, d, D) = 0$ because if $|D| = f_w$ consequently

$\frac{|D|}{f_w} = 1$. As a consequence of $\ln(1) = 0$, $\text{idf}(w, D) = 0$, which means that $\text{tf-idf}(w, d, D) = 0$. In order to avoid this effect, it is possible to sum a constant at the idf equation. In this study, we adopt the constant as 1, as can be seen as follows

$$\text{idf}^*(w, D) = 1 + \ln \left(\frac{|D|}{f_w} \right). \quad (2.24)$$

By considering tf^* and idf^* , the term frequency-inverse document frequency can also be defined as

$$\text{tf-idf}^*(w, d, D) = \text{tf}^*(w, d) \times \text{idf}^*(w, D). \quad (2.25)$$

This variation is implemented in the *Scikit-Learn* (PEDREGOSA *et al.*, 2011; BUITINCK *et al.*, 2013), which is a Python library. In this thesis both variations of tf-idf were employed.

2.6.2.2 Cosine similarity

By considering vectors of word frequency, similarity measurements have been employed to make comparisons between texts (HUANG, 2008). When comparing two different texts, it is possible to have many different words, consequently many of the vectors can have frequency equal to zero. So, many of the traditional measurements of distance and similarity cannot be employed. In this case, two vectors with lots of zeros could be considered similar (HAN, 2005). The cosine similarity ignores all the zeros that take part in the same position in both of the compared vectors. Such a measurement is defined from the inner product of two vectors, as follows

$$A \cdot B = \|A\| \|B\| \cos(\theta), \quad (2.26)$$

where A and B are vectors and θ is the angle between them. By isolating the term $\cos(\theta)$, it is found that

$$\cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}. \quad (2.27)$$

The cosine similarity is the cosine of the angle between the vectors A and B . Therefore, because we are considering only vectors with positive values, $\theta \in [0, \pi/2]$ and, consequently the calculated similarity is in the interval $[0, 1]$ and is defined as

$$\text{sim}(A, B) = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}. \quad (2.28)$$

2.6.2.3 Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a technique that can be applied to group texts and to find the key-words (MANNING; RAGHAVAN; SCHÜTZE, 2008). As the first

step, the pre-processing steps can be applied. From the set of documents D , a matrix M is created, with dimension n_w against $|D|$, where n_w is the total number of words of all documents and $|D|$ is the number of documents. One optional step is to consider M as $\text{tf-idf}(w, d, D)$ instead of the word counts. In this study, we opted to consider the tf-idf (without employing the variation shown in Section 2.6.2.1). By considering M , the singular-value decomposition (SVD) is computed, which is a factorization of M and is computed as follows

$$M = U\Sigma V^*, \quad (2.29)$$

where U represents the coordinates of each word in the new space, V^* is a matrix that represents the coordinates of texts, and Σ the matrix of singular values. The squares of Σ are employed to find the contributions and to approximate the matrix. In order to project LSA data in a n -dimensional space, the n first columns of U and V^* are employed to visualize words and texts, respectively. We plot only two dimensions because in this thesis the LSA approach was employed only to visualize the results.

2.6.3 Bag-of-character

In addition to the methods that use BOW, it is possible to consider the Bag Of Characters (BOC) (FELDMAN; SANGER, 2007), which are techniques that consider the text characters instead of the words. Characters are defined as letters of the alphabet, digits, punctuation, spaces, and any other symbol. Due to the reduced number of characters in a text, this approach can grasp low information (FELDMAN; SANGER, 2007). So, another possibility, still considering the characters, is to count the number of distinct n -grams, where n is the number of considered characters. The same methodologies that are applied to BOW can be employed for BOC, by considering the vectors of n -grams counts. However, in BOC it is possible to consider the texts without the pre-processing step because all the characters including digits and punctuation can be considered.

2.7 Complex network-based models

In this section we present some of the proposed approaches to represent texts as complex networks, including the co-occurrence networks and a simple example that illustrates all the steps from the text to create the network. Furthermore, we describe a methodology of network-based applications for text segmentation.

2.7.1 Word networks

Many different approaches have been proposed to model texts into complex networks. One example is the syntactic dependency network, in which the nodes represent words and the connections are created according to the syntactical dependency between

them. More specifically, the words that are in the same sentence can be connected if the pair of words has a syntactical dependency (FERRER I CANCHO; SOLÉ; KÖHLER, 2004). Another network approach that also represents the syntactical relationship among the words is the co-occurrence network (BREDE; NEWTH, 2008). In these networks the nodes represent words, and the connections are created between the adjacent words. By considering this approach, many analyses were described in the literature. For example, the centrality measurements were employed as features, which were used in classification problems (MASUCCI; RODGERS, 2006; SHENG; LI, 2009; AMANCIO *et al.*, 2012a; AMANCIO *et al.*, 2011). Many of these measurements are shown in Section 2.2.

Usually, the pre-processing steps are applied to the texts before creating the networks. The first step is the detection of tokens (and their respective POS tags), followed by the deletion of numbers and punctuation. The upper-case words are converted into lower-case, the stopwords are removed, and the text is lemmatized. The process of stemming can also be applied, but in this thesis we avoided this step. Additionally, because each study aims to grasp different information from the text, the employed steps could vary. We show an example of text, obtained from the novel *Emma*, by Jane Austen¹, the respective pre-processing steps are illustrated in Chart 1 and the respective co-occurrence network is shown in Figure 19.

2.7.2 Topic segmentation

Topic segmentation is defined as the task of partitioning a text in coherent parts, based on its topics (MIHALCEA; RADEV, 2011). There are some possibilities to detect topics by employing complex network tools, such as to use a network in which each node represents a sentence, and the connections are formed by the similarities between them (MIHALCEA; CORLEY; STRAPPARAVA, 2006). By using this network, many different methodologies can be applied. In order to segment a document, a method of text segmentation was proposed (MALIOUTOV; BARZILAY, 2006). A more recent approach, devoted to large texts, was developed by Lancichinetti *et al.* (2015). This methodology consists in employing methods of community detection together with an adaptation of Latent Dirichlet Allocation. Latent Dirichlet Allocation is the state-of-art method of topic modeling domain.

In this thesis, we give more emphasis on the method proposed by Martinez-Romo *et al.* (2011), which is based in co-occurrence of words in sets. By considering each word as a node, and N sets of words (e.g., documents), the importance of each of the co-occurrences is computed between the pairs of words, which are represented by the edges. This importance is compared to a null model, which consists of a model with randomly generated samples. For two words w_i and w_j , one edge is kept only if the frequency of

¹ Obtained from the Gutenberg project (<http://www.gutenberg.org>)

Chart 1 – Example of the pre-processing steps by the novel *Emma* by Jane Austen.

Step	Text
Original text	The intermediate month was the one fixed on, as far as they dared, by Emma and Mr. Knightley. –They had determined that their marriage ought to be concluded while John and Isabella were still at Hartfield, to allow them the fortnight’s absence in a tour to the seaside, which was the plan. –John and Isabella, which was the plan.–John and Isabella, and every other friend, were agreed in approving it. But Mr. Woodhouse–how was Mr. Woodhouse to be induced to consent? –he, who had never yet alluded to their marriage but as a distant event.
Tokens (without caracteres of punctuation) and numbers	’The’, ’intermediate’, ’month’, ’was’, ’the’, ’one’, ’fixed’, ’on’, ’as’, ’far’, ’as’, ’they’, ’dared’, ’by’, ’Emma’, ’and’, ’Mr’, ’Knightley’, ’They’, ’had’, ’determined’, ’that’, ’their’, ’marriage’, ’ought’, ’to’, ’be’, ’concluded’, ’while’, ’John’, ’and’, ’Isabella’, ’were’, ’still’, ’at’, ’Hartfield’, ’to’, ’allow’, ’them’, ’the’, ’fortnight’, ’absence’, ’in’, ’a’, ’tour’, ’to’, ’the’, ’seaside’, ’which’, ’was’, ’the’, ’plan’, ’John’, ’and’, ’Isabella’, ’and’, ’every’, ’other’, ’friend’, ’were’, ’agreed’, ’in’, ’approving’, ’it’, ’But’, ’Mr’, ’Woodhouse’, ’how’, ’was’, ’Mr’, ’Woodhouse’, ’to’, ’be’, ’induced’, ’to’, ’consent’, ’he’, ’who’, ’had’, ’never’, ’yet’, ’alluded’, ’to’, ’their’, ’marriage’, ’but’, ’as’, ’a’, ’distant’ e ’event’.
Text without stopwords	’intermediate’, ’month’, ’one’, ’fixed’, ’far’, ’dared’, ’emma’, ’mr’, ’knightley’, ’determined’, ’marriage’, ’ought’, ’concluded’, ’john’, ’isabella’, ’still’, ’hartfield’, ’allow’, ’fortnight’, ’absence’, ’tour’, ’seaside’, ’plan’, ’john’, ’isabella’, ’every’, ’friend’, ’agreed’, ’approving’, ’mr’, ’woodhouse’, ’mr’, ’woodhouse’, ’induced’, ’consent’, ’never’, ’yet’, ’alluded’, ’marriage’, ’distant’ e ’event’.
Lemmatized text	’intermediate’, ’month’, ’one’, ’fix’, ’far’, ’dare’, ’emma’, ’mr’, ’knightley’, ’determine’, ’marriage’, ’ought’, ’conclude’, ’john’, ’isabella’, ’still’, ’hartfield’, ’allow’, ’fortnight’, ’absence’, ’tour’, ’seaside’, ’plan’, ’john’, ’isabella’, ’every’, ’friend’, ’agree’, ’approving’, ’mr’, ’woodhouse’, ’mr’, ’woodhouse’, ’induce’, ’consent’, ’never’, ’yet’, ’allude’, ’marriage’, ’distant’ e ’event’.

co-occurrence (f) if higher than the expected value of a null model. The significance of the frequency f can be computed by the probability between two words that co-occur, $p(f)$ as

$$p(f) = \frac{\binom{N}{f, n_1-f, n_2-f}}{\binom{N}{n_1} \binom{N}{n_2}}, \quad (2.30)$$

where

$$\binom{x}{y_1, \dots, y_n} \equiv \frac{x!}{(x_1! \dots x_n!)} \frac{1}{(x - y_1 - \dots y_n)!}. \quad (2.31)$$

Equation 2.30, $p(k)$, can be rewritten as

$$\{a\}_b \equiv \prod_{i=0}^{b-1} (a - i), \quad (2.32)$$

Figure 19 – Network representing a paragraph extract from the novel Emma. Each node is a lemma in the text.

Source: Elaborated by the author.

which is adopted for $a \geq b$. In this case, the likelihood $p(f)$ can be written as

$$\begin{aligned}
 p(f) &= \frac{\{n_1\}_f \{n_2\}_f \{N - n_1\}_{n_2 - f}}{\{N\}_{n_2} \{f\}_f} = \frac{\{n_1\}_f \{n_2\}_f \{N - n_1\}_{n_2 - f}}{\{N\}_{n_2 - f} \{N - n_2 + f\}_k \{k\}_f} \\
 &= \prod_{j=0}^{n_2 - f - 1} \left[\frac{N - j - n_1}{N - j} \right] \prod_{j=0}^{f - 1} \frac{(n_1 - j)(n_2 - j)}{(N - n_2 + f - j)(f - j)}. \quad (2.33)
 \end{aligned}$$

If in a given set the number of co-occurrences of two words is r , the p -value p associated to r can be computed as

$$p(f \geq r) = \sum_{f \geq r} p(r), \quad (2.34)$$

where $p(r)$ is computed according to equation 2.33. Now, using equation 2.34, the most significant edges can be selected.

2.7.3 Applications of complex networks to text mining problems

Complex networks have been used to describe characteristics of natural languages (CONG; LIU, 2014; FERRER I CANCHO; SOLÉ, 2003). For example, co-occurrence networks have been used to model structural features of languages, which are related to syntax (FERRER I CANCHO; SOLÉ; KÖHLER, 2004). Additionally, characteristics of

real systems were found in the structure of co-occurrence networks, such as the small-world effect (FERRER I CANCHO; SOLÉ, 2001). According to Gao *et al.* (2014), by representing networks through weighted co-occurrence networks, some languages were found to have similar characteristics. Remarkably, the languages that have similar origins tend to have comparable values of network measurements. Additionally, when networks formed by syntactic links are considered, the same type of patterns are observed (FERRER I CANCHO; SOLÉ; KÖHLER, 2004; AMANCIO *et al.*, 2012b). Furthermore, in order to better understand the complexity of the languages, Wu, Zhang and Ren (2017) employed a complex network approach and found that there is a strong relationship between the complexity of the network structure and textual cognitive complexity.

Several graph-based approaches also employ topological information of text-based networks to perform some type of classification (FERRER I CANCHO; SOLÉ, 2001; LIU; CONG, 2013; ERKAN; RADEV, 2004; ANGELOVA; WEIKUM, 2006; JIN; SRIHARI, 2007; ZHAO; MAO; LU, 2018; METZLER; CROFT, 2004; MASUCCI *et al.*, 2011). For instance, the quantification of textual complexity (LIU; XU, 2011; AMANCIO *et al.*, 2012). As another task, dynamical properties of complex networks have been used to grasp stylistic characteristics of networks (AMANCIO, 2015c). In many studies co-occurrence networks have been used to capture authors' styles (AMANCIO *et al.*, 2011; AMANCIO, 2015a; MEHRI; DAROONEH; SHARIATI, 2012), and the measured network features can be employed in the task of authorship attribution. Sometimes these network-based features do not result in the best accuracies but can be used as a complement of standard methodologies, such as bag-of-words and embeddings techniques.

Also considering co-occurrence networks, Amancio *et al.* (2013), Montemurro and Zanette (2013) analyzed characteristics of a mysterious manuscript, called Voynich, which consists in unknown texts with symbols that apparently form words. Interestingly, both studies showed that the Voynich syntax seems to be coherent with natural languages. Also considering characteristics of real and fake texts, Masucci and Rodgers (2009) compared between the properties of real and shuffled texts and proposed a measurement to distinguish between these two types of networks. In this case, the employed co-occurrence networks include words and punctuation.

Other studies comprise different types of networks. In the summarization context, Salton *et al.* (1997) proposed an approach that took into consideration the semantical relationship between the paragraphs of the text. Another approach has been used in word sense disambiguation problem (AGIRRE; SOROA, 2009). In order to measure how ambiguous the words are, the authors employed the WordNet. By employing a modified version of the page rank algorithm to identify the most central synset in a given context, the authors found that it is possible to identify the sense of words that are being conveyed. The same problem was addressed using a structural approach in (SILVA; AMANCIO,

2012).

In other to address other characteristics of texts, tools of text processing are employed to add information to the network analysis. One example of this application is in the area of science of science, which consists of studies regarding the mechanisms of science (FORTUNATO *et al.*, 2018). For example, the approaches that use textual features combined with a co-citation network, in which the nodes represent papers and the connections are created when one paper cites another (YU *et al.*, 2017; SILVA *et al.*, 2016a). In (YU *et al.*, 2017; SILVA *et al.*, 2016a), the topics of the papers were detected through modularity-based community detection. Interestingly, by considering the found communities and the text mining tools, in (SILVA *et al.*, 2016a) the keywords of the papers were also extracted.

NETWORK-BASED APPROACHES TO TEXT MINING

In this Chapter, we present the proposed methodologies that are related to machine learning and text mining, which are organized according to the analysis scales of the texts. First, we present an approach that employs the co-occurrence networks, in which the nodes represent words and are connected locally. In the following, we propose a network-based methodology to detect subjects in texts. In these networks, the nodes also represent words, which are connected according to the neighborhoods of words in the text. As another approach, we propose a network that is based on paragraphs, which represents nodes, and the connections are created through semantical similarity. Finally, we employ an approach that considers groups of paragraphs as nodes, and the connections are also computed through semantical similarities. This network is used to represent the unfolding of the story over time. By considering the didactical purpose, we published a text concerning the key-points of the methodologies presented in this chapter ([ARRUDA; COSTA, 2018](#)).

3.1 Databases

In this section, we present all the datasets employed to the machine learning-related studies of this thesis. These databases consists in real and artificial documents.

3.1.1 *Brown Database*

The database “A Standard Corpus of Present-Day Edited American English, for use with Digital Computers” ([KUCERA; FRANCIS, 1979](#)), known as *Brown Corpus*, has 500 documents which consist of two main groups imaginative and informative texts. All the texts were published in 1961, in The United States of America. Furthermore, the documents are organized into subgroups, as follows:

- **Informative prose** (374 documents):
 - **Press: Reportage** (44 documents): political, sports, society, spot news, financial, and cultural;
 - **Press: Editorial** (27 documents): institutional, personal, and letters to the editor;
 - **Press: Reviews** (17 documents): theatre, books, music, and dance;
 - **Religion** (17 documents): books, periodicals, tracts;
 - **Skills and Hobbies** (36 documents): books and periodicals;
 - **Popular Lore** (48 documents): books and periodicals;
 - **Belles Lettres, Biography, Memoirs, etc.** (75 documents): books and periodicals;
 - **Miscellaneous** (30 documents): government documents, foundation reports, industry reports, college catalog, and industry house organ;
 - **Learned** (80 documents): natural sciences, medicine, mathematics, social and behavioral sciences, political science, law, education, humanities, and technology and engineering.
- **Imaginative Prose** (126 documents):
 - **General Fiction** (26 documents): novels and short stories;
 - **Mystery and Detective Fiction** (24 documents): novels and short stories;
 - **Science Fiction** (6 documents): novels and short stories;
 - **Adventure and Western Fiction** (29 documents): novels and short stories;
 - **Romance and Love Story** (29 documents): novels and short stories;
 - **Humor** (9 documents): novels, essays, etc.

3.1.2 Gutenberg

Project Gutenberg¹ is a volunteer project created by Hart (1971) that aims to disseminate books in digital format. This dataset comprises more than 57.000 books of the free domain, which are written in the following languages: English, German, Portuguese, France, and Italian. Additionally, the project has books in some different files extensions, e.g., “txt” (UTF-8) and “html” formats. The majority of the Gutenberg books have copyright already expired, and no fee is necessary to download them.

¹ <https://www.gutenberg.org/>

3.1.3 Holy Bible dataset

In this thesis, we also considered part of the Holy Bible text. This dataset comprises three books obtained from the New Testament: Matthew, Mark, and Luke. Furthermore, we used these texts of 16 different languages: Arabic, Basque, English, Esperanto, German, Greek, Hebrew, Hungarian, Korean, Latin, Maori, Portuguese, Russian, Swahili, Vietnamese, and Xhosa. In order to have a larger text, we concatenate these three books into a single document. We employ the same procedure as (AMANCIO *et al.*, 2013). As an additional step, we manually identified the Holy Bible paragraphs, by using a Portuguese version of the book. For all considered languages, the paragraphs comprise the same verses, which are 658.

3.1.4 Voynich Manuscript

Another considered text consists of the Voynich manuscript, which is known to be a mysterious document. Many different aspects of this text have been studied for several years (REDDY; KNIGHT, 2011). For example, studies take into consideration textual analysis (REDDY; KNIGHT, 2011). Other researchers have been employing complex networks techniques to better understand the manuscript properties (AMANCIO *et al.*, 2013; MONTEMURRO; ZANETTE, 2013).

Because the original version of the Voynich manuscript is written in an unknown language and uses an unknown set of characters, these characters were translated into a set of known symbols. Here, we considered the European Voynich Alphabet (EVA) (ZNADBERGEN, 2018). This version contains the entire book translated into European characters, which were manually identified. In order to have better quality, and avoid some possible mistakes, for each line of the manuscript different translations were done. In this thesis, we considered the texts as the most frequent character of each position for all different translations. Because here the information about paragraphs is also necessary, we visually inspected the original manuscript and detected all paragraphs.

3.1.5 Artificial texts

Apart from the databases of real texts, we propose a methodology to generate artificial texts, in which we labeled the subject of each paragraph. These texts were created from real texts obtained from *Wikipedia: The Free Encyclopedia*², which is an on-line encyclopedia, where the users collaborate with the texts. Moreover, the content is with free domain and can be freely used.

The first step to create artificial texts was to select some wikipedia articles (written in English), which are classified in 8 distinct topics and, for each topic, 5 subtopics. From

² <https://www.wikipedia.org/>

each of the articles, we extracted a set of 10 paragraphs. The selected topics and subtopics are presented as follows:

1. **Sports:** football, basketball, golf, soccer, and swimming;
2. **Animals:** bird, cat, dog, lion, and snake;
3. **Food:** bean, cake, ice cream, pasta, and rice;
4. **Musical genres:** classical, funk, jazz, rock, and pop;
5. **Cities:** Barcelona - Spain, Budapest - Hungary, London - England, Prague - Czech Republic, and Rome - Italy;
6. **Scientists:** Albert Einstein, Gottfried Leibniz, Linus Pauling, Santos Dumont e Alan Turing;
7. **Actors:** Five distinct actors;
8. **Soccer player:** Five distinct soccer players.

The names of the actors and soccer players were omitted because the articles refer to people who are alive.

In order to build artificial texts, the paragraphs are randomly chosen by considering two parameters: the desired number of subtopics, n_s , and the desired number of paragraphs for each subtopic, n_p . For each pair (n_s, n_p) , 200 documents were created. Aiming to create texts that have two levels of granularity regarding their subtopics, the selection of paragraphs were done by considering a two-fold way. The first comprises the Coarse-Grained Subtopics, henceforth called CGS dataset, in which the n_s subtopics take part in distinct topics. The second approach refers to a Fine-Grained Subtopics, called FGS dataset, and the texts are obtained from distinct subtopics of the same major topic. By employing this model the artificial text is not consistent because the original documents were written by distinct people that have a diverse writing style. However, there are also positive aspects of this approach, as the possibility to compare different levels of granularity and the clear distinction between each of the topics and subtopics.

3.2 Classification between imaginative and informative texts

In this section, we employ network-based approaches to classify texts between imaginative and informative. So, we classify the texts according to the two main groups of Brown database: informative and imaginative texts. In order to have the classes with

the same number of samples, the texts were organized into groups, where each one has 126 texts. The informative texts were reduced to 80 Learned, 30 Miscellaneous, and 16 Belles Lettres, Biography, Memoirs, etc. In the set of imaginative texts, all the texts were considered. The methodology and results presented in this section are published in (ARRUDA; COSTA; AMANCIO, 2016b).

3.2.1 From texts to co-occurrence networks

Because the network size influences many of the complex network measurements, all the texts employed in this analysis shall have the same size. In order to deal with this problem, segments of the texts were cut away, by keeping the same number of words as the small text. More specifically, by maintaining the order of words of the document, the first x words are kept, in which x is the number of words of the smallest of the considered texts. Before this process, all the documents are pre-processed, where the numbers and punctuation are removed, and words were lemmatized. Two network approaches were considered, with or without stopwords. For both types, we employ some network measurements, as follows: number of nodes ($|V|$), degree (k), neighborhood connectivity (N), clustering coefficient (cc), betweenness centrality (B), closeness centrality (C), eccentricity (E), eigenvector centrality (Ec), PageRank (Pr), accessibility (α) (for $h = \{2, \text{and } 3\}$), generalized accessibility (α^∞), backbone and merged symmetry (S) (for $h = \{2, 3, \text{and } 4\}$), and modularity (Q).

By considering these texts, three different network-based strategies were employed:

- **Global strategy without stopwords (GS):** In this strategy, by employing the preprocessed texts with stopwords, the co-occurrence networks are created. All the measurements of nodes were applied. For a vector X , which represent a given measurement, the following statistics are applied:
 - Average: $\langle X \rangle$;
 - Standard deviation: $\sigma(X)$;
 - Median: X ;
 - Maximum: $\max(X)$;
 - Minimum: $\min(X)$.

In addition to the statistics of measurements of nodes, we also employed the modularity measurement, Q , which considers the network globally;

- **Local strategy without stopwords (LS):** As in the previous approach, the stopwords were also removed, and the node-based measurements obtained from the

co-occurrence networks are considered. In this case, the values of the measurement vector, X , are considered as characteristics (representing the words);

- **Local strategy with stopwords (LSS):** This method is the same as LS, but here the network is created by keeping the stopwords. We opted to considering stopwords because it has been useful in the task of style identification (SEGARRA; EISEN; RIBEIRO, 2013).

3.2.2 Pattern recognition and evaluation

For all of the employed approaches, we used the Information Gain, IG, attribute selection (BRAMER, 2007) (IG is described in Appendix E), and for the classification step, we employed three classifiers: Naive Bayes (JOHN; LANGLEY, 1995), KNN (KANTARDZIC, 2011), and CART (BREIMAN *et al.*, 1984). We chose these classifiers because each one represents a different type of methodology, which can grasp different information from the employed data. For more information regarding the classifiers see Appendix C. All the methods were executed with the default parameters. It is important to highlight that this classification aims to test the quality of the network-based features regarding the task of classifying between informative and imaginative texts. So, more sophisticated classifiers were not examined here. In addition to the network-based approaches, we compared our results to an approach based on BOW, in which we considered the counts of stop words and a BOC (bag of characters) approach. In this case, we regarded as counts of 2-grams of characters. In order to apply BOC the texts were not pre-processed.

The results are presented in two parts. In the first step, the methods that use complex networks are employed (GS, LS, and LSS), and in the second part, we compare our results with a traditional approach. Because the employed problem do not demand much computational resources, we opt to use *leave one out* cross-validation method (see Appendix C). The chosen parameter of KNN classifier was $K = 3$. For all of the classifiers, we ordered the most relevant features, according to IG. The results refers to the better classification, by employing all combinations of the selected features. Furthermore, by considering the method that provides the better results, we employed a method of quantifying feature relevance, which is better described in Appendix C.5.

3.2.3 Results and discussion

Before presenting the proposed analysis, we employed the LSA projection to contrast our results with the following results (LSA is shown in Figure 20). The styles of writing are grouped according to words with different nature. In the case of informative documents it is possible to see words related to specific subjects, such as *state* and *system*. In the group of informative texts, there are words related to someone, or actions, such

as *mr* and *say*. Additionally, in LSA the imaginative texts take part in a less widespread group. This organization can be generated by the use of a fewer number of words in the imaginative texts.

Figure 20 – LSA visualization of informative and imaginative texts. The ten most frequent words of the texts are employed as features. It is possible to identify two groups. The contribution of the axis D1 and D2 are 0.45 and 0.24, respectively.

Source: Elaborated by the author.

GS strategy was executed by considering the co-occurrence networks of each document (preprocessed and without stopwords), and the centrality measurements and their respective statistics were computed. IG attribute selection was employed and the following features were selected:

- Vocabulary size (number of network nodes): V ;
- Degree: $\langle k \rangle$;
- PageRank: $\tilde{Pr}$, $\langle Pr \rangle$, $\sigma(Pr)$ e $\max(Pr)$;
- Clustering Centrality: $\sigma(cc)$ e $\langle cc \rangle$;
- Closeness centrality: $\min(C)$, $\sigma(C)$, $\langle C \rangle$ e $\tilde{C}$;
- Generalised Accessibility: $\langle \alpha^\infty \rangle$;
- Betweenness centrality: $\langle B \rangle$ e $\tilde{B}$.

Classification tests were employed by considering the selected characteristics. The results of all classifiers are shown in Table 1. The highest accuracy of GS was obtained with

Table 1 – Accuracy obtained from the three different approaches and the three classifiers. The highest accuracy was obtained by using LSS.

	KNN	CART	Naive Bayes
Global strategy without stopwords - GS	72%	78%	75%
Local strategy without stopwords - LS	92%	92%	92%
Local strategy with stopwords - LSS	95%	95%	95%

CART classifier, and with accuracy of 78%. Furthermore, the p -value of this classification was $p < 10^{-10}$, which means that the results are statistically significant.

In the following, we analyze the local strategies, the first one is LS, which do not consider stopwords of the texts. For this classification test, 15 characteristics were selected also considering IG criteria. In contrast to the previous test, the local features were only based on the nine measurements as follows:

- Symmetry *backbone*: $Sb^{(h)}$, for $h = \{2, 3, 4\}$;
- Symmetry *merged*: $Sm^{(h)}$, for $h = \{2, 3, 4\}$;
- Accessibility: $\alpha^{(h)}$, for $h = \{2, 3\}$;
- Generalized accessibility: α^∞ .

We opt for considering only symmetry and accessibility measurements because they are not correlated with the frequency of words in documents, as reported in (AMANCIO; SILVA; COSTA, 2015). When LS was employed, for all of the considered classifiers, the obtained accuracies were significantly better than GS, as can be seen in Table 1. The highest difference was found for KNN, in which the accuracy was 20% better.

The local strategy with stopwords (LSS) was the approach that provided the best results, as also shown in Table 1. The highest accuracy was 95%, which was obtained by using all of the employed classifiers. Figure 21 shows a PCA visualization of the results, which shows that there is a clear separation between the two considered classes. Furthermore, to understand the complex network measurements and words (represented by the nodes) that provided the best results, we analyzed the features according to the quantification of feature relevance (see Appendix C.5). These features were computed by using CART classifier, and are ordered as follows:

- Symmetry *merged*: $Sm^{(h=2)}(the)$;
- Symmetry *merged*: $Sm^{(h=3)}(by)$;

- Symmetry *backbone*: $Sb^{(h=4)}(by)$;
- Symmetry *merged*: $Sm^{(h=3)}(an)$;
- Generalized accessibility: $\alpha^\infty(have)$;
- Symmetry *merged*: $Sm^{(h=4)}(by)$;
- Generalized accessibility: $\alpha^\infty(it)$;
- Generalized accessibility: $\alpha^\infty(by)$.

Figure 21 – PCA of the network with stopwords, LSS, where it is possible to see a clear separation between the groups.

Source: Elaborated by the author.

Although the stop words take part in both classes, informative and imaginative, the connection patterns of the nodes that represent these words are different when comparing these two classes. An example of a stop word is the word “the”, which is one of the measurements that better discriminate between the classes singly, as shown in Figure 22. Furthermore, the most relevant centrality measurement is the symmetry, which is applied to specific nodes. Interestingly, this result agrees with the results found by [Amancio, Silva and Costa \(2015\)](#), which shows the capability of these measurements to be more discriminative than the traditional measures.

In order to compare the network-based approaches with traditional methods, we employed two different types of measurements of writing style. In contrast to the LSS approach, the first comparison was made by employing counts of stop words. The most informative features were found for IG attribute selection. The best result was obtained through the *KNN* classifier, and the resultant accuracy was 97%. The PCA projection of

Figure 22 – Density of the backbone symmetry ($Sb^{(h=3)}$) of the word “the”. In the imaginative texts, $Sb^{(h=3)}$ tend to be higher.

Source: Elaborated by the author.

this method is shown in Figure 23. The other approach, based on the bi-grams count of characters was similar to the first and resulted in an accuracy of 98%, also considering KNN classifier (see the PCA projection in Figure 24). By comparing the PCA projections of BOW, BOC, and LSS, in LSS, PCA representing the informative class is more scattered than the imaginative class, but the opposite happens in BOW, and BOC approaches are computed.

Figure 23 – PCA of the features of stopwords count, by employing the bag-of-words approach. In this plot there is a clear separation between the two groups.

Source: Elaborated by the author.

Figure 24 – Results obtained from the approach based on bag-of-characters. It is possible to see a clear separation between the groups.

Source: Elaborated by the author.

3.2.4 Conclusions

In this section, we described and compared a three-fold method based on complex networks to characterize and classify texts between informative or imaginative. By considering the pre-processed texts without stopwords, we started from a global strategy (GS), which is based on global characteristics of the networks. In the following, we considered the same networks and proposed a local strategy (LS), in which the features are computed locally in the network and represent the patterns created by specific words in the texts. We also considered the local strategy with stopwords (LSS), which is similar to the previous strategy, but the networks were computed by considering all the words of the text. For all of the strategies three different classifiers were considered (Naive Bayes, KNN, and CART), and the features that provided the better results were analyzed by a method that quantifies the features relevance. In order to compare the results obtained with the network-based approaches, two traditional methods were employed a BOW, and a BOC approaches.

The measurements selected in the global analysis, GS, were not used in the local investigation, except generalized accessibility α^∞ , which take part in both local approaches, LSS and GC. However, this measurement was not selected in LS analysis. The networks employed in GS and LS were the same, but the analysis were different. So, we can infer that some local patterns can be mitigated in global strategies. The connection patterns of the nodes that represent a few number of words in the texts have information that can differentiate both classes. Moreover, the majority of the words devoted to studying texts by using complex network tools focus on the global properties of the networks. This study

revealed that valuable information could be obtained in specific regions of the network. Consequently, it is possible that local approaches can improve the quality of other related procedures.

3.3 Network-based subject detection

By considering the subjects of the texts, a novel approach of text-based representation was proposed. This methodology is based on the co-occurrence networks and community detection, in which the communities grasp information of the text subjects. The approach considers the text in a higher scale than the traditional co-occurrence networks. The methodologies described in this section are published in (ARRUDA; COSTA; AMANCIO, 2016a).

3.3.1 Network approaches

We propose three adoptions based on the co-occurrence networks and community detection to extract subjects of texts. For each of the proposed adoptions, we test the subject detection in both datasets CGS (coarse-grained subtopics) and FGS (fine-grained subtopics). For more information, see Section 3.1.5. Before creating the networks, the pre-processing steps were performed, and the words were lemmatized. As an additional step, only the verbs and nouns were maintained, and all other words are removed. The three network approaches are described as follows:

- **Extended Co-occurrence (EC):** As in the co-occurrence networks, each node represents a word in the text, but the connections are not created only between adjacent words. The edges are created between words that have a distance equal or less than $d = (\omega - 1)$ words. For example, for $\omega = 2$ and the pre-processed text having the words “ $w_1 w_2 w_3 w_4$ ”, the set of edges is $\mathcal{E} = \{w_1 \leftrightarrow w_2, w_2 \leftrightarrow w_3, w_3 \leftrightarrow w_4, w_1 \leftrightarrow w_3, w_2 \leftrightarrow w_4\}$. Furthermore, it is important to highlight that words taking part in distinct paragraphs and with distance equal or less to d are also connected. If $\omega = 1$, the traditional co-occurrence model is recovered;
- **Paragraph-Based (PB):** In this model, each node of the network also represents a word, and the edges are created connecting all the words of a paragraph in a clique, in which clique is defined as a fully connected network. This approach is related to the principle that the paragraph structure represents sub-structures of an organized text because all its words are semantically related (VÉRONIS, 2004);
- **Adapted Paragraph-Based (APB):** This model is based on the approach proposed by Martinez-Romo *et al.* (2011), presented in Section 2.7.2. In APB, the PB method is applied, and the edges with lower statistical significance are erased.

Figures 25a, 25b and 25c represent examples of EC, PB, and APB networks, respectively, obtained from a Wikipedia text about cars. For each of the networks, we employ Fast Greedy algorithm of community detection. For each model a different number of communities is found. Furthermore, in APB approach, the community organization is not clear as in the other two approaches.

Figure 25 – Examples of EC, PB and APB networks generated from a text about cars. The colors of the nodes represent the respective communities obtained by using Fast Greedy method.

Source: Elaborated by the author.

3.3.2 Detection of subjects via community detection in networks

By considering the presented network models (EC, PB, and APB), we apply a methodology based on community detection to find the text subjects. As in the first step, the communities are computed on the network (here we used Fast Greedy algorithm). From the initial communities, we propose a methodology that finds the desired number of subjects n_s , which equals to the number of communities (the user defines this number). By considering the set of communities $\mathcal{C} = \{c_1, c_2, \dots\}$, the set of paragraphs $\Pi = \{\pi_1, \pi_2, \dots\}$ of the text, where i is an index of a given node. Starting from the number of obtained communities, n_c , if $n_c = n_s$, labels equivalent to all nodes are defined as taking part of $\mathcal{C}$. In this way, each paragraph is represented by a set of labels $\mathcal{L}(\pi_j) = \{l_1^{(\pi_j)}, l_2^{(\pi_j)}, \dots\}$, where $l_i^{(\pi_j)} \in \mathcal{C}$ is the label associated to i th word that take part in the j th paragraph (π_j). The number of times a given label takes part in a paragraph is computed by

$$f(c_i, \pi_j) = \sum_{l \in \mathcal{L}(\pi_j)} \delta(c_i, l), \quad (3.1)$$

where $\delta(c_i, l)$ is the Kronecker delta. The subject associated to paragraph π_j is defined as

$$\tilde{s}(\pi_j) = \arg \max_{c_i \in \mathcal{C}} f(c_i, \pi_j). \quad (3.2)$$

Through equation 3.2, the most frequent label is chosen as the paragraph label.

The method described by equation 3.2 can be applied only when $n_c = n_s$. Otherwise, it is not possible to associate the communities with subjects. So, we propose an adoption to decrease the number of communities in the network, by reducing the number of nodes. The proposed methodology consists in erasing the nodes of the community that have more overlap with subjects while $n_c > n_s$. This overlap is defined for each of the communities $\sigma(c_i)$. First, the paragraphs are associated with communities, in which we consider, for all words of the paragraph and the respective labels, the community is assigned with the same label as the most frequent community label c_i . The paragraphs are represented by the subset $\Pi_D(c_i) = \{\pi \in \Pi \mid \tilde{s}(\pi) = c_i\}$. As the next step, the words $\Pi_D(c_i)$ that are associated with the community c_i are counted. The overlap, $\sigma(c_i)$, is defined as the number of words in $\Pi_D(c_i)$, associated with the community $c_j \neq c_i$. Mathematically, this measurement is defined as

$$\sigma(c_i) = 1 - \frac{\sum_{\pi \in \Pi_D(c_i)} \sum_{l \in \mathcal{L}(\pi)} \delta(c_i, l)}{\sum_{\pi \in \Pi_D(c_i)} \sum_{l \in \mathcal{L}(\pi)} 1}. \quad (3.3)$$

The method can be seen in Algorithm 1. An example of the detection of subjects is shown in Figure 26. By using the same networks variations shown in Figure 25, the proposed method was applied considering $n_c = 2$, and the results are shown in Figure 27. The obtained communities are related to the text content because the subjects of the first

paragraphs are about definitions and the invention of the car, and the others represent their parts and characteristics.

Algorithm 1 – Algorithm of subject detection by using network communities.

- 1: **Data:** n_c , is the number of communities and n_s is the number of subjects;
 - 2: **while** ($n_c > n_s$) **do**
 - 3: $c_k \leftarrow$ label of the community with higher overlap;
 - 4: erase all the nodes with label c_k ;
 - 5: detect the network communities again;
 - 6: $n_c \leftarrow$ number of communities of the networks;
 - 7: **end while**
 - 8: **Result:** The found communities are associated to the subjects.
-

Figure 26 – Example of the employed subject detection method, in which three communities were found, and we aim at finding only two subjects.

Source: Elaborated by the author.

3.3.3 Modularity analysis

The proposed models to find topics in texts are related to the organization of communities and depend on how well defined they are. So, we studied the relationship between the network modularity and the variation of the parameter ω , which defines the maximum distance between words to create the connections. Even knowing that there is the problem of resolution limit in modularity, here we opt to use this measurement because this characteristic leads our method to avoid subjects that are much smaller than the others. Note that ω controls the number of edges, which is directly related to the network degree. It is known that the modularity measurement is also related to the degree (FORTUNATO, 2010), in which networks with higher average degrees tend to have

(a) EC.

(b) PB.

lower modularity. For this reason, we define a normalized version of modularity Q_N , as follows

$$Q_N = Q - \langle Q_S \rangle, \quad (3.4)$$

where $\langle Q_S \rangle$ is the average modularity obtained from 30 samples of random networks with the same number of nodes and edges as the original network. This methodology is applied in order to find the parameter ω to create the networks.

We define the parameter ω employed to the extended co-occurrence (EC) approach, and consequently, we justify the way the connections occur in the remaining methods. The range of different numbers of connections of the nodes of EC was employed, and the normalized modularity was computed Q_N . Figure 28a shows the computed values for Q , Q_N and Q_S against the number of the nearest connected words ω . By considering this test in the remaining models (see Figures 28b and 28c), we also introduce the parameter ω (that is not defined in Section 3.3.1). In the case of PB and APB, ω limits the number of connections that are created only between nodes that represent words of the same paragraph. This analysis was executed in both with both CGF and FGS datasets, but here we present only the measurement of FGS because the results are similar. For all the employed models, for lower values of ω , high values of modularity, Q , were found. This effect probably happens due to the variation of the network average degrees.

As the values of ω increase, the network connections tend to become denser, and, consequently, Q decreases. Similarly, the modularity of the random versions of the network, $\langle Q_S \rangle$, also decreases when ω increases. The normalized modularity, Q_N , presents a different behavior. This measurement increases approximately up to $\omega \leq 20$, and, for higher values, tend to be constant. Moreover, for the network model EC (Figure 28a), Q_N reaches the maximum value approximately when $\omega = 20$. So, we opted to use the parameter $\omega = 20$ in EC model. The same results also support the choice of using the clique (connection between all words of the paragraph), instead of employing the parameter ω to the PB and APB models.

3.3.4 Results and discussion

In order to compare the performance of the proposed methods, all the variations of the network (EC, PB, and APB) were employed to both text types, CGF and FGS datasets described in Section 3.1.5. For each of the methods and each network model, we generate 200 texts. Additionally, aiming to compare the found results with traditional measurements, we use two techniques of data clustering, K-means (BOCK, 2007), and Expectation Maximization (MOON, 1996), named BOW-K and BOW-EM, respectively. These clustering methods are described in Appendix C. In order to measure the performance of the clustering methods, we use Normalized Mutual Information (NMI), which is described in Appendix E.2.

Figure 28 – Example of the evolution of modularity measurement and the respective normalization for the methods EC, PB, and APB. The presented plots were obtained by employing the FGS database. There are no significant changes of modularity value with $\omega > 20$.

Source: Elaborated by the author.

Figure 29 shows the results obtained for the tests by considering the coarse-grained subtopics (CGS) texts (see Section 3.1.5). At least two network-based methods provided a better value of NMI than BOW-K and BOW-EM. The single networked methodology that provides results worse than BOW-K and BOW-EM is ABP, as shown in Figures 29g, 29h and 29i. By comparing the network-based methodologies, APB provides the better results when $n_s = 2$, and for $n_s = 3$ and $n_s = 4$, the better results are obtained by using EC.

In the case of the fine-grained subtopics (FGS) dataset, the computed NMI values are lower than in the first case. Such results are shown in Figure 30. As in the previous case, the results obtained from network-based results applied to FGS texts are better than BOW-K and BOW-EM for almost all cases. The systematical comparison among the methods revealed that the obtained performances (measured by NMI) vary according to

the employed n_p and n_s parameters. However, in general, the better results were obtained for $n_s = 3$ and $n_s = 4$ and using EC approach, as shown in Figures 30d, 30e and 30f. By comparing both datasets, the differences between network-based models and BOW are similar, which means that the proposed methods increased the performance in a comparable percentage for both difficult levels.

3.3.5 Conclusions

In this section, we present a three-fold methodology for extracting subjects of texts, which are based in co-occurrence networks and community detection. These methodologies were compared among themselves, and with traditional methodologies, by considering two levels of difficulty. Such levels were given by coarse-grained subtopics (CGS) and fine-grained subtopics (FGS) dataset. In order to compare the results of the found labels to the patten of the text, we employ normalized mutual information NMI. For all the results, there is a tendency of the network-based methods to provide better results than the bag-of-words variations. By comparing the two employed datasets, in general, the method that provided better results is EC. Interestingly, this is the most straightforward approach among the proposed variations.

Figure 29 – NMI computed between the found topics against the real topics and subtopics of CGF texts.

(a) $n_s = 2$ e $n_p = 3$.

(b) $n_s = 2$ e $n_p = 4$.

(c) $n_s = 2$ e $n_p = 5$.

(d) $n_s = 3$ e $n_p = 3$.

(e) $n_s = 3$ e $n_p = 4$.

(f) $n_s = 3$ e $n_p = 5$.

(g) $n_s = 4$ e $n_p = 3$.

(h) $n_s = 4$ e $n_p = 4$.

(i) $n_s = 4$ e $n_p = 5$.

Source: Elaborated by the author.

Figure 30 – NMI computed between the found topics against the real topics and subtopics of FGS texts.

(a) $n_s = 2$ e $n_p = 3$. (b) $n_s = 2$ e $n_p = 4$. (c) $n_s = 2$ e $n_p = 5$. (d) $n_s = 3$ e $n_p = 3$. (e) $n_s = 3$ e $n_p = 4$. (f) $n_s = 3$ e $n_p = 5$. (g) $n_s = 4$ e $n_p = 3$. (h) $n_s = 4$ e $n_p = 4$. (i) $n_s = 4$ e $n_p = 5$.

Source: Elaborated by the author.

3.4 Paragraph-based networks

By considering a higher scale than the previous network representations, here we propose a novel approach that takes into consideration the paragraphs of the texts. More specifically, in this approach, the network nodes are represented by paragraphs, and the edges are weighted by the semantical similarities between the pair of paragraphs. By using this network model, we aim to represent the relationship among all the paragraphs. So, it is possible to contrast networks created from real and shuffled texts. From this methodology, we analyze the unknown manuscript called Voynich, which is written in an unknown language and is about an obscure subject. The subject of the present section was published in (ARRUDA *et al.*, 2019).

3.4.1 Paragraph-based network approach

In order to represent the semantical relationship among short contexts (i.e., paragraphs) of a given document, we propose a novel paragraph-based network. A previous model of paragraph-based network approach was proposed by (SALTON *et al.*, 1997), in which the authors studied in the summarization task. In this thesis, to create a paragraph-based network, the raw text (as shown in Figure 31(a)) is divided into paragraphs, which are considered as network nodes. The links between paragraphs are created between all pairs of nodes, which is computed as the cosine similarity between vectors representing word counts of the paragraphs (cosine similarity is better explained in Section 2.6.2.2). For that, we considered the term frequency-inverse document frequency (tf-idf) weighting map (see details in Section 2.6.2.1), as shown in Figure 31(b). We considered each paragraph as a document d and the entire text represents the set of documents D of documents. The created network is shown in Figure 31(c).

We characterize the paragraph-based networks by employing complex networks measurements. Because many of these measurements are defined only for unweighted edges, we removed the edges with lower values using a given threshold T . For the considered networks, we chose a threshold that keeps 5% of the network edges. This parameter was empirically chosen by considering that for all the tested networks almost all the nodes are connected into the highest connected component. This is an essential step since it is known that many of the network measurements are very sensitive to both size and density of connections (COSTA *et al.*, 2007a; AMANCIO *et al.*, 2013). Furthermore, in preliminary tests, we observed that a perturbation of T does not alter the conclusions of our experiments. The effect of the thresholding step is illustrated in Figure 31(d).

Figure 31 – Example of the paragraph-based network for a text with five paragraphs, which is shown in item (a). Item (b) represents the tf-idf of the paragraphs, and (c) and (d) are the weighted and unweighted versions of the network. In the unweighted network the threshold T provides an average degree of two.

Source: Elaborated by the author.

3.4.2 Network variations

In order to compare real and shuffled texts by considering two possibilities of shuffling, we considered three types of paragraph-based networks:

- **Real Texts (RT)** – obtained directly from the pre-processed texts;
- **Shuffled Words (SW)** – the words of the texts are shuffled, and new paragraphs are defined having the same number of words as the original paragraphs;
- **Shuffled Sentences (SS)** – similarly to SW, in SS the sentences are shuffled, and the paragraphs are defined having the same number of words as the original paragraphs.

Some examples of the presented versions obtained from an extract of the book “Alice’s Adventures in Wonderland”, by Lewis Carroll are shown as follows:

1. **Real Texts (RT) version:** The Fish-Footman began by producing from under his arm a great letter, and this he handed over to the other, saying, in a solemn tone, ”For the Duchess. An invitation from the Queen to play croquet.” The Frog-Footman repeated, in the same solemn tone, ”From the Queen. An invitation for the Duchess to play croquet.” Then they both bowed low and their curls got entangled together;
2. **Shuffled Words (SW) version:** to got bowed Fish-Footman play An low in letter handed tone other Duchess . The . ” great croquet over a tone Frog-Footman ; ,

Duchess began . to ” by the the from , saying for in , An this solemn play solemn entangled , croquet Then arm together The both a invitation and the curls , from , the repeated their same under the he ” ” they For producing his to the From Queen invitation and . Queen;

3. **Shuffled Sentences (SS) version:** An invitation for the Duchess to play croquet. An invitation from the Queen to play croquet. ” The Frog-Footman repeated, in the same solemn tone, ”From the Queen. The Fish-Footman began by producing from under his arm a great letter, and this he handed over to the other, saying, in a solemn tone, ”For the Duchess.” Then they both bowed low and their curls got entangled together.

3.4.3 Informativeness analysis

By considering the paragraph-based representation of texts, complex network measurements are computed to characterize these representations. We evaluate if these measurements are informative according to the differences between the measurements obtained from real books and their respective shuffled versions. So, the measurements that comply with this condition can be considered able of discriminating between real and random manuscripts. The informative measurements are valuable to verify if the Voynich manuscript is a real book. In the following, we present the test of informativeness:

- *Criterion A:* by considering this criterion, we verify if the values computed from the set of all shuffled texts and all real texts can be discriminated between themselves. For this analysis, we considered the total number of books in the RT, N_{RT} , and the number of shuffled versions generated for each book in RT, N_S . The comparison was performed between N_{RT} values with $N_{RT} \cdot N_S$.
- *Criterion B:* for each measurement, this criterion consists in comparing the value obtained for a given real text (RT) with the values obtained from its respective shuffled version. For a given measurement X , the distance between a real text and the respective shuffled versions can be computed by calculating the *z-score*, also known as standard score:

$$z(X) = \frac{x - \langle X^{(R)} \rangle}{\sigma(X^{(R)})}, \quad (3.5)$$

where x is the measurement obtained from the real text, $X^{(R)}$ is the set of values obtained from the N_S shuffled versions (SW or SS), $\langle \dots \rangle$ is the average, and σ is the standard deviation of the distribution, respectively. We consider the measurement X as informative when $z(X) > 1$. When z-score is calculated, the difference between x and $\langle X^{(R)} \rangle$ is given in terms of the standard deviation of the distribution. So, we

consider the measurement as informative when the real value x is found to be at least with the distance $\sigma(X^{(R)})$ from the distribution $X^{(R)}$.

For both criteria, we employed the number of samples $N_S = 30$ (for SW and SS) for each real text.

3.4.4 Dependency with language and semantics

In order to analyze if the measurements obtained from paragraph-based networks can extract syntactical and/or semantical characteristics of the represented texts, we compute many network measurements. A similar analysis has been undertaken by (AMANCIO *et al.*, 2013). In this analysis, we considered two datasets, the Holy Bible dataset, in which the texts are about the same subject (a fixed semantics), and a dataset of books obtained from the Project Gutenberg, in which we considered books that have a varied range of subjects and languages (by varying semantics and syntaxes). For a given measurement, $X_{t=\pi,l}$ is the set of computed measurements obtained for X in a dataset comprising the same book π but in different languages (l) (Bible dataset). In a similar way, $X_{t,l=\lambda}$ represents the set of values obtained for the measurement X in a dataset comprising different texts (t) written in the same language $l = \lambda$. Observe that if a measurement is more dependent on the language (i.e., the syntax) than on the subject (i.e., the semantics), it is expected that variability of the distribution of $X_{t=\pi,l}$ is larger than the variability of $X_{t,l=\lambda}$. On the other hand, if X is more dependent on semantics, it is expected that the variability of $X_{t,l=\lambda}$ is larger than the variability of $X_{t=\pi,l}$ (AMANCIO *et al.*, 2013). We calculated the variability of the distributions by employing the coefficient of variation (CV) of the distribution, which is computed as

$$CV(X) = \frac{\sigma(X)}{\langle X \rangle}. \quad (3.6)$$

3.4.5 Results and discussion

In these experiments, we employed two datasets, the Holy Bible dataset (described in Section 3.1.3), and books obtained from the Project Gutenberg (described in Section 3.1.2). The English books were chosen to study characteristics of the network that varies in the same language. However, for different subjects, to better understand the variations among language and topics, we considered: English, French, German, Italian and Portuguese. These books are listed as follows:

1. **English:** The Adventures of Sherlock Holmes (Arthur Conan Doyle), The Tragedy of the Korosko (Arthur Conan Doyle), The Valley of Fear (Arthur Conan Doyle), Uncle Bernac - A Memory of the Empire (Arthur Conan Doyle), Dracula's Guest (Bram Stoker), The Lair of the White Worm (Bram Stoker), The Jewel Of Seven

Stars (Bram Stoker), The Man (Bram Stoker), The Mystery of the sea (Bram Stoker), A Tale of Two Cities (Charles Dickens), Barnaby Rudge: A Tale of the Riots of Eighty (Charles Dickens), American Notes (Charles Dickens), Great Expectations (Charles Dickens), Hard Times (Charles Dickens), The Works of Edgar Allan Poe – Volumes 2 and 4 (Edgar Allan Poe), Beasts and Super-Beasts (Hector H. Munro), The Chronicles of Clovis (Hector H. Munro), The Toys of Peace (Hector H. Munro), The Girl on the Boat (P. G. Wodehouse), My Man Jeeves (P. G. Wodehouse), Something New (P. G. Wodehouse), The Adventures of Sally (P. G. Wodehouse), The Clicking of Cuthbert (P. G. Wodehouse), A Pair of Blue Eyes (Thomas Hardy), Far from the Madding Crowd (Thomas Hardy), Jude the Obscure (Thomas Hardy), The Mayor of Casterbridge (Thomas Hardy), The Hand of Ethelberta (Thomas Hardy), Barry Lyndon (William M. Thackeray), The History of Pendennis (William M. Thackeray), The Virginians (William M. Thackeray), and Vanity Fair (William M. Thackeray).

2. **French:** Le fils du Soleil (Gustave Aimard), Face au Drapeau (Jules Verne), Pierre de Villerglé (Louis Amédée Achard), Les Idoles d'argile (Louis Reybaud), and Han d'Islande (Victor Hugo).
3. **German:** Die Wahlverwandtschaften (Goethe), Der Moloch (Jakob Wassermann), Königliche Hoheit (Thomas Mann), and Lichtenstein (Wilhelm Hauff).
4. **Italian:** Il Peccato di Loreta (Alberto Boccardi), La Montanara (Anton Giulio Barrili), Alla Finestra (Enrico Castelnuevo), Sciogli la treccia, Maria Maddalena (Guido da Verona), and La Pergamena Distrutta (Virginia Mulazzi).
5. **Portuguese:** Amor de Perdição (Camilo Castelo Branco), A Cidade e as Serras (Eça de Queirós), Os Bravos do Mindello (Faustino da Fonseca), Transviado (Jaime de Magalhães Lima), and Uma Família Inglesa (Júlio Dinis).

By employing these datasets, we analyzed some properties of the metrics extracted from the paragraph-based networks. The two main focus of this analysis are given by: characterization concerning informativeness and the ability of the metrics to capture syntactical and/or semantic textual characteristics. As a case of sample, we considered the found results to examine Voynich manuscript.

3.4.5.1 Informativeness analysis

In this section, we apply both of the presented approaches to compute informativeness (described in Section 3.4.3). First, we consider Criterion A, in which a measurement X is considered informative if the value of X in a real (RT) text differs from the values of X obtained from shuffled texts (SW and SS). Here, we considered both datasets (the

Holy Bible and the entire Books dataset). The results are presented in Table 2. In order to compare measurements in distinct intervals, for each measurement and each of the datasets, we standardized the data by considering the three types of texts (RT, SS, and SW). As a result, the average value of each standardized measurement shown in Table 2 is zero (for a given measurement $RT + SS + SW = 0$) and the standard deviation is one.

Table 2 – Measurements obtained for RT, SS, and SW networks by considering Holy Bible and Books dataset. To be possible to compare different measurements, all the presented data is standardized.

X	Holy Bible dataset			Books dataset (all)		
	RT	SS	SW	RT	SS	SW
$\langle k \rangle$	-0.12 ± 1.14	$+0.13 \pm 0.87$	-0.01 ± 0.96	-0.18 ± 1.02	$+0.05 \pm 0.98$	$+0.13 \pm 0.98$
$\sigma(k)$	-0.59 ± 0.52	$+0.56 \pm 1.15$	$+0.02 \pm 0.86$	-0.78 ± 0.65	$+0.47 \pm 0.92$	$+0.31 \pm 0.90$
$\langle B \rangle$	$+0.48 \pm 0.90$	-0.23 ± 1.02	-0.24 ± 0.90	-0.05 ± 1.07	$+0.05 \pm 1.03$	$+0.00 \pm 0.89$
$\sigma(B)$	-0.22 ± 0.89	$+0.23 \pm 1.11$	-0.01 ± 0.93	-0.66 ± 0.68	$+0.52 \pm 0.94$	$+0.13 \pm 0.97$
$\langle cc \rangle$	$+0.74 \pm 0.60$	-0.58 ± 1.09	-0.16 ± 0.74	$+0.31 \pm 0.90$	-0.17 ± 1.06	-0.15 ± 0.96
$\sigma(cc)$	$+0.52 \pm 0.45$	-0.32 ± 1.37	-0.20 ± 0.70	-0.28 ± 0.64	$+0.27 \pm 1.21$	$+0.01 \pm 0.98$
$\langle N \rangle$	-0.14 ± 0.74	-0.09 ± 1.33	$+0.23 \pm 0.76$	$+0.31 \pm 0.83$	-0.13 ± 1.14	-0.18 ± 0.92
$\sigma(N)$	$+0.39 \pm 0.76$	-0.18 ± 1.28	-0.21 ± 0.75	-0.28 ± 0.85	$+0.08 \pm 1.13$	$+0.20 \pm 0.94$
$\langle Ecc \rangle$	$+0.64 \pm 1.04$	-0.42 ± 0.91	-0.22 ± 0.67	$+0.01 \pm 1.08$	-0.09 ± 1.03	$+0.07 \pm 0.87$
$\sigma(Ecc)$	$+0.11 \pm 1.02$	$+0.01 \pm 1.09$	-0.13 ± 0.86	-0.38 ± 1.04	$+0.18 \pm 0.93$	$+0.20 \pm 0.92$
$\langle EC \rangle$	$+0.73 \pm 1.22$	-0.54 ± 0.49	-0.19 ± 0.64	-0.18 ± 1.13	$+0.06 \pm 0.92$	$+0.12 \pm 0.91$
$\sigma(EC)$	$+0.33 \pm 1.09$	-0.13 ± 0.96	-0.20 ± 0.85	-0.53 ± 0.82	$+0.32 \pm 1.01$	$+0.21 \pm 0.93$
$\langle C \rangle$	-0.64 ± 0.79	$+0.32 \pm 1.02$	$+0.33 \pm 0.84$	-0.02 ± 1.08	$+0.00 \pm 1.03$	$+0.02 \pm 0.89$
$\sigma(C)$	-0.42 ± 0.55	$+0.43 \pm 1.28$	-0.01 ± 0.83	-0.87 ± 0.59	$+0.44 \pm 0.89$	$+0.43 \pm 0.84$
$\langle Sb^{(2)} \rangle$	-0.09 ± 0.67	-0.09 ± 1.26	$+0.18 \pm 0.96$	$+0.48 \pm 0.87$	-0.40 ± 0.97	-0.09 ± 0.96
$\sigma(Sb^{(2)})$	$+0.13 \pm 0.53$	$+0.06 \pm 1.39$	-0.19 ± 0.84	-0.45 ± 0.69	$+0.29 \pm 1.13$	$+0.16 \pm 0.96$
$\langle Sm^{(2)} \rangle$	$+0.12 \pm 0.37$	-0.33 ± 1.52	$+0.21 \pm 0.63$	$+0.55 \pm 0.83$	-0.34 ± 1.06	-0.20 ± 0.85
$\sigma(Sm^{(2)})$	$+0.18 \pm 0.60$	-0.02 ± 1.20	-0.16 ± 1.08	$+0.37 \pm 0.62$	-0.43 ± 1.24	$+0.06 \pm 0.87$
$\langle Sb^{(3)} \rangle$	$+0.13 \pm 0.65$	-0.20 ± 1.29	$+0.06 \pm 0.93$	$+0.41 \pm 0.90$	-0.26 ± 1.02	-0.14 ± 0.94
$\sigma(Sb^{(3)})$	-1.09 ± 0.68	$+0.61 \pm 0.57$	$+0.48 \pm 0.65$	-0.52 ± 0.96	$+0.21 \pm 1.00$	$+0.31 \pm 0.82$
$\langle Sm^{(3)} \rangle$	-0.48 ± 0.72	$+0.51 \pm 1.16$	-0.03 ± 0.81	-0.38 ± 1.01	$+0.15 \pm 1.01$	$+0.23 \pm 0.85$
$\sigma(Sm^{(3)})$	-0.66 ± 0.96	$+0.53 \pm 0.89$	$+0.13 \pm 0.74$	-0.46 ± 1.05	$+0.15 \pm 0.94$	$+0.30 \pm 0.83$
$\langle Sb^{(4)} \rangle$	$+0.30 \pm 0.61$	-0.02 ± 1.39	-0.28 ± 0.72	-0.28 ± 0.90	$+0.11 \pm 1.10$	$+0.17 \pm 0.93$
$\sigma(Sb^{(4)})$	$+0.42 \pm 0.62$	-0.25 ± 1.29	-0.18 ± 0.83	-0.30 ± 0.91	-0.01 ± 1.11	$+0.31 \pm 0.87$
$\langle Sm^{(4)} \rangle$	$+1.00 \pm 0.71$	-0.69 ± 0.60	-0.30 ± 0.76	$+0.35 \pm 1.10$	-0.28 ± 0.93	-0.07 ± 0.84
$\sigma(Sm^{(4)})$	-0.23 ± 0.95	$+0.24 \pm 1.05$	-0.01 ± 0.94	-0.57 ± 1.07	$+0.29 \pm 0.83$	$+0.28 \pm 0.82$
$\langle \alpha^{(\infty)} \rangle$	-0.22 ± 1.10	-0.08 ± 1.02	$+0.30 \pm 0.78$	$+0.31 \pm 0.95$	-0.25 ± 1.04	-0.06 ± 0.93
$\sigma(\alpha^{(\infty)})$	-0.20 ± 1.00	-0.25 ± 0.86	$+0.45 \pm 0.98$	-0.86 ± 0.71	$+0.37 \pm 0.85$	$+0.49 \pm 0.82$
$\langle \alpha^{(2)} \rangle$	-0.50 ± 0.52	$+0.12 \pm 1.28$	$+0.38 \pm 0.82$	$+0.19 \pm 0.96$	-0.17 ± 1.05	-0.02 ± 0.96
$\sigma(\alpha^{(2)})$	-0.04 ± 0.81	-0.15 ± 1.16	$+0.19 \pm 0.97$	-0.79 ± 1.05	$+0.20 \pm 0.65$	$+0.60 \pm 0.68$
$\langle \alpha^{(3)} \rangle$	$+0.02 \pm 0.51$	-0.24 ± 1.38	$+0.22 \pm 0.86$	$+0.48 \pm 0.78$	-0.33 ± 1.05	-0.15 ± 0.96
$\sigma(\alpha^{(3)})$	$+1.03 \pm 0.56$	-0.78 ± 0.68	-0.26 ± 0.69	$+0.77 \pm 1.01$	-0.52 ± 0.78	-0.25 ± 0.68
Q	$+1.32 \pm 0.33$	-0.91 ± 0.22	-0.41 ± 0.31	$+1.24 \pm 0.65$	-0.64 ± 0.38	-0.60 ± 0.37

Taking into consideration the Holy Bible dataset, the measurement that best discriminated real from shuffled texts was modularity (Q). In the comparison between real networks and their respective shuffled versions, Q calculated from RW network differs 2.24 and 1.73 from the SW and SS versions, respectively. Because Q quantifies how well defined the communities of the network are, this result suggests that the community structure

is better defined in real networks than in their respective shuffled versions. This result might be a consequence of the organization of topics present in real texts.

In addition to the modularity, other measurements were also found to be informative. When comparing RT and SW, the largest differences were found for: the accessibility ($\sigma(\alpha^{(3)})$), symmetry ($\sigma(Sb^{(3)})$, $\langle Sm^{(4)} \rangle$), and the clustering coefficient ($\langle cc \rangle$). In the comparison between RT and SS, the measurements that best discriminate were found for: symmetry ($\sigma(Sb^{(3)})$ and $\langle Sm^{(4)} \rangle$), accessibility ($\sigma(\alpha^{(3)})$), and closeness ($\langle C \rangle$). Many of the measurements that are able to distinguish between real and shuffled texts, also distinguish both shuffled texts.

As in the previous case, by considering the Books dataset, modularity, Q , was found to be the measurement that best distinguishes RW from SS and SW. As a consequence of the higher values of Q for RW networks than for the shuffled versions (SW and SS), the community structure also plays an important role when we considered this dataset. This result means that the informativeness is a characteristic that seems to be independent on syntax and semantics. Other measurements were also found to discriminate RW from SS and SW: clustering coefficient ($\sigma(C)$), degree ($\sigma(k)$) and accessibility ($\alpha^{(\infty)}$, and $\sigma(\alpha^{(2)})$ and $\sigma(\alpha^{(3)})$).

In the following, we consider Criterion B (described in Section 3.4.3). In this analysis, we use z-score to compare a real texts with their corresponding shuffled versions. We do not consider the entire books dataset, but only the English part. So, it is possible to better understand the measurements that are more related to language (Holy Bible dataset), or semantics (English part of books dataset). Table 3, shows for each measurement (X) the percentage of documents with significant difference between real and shuffled texts ($z(X) > 1$).

As in the previous tests, Q implies an important role for both of the considered datasets, and reached 100% of informativeness for all tests. Other measurements that also resulted in similar values for both datasets are: $\sigma(k)$, $\langle Sm^{(3)} \rangle$, and $\langle cc \rangle$. These measurements were found to be informative for approximately 50% of the samples, but for many other cases, the informativeness level varied according to the dataset. Some measures are more informative in the Holy Bible dataset, $\sigma(Sb^{(3)})$, and $\sigma(\alpha^{(3)})$, while others are more informative in the Books dataset, $\sigma(C)$ and $\langle Sm^{(2)} \rangle$.

Except for Q , the results suggest that it is essential to analyze the characteristics of the dataset to decide if network measurements extracted from paragraph networks can be classified as informative. The found results agree with the assertion that paragraph-based networks can be less informative than other types of text networks (AMANCIO, 2015d). In the case of co-occurrence networks, the majority of the network measurements were considered informative even though different datasets were employed.

Table 3 – Percentage of documents in each dataset where the difference between a real text (English part) and the corresponding shuffled version was found to be significant. Apart from the modularity, the informativeness seems to depend on the type of dataset used.

Measurements		SW		SS	
		Holy Bible	Books	Holy Bible	Books
Degree	$\langle k \rangle$	31.25%	30.30%	18.75%	36.36%
	$\sigma(k)$	56.25%	78.79%	31.25%	72.73%
Betweenness	$\langle B \rangle$	37.50%	48.48%	50.00%	57.58%
	$\sigma(B)$	6.25%	66.67%	6.25%	60.61%
Clustering	$\langle cc \rangle$	50.00%	27.27%	50.00%	39.39%
	$\sigma(cc)$	12.50%	30.30%	62.50%	30.30%
Neighborhood	$\langle N \rangle$	6.25%	42.42%	6.25%	63.64%
	$\sigma(N)$	25.00%	39.39%	37.50%	60.61%
Eccentricity	$\langle Ecc \rangle$	37.50%	30.30%	43.75%	45.45%
	$\sigma(Ecc)$	37.50%	51.52%	56.25%	54.55%
Eigenvector	$\langle EC \rangle$	62.50%	57.58%	62.50%	57.58%
	$\sigma(EC)$	31.25%	78.79%	37.50%	66.67%
Closeness	$\langle C \rangle$	43.75%	39.39%	62.50%	51.52%
	$\sigma(C)$	12.50%	90.91%	25.00%	93.94%
Symmetry	$\langle Sb^{(2)} \rangle$	6.25%	54.55%	25.00%	51.52%
	$\sigma(Sb^{(2)})$	6.25%	42.42%	12.50%	42.42%
	$\langle Sm^{(2)} \rangle$	0.00%	57.58%	6.25%	63.64%
	$\sigma(Sm^{(2)})$	6.25%	30.30%	12.50%	42.42%
	$\langle Sb^{(3)} \rangle$	18.75%	45.45%	18.75%	48.48%
	$\sigma(Sb^{(3)})$	93.75%	36.36%	93.75%	48.48%
	$\langle Sm^{(3)} \rangle$	56.25%	57.58%	25.00%	66.67%
	$\sigma(Sm^{(3)})$	75.00%	57.58%	68.75%	57.58%
	$\langle Sb^{(4)} \rangle$	0.00%	54.55%	37.50%	66.67%
	$\sigma(Sb^{(4)})$	18.75%	63.64%	37.50%	66.67%
	$\langle Sm^{(4)} \rangle$	93.75%	54.55%	87.50%	54.55%
	$\sigma(Sm^{(4)})$	31.25%	54.55%	31.25%	66.67%
Accessibility	$\langle \alpha^{(\infty)} \rangle$	0.00%	54.55%	37.50%	54.55%
	$\sigma(\alpha^{(\infty)})$	12.50%	93.94%	18.75%	93.94%
	$\langle \alpha^{(2)} \rangle$	12.50%	42.42%	50.00%	42.42%
	$\sigma(\alpha^{(2)})$	18.75%	51.52%	12.50%	81.82%
	$\langle \alpha^{(3)} \rangle$	0.00%	54.55%	25.00%	51.52%
	$\sigma(\alpha^{(3)})$	100.00%	66.67%	87.50%	63.64%
Modularity	Q	100.00%	100.00%	100.00%	100.00%

3.4.5.2 Dependency with language and semantics

In this section, we analyze the dependency with language and semantics. More specifically, the variability was considered for two scenarios: (i) semantics (text) is constant and the language (syntax) is varied (Holy bible dataset); and (ii) language is the same, and the semantics varies (English part of Books dataset). We employ the English language because in our dataset there is a larger number of English texts than other languages. First, we identified measurements that are more related to characteristics of syntax (the Holy Bible dataset). The measurements that revealed more significant variability were: accessibility ($\langle \alpha^3 \rangle$ and $\sigma(\alpha^2)$), degree ($\langle k \rangle$), eccentricity ($\sigma(Ecc)$), symmetry ($\langle Sb^2 \rangle$, $\langle Sb^3 \rangle$, $\sigma(Sm^4)$ and $\sigma(Sb^2)$), neighborhood ($\langle N \rangle$) and betweenness ($\sigma(B)$).

In the case of semantic changes identification (English Books dataset), the measurements that were found to have highest coefficients of variation are: eccentricity ($\langle Ecc \rangle$), closeness ($\langle C \rangle$), symmetry ($\langle Sm^4 \rangle$ and $\langle Sb^4 \rangle$), betweenness ($\langle B \rangle$), degree ($\langle k \rangle$), eigenvector centrality ($\langle EC \rangle$), accessibility ($\langle \alpha^2 \rangle$ and $\langle \alpha^\infty \rangle$) and neighborhood ($\langle N \rangle$). Interestingly, some measurements were found to be dependent on both syntax and semantics (e.g., $\langle N \rangle$). Alternatively, for both symmetry and accessibility, the measurements can capture information regarding syntax or semantics when different hierarchical levels are considered.

In order to probe which is more prevalent between (i) the ability to detect changes in syntax and (ii) the ability to detect changes in semantics, we analyzed the measurements by comparing the coefficient of variation of the considered datasets. The results are shown in Figure 32, in which, the top sub-panels illustrate the results obtained for $\sigma(k)$, Q , and $\sigma(cc)$. In the present cases, there is no significant difference between these values, which means that both, syntax and semantics, are captured.

3.4.5.3 Classification tests

In order to illustrate an application of the paragraph-based network, we considered two classification tasks. In the first case, we studied the problem of classifying a manuscript into two classes: (i) shuffled texts and (ii) meaningful documents. In the second classification problem, we employed the framework that is proposed in the first case to analyze an unknown text (Voynich manuscript). More specifically, we probe if Voynich manuscript can be considered compatible with real texts, by considering the structure of the relationship among its paragraphs.

3.4.5.3.1 Discriminating between real and shuffled texts

In order to illustrate the capabilities of paragraph-based networks to characterize texts for real application, we apply our methodology to distinguish between real and shuffled texts. For each document of the books dataset, we created the three versions

Figure 32 – Comparisons of the measurements that provided lowest values CV, by considering the average between CV computed for Books and Bible datasets. The error bars represent the confidence interval of a mean for the interval of 95% of confidence.

Source: Elaborated by the author.

of the paragraph-based networks (RT, SW, and SS). We considered the network measurements as features for our classification. Furthermore, we employed only the 10 most informative measurements (shown in Table 2), which are: Q , $\sigma(C)$, $\sigma(\alpha^\infty)$, $\sigma(k)$, $\sigma(B)$, $\sigma(\alpha^3)$, $\sigma(\alpha^2)$, $\langle Sm^{(2)} \rangle$, and $\sigma(EC)$. This procedure was employed as an attribute selection method. As the next step, all the features were standardized.

For the classification task, we employ the live-one-out cross-validation method, and the SMO (Sequential Minimal Optimization) classifier, which is an SVM implementation available in Weka (WITTEN *et al.*, 2016; HALL *et al.*, 2009) (This method is better explained in Appendix C). For this test, we considered the standard parameters, which are also chosen according to the procedure defined in (AMANCIO *et al.*, 2014). By considering the three classes (RT, SW, and SS), the accuracy is 76.08%, the true positive rate is 0.98 for the RT samples and 0.71 and 0.50 for SW and SS, respectively, and false negative rates are 0.02, 0.24, and 0.14 for RT, SW, and SS, respectively. These results indicate that the proposed framework is able to differentiate between real and the two classes of shuffled texts. However, when comparing the two classes of shuffled documents, more classification mistakes happens.

We also evaluate the classification considering only two classes. For this test, the shuffled class contains both shuffled versions (SS and SW). The classification reached 98.72% of accuracy, and the false positive rate of RT networks is zero. The two classi-

Figure 33 – Probability density function (pdf) of the first component of the linear discriminant analysis projection of the selected features which were employed in classifying the texts into two classes. In his analysis we compared two classes, which are *Real* against *Rhuffled* (SW and SS) texts.

Source: Elaborated by the author.

fication mistakes happened for real texts classified as shuffled texts. This result means that when a sample is classified as real text, the classification tends to be right. In order to visualize the results, we projected the features into a Linear Discriminant Analysis (LDA) (FRIEDMAN; HASTIE; TIBSHIRANI, 2001), which illustrates the separation between the two classes into a single dimension (see Figure 33). Due to the importance of modularity (Q) in the informativeness analysis, we also evaluate the performance of the classification by employing Q as a single feature. The found accuracy rate reached 96.79%, which sustain the importance of Q in discriminating real and shuffled texts.

We compare the obtained results with a *doc2vec* based technique (we considered the version implemented in *Gensim* library³). This approach is based on the embedding model proposed by Le and Mikolov (2014) and in the negative sampling method (MIKOLOV *et al.*, 2013b). We considered the distributed memory version and the mean of the context word vectors (DAI; OLAH; LE, 2015). As the method parameters, we set the vector as 200 elements, windows size with 8 positions, the number of iterations equals to 10, and we removed the words that occur less than 19 times. In order to create a train our *doc2vec* vectors, we employ the entire English version of Wikipedia⁴. In this test, we considered only the English documents. Because the number of paragraphs is not a limitation to this approach, we added more books, as listed as follows: *The Book of Snobs* (by William Makepeace Thackeray), *The Unbearable Bassington* (by Saki), *The Works of Edgar Allan*

³ <https://radimrehurek.com/gensim/>

⁴ [<https://www.wikipedia.org>](https://www.wikipedia.org)

Poe – Volumes 1, 3 and 5 (by Edgar Allan Poe), Through the Magic Door (by Arthur Conan Doyle), and When William Came (by Saki). These books were not considered in the network-based tests because they have a fewer number of paragraphs than Voynich manuscript. So, it would not be possible to employ the proposed methodology.

By considering the pre-trained dataset, for each novel and for each version (RT, SW, and SS), we created one vector. This result is illustrated in Figure 34, which shows the respective PCA projection. Also considering these vectors, we test the capability of this approach to classify the texts as RT, SW, or SS. Using a SVM classifier, with the same parameters employed before, the accuracy of 70.8% was found. This methodology was found to be able to classify correctly all SW samples (with 100% of true positives), which are texts that do not follow a grammar. Nevertheless, 57.5% of the RT samples were classified as SS. In the comparison between two classes, Real (RT) and the shuffled versions (SW and SS), only 63,3% were correctly classified, in which 67.5% of the samples of RT were incorrectly classified as shuffled texts.

Figure 34 – PCA projection computed from the *doc2vec* features. There are two groups, where one represents SW, and the other group represents RT and SS.

Source: Elaborated by the author.

3.4.5.3.2 Case Example: Voynich manuscript

By considering the characteristics obtained from paragraph-based networks, we analyzed if Voynich manuscript is compatible with real or shuffled texts. Voynich manuscript is an incomprehensible manuscript. Characteristics of Voynich manuscript are described in Section 3.1.4. When comparing Voynich manuscript with shuffled texts, we did not

consider SS versions of Voynich manuscript because in this manuscript it is not possible to define sentences.

Many researchers have been studying if this mysterious text is meaningful (BELFIELD, 2007). Figure 35 shows the paragraph-based network of Voynich manuscript. In order to associate the unknown semantics to a possible content of the book, for each paragraph, we assigned the picture that takes part in its respective page. For paragraphs in pagers without pictures we assigned as *text*. In the visualization, it is possible to note a well-defined community structure, with two main groups. Moreover, this structure seems to capture characteristics of the organization of Voynich manuscript, in which there is a separate group of paragraphs with figures of plants. Because modularity was found to be informative in the previous analysis, the characteristics of communities shown in Figure 35 suggests that Voynich manuscript can be compatible with real texts. Interestingly, a similar conclusion has been reported for other researchers that took into consideration the different characteristics of the manuscript (AMANCIO *et al.*, 2013; MONTEMURRO; ZANETTE, 2013; REDDY; KNIGHT, 2011).

Since modularity, Q , was found to be much higher for the set of real texts (RT) than to the set of shuffled texts (SS and SW), we compare this measurement to networks created from Voynich manuscript. Figure 46 shows the distribution of Q for the three versions of paragraph-based networks (RT, SS, and SW) created from the entire Books dataset. Interestingly, the modularity of the Voynich network is compatible with real texts, and for all versions of Voynich as SW, Q is compatible with the modularity extracted from SS and SW. Additionally, we employed the same classifier as in the previous section to analyze Voynich manuscript. In this analysis, we use the entire Books dataset as a training set. The manuscript was classified as real text, and a set of 30 SW networks of Voynich were also classified with an accuracy of 100%. Our results indicate that Voynich manuscript is compatible with natural languages at the paragraph level.

3.4.6 Conclusions

In this section, we presented a novel paragraph-based network, which is created by considering the semantical similarities between paragraphs. By considering the proposed model, many network measurements were extracted, and two main properties were analyzed, the dependency with language and/or semantics. Additionally, this framework was applied in a machine learning task, of discriminating between real and shuffled texts. The measure that better described the real networks is modularity, Q , which means that the networks obtained from real texts tend to have a structure of communities. This characteristic reflects the information concerning how writers tend to organize the subjects in a text. By considering the entire machine learning framework, we analyzed Voynich manuscript, which is an unknown text written with obscure words.

Figure 35 – Paragraphs-based network of Voynich manuscript. The labels consist in topics that were manually defined according to the manuscript pictures, as follows: (i) *text*, when there are no images; (ii) *plants*; (iii) *bath*, with figures of women and bath-like shape; (iv) *astronomy*, with spatial-like figures (e.g., stars); and (v) *vase*. The employed software of visualization was implemented by [Silva \(2015\)](#).

Source: Elaborated by the author.

Regarding the paragraph-based networks, Voynich manuscript was found to be compatible with real texts. In this manuscript the order of the pages is unknown. Consequently, the order of paragraphs is also unknown. So, the proposed approach is suitable for this task because the proposed method is not dependent on the paragraph order. Our results regarding Voynich manuscript agrees with previous studies that considered word networks. These works found that the manuscript syntax is coherent with natural languages ([AMANCIO *et al.*, 2013](#); [MONTEMURRO](#); [ZANETTE, 2013](#)). Recently, another study considered several natural languages and found that Hebrew is the most likely language of the manuscript ([HAUER](#); [KONDRAK, 2016](#)). The results presented in previous studies and this section indicate that the attempts to decipher the manuscript are relevant since the manuscript could contain a meaningful text. The paragraph-based approach also paves the way for many other tasks. For example, the paragraph-based features can be

Figure 36 – Distributions of Probability Density Function, PDF, by considering the modularity measurement of RT, SW, and SS of networks. The modularity of the two versions of the Voynich manuscript, RT and SW, are represented by blue and orange circles, respectively.

Source: Elaborated by the author.

employed as a complement to other features. Some examples of applications are: authorship and plagiarism identification ([AMANCIO, 2015b](#)), automatic identification of literary movements ([AMANCIO; OLIVEIRA JR; COSTA, 2012a](#)), fraud detection ([WANG; XU, 2018](#)), among others ([SALLOUM *et al.*, 2017](#)).

3.5 Mesoscopic analysis of texts

In this section, we present a novel representation of texts as complex networks that aims to capture information of the unfolding of the story along time. This methodology consists of a network model based on similarities between chunks of texts. So, here we consider a higher scale level than the paragraph-based networks. We applied this representation to two different tasks: classification between real and shuffled texts and authorship attribution. For the second task, we explore the information obtained from network visualization. The methodology and results presented in this section were published in ([ARRUDA *et al.*, 2018](#); [MARINHO *et al.*, 2017](#); [ARRUDA *et al.*, 2018](#)).

3.5.1 Mesoscopic network approach

An essential characteristic to create the mesoscopic networks is that the book shall be organized. In other words, the paragraphs shall be in the same order as the original text. This approach starts with a pre-processed document, in which we considered all steps presented in Section 2.6.1. By considering an organized and pre-processed text O ,

Because in this type of network the similarity measurement (cosine similarity) depends on the content of each document, the edge weights are not comparable among different texts. So, it is not appropriate to consider a fixed threshold of T for all the texts. Another limitation of employing the threshold is that a low T value could lead to removing almost all edges, and when T value is too high, almost all nodes are kept. In order to deal with both problems, we define T to keep a given percentage of the edges, with the highest weights. A similar methodology was used in Section 3.4.1. In contrast to the previous methodology, here we define a set of thresholds as $\{T_{5\%}, T_{10\%}, T_{15\%}, \dots, T_{95\%}\}$ where $T_{5\%}$ is the threshold T that keeps only 5% of the edges with the highest weights, $T_{10\%}$ is the threshold T for 10%, and so on. The classification takes into consideration all these versions of networks and their respective network measurements.

In this analysis, we considered the entire books. So, the number of nodes varies, which can influence many complex network measurements. In order to deal with this problem, in this analysis, we considered only two local measurements: clustering coefficient (cc) and matching index (μ). These measurements were chosen to include distinct information of the network, in which cc is computed from the network nodes, and μ is computed from the network edges. As additional information the measurements were sorted according to the text sequence, giving rise to a time series. In the case of μ , we created the time series by considering the order of edges:

$$\{\mu_{0,0}, \mu_{0,1}, \dots, \mu_{0,n-1}, \mu_{1,0}, \mu_{1,1} \dots \mu_{1,n-1}, \dots, \mu_{n-1,n-1}\}. \quad (3.7)$$

If there is no edge linking two nodes, there is no corresponding value in the time series.

3.5.3 Image-based features

Taking into consideration that the visualizations of mesoscopic networks seem to be a kind of “calligraphy” of the authors, we consider the analysis with image-based features (see Figure 38). More details regarding visualizations and the visualizations are shown in section 3.5.4.3. The mesoscopic networks were visualized in two dimensions by using the version of a force directed algorithm implemented by Silva *et al.* (2016a) (see section 2.4). Before measuring features employed in the visualization of the networks, we convert the images to monochromatic (also called binary images).

As the first step, a preprocessing was employed. This step consists in dilating and eroding (EFFORD, 2000) the images to remove small holes. In order to keep the same parameter of preprocessing to find the dilation and erosion kernel, we considered the object size. We use the higher eigenvalue (λ_1) of the the Principal Component Analysis (PCA) (JOLLIFFE, 2002), and compute the kernel size as follows

$$c_{dim} = c\lambda_1, \quad (3.8)$$

where c is a constant, set as $c = 3 \cdot 10^{-4}$. In other words, the larger the object is, the larger the kernel size. The following features were extracted from the preprocessed images: *Area*, *Perimeter*, *Euler number*, *Minimum enclosing circle* (SKYUM, 1991), *Convex hull* (SKLANSKY, 1982), *Elongation* (COSTA; CESAR, 2000), *Lacunarity* (RODRIGUES; BARBOSA; COSTA, 2005), and *Fourier-based features*. (COSTA; CESAR, 2000). The image-based features are described in Appendix D.

Figure 38 – Scheme illustrating the mapping of texts into networks and the classification process. Initially, documents are mapped into networks. Then, two types of features are calculated: network features and attributes derived from network visualizations. The classification can then be performed using both network and image features.

Source: Elaborated by the author.

In order to compare the results obtained from the image-based features, we also employed some centrality measurements (see Figure 38). Because the network size profoundly influences many of these measurements, here we employed only local measurements, as follows: accessibility (for $h = \{2, 3\}$), degree and assortativity, backbone and merged symmetry (for $h = \{2, 3, 4\}$), the average degree of neighbors, and clustering coefficient. Since almost all measurements are computed for single nodes, some statistics were considered: average, standard deviation, and skewness. These statistics were employed as features in our classifier.

3.5.4 Results and discussion

Here, we present the results belonging to the methodology presented in this section. First, we present a case study, in which we analyze the network visualization in created from the novel “Alice’s Adventures in Wonderland”, by Lewis Carroll. In order to probe the characteristics of grasping story information, we compare the networks created from real books and shuffled stories. These networks were also employed in the authorship

attribution task. For that, we considered features obtained from image analysis, through visualizations, and network-based features.

3.5.4.1 Case study: Mesoscopic analysis of “Alice’s adventures in wonderland”

We investigate the mesoscopic network created from the novel “Alice’s Adventures in Wonderland”, by Charles Lutwidge Dodgson, under the pseudonym Lewis Carroll. This book narrates a fiction about a little girl called Alice. She follows a rabbit, fall in a hole, and arrives in a fantasy world. This book has 12 chapters, which are shown in Chart 2. By considering this book, two versions of the mesoscopic network were created, $\mathcal{G}_1$ and $\mathcal{G}_2$, in which two distinct thresholds were employed, $T_1 = 0.31$ and $T_2 = 0.18$, respectively. Furthermore, we set the window size as $\Delta = 20$.

Chart 2 – Chapters of Alice’s Adventures in Wonderland, by Lewis Carroll.

Number	Title
I	Down the Rabbit-Hole
II	The Pool of Tears
III	A Caucus-Race and a Long Tale
IV	The Rabbit Sends in a Little Bill
V	Advice from a Caterpillar
VI	Pig and Pepper
VII	A Mad Tea-Party
VIII	The Queen’s Croquet-Ground
IX	The Mock Turtle’s Story
X	The Lobster Quadrille
XI	Who Stole the Tarts?
XII	Alice’s Evidence

As the first analysis, we investigate the properties of $\mathcal{G}_1$ (see Figure 39), which is simpler than $\mathcal{G}_2$. There are connections between the nearest nodes, which creates a chain-like structure. This effect happens because of the windows of paragraphs, which share many of their content. Such a characteristic is better shown in Figure 39 (a), where the colors represent the paragraphs order in the text. In contrast to this characteristic, connections between nodes created from distinct regions of the text represent high similarity between the respective chunks of text. In this case, there are no overlaps of paragraphs.

In order to better understand the relationship between the mesoscopic structure of the network and the book story, we generate a network visualization in which the nodes are colored according to the chapters. This visualization is shown in Figure 39 (b)). Additionally, relationships between the text and the network are presented as follows:

- There are no strong connections from nodes of Chapter 1 to nodes of other chapters, except for chapter 2 and 3, which is explained by the overlap between subsequent paragraphs. The lack of long-range connections between some nodes of the first

Figure 39 – Visualization of the $\mathcal{G}_1$ network, which represents the book “Alice’s Adventures in Wonderland”, by Lewis Carroll. Such a visualization was projected by considering the threshold $T_1 = 0.31$, and the nodes represent sets of 20 paragraphs. The paragraphs order are shown in item (a), and item (b) represents the book chapters. The node is assigned to the paragraph that is in the middle of the set of paragraphs. In this example is 10. The employed software of visualization was implemented by [Silva \(2015\)](#).

Source: Elaborated by the author.

chapter and other parts of the story probably happens because the first chapter of the novel is substantially different from almost every other part. In Chapter 1, the story starts in a more realistic scenario, in which there are fewer descriptions of the fantasy locations and creatures, except for the presence of the Rabbit. On the other hand, in the rest of the book fantasy world, characters, and situations are described;

- Network nodes created from Chapters 2, 3, and the beginning of Chapter 4 are connected among themselves. This part of the book evidenced Alice’s feelings, by including some negative words, such as *poor*, *hopeless*, *cry*, *tears*, *tired*. Chapters 3 and 4 are not so emotional as Chapter 2, but negative words are still used referring to past events. It may be a consequence of the fact that this story part described the period when Alice was frightened about the world she arrived. All these chapters mentioned when she cried and a pool of tears was formed. Moreover, she always remembers things about her cat, *Dinah*, and how good her pet was at catching other animals;
- In chapter 5, there are connections with a high edge weight between regions of the same chapter. These edges probably happens because of a long conversation between

Alice and the *Caterpillar*. They talk about some topics, discussing the many sizes she had during that day and how confused she was about herself;

- The connection between chapters 7 and 11 can be related to the character *Hatter*, who is drinking tea in both chapters. In Chapter 7, this character is in the tea party, and in Chapter 11, he is in a trial. Furthermore, in both chapters he talks about specific kinds of food related to the tea party, e.g., bread and butter, and include the characters *Alice*, *March Hare*, *Dormouse* and the *Hatter*;
- A highly connected group of nodes can be seen at the end of Chapter 9 and the beginning of chapter 10. In the last paragraphs Alice is introduced to *The Mock Turtle*, and their conversation ended only in chapter 10. Additionally, both chapters are related to school. At the end of Chapter 9, the *The Mock Turtle* explains to Alice his days at school, and the lessons he used to have. In Chapter 10, some characters asked Alice to repeat lessons. As a result, she felt like she was at school.

In contrast to the visualization displayed in Figure 39, Figure 40 shows a visualization of $\mathcal{G}_2$, which was computed by considering a lower threshold value, $T_2 = 0.18$. In this case, the network preserved more edges than in the first case. This parameter was explored in order to illustrate the potential of our method to map the story in a multi-scale fashion. In Figure 40(a) it is possible to observe that the network still has a chain-like structure similar to $\mathcal{G}_1$ version. However, in this network, there are more connections between nodes from different parts of the book. Because of this higher number of edges, $\mathcal{G}_2$ network captures more fine-grained information about the relationships between chunks of paragraphs. By comparing $\mathcal{G}_1$ and $\mathcal{G}_2$ versions, while Chapter 1 is connected only with nodes that represent Chapter 2 and 3 in $\mathcal{G}_1$, in $\mathcal{G}_2$ these nodes also connect with chapters 4 and 7. Apart from this advantage, the fine-grained networks may present some disadvantages because they tend to incorporate more local characteristics, which can include some noise and relationships provided from weak related contextual content.

3.5.4.2 Discriminating real from shuffled texts

In order to illustrate the potential of the mesoscopic networks, we classified the texts by considering real (RT) and two versions of shuffled texts (SW and SP). The features were computed by employing the methodology presented in Section 3.5.2, which consists in a temporal series-based analysis of clustering coefficient (cc) and matching index (μ). Figure 41 illustrates the importance of analyzing cc in terms of temporal series for both versions of the network, $\mathcal{G}_1$ and $\mathcal{G}_2$. There are complex patterns of periodic measurements of low and high values of cc . This pattern can be seen along time, and low values of cc characterize regions with long-range connections. A similar pattern was found for μ , which also captures the presence of long-range links. Because μ quantifies

Figure 40 – Visualization of the $\mathcal{G}_2$ network, which represents the book “Alice’s Adventures in Wonderland”, by Lewis Carroll. Such a visualization was projected by considering the threshold $T_1 = 0.31$, and the nodes represent sets of 20 paragraphs. The paragraphs order are shown in item (a), and item (b) represents the book chapters. The node is assigned to the paragraph that is in the middle of the set of paragraphs. In this example is 10. The employed software of visualization was implemented by [Silva \(2015\)](#).

Source: Elaborated by the author.

the number of shared neighbors of an edge, this measurement can be used to identify long-range connections. Observe that the nodes of mesoscopic networks normally share many neighbors since nodes represent overlapping sets of paragraphs. On the other hand, nodes representing distant regions share fewer neighbors than other nodes.

In Figure 42, we also considered the book “Alice’s Adventures in Wonderland”. Here we compare cc of the three versions of the network RT, SW, and SP. The average values of cc are the same for all versions of the same network, in this case, all three versions fluctuates around $cc \simeq 0.78$, but the fluctuations are markedly dissimilar. The most significant variations were found for the real version of the book (RT), while for both shuffled versions (SW and SP) the variations consist in more extensive regions of weak fluctuations (e.g., see nodes from 180 to 280 in Figure 42(b)). As in the previous case, a similar pattern was obtained for μ .

The dataset considered in this analysis consists of the same dataset used for the classification of the paragraph-based networks (Section 3.4.5). Because in the previous analysis we took into account only books having at least the same number of paragraphs as Voynich manuscript, here we included some other books that have a smaller number of paragraphs, which are:

Figure 41 – Visualization of the networks representing “Alice’s Adventures in Wonderland”. Item (a) represents the network $\mathcal{G}_1$ ($T_1 = 0.31$) and item (b) represents the network $\mathcal{G}_2$ ($T_2 = 0.18$). Clustering coefficient cc is indicated by the node colors, in which nodes with the highest values are represented in orange. There is a non-trivial pattern of clustering coefficient along the network nodes. The employed software of visualization was implemented by Silva (2015).

Source: Elaborated by the author.

- **Edgar Allan Poe:** *The Works of Edgar Allan Poe* (Volumes 1, 3 and 5);
- **Hector H. Munro (Saki):** *When William Came* and *The Unbearable Bassington*;
- **William M. Thackeray:** *The Book of Snobs*;
- **Robert Walser:** *Geschwister Tanner* (written in German).

As in the paragraph-based analysis, we divided the dataset into two groups, according to their languages: (i) English and (ii) Other languages.

In order to analyze the measurements as time series, we employ the coefficient of variation (DAS, 2008) to the set of observations X , in which X represents the ordered set of values of cc or μ . The coefficient of variation $CV(X)$ is computed as presented in equation 3.6. For a given a window size, δ , and for a given subsequence of X , $\mathcal{X}_k^\delta = \{x_k, x_{k+1}, \dots, x_{k+\delta-1}\}$, the coefficient of variation, $CV(\mathcal{X}_k^\delta)$, is calculated. By considering this measurement in a multi-scale, we considered a set of window sizes as $\delta = \{3, 5, 7, 10, 15, 20, 25, 30, 35, 40, 50\}$. For each value of δ , we summarize the values of fluctuations by averaging over all $c_v(\mathcal{X}_k^\delta)$ as follows:

$$\mathcal{C}_v^\delta(X) = \frac{1}{N} \sum_{k=1}^{n-\delta+1} c_v(\mathcal{X}_k^\delta). \quad (3.9)$$

Figure 42 – Clustering coefficient, cc , measured from all network nodes of real and shuffled versions created from the book “Alice’s Adventures in Wonderland” (RT, SW, and SP). This example was measured for $T_1 = 0.31$, in order to select the strongest semantical links.

(a) RT

(b) SP

(c) SW

Source: Elaborated by the author.

In order to characterize each of the networks, we employ $\mathcal{F} = \{\mathcal{C}_v^{\delta=3}, \mathcal{C}_v^{\delta=5}, \mathcal{C}_v^{\delta=7}, \dots\}$ as the set of features, where X represents vectors of values of clustering coefficient and matching index.

By considering these features, we visualized all the books measurements, by considering the three classes of texts (RT, SW, and SP) in a PCA projection, as shown in Figure 43(a). The two first components account for approximately 76% of the projection variance. The networks that take part in the same class tend to be close in groups. This effect is confirmed by the average distance between the classes, as shown in Table 4.

In order to compare the results obtained from the three types of mesoscopic networks, we performed a similar analysis by considering a traditional approach based on co-occurrence networks. Since the pre-processing step can vary among different languages, only the English dataset was considered. Because the number of network nodes varies in mesoscopic networks, these networks were trimmed to have the same number of words. Additionally, the features were extracted by considering the same measurements as we employed in Section 3.2.1, in which, apart from modularity, the following statistics were used for all other measurements: maximum value ($\max(X)$), median ($\tilde{X}$), minimum value ($\min(X)$), and standard deviation $\sigma(X)$. The obtained data was also projected into a PCA, as shown in Figure 43(b), in which the sum of the two main PCA components accounts for 70% of the projection. In contrast to the results obtained from mesoscopic networks, here the group of networks from RT and SP are not distinguishable. It happens because for both network types (RT and SP), the networks are similar and the differences happen only when words in different paragraphs are connected. This effect happens because in SP the paragraphs are the same as in RT; the difference is due to the order of paragraphs. Also, co-occurrence networks are devoted to grasping linguistic/syntactical features. In mesoscopic representation, language structure is kept, and only the mesoscopic structure is changed (in SP texts). Table 5 confirms those results by considering the distances, as also shown in the densities presented in Figure 43(b).

Table 4 – Average distance among mesoscopic networks from the same class. When using mesoscopic networks, it is possible to discriminate real texts from those generated by both shuffled words and paragraph.

	RT	SW	SP
RT	0.00	13.67	10.98
SW	13.67	0.00	13.21
SP	10.98	13.21	0.00

We also analyze the database by using an unsupervised approach based on the K-means algorithm (FRANK *et al.*, 2009) (more details about this method are presented in Appendix C). We employed all documents of the datasets, divided into real texts and the two classes of shuffled texts. In this analysis, we employed the 6 principal components

Figure 43 – PCA projections of the networks generated from Real Texts (RT), Shuffled Paragraphs (SP) texts, and Shuffled Words (SW) texts. The projections (a) and (b) represent the mesoscopic and the co-occurrence networks, respectively. The circles and stars represent English and Other languages datasets. In co-occurrence analysis, only the English texts were considered. These plots also show the densities of the samples according to the PCA components, in which the continuous and dashed lines represent English and Other languages datasets, respectively.

(a) Mesoscopic networks.

(b) Co-occurrence networks.

Source: Elaborated by the author.

Table 5 – Average distance among co-occurrence networks from the same class. If co-occurrence networks are used, real texts and texts of shuffled paragraphs cannot be discriminated.

	RT	SW	SP
RT	0.00	7.18	0.12
SW	7.18	0.00	7.08
SP	0.12	7.08	0.00

as features. The choice of the number of components was empirically found aiming to provided better results. First, we considered PCA components computed from all measurements. Only 8.9% of instances were incorrectly clustered, and 0.02% of false negatives were found for SP networks. Furthermore, when we considere the features separately, the results obtained only for clustering coefficient-based features outperform the results obtained from matching index. The percentage of incorrectly assigned instances are 11.7% and 16.7%, when clustering and matching index are used, respectively.

In order to measure the quality of the clusters, we employed two measurements, accuracy the Adjusted Rand Index (ARI) (HUBERT; ARABIE, 1985). The same methodology was applied to the traditional approach that considers network measurements of co-occurrence networks. We used the same measurements in the first analysis considering co-occurrence networks. In this case, the ten principal components of PCA were employed as features, which was also empirically chosen. These cluster quality indexes measured for both types of networks (mesoscopic and co-occurrence) are shown in Table 6. As expected from the analysis of Figure 43(b), co-occurrence networks are not able to distinguish between RT from SP documents. As a consequence, 72.5% of SP texts were incorrectly classified as RT, which reflects in accuracy and in ARI, which are much lower in co-occurrence networks. On the other hand, the results obtained from mesoscopic network indicate that this type of network can distinguish the texts among the tree classes.

Table 6 – Comparison of different network approaches, by considering the K-means clustering performance. Two different measurements were applied: Adjusted Rand Index (ARI) and Accuracy. In both cases, 1 indicates that all instances are correctly classified, and 0 indicates the opposite.

	ARI	Accuracy
Mesoscopic (Clustering)	0.679	0.883
Mesoscopic (Matching Index)	0.576	0.833
Mesoscopic (all features)	0.749	0.911
Co-occurrence	0.268	0.575

Another characteristic of the mesoscopic model that can distinguish real from shuffled texts is the existence of long-range connections. More specifically, we define a long-range connection as a link that connects two nodes that are far apart in the document. As shown in Section 3.5.4.1, this type of connection is created when a previous subject or con-

text repeats in the book story. This type of links can be a consequence of the long-range correlation effect (EBELING; NEIMAN, 1995), which is essential for mapping a multidimensional conceptual space into a smaller dimensional space (ALVAREZ-LACALLE *et al.*, 2006). This characteristic can be related to the nature of the book (e.g., scientific writing, tales, or novels) or the writing style. By considering the three versions of “Alice’s Adventures in Wonderland” (RT, SW, and SP), we analyze the presence of long-range links in terms of edge weights, as shown in Figure 44. More specifically, the scatterplots shows all edges weights against the time difference between linked nodes, in which time corresponds to the order of paragraphs in the document. For all of the document versions, higher values of weight are found for smallest time differences. This effect is imposed by the construction rules of mesoscopic networks, in which adjacent nodes share chunks of text. As highlighted in the insets of Figure 44, real texts tend to present stronger long-range connections than shuffled texts. The presence of strong long range connections only in the RT case can be related to the contextual relationships along the story.

Figure 44 – Example of time difference between linked nodes against their respective edges weights, by considering “Alice’s Adventures in Wonderland” by Lewis Carroll. This measurement was computed for all network edges and the inset represents a region of long-range links. Comparing the different classes of texts, it is evident that strong long-range connections are more likely to appear in real networks.

Source: Elaborated by the author.

3.5.4.3 Use case: Authorship attribution

In the following, we provide an analysis of the visual properties of the mesoscopic networks. We also discussed the accuracy of the authorship attribution results obtained from experiments considering all authors and experiments considering only pairs of authors. In order to test our approach in the authorship attribution task, we employed the following list of 10 authors in a total of 50 books:

1. **Herman Melville:** “Moby Dick, Or, The Whale”, “The Confidence-Man: His Masquerade”, “The Piazza Tales”, “Typee: A Romance of the South Seas”, and “White

- Jacket, Or, The World on a Man-of-War”;
2. **B. M. Bower:** “Cabin Fever; Lonesome Land; The Long Shadow; The Lookout Man; The Trail of the White Mule”;
 3. **Jane Austen:** “Emma”, “Mansfield Park”, “Persuasion”, “Pride and Prejudice”, and “Sense and Sensibility”;
 4. **Mark Twain:** “A Connecticut Yankee in King Arthur’s Court”, “Adventures of Huckleberry Finn”, “The Adventures of Tom Sawyer”, “The Prince and the Pauper”, and “Roughing It”;
 5. **Charles Darwin:** “Coral Reefs”, “Geological Observations on South America”, “The Different Forms of Flowers on Plants of the Same Species”, “The Expression of the Emotions in Man and Animals”, and “Volcanic Islands”;
 6. **Charles Dickens:** “American Notes”, “A Tale of Two Cities”, “Barnaby Rudge: A Tale of the Riots of Eighty”, “Great Expectations”, and “Hard Times”;
 7. **Edgar Allan Poe:** “The Works of Edgar Allan Poe” (Volume 1 - 5);
 8. **Hector H. Munro (Saki):** “Beasts and Super Beasts”, “The Chronicles of Clovis”, “The Toys of Peace”, “The Unbearable Bassington”, and “When William Came”;
 9. **Thomas Hardy:** “A Changed Man and Other Tales”, “A Pair of Blue Eyes”, “Far from the Madding Crowd”, “Jude the Obscure”, and “The Hand of Ethelberta”;
 10. **Henry James:** “The Ambassadors”, “The American”, “The Portrait of a Lady - Volume 1”, “The Real Thing and Other Tales”, and “The Turn of the Screw”.

As illustrated in Section 3.5.4.1, the visualizations of mesoscopic networks represent the unfolding of the story along time, in which the long-range connections represent periods of the story when a given subject is repeated. Interestingly, the visualization of the networks created from texts of the same author seems to be similar. To illustrate some of the visual specificities present in these representations, we employed the visualization of three different authors, which are: Edgar Allan Poe, Thomas Hardy, and Mark Twain. These representations are shown in Figure 45. In order to better visualize the narrative flow, the color was added to the images, in which blue and green nodes represent the beginning and the end of the story, respectively.

The procedure employed here aims to probe that visual characteristics of mesoscopic networks reflect features of writing style regarding the story organization. We calculated all the image features, IF, (shown in Section 3.5.3), which were selected and ordered according to SVM’s (Support Vector Machine) attribute selection (GUYON *et al.*,

Figure 45 – Examples of the visualization of networks created for texts of three different authors: Edgar Allan Poe, Thomas Hardy, and Mark Twain. The visualizations of each author seem to follow a specific type of pattern. The employed software of visualization was implemented by [Silva \(2015\)](#).

(a) The Works of Edgar Allan Poe (Volume 1), by Alan Poe.

(b) The Works of Edgar Allan Poe (Volume 2), by Alan Poe.

(c) The Works of Edgar Allan Poe (Volume 3), by Alan Poe.

(d) A Changed Man and Other Tales, by Thomas Hardy.

(e) A Pair of Blue Eyes, by Thomas Hardy.

(f) Far from the Madding Crowd, by Thomas Hardy.

(g) Adventures of Huckleberry Finn, by Mark Twain.

(h) The Adventures of Tom Sawyer, by Mark Twain.

(i) The Prince and the Pauper, by Mark Twain.

Source: Elaborated by the author.

2002a). This method is described in Appendix C. Because in this analysis we have 10 authors (classes) and only 5 books per author, we employed two unsupervised classification methods: Expectation Maximization (EM) (LLOYD, 1982) and KMeans (DEMPSTER; LAIRD; RUBIN, 1977) (see Section 3.5.3). In order to find the best number of features that lead to the higher accuracy, we tested the n first features of the ordered ranking, by varying n from 1 to the total number of features (see Table 7). The best classification accuracy was found for EM, 54%, in which the number of features was $n = 14$. This result supports the hypothesis that there are visual differences among the visualizations of the mesoscopic networks, which are captured by the image features. This procedure was employed because we aimed to analyze the features, and not to find the best classifier to the problem of authorship attribution.

Table 7 – Classification accuracy and number of used features considering different measurements.

	EM		KMeans	
	Accuracy	Features	Accuracy	Features
IF	54%	14	48%	13
NF	54%	15	50%	3
IF+NF	58%	19	54%	34

The books can also be classified according to features obtained from the network topology. For that purpose, we employed of network-based features (NF). The set of employed features is described in Section 3.5.3. The same machine methodology that we applied to the image features was used here. The obtained accuracy of 54% is the same found for by employing image-based features. However, here the number of used features was $n = 15$. We also test the combination of all features (IF + NF) and also used the same classification methodology. A better accuracy rate was found for 58%, with $n = 19$. In order to illustrate the capability of this features to be employed in authorship attribution, we plot a Linear Discriminant Analysis (LDA) projection (see Figure 46). In this projection, we considered the features that provided the better classification results (IF + NF). The books of the same author tend to be grouped.

3.5.5 Conclusions

We proposed an approach that considers the semantical similarity between textual segments that represent network nodes, which consist of sequences of adjacent paragraphs. This representation grasps to capture long-range correlations and the temporal unfolding of documents. For that purpose, two parameters were defined: (i): a number of consecutive paragraphs in each window and (ii) a threshold used to prune connections among nodes with low contextual similarity. Additionally, the visualizations of mesoscopic networks can also be employed usefully in the identification of characteristics of novels and writers.

Figure 46 – LDA projection by considering all measurements obtained from the selected network measurements and image-based measurements. Each color represents a writer. The plots also show the densities of samples according to the respective classes (writers). Points that represent books of the same writer seem to be grouped on the scatter plot.

Source: Elaborated by the author.

In order to probe the capabilities of our method, some analyses were employed. As the first step, we considered the novel “Alice’s Adventures in Wonderland”, by applying network visualization techniques to the generated mesoscopic network. We identified the characteristics of the story and the implications in the visualization. For example, the identification of long-range connections were created because of a specific character that takes place in two distinct chapters. As another way to probe that our methodology can grasp information regarding the story, we compare the networks created from real texts with networks built from shuffled texts. By considering network measurements and an analysis based on temporal series employed to a classifier, the better result was found for 0.91. This analysis was also used in a set of different languages, which indicates that the method can also be employed to texts provided from other languages.

This methodology was also employed in the task of authorship identification. First of all, by considering a visual inspection, we noted that the books of the same author seem to have a kind of mesoscopic “calligraphy”. In other words, the shapes of the visualizations seem to follow specific patterns according to the writing style. In order to probe that the writers tend to follow the same calligraphy, we analyzed image-based features. These features were employed to unsupervised classifiers, and the found result was compatible with the classification considering network-based features. Possible extensions of the use

of mesoscopic networks include other classification tasks, such as authorship attribution task employed to bigger sets of novels and analysis concerning literary periods. Moreover, the mesoscopic networks can be created from any temporal series. For instance, data obtained from the stock market and weather.

ANALYSIS OF LANGUAGE AND KNOWLEDGE VIA NETWORK DYNAMICS

In this chapter, we present the methodologies and the respective dynamics related to language and knowledge analysis. Knowledge acquisition is intrinsically connected to language because human beings communicate and transmit their discoveries in spoken and written word. Furthermore, there is a relationship between the dynamics of the evolution of languages and knowledge acquisition, where both share many aspects, and can be considered complementary ([DICKMANN, 2011](#)).

We study representations of knowledge, starting from a network whose nodes and edges represent papers and citations, respectively. In the following, we simulated the knowledge acquisition by proposing a multi-agent dynamics. As another dynamics, we considered an engine that mounts a puzzle, which is similar to a person to learn a concept. We also consider the transmission of information, by using tools of information theory (see [Appendix F](#)) and comparing variations of random walk dynamics. As a case example, we simulated the transmission of language, considering information theory, which is employed to improve the transmission quality.

4.1 Datasets

In this section, we present all the dataset we employ in this thesis. First, we show the American Physical Society (APS) dataset, which comprises pieces of information regarding articles published by APS. We also present two networks of knowledge. Additionally, we present datasets of the frequency of syllables in a set of languages.

4.1.1 APS dataset

APS dataset¹ consists of the set of all articles published by the American Physical Society. In this dataset, each article is represented by its own Digital Object Identifier (DOI), and the citations are organized as pairs of DOIs. For example, if for a given pair of papers, *A* cites *B*, the pair of DOIs for *A* and *B* are presented in the same line of the file. Furthermore, metadata are included, such as the journal name and the year of publication. The list of these pairs of DOIs can be considered as an adjacency list that represents a network of citations. In order to consider the directed version of this network, we adopt the directions followed from the citing to the cited articles.

4.1.2 Knowledge representation

In order to represent knowledge, we employed two real-world networks representing knowledge. The first is Wikipedia (SILVA *et al.*, 2011), henceforth called WIKI network, in which each node represents an article and the edges are created when there is a hyper-link from an item to another. In this case, we used a subset of Wikipedia² articles that comprise only two main categories: Biology and Mathematics. As the second knowledge representation, we considered a network of citations of papers obtained from the Web of Science (SILVA *et al.*, 2016a) (WOS). In this network, each node represents a paper, and the edges are created between two articles when one cites the other. Furthermore, we considered only documents in the complex networks field. Both networks were considered as being undirected.

4.2 Integration between theoretical and applied physics

In this section, we employ a network-based analysis on a citation network to quantify the integration between theoretical and applied physics. We considered the APS dataset, which provides information concerning physics related articles from 1893 to 2016. More details on APS dataset and the respective network is presented in Section 4.1.1. Our approach consists of analyzing groups of papers, represented by network nodes, and better understanding the relationship between groups. This analysis comprises the higher scale we consider in this thesis. The methodology and results presented in this section are published in (ARRUDA; COMIN; COSTA, 2018a).

4.2.1 Methods

We started our analysis by classifying the article into three groups: *applied physics*, *theoretical physics*, and *control*. In the first test, we considered as the two first groups the

¹ <<https://journals.aps.org/datasets>>

² <<https://www.wikipedia.org>>

papers taking part in *Phys. Rev. Applied* and *Phys. Rev. D* journals, respectively, and the papers from the remaining journals were used to define the control group, henceforth called Control (A). *Phys. Rev. D* includes works of more academic-related areas, such as field theory, and quantum field theory, and *Phys. Rev. Applied* publishes articles related to applications, including engineering and technology. Because *Phys. Rev. Applied* is recent and consequently the number of published papers is much lower than the other journals, we also considered *Phys. Rev. B* as the applied physics group, which publishes more application-oriented papers. In this case, the remaining of the papers were also considered as the control group, henceforth called Control (B). The organization of the groups of papers and their respective pieces of information is shown in Chart 3. These two sets of articles were considered in experiments that were called *Case Study #1*, when *Applied (Rev. App.)* was employed, and *Case Study #2*, when *Phys. Rev. B* was considered.

Chart 3 – The adopted groups of articles and some of their properties. The used dataset comprises articles from 1893 to 2016. Control (A) and Control (B) represent the Control group of articles when, respectively, *Phys. Rev. Applied* and *Phys. Rev. B* is employed.

Group	Number of articles	Years
Applied (Rev. App.)	703	2014–2016
Applied (Rev. B)	176,381	1970–2016
Theoretical (Rev. D)	80,427	1970–2016
Control (A)	509,325	1893–2016
Control (B)	333,647	1893–2016

We considered the hypothesis that there is the presence of multiple shortest paths, with low lengths, connecting two sets of articles, which reflects integration and the relationship between them. The distances between papers were quantified concerning topological distances of network nodes. For that purpose, we use the distribution of shortest paths by considering sets of 705 papers randomly selected from the three groups. This number of samples was adopted because it is approximately the number of articles in *Phys. Rev. Applied*, which is the journal with the lower number of papers. Pairs of nodes without a path connecting them were removed from our statistics. In order to understand this measurement in a more intuitive way, we employed the measurement $1/d_{i,j}$, in which $d_{i,j}$ is the shortest path length between the nodes i and j , which we called *cohesion*. When cohesion is higher for a given pair of nodes, the distances between these nodes are lower. So, more related papers tend to have cohesion next to 1, which is the highest possible value, and as a complement cohesion next to 0 indicates that the articles tend to be unrelated.

As a complement of computing cohesion, we considered Betweenness Centrality, which is related to the network flow. In other words, this measurement accounts for the shortest paths that go through the nodes that represent a given journal. This measurement is better described in Section 2.2.2, and has been employed to similar studies, such as the identification of emerging fields (LEE, 2008), analysis of interdisciplinary (LEY-

DESDORFF, 2007), and scientific impact (YAN; DING, 2009).

4.2.2 Results and discussion

In this subsection, we presented the results into two case studies, by considering the papers of *Phys. Rev. Applied* and *Phys. Rev. B*, as the applied group. For both cases, we discussed the analysis concerning directed and undirected versions of the network.

4.2.2.1 Case Study #1 - *Phys. Rev. Applied*

In this case study, we considered *Phys. Rev. Applied*, *Phys. Rev. D*, and Control (A) as applied, theoretical, and control groups, respectively. First, we analyze the directed version of the network. In order to better understand the relationship between the areas, we presented the results as a digraph, as shown in Figure 47. In this figure, for each edge, we presented the distribution and average values of cohesion, which is also represented by the edge widths. Additionally, we plot the histograms of betweenness centralities with the respective average values. The connections that are too small were removed from the digraph. There is a little difference among the average values of cohesion, which indicates that there is uniformity among the connections between the three considered groups.

Figure 47 – Cohesion (blue), defined as the reciprocal of the shortest path length, and betweenness centrality (green) relationships between *Phys. Rev. D*, *Phys. Rev. Applied* and Other APS journals for the directed network. The legend at the upper right corner indicates the width range of the links. Observe that all links have similar width, suggesting symmetric relationships between all groups.

Source: (ARRUDA; COMIN; COSTA, 2018a).

Figure 47 also displays the betweenness centrality distribution of the nodes, and similar results were also obtained. However, the group that represents papers devoted to applied studies presents a less symmetric histogram than the other cases. This effect could be caused as a consequence of the size of the applied group, which is smaller than the remaining two groups. Similar results were found for the undirected version of the network, as shown in Figure 48. In this case, the distributions of betweenness centrality are similar for all groups. Also, in the case of the interconnections between groups, cohesion resulted in comparable values.

Figure 48 – Cohesion (blue) and betweenness centrality (green) relationships between Phys. Rev. D, Phys. Rev. Applied and Other APS journals for the undirected network. The legend at the upper right corner indicates the width range of the links.

Source: (ARRUDA; COMIN; COSTA, 2018a).

4.2.2.2 Case Study #2 - Phys. Rev. B

Here, the group of applied articles is represented by *Phys. Rev. B*, and *Phys. Rev. D* and Control (A) are represented by theoretical and control groups, respectively. As in the first case study, we summarize the results into two figures, in which we present information regarding the cohesion between groups of papers, and betweenness centrality for the groups. The directed results are presented in Figure 49. Similarly to the results presented in Figure 47, in this case, the values of cohesion histograms are similar in terms of their averages for all cases, which also indicates that there is uniformity among the connections between the three considered groups. Observe that, as a consequence of the larger number of articles in *Phys. Rev. B* than in *Phys. Rev. Applied*, here there are

more edges. These results indicate that citations between papers of these journals have no preferred direction. The distributions of betweenness centrality also presented similar behavior for all groups. When we considered the undirected version of the APS network, the distributions and average values of cohesion and betweenness histograms resulted similar for the three considered groups, as shown in Figure 50.

Figure 49 – Cohesion (blue) and betweenness centrality (green) relationships between Phys. Rev. D, Phys. Rev. B and Other APS journals for the directed network. The legend at the upper right corner indicates the width range of the links.

Source: (ARRUDA; COMIN; COSTA, 2018a).

4.2.3 Conclusions

In this section, we employed concepts developed as tools of complex networks (e.g., shortest paths, and betweenness centrality) to quantify the relationship between different areas of physics. These areas were defined according to the scope of physics-related journals, which were derived from the APS dataset. As a theoretical journal, we considered *Phys. Rev. D*, which we believe that is the most theoretical-related option. In order to have a more consistent result, as an applied journal, we considered *Phys. Rev. Applied* and *Phys. Rev. B*. All in all, the results indicate that the relationships between the studied groups are similar. Thus, papers of theoretical and applied journals of APS dataset tend to be equally integrated.

Figure 50 – Cohesion (blue) and betweenness centrality (green) relationships between Phys. Rev. D, Phys. Rev. Applied and Other APS journals for the directed network. The legend at the upper right corner indicates the width range of the links.

Source: ([ARRUDA; COMIN; COSTA, 2018a](#)).

4.3 Simulating knowledge acquisition

There are many different dynamics executed on complex networks ([GUARDIOLA *et al.*, 2000](#); [MOTTER; LAI, 2002](#); [MORENO; NEKOVEE; PACHECO, 2004](#); [BARRAT; BARTHLEMY; VESPIGNANI, 2008](#)). Among these studies, some authors simulate the language evolution ([GOEL *et al.*, 2016](#); [CARRO; TORAL; MIGUEL, 2016](#)) and, in a more embracing attempt, some researches investigate the knowledge acquisition ([SILVA *et al.*, 2010](#)). Among these studies, some researchers try to describe how researchers discover new knowledge ([BATISTA; COSTA, 2010](#); [SILVA *et al.*, 2010](#)), such as how knowledge acquisition happens ([FOSTER; RZHETSKY; EVANS, 2015](#); [COKOL *et al.*, 2005](#)). Another possibility proposes a model that aims to improve the collective discovery ([RZHETSKY *et al.*, 2015](#)). Observe that some of these studies were developed by using complex networks ([COSTA, 2006](#); [SILVA *et al.*, 2010](#)). In this study, we proposed a novel model that aims at simulating collective discovery, in which authors attract others to study related subjects. The methods and results presented in this section are published in ([ARRUDA *et al.*, 2017](#)).

4.3.1 Problem definition

In this study, we proposed a dynamics that is executed on complex networks. So, the vertices represent concepts and the relationship between these concepts are indicated by the edges. In order to represent the knowledge discovery, we adopt a True Self-Avoiding Walk (TSAW) dynamics (KIM; PARK; YOOK, 2016). The probability of a node j to be visited from a node i varies according to a $\psi(i, j)$ function, which is computed as follows

$$\psi(i, j) = \Omega^{n_v(j)}, \quad (4.1)$$

where Ω is a constant (here we fixed in 0.5) and $n_v(j)$ is the number of times node j was already visited. Observe that when the dynamics starts, n_v is equals to zero for all nodes. Because $\psi(i, j)$ is computed for all $j \in J$ neighbors of i , the transition probabilities are computed as

$$P_w(i, j) = \frac{\psi(i, j)}{\sum_{j \in J} \psi(i, j)}. \quad (4.2)$$

We considered more than one agent in our dynamics. So, this equation is defined for each agent separately. In other words, the probability $P_w(i, j)$ considers only the notes where a given agent a visited.

In order to consider the interaction among the researchers, we employ the concept of stochastic flight, also known as jump. The dynamics simulation represent researchers discovering new concepts, and when they jump, they change their field of research. For that, we defined a probability of jump, γ , for each iteration of the dynamics. As a complement, we modeled a field of attraction among the researchers, which acts only when an agent jumps. So, the researcher has a higher probability to move to a node next to other researchers. This field of attraction is computed for each agent a and for all node i , according to

$$E_a(i) = \eta_a \exp(-d(i, j)\tau), \quad (4.3)$$

where j is the current position of the agent a , τ is a constant of function decay, η is a vector of fitness of the researchers, and $d(i, j)$ is the geodesic distance between the nodes i and j . The resultant field is computed by

$$\varepsilon(i) = \sum_{a \in A} E_a(i), \quad (4.4)$$

where A is the set of all agents. In order to compute the probability of a given agent, a , to arrive to a node i , we employ the following equation

$$P(i) = \frac{\varepsilon(i)}{\sum_{j \in V} \varepsilon(j)}. \quad (4.5)$$

The proposed dynamics was executed for a set of networks, as shown in Chart 4. Moreover, many possibilities of the parameters η , τ , and γ were tested. So, it was possible

to better understand how the parameters influence the knowledge discovery. For each execution of the dynamics, 1500 iterations were computed, and for each network, the simulation was repeated 300 times. In the case of η , we considered only a vector with zeros and ones that represent influent and not influent researchers, respectively. We opt for this adoption to simplify our tests and to reduce the number of parameters. As a measurement of the quality of the dynamics, we considered the total number of explored nodes, TE . More specifically, this measurement consists of the number of nodes that were visited at least once for any of the agents.

Chart 4 – Informations concerning the networks used to execute knowledge acquisition dynamics.

Network	Description	N	k	Parameters
LA	Bidimensional lattice	10k	3.96	–
TLA	Bidimensional toroidal lattice	10k	4.00	–
WS-1	Watts-Strogatz model	10k	4.00	$p = 0.001$
WS-2	Watts-Strogatz model	10k	4.00	$p = 0.005$
WAX	Waxman model	$\approx 10k$	6.02	$\alpha = 1$, $\beta = 0.015$, and $L = 1$
BA	Barabasi-Albert model	10k	6.00	
CN	Community Network Lancichinetti Benchmark (LB)	10k	5.63	
WIKI	Subset of wikipedia encompassing articles from Biology and Mathematics	$\approx 12k$	7.29	–
WOS	Citation network obtained for the query "Complex Networks" on the Web of Science	$\approx 11k$	17.08	–

Apart from using TE , which considers the performance from a global point of view, we also employed a performance measurement that considers regions of the network. First of all, these regions were defined according to accessibility, which is known to detect borders in networks (TRAVENTÇOLO; COSTA, 2008). As a parameter of this measurement, we empirically used $h = 3$. We compute $\alpha^{(3)}$ for all network nodes, and the values were sorted in ascending order. We define windows of size n_R , and here we use a value of n_R that represents 20% of the number of network nodes. In other words, for a given interval of values, the respective nodes were selected. The windows of nodes were defined for a range between two consecutive windows, henceforth called step, which we adopt the step size as 5% the total number of network nodes. Observe that there are overlaps of nodes for some distinct windows. Because in the analysis of regions we aimed to describe how fast the region is explored, we use the measurement TE by considering only the first 1000 iterations. Such a parameter was defined according to the results obtained in the global analysis, presented in the next section.

4.3.2 Results

In this subsection, we present the knowledge acquisition results by considering two distinct characteristics: (i) Performance regarding *global* discovery and (ii) Performance regarding *local* discovery. For both analyses, we discussed a varied range of parameters and all networks presented in Chart 4. The number of researchers was fixed in 20 for all

tests, and the analyses consist of 300 realizations of the dynamics for each combination of parameters. Furthermore, we fixed the parameters as $\eta_i = 1$, for $i = 1, 2, \dots, |A|$, where A is the set of walkers, $\tau = 1$, and $\gamma = 0.1$. For the three parameters, for each analysis one parameter is chosen to vary.

4.3.2.1 Global analysis

In the first analysis, we considered both versions of the lattice (LA and TLA) because of their simplicity. Since the results obtained from LA and TLA regarding $\langle TE \rangle$ are similar for all parameters (η , τ , and γ), we opt to show only the LA results. First, we analyze the variations of η parameter, which resulted in similar values of $\langle TE \rangle$, as shown in Figure 51 (a). For all tests concerning the vector η , we define percentages of nodes with zeros and ones that represent influent and non influent researchers, respectively. These defined percentages were: $\{5\%, 10\%, 15\%, 20\%\}$. The found results indicated that the number of agents with high influence did not change the dynamics performance significantly. By considering the variation of τ , which defines the locality of the field, and the probability of jump, γ , the dynamics performance, $\langle ET \rangle$, was greatly affected and a higher standard deviation was found. The evaluated performances of τ and γ , are shown in Figure 51 (b) and Figure 52 (a), respectively. Interestingly, we found that the performance decreases when η or τ increases. Furthermore, the results computed for the spatial-related models, WS-1, WS-2, and WAX are similar to those obtained by executing the dynamics on LA.

Figure 51 – Average of the total number of explored nodes ($\langle TE \rangle$) and the corresponding standard deviations, represented by the shadow regions, executed in LA.

Source: Elaborated by the author.

Taking into consideration γ parameter, we compared similarities among the curves of $\langle TE \rangle$ obtained for the networks LA, WS-1, and WS-2, as shown in Figure 52 (a) (b) and (c). The standard deviations of finding in TE are similar among all network models, and when γ increases the dynamics performance changes significantly for all of these networks. An-

Figure 52 – Average of the total number of explored nodes ($\langle TE \rangle$) and the corresponding standard deviations, represented by the shadow regions, computed by varying the jump probability (γ).

Source: Elaborated by the author.

other interesting characteristic is that the influence of γ variations diminished when the probability of rewiring rises for WS model. LA also corresponds to a WS model with rewiring probability equals to zero, which indicates that $\langle TE \rangle$ tends to vary more according to γ when the network is more spatial or organized. However, this effect is not only a consequence of the spatiality of networks because the results obtained for the WAX presented a different behavior, as shown in Figure 53. In this case, WAX network also has a spatial structure but presented a low variation with the increase of $\langle TE \rangle$ when γ varied. For both cases, the simulation resulted in a better exploration performance when the agent did not jump. Furthermore, similar results were found for the parameter τ .

Figure 53 – Average of the total number of explored nodes ($\langle TE \rangle$) and the corresponding standard deviations of our dynamics, obtained by varying the values of γ when executed on WAX network.

Source: Elaborated by the author.

By considering the jump probability (γ) for WOS, WIKI and CN networks, in contrast with the results obtained for spatial networks, we found that employing networks the increasing of γ improves the performance of the dynamics. See Figure 52(d-f) and Figure 54(a). The structures of these networks are more related to knowledge structures. However, the effects caused when γ is changed were found to be much lower than for the networks with spatial characteristics. Additionally, the performance obtained for the dynamics when executed on WIKI network exhibited, in general, performances considerably lower than the other models. Interestingly, this network also differs from CN network, which reproduces the characteristics of community structure and node degree distribution that also take part in WIKI network. The other parameters η and τ , did not cause a significant variation of the dynamics performance ($\langle TE \rangle$). For both cases, the results are similar to those obtained for the BA network (shown in Figures 54(b) and (c)).

4.3.2.2 Analysis by regions

In the following, we analyze the dynamics performance ($\langle TE \rangle$) concerning the network regions. For that, we considered the average of the total number of explored nodes after 1000 iterations of the dynamics, henceforth called $\langle TE_{1000} \rangle$. We choose this

Figure 54 – Average of the total number of explored nodes ($\langle TE \rangle$) and the corresponding standard deviations when executed on BA network. The influence of η and τ parameters were not found to be significative. However, higher values of jump probability, γ , results in a more efficient exploration.

Source: Elaborated by the author.

number of iterations because by choosing a very small number, the exploration can be local, resulting in similar values for $\langle TE_{1000} \rangle$ all networks. On the other hand, if we consider high values, some regions could be explored entirely also resulting in similar performances.

Similarly to the first test, we considered the LA network. Here we did not consider the TLA because in this case, it is not possible to define regions. For LA, in order to define regions, we compute the Chebyshev distance (HEIJDEN *et al.*, 2004), known as maximum value distance, instead of computing accessibility. We considered the distance between the node position and the geographical center of the network. The obtained results for the dynamics simulation are presented in Figure 55. We did not present the graph obtained for different values of η because no significant variations among $\langle TE_{1000} \rangle$ were found. In case of jump probability, γ , no significant changes were found when we compare a fixed value γ in different regions (shown in Figure 55(a)). A similar behaviour was found for a range of η values, when we adopt $\tau \leq 0.01$ (see Figure 55(b)). However, for the remaining values, lower distances to the center of the network provided better results.

Figure 55 – $\langle TE_{1000} \rangle$ performances and the respective standard deviations computed from the dynamics execution for each network region, defined by Chebyshev distance.

Source: Elaborated by the author.

We also analyze the regions computed for accessibility measurement, $\alpha^{(3)}$. The variations of η vectors have distinctly resulted in $\langle TE_{1000} \rangle$ for each network region and each network. The comparison among the executions of the dynamics in different networks is shown in Figures 56. In the majority of the cases, variations of η vectors did not result in different performance variations for any of the network regions. This characteristic was found for the networks that represent a knowledge structure (CN, WOS, WIKI, and BA), see Figures 56(c), (d), (e), and (f). Furthermore, when comparing among regions in the same network, for all cases, the most central regions tend to have higher values of $\langle TE_{1000} \rangle$. In contrast with distance, in the case of accessibility, the higher the measurement value, the more central the region is. In summary, in the case of knowledge structure, the number of visible agents, η , did not significantly impact on $\langle TE_{1000} \rangle$. However, BA network is an exception (as shown in Figures 56(f)) because there are no significant variations of $\langle TE_{1000} \rangle$ measured from distinct network regions. We did not show results of τ because it is similar to those obtained for η .

In case of WAX, no significant differences of $\langle TE_{1000} \rangle$ were found for different combinations of η for regions with higher values of $\alpha^{(3)}$. However, the differences between the curves and standard deviations of $\langle TE_{1000} \rangle$ were found to be higher in the network borders, as shown in Figure 56(b). The variations of the percentage of influent agents, η , in simulation executed on WS networks were found to be directly related to $\langle TE_{1000} \rangle$, see Figure 56(a).

Figure 57 shows the dynamics performance $\langle TE_{1000} \rangle$ involving variations of γ . For the majority of the cases, the more central a region is, the higher $\langle TE_{1000} \rangle$ was found to be, as shown in Figure 57. The two exceptions were WOS and BA networks, as can be seen in Figure 57(d) and (f). In the case of WOS network, for a particular range of accessibility values (around 1700), the performance decreased significantly, which is different from

Figure 56 – Performances, $\langle TE_{1000} \rangle$, computed for network regions for distinct values of η . These regions are defined according to accessibility measurement with $h = 3$.

Source: Elaborated by the author.

all other considered models. The performances measured for BA network also presented distinctive results. Similarly to the performance measured for η parameter, simulations on BA network by considering variations of γ does not result in $\langle TE_{1000} \rangle$ values that vary substantially along the network regions. Additionally, in contrast with the results of all models, the performance measured for regions of BA network that presents low accessibility were found to result in the highest $\langle TE_{1000} \rangle$ values.

Furthermore, the standard deviations of TE_{1000} measured for simulations executed on knowledge networks present lower values when compared to the results obtained for spatial networks. According to these results, the unexpected results obtained for WOS networks were possibly not caused by statistical fluctuations, but a consequence of the network structure. In a general view, the results measured from the knowledge structures (WOS, WIKI, CN, and BA) provided less variation than the remaining networks (LA, WS-2, and WAX).

4.3.3 Conclusions

Studies regarding the learning process have been proposed in literature (KOPO-NEN; NOUSIAINEN, 2014; PARESCHI; TOSCANI, 2014; BATISTA; COSTA, 2010). In this section, we introduced a novel approach that simulates researchers in the process of knowledge acquisition. Here, knowledge was represented by complex networks and the process of discovery was performed by a multi-agent dynamics, in which some researchers attract themselves. We also addressed the characteristics of memory, which is represented by true self-avoiding dynamics. These dynamics were executed on some network models and two real networks (WIKI and WOS), and we found that the resultant performances can be distinctly maximized for different characteristics of the networks. For the geographically related networks, by increasing the jump probability (γ), the speed of learning decreases. However, for the remaining tested models, the opposite behavior is found. This effect can be due to the presence of hubs, which can let the agent trapped because the hub is crossed for many paths that isolate groups of nodes.

In order to better understand the performance of the dynamics in groups of nodes, we compared the performances obtained for different network regions (e.g., border and center). For the majority of cases, higher performances were found for the core than for the borders. In the case of networks with a structure related to knowledge, by varying all the dynamics parameters, the behavior of performance curves did not change substantially. One exception was found for jump probability (γ) when the dynamics are executed on WOS network. These results indicate that the resultant performance is more dependent on the network region than on the parameter variations.

Figure 57 – Performances, $\langle TE_{1000} \rangle$, computed for network regions for distinct values of γ . These regions are defined according to accessibility measurement with $h = 3$.

Source: Elaborated by the author.

4.4 Intelligent complex networks

Here, we propose and simulate a machine with the ability to solve problems. This machine is henceforth called *reasoning engine* (*RE*), which is represented by a network, N_{RE} . As a case study, the task that *RE* develops is to mount a puzzle, which is also represented by a network N_P . This process is illustrated in Figure 58. This proposed framework is not only related to assembling puzzles but a generic way to deal with problems obtained from the real world. Observe that this dynamics represents a lower scale than the study developed in the previous section, in which we simulated knowledge acquisition by considering collective discovery. The methods and results presented in this section are published in (ARRUDA; COMIN; COSTA, 2018b).

Figure 58 – Scheme of how the dynamic works. The first part represents the set of nodelets, which are disconnected, and the second part represents the network. The connections are created only when there is a correct fit between a pair of nodelets. Additionally, the arrow represents *RE*.

Source: Elaborated by the author.

4.4.1 Dynamics definitions

The proposed approach aims to mount a puzzle (N_P network) from a set of disconnected parts. Each of the parts is represented by a N_P node and its respective tags of

connections. These structures are henceforth called *nodelets*. So, there is a single possibility of association for each edge (see the first part of Figure 58). The reasoning engine N_{RE} , also represented by a network, receives the nodelets and assembles the puzzle. More specifically, each N_{RE} node stores a set of N_P nodelets. For each iteration of the dynamics, N_{RE} nodes send copies of their N_P nodelets to their respective neighbors and connect N_P nodelets by considering the respective interconnectivity tags. This process is repeated for a given number of iterations. Observe that all N_{RE} nodes try to mount the entire N_P .

Two characteristics were added to this model. The first consists of different strategies of initially distributing the nodelets among the N_{RE} nodes. By considering a given centrality measurement X of N_{RE} and for a N_P network with n nodelets, X is sorted, and the nodes that represent the first n positions receive a N_P nodelet, which is randomly selected. A N_{RE} node can receive a single nodelet, and all nodelets are distributed on N_{RE} . For the assigned priority, we considered the following measurements:

- *Random*: The nodelets are assigned to random positions, which is the most straightforward manner;
- *k*: The positions are assigned according to degree measurement. In this case, we considered the positions of higher and lower positions, henceforth called *k*-lower and *k*-higher, respectively;
- *EC*: The positions are assigned according to eigenvector centrality (*EC*) measurement. We also considered the positions of higher and lower positions, henceforth called *EC*-lower and *EC*-higher, respectively.

The second characteristic consists of defining the memory size of N_{RE} nodes. In other words, for each N_{RE} node a buffer is defined, which can store a given number of N_P connected components. By considering this characteristic, two case studies were defined: (i) Infinite buffer size and (ii) limited buffer size.

4.4.2 Results

In this subsection, we present both case studies. In the first, with infinite buffer size, we tested three possibilities of networks as N_{RE} and N_P , which are ER, BA, and WS. As an additional experiment, we employed a variation of WAX network to better understand the relationship between characteristics of shortest path lengths and the process of mounting puzzles. In the case of limited buffer size, by considering the same set of network models (ER, BA, and WS), we tested two possibilities: (i) Fixed buffer size and (ii) Buffer size that varies according to node degree.

4.4.2.1 Case 1: Infinite Buffer Size

As mentioned before, this test consists of executing our dynamics by considering a buffer with infinite size. In the first tests, we considered N_{RE} and N_P networks as ER, BA, and WS, with the configurations presented in Table 5. Also, N_{RE} and N_P networks were created with approximately 10000 and 1000 nodes, respectively. The simulations were executed 200 times, and for each execution, a new network was created by considering the fixed parameters. Observe that, for both networks, we considered only the largest connected component. Because the maximum possible number of iterations to mount the entire puzzle is the network diameter (the largest shortest path between each pair of nodes), we normalize the time to mount the puzzle by the N_{RE} diameter, henceforth called normalized time. So it is possible to compare the results among different versions of N_{RE} .

Chart 5 – Configurations of N_P and N_{RE} networks employed to Case 1.

Network model	N_{RE}	N_P
WS	$p = 0.05$	$p = 0.02$
ER	$p = 0.0004$	$p = 0.004$
BA	$m = 2$	$m = 2$

In order to compare among the versions of N_{RE} and N_P , and the different strategies of assigned priority, for each combination we plot the fraction of mounted N_P against normalized time. These results are shown in Figure 59. By comparing among the different versions of N_{RE} networks, the version that assembles the puzzles faster is BA network, followed by the WS and ER. However, in this test, the choice of N_P was not decisive to interfere in the resultant performances. Another variable that is essential to better understand these dynamics is the choice of assigned priority. The assignment priorities that sooner increased the fraction of mounted puzzles are given, in increasing order, by higher values of EC , higher values of k , random nodes, lower values of k , and lower values of EC . Interestingly, the same order was found for all employed tests. Additionally, in the majority of the cases, the results measured between the higher values of EC and k , and between the lower values of EC and k were found to be similar, except when we employed $N_{RE} = BA$. In this case, when assignment priority was chosen as random or lower values of k , the resultant curves were found to be similar.

An important characteristic that can be observed from this dynamics is the relationship between time to mount puzzles and the network diameter. For that, we proposed here a variation of Waxman model (WAX), in which we can fix the node degree and control the network diameter. More information on the original version of WAX is shown in Section 2.3.5. By considering that the nodes are randomly spread in a 2D space, WAX consists in giving the probability of connection between nodes according to an exponential

Figure 59 – Fraction of assembled N_P nodes as a function of normalized time. Here we plot all combinations between N_P and N_{RE} . Observe that for each type of N_P a different behavior was found, but N_{RE} did not change the results significantly.

Source: Elaborated by the author.

decrease of geographical distances, as follows

$$p = \alpha \exp(-d/\beta), \quad (4.6)$$

where d is the geographical distance and α , and β are constants. Here we fixed $\alpha = 1$ and varied β (0.01, 0.02, and 0.03) to generate networks with different shortest path lengths, and consequently different diameters. Observe that the shortest path lengths decrease if β increases. In order to create networks with a fixed average node degree, the network edges are randomly created until the desired average degree is reached. The average employed degree was fixed as 4, and the networks have approximately 10000 nodes.

Here we present the results concerning the time to mount the first puzzle (T_F), as shown in Figure 60, in which we show results of N_P as ER network because T_F depends only on the number of N_P nodelets. For all versions of N_{RE} , the best results were found

when the assignment priority is proportional to EC . By comparing β values, we found that as β increases, the time to mount the puzzles becomes lower.

Figure 60 – By considering $N_P = ER$, and N_{RE} networks as a variation of the WAX model, we plot the average and standard deviation of the time to mount an entire puzzle (T_F). Observe that the lowest values of T_F were found when the assigned priorities of N_P nodelets on N_{RE} were considered as the higher values of EC .

Source: Elaborated by the author.

4.4.2.2 Case 2: Limited Buffer Size

In the following tests, we considered another characteristic that is present in real machines, the limited buffer size, as presented in Section 4.4.1. Furthermore, we include a feature based on humans memory, which can better store concepts that are connected. Here we considered a range of different buffer sizes (from 1 to 10). We employ the same buffer size for all nodes of a given N_{RE} network. We employed N_{RE} and N_P networks with approximately 1000 and 289 nodes, respectively. This modification was used since this variation of the dynamics demands a higher computational cost than the first tests. As in the first tests, we employed the same possibilities of assigned priority, and a new version of each network was generated for each execution, which was 200 times. The network configurations are shown in Table 6.

Chart 6 – Configurations of N_P and N_{RE} networks employed to Case 2.

Network model	N_{RE}	N_P
WS	$p = 0.45$	$p = 0.1$
ER	$p = 0.004$	$p = 0.0135$
BA	$m = 2$	$m = 2$

As well as in the first test, we analyze the impact of buffer size on the fraction of executions in which puzzles were completed assembled. Figure 61 shows the comparison of the assigned priorities concerning the fraction of dynamics executions that were able to assemble the puzzle entirely. In this figure, we presented only BA as N_{RE} and N_P networks

because the results were found to be qualitatively similar for the other cases. In summary, for all cases, as the buffer size increases, the fraction of N_{RE} that entirely mounted N_P also increases. Despite the similarities between the configurations, some differences were found. For example, when we employ $N_{RE} = BA$, the highest fractions of completed puzzles were observed for all choices of N_P . However, in the case of other N_{RE} possibilities, the choice of N_P influences the fraction of completed puzzles. When N_P is set as BA, this number was found to be higher for lower values of buffer size. The different assigned priorities changed the behavior of the curves, as shown in Figure 61. By considering $N_{RE} = BA$, the curves were found to be separated, except when comparing the k -lower and random strategies. For the remaining cases, similar results were observed when the dynamics were performed for the assigned priorities of EC -lower, k -lower, and random.

Figure 61 – Average and standard deviation of the fraction of completed puzzles against buffer size. In this case, we considered $N_{RE} = BA$ and $N_P = BA$. Due to the high standard deviations, here we represent only 20% of this measurement. For some values of buffer size, the assigned priority of k -lower is more effective than the random choice.

Source: Elaborated by the author.

The puzzle (N_P) is seldom completely assembled when we employ buffer sizes of 2 and 3. By considering buffer size of three positions, the networks were assembled in more than 84% of the cases when assigned priority was set as Ec -higher and N_{RE} was BA. Also considering $N_{RE} = BA$, by employing buffer sizes of 3 and assigned priority set as Ec -higher, the networks were assembled in more than 84% of the cases. In case of assigned priority as k -higher and N_P as BA, the percentages of entirely assembled N_P are 43%, 52%, and 19% for N_{RE} as WS, ER, and BA, respectively. A more detailed analysis was done by considering the buffer size of 4 components. These results are shown in Table 8. The higher percentages of completely assembled puzzles are found for k -higher and Ec -higher, which was found for the remaining considered buffer sizes. Additionally, the better option of N_{RE} is BA network, and the puzzle that was found to be easily assembled is also BA network.

As another analysis, we considered different sizes of the buffer of each node in the

Table 8 – Fractions of N_P mounted entirely, in which the dynamics was executed, by considering the buffer size of 4. Observe that rows and columns represent N_{RE} and N_P , respectively.

	Random			k -lower			k -higher			Ec -lower			Ec -higher		
	WS	ER	BA	WS	ER	BA	WS	ER	BA	WS	ER	BA	WS	ER	BA
WS	0.00	0.00	0.32	0.01	0.00	0.42	0.40	0.11	0.95	0.00	0.00	0.26	0.16	0.08	0.76
ER	0.00	0.00	0.23	0.00	0.00	0.29	0.51	0.23	0.94	0.00	0.01	0.21	0.24	0.13	0.72
BA	0.04	0.03	0.14	0.03	0.04	0.10	0.33	0.42	0.63	0.20	0.23	0.39	0.98	0.93	0.97

same N_{RE} . We employ the buffer size following the node degrees, see the results in Table 9. Because the average degree of the employed N_{RE} is approximately four, the average buffer size of the nodes is also four. When we compared the results obtained from degree-based buffer sizes with the results shown in Table 8, for all combinations of N_P and N_{RE} networks, the respective fractions of entirely mounted puzzles were found to be highest. In the case of N_{RE} following BA model, this strategy provided almost all N_P wholly assembled.

Table 9 – Fractions of N_P mounted entirely, in which the dynamics were executed, by considering the buffer size equal to the degree of the nodes in N_{RE} . The rows and columns represent N_{RE} and N_P , respectively. Observe that the average node degree is four, which means that the average buffer sizes are the same as the results presented in Table 8.

	Random			k -lower			k -higher			Ec -lower			Ec -higher		
	WS	ER	BA	WS	ER	BA	WS	ER	BA	WS	ER	BA	WS	ER	BA
WS	0.10	0.03	0.34	0.01	0.02	0.23	0.81	0.62	0.99	0.01	0.04	0.12	0.83	0.58	0.99
ER	0.04	0.07	0.21	0.04	0.01	0.20	0.94	0.81	1.00	0.03	0.03	0.07	0.93	0.86	1.00
BA	0.98	0.98	0.99	1.00	0.97	0.99	0.99	0.98	1.00	0.99	1.00	0.99	1.00	1.00	1.00

In the following, we studied another characteristic of the proposed dynamics that consists in the time to mount a puzzle, N_P . Since in many cases the puzzle is not entirely assembled, we measured the time to mount 90% of N_P . Furthermore, here we compared both strategies: (i) all nodes with the same buffer (see Figure 62) and (ii) buffer size varying according to the N_{RE} node degree (as shown in Figure 63). Observe that, as well as in the previous tests, here the second strategy always provided faster results than the first strategy (fixed buffer size). In a similar behavior to Case 1 (infinite buffer size), the increasing order of the most efficient N_{RE} is: BA, ER, and WS. The normalized difference between the time to mount 90% of N_P presented in Figures 62 and 63 are shown in Figure 64. This normalization consists of dividing the subtraction by times indicated in Figure 62, which consists of the higher values. In the majority of the cases, the largest differences are obtained when $N_{RE} = BA$ network. One exception was found for assigned priority as k -higher, and $N_P = BA$, which means that the time difference also depends on N_P .

4.4.3 Conclusions

The dynamics proposed here presents a possibility to use complex networks to create a system that processes artificial intelligent tasks. Here, we employ a study case to test our approach in the task of assembling a puzzle, which can also be represented

Figure 62 – Comparisons of times to mount at least 90% of the nodelets of one N_P among the assigned priorities. For all nodes, we considered buffer size of four positions. The colors of the plots represent the different versions of N_{RE} .

Source: Elaborated by the author.

Figure 63 – Comparisons of times to mount at least 90% of the nodelets of one N_P among the assigned priorities. For all nodes, we considered buffer size equal to the respective node degree. Consequently, the average buffer size is four. The colors of the plots represent the different versions of N_{RE} .

Source: Elaborated by the author.

Figure 64 – Normalized time computed between the time to mount at least 90% of the nodelets of N_P . Observe that the colors represent the different versions of N_{RE} .

(a) Random.

(b) k -lower.

(c) k -higher.

(d) EC -lower.

(e) EC -higher.

Source: Elaborated by the author.

as networks. An important characteristic that was incorporated into our dynamics is the memory of the nodes, also called buffer. Some possibilities were tested into cases, consisting in infinity buffer size, and limited buffer size. For the second case, we test a range of buffer with a fixed size. In other words, all the network had the same buffer size. Furthermore, we considered buffers with varied sizes, which were defined according to the node degrees. In order to measure the performance of our approach, we considered the characteristics of the time to assemble puzzles. For the tests that consider the infinity buffer size, BA network is found to provide the shortest processing time for all of the considered puzzles. Another characteristic that is decisive to the speed of the solution is the adopted assigned priority, which is the initial distribution of the puzzle pieces. For all cases and all tests, Eigenvector Centrality is found to be the most effective assigned priority. Another interesting finding is that different buffer sizes, having the size proportional to the network node degrees, are found to improve the puzzle assembly performance when the results are compared to the approach that employs the same buffer equal to the average degree.

4.5 Connecting network science and information theory

In this section, we study a framework that simulates network transmission. One previous study that considered information science and complex networks is proposed by [Rosvall and Bergstrom \(2008\)](#), which uses information theory to detect communities in networks. In this thesis, a temporal series is created from a given network, which consists of a given type of random walk that acquires symbols representing the visited nodes. These symbols are then converted to a temporal series that can be transmitted. For the transmission channel, we considered the original time series and its respective compacted version. Observe that the efficiency of this framework can be improved when the transmission channel considers compacted data. By considering this data, a new network is created. In summary, we present an integrated framework, which consists of a dynamics that acquires information from a network, a communication channel transmits the information, and the data used to create a new network. More information regarding compaction and information theory are presented in [Appendix F](#). The methodologies and results presented here are published in ([ARRUDA *et al.*, 2019](#)).

4.5.1 Problem definition

The proposed procedure is illustrated in [Figure 65](#). With basis on a given dynamics, we propose a framework that creates a time series from a given complex network. We employed some variations of the random walk dynamics and the respective one-dimensional time series (more information on the employed random walks are presented in [Section 4.5.2](#)). Each time series is obtained by an agent that walks on the

network and, for each visited node, a symbol is assigned, see Figure 65(a). This time series can be considered as a one-dimensional projection of the original network and when transmitted, can be employed to reconstruct a model of the original network, as shown in Figure 65(b). The pair of consecutive symbols of the time series is used to create the edges of the reconstructed network. Another feature proposed here is the compression of the time series, which can implicate in a more efficient transmission. For that, we employ the Huffman coding (CORMEN *et al.*, 2001), which is described in Appendix F. Observe that we set the Huffman dictionary being proportional to the node degrees of the network. So, the information concerning the network topology was used to compress the generated time series.

Figure 65 – Scheme representing the proposed framework. Items (a), (b), and (c) represent the original network, the obtained time series, and the reconstructed network, respectively.

Source: Elaborated by the author.

4.5.2 Random walk dynamics

In this study, we employed three different types of random walks. The first model employed is the standard Random Walk (RW), in which the agent goes from a node to its neighbors by considering the same probability for all of them (more details are presented in Section 2.5). The first variations consist of two types of biased random walks, the transition probabilities between a node i and its neighbors being proportional to the neighbor degrees (RWD) (BONAVENTURA; NICOSIA; LATORA, 2014; COMIN, 2016). The biased version considers the probabilities as the inverse of the node degrees (RWID) (BONAVENTURA; NICOSIA; LATORA, 2014). Because the probability p_{ij} for the agent to go from a given node i to j depends on the degree of the i th neighbors, the transition probability can be calculated as

$$p_{ij} = \frac{k_j^\alpha}{\sum_{l \in \Gamma_i} k_l^\alpha}, \quad (4.7)$$

where Γ_i is the set of neighbors of node i . For $\alpha = 1$ RWD dynamics is described, for $\alpha = -1$ we have the RWID case, and for $\alpha = 0$ the probabilities of RW dynamics are obtained. It is known that when RW is executed on a undirected network, the frequency of visits of the nodes are directly proportional to their respective node degrees in the steady state (COSTA *et al.*, 2007b).

As another possibility, we employ a variation of the true self-avoiding random walk (TSAW) (KIM; PARK; YOOK, 2016). These dynamics were better described in Section 4.3.1. The variation considered here is that the TSAW memory is associated with the network edges, in contrast with the methodology presented in Section 4.3.1. So, the probabilities of visiting the already visited edges decrease. As in the previous case, we fixed the constant Ω in 0.5. Observe that this variation was considered because to have a perfect representation of the original network, it is necessary to have all the connected nodes as pairs in the obtained time series. In other words, the walker shall cross all network edges.

4.5.3 Analysis of network compression

The simplest way to quantify the performance of the time series is to measure the time series length according to a given percentage of reconstruction. Here we considered 90% of the original network (measured concerning the number of edges). We referred this time as T_{90} . In order to measure the performance regarding compressed time series, we considered the size of the compressed time series, which is henceforth called T_{90}^C . To be possible to compare both representations, original and compacted time series, we considered that the symbols of the original time series have the size of 10 bits. This value

was chosen because it is the minimum size that can store all symbols of the employed networks.

We also considered a measurement that accounts for the long term transmission ratio, defined as

$$R_L = \frac{T_L}{T_L^C}, \quad (4.8)$$

where T_L is a sufficiently large time and T_L^C is the compressed time series size, in which T_L . Here we defined T_L as a million of iterations having symbols with 10 bits. The same equation was employed to define R_{90} . In this case, we consider the ratio between T_{90}^C and T_{90} .

Another measurement consists of analyzing the relationship between the topological and dynamical measurements. It is computed as the Pearson coefficient between the frequencies of visits and the network node degrees, and this measurement is henceforth called *steering coefficient* S . Additionally, S is calculated for both versions of the measured time series, S_{90} and S_L , which represent S for a time series that provides 90% of network reconstruction and a substantial time of exploration, respectively. These measurements are intrinsically related to T_{90} and T_L .

4.5.4 Results and discussion

Here, we present the results regarding the framework proposed in this section. We considered four approaches of random walk (introduced in Section 4.5.2), and six different complex networks, BA, ER, WS1 (a WS network with $p = 0.01$), WS2 (a WS network with $p = 0.005$), Wax, and Geo. For this analysis, we considered 30 simulations for each combination of model and dynamics. We also compare simulations concerning undirected and directed cases. For the undirected case, the number of network nodes was set as 1000 with average degree approximately 8.

The results concerning the dynamics executed in undirected networks are presented as parallel coordinates plots, as shown in Figure 66. The lowest values of T_{90} and T_L were found when we employ TSAW dynamics. By comparing the topologies, the results for the BA network were different from the others, mainly for T_{90} . RWD dynamics provided a much higher value of T_{90} (see Figure 66(a)). These phenomena probably happen since the moving agent tends to alternate between hubs. As a consequence, the nodes with small degrees tend to be avoided. A similar result was found when T_{90}^C was measured, but in this case, the values of T_{90}^C for the RWD dynamics in the BA model are not as distinguished as those obtained for the first measurement (as shown in Figure 66(b)). Interestingly, the results obtained for both versions of the geographic-related networks (WAX and GEO) tend to differ significantly.

Figure 66 – Results regarding all the employed network models and varying the random walk dynamics for undirected networks. The parallel coordinates present the average and standard deviation. Items (a), (b), (c), (d), (e), and (f) represent T_{90} , T_{90}^C , S_{90} , S_L , R_{90} , and R_L .

Source: Elaborated by the author.

By analyzing the steering coefficient, similar values of S_{90} were obtained for the RW, RWD and TSAW dynamics, as shown in Figure 66(c). In the case of RWID dynamics, we found that the network degrees are not a good predictor of the frequency of visits. In case of WS1 and WS2 networks, the dynamics executions lead typically to the lowest values of S_{90} , which is probably related to the fact that in WS networks the majority of the nodes have the same degree. Consequently, this characteristic leads to a similar frequency of visits. Furthermore, BA network resulted in the most significant values of S_{90} . This effect can be a consequence of the power-law degree distribution of the BA model, which can intensify the number of visits in the nodes with higher degrees. Observe that, in the majority of the cases, degree is a useful method to predict the symbol probabilities for using in Huffman compaction. The other version that considers the long-term steering coefficient values (S_L) is shown in Figure 66(d). Observe that TSAW dynamics led to the highest values of S_L , followed by RWD. However, different behaviors were found for RWID for S_L , and smaller values than those obtained for the other dynamics were found.

The other analyzed characteristic is the compression ration. The results for R_{90} and R_L are shown in Figures 66(e) and (f), respectively. Similar results are obtained for almost all network models, except for BA network. Interestingly, GEO and WAX networks provided divergent results for T_{90} and T_{90}^C , but implied similar values of R_{90} and R_L .

Regarding the directed networks, the directionality of the networks was defined according to the reciprocity of the edges (GARLASCHELLI; LOFFREDO, 2004). This measurement is defined for each node as

$$r = \frac{Ne^{\leftrightarrow}}{Ne}, \quad (4.9)$$

where $Ne^{\leftrightarrow}$ is the number of edges that have both directions and Ne is the total number of edges. For the directed networks, we set the average values of $Ne^{\leftrightarrow}$, which preserves the original network size and includes strong directionality. In order to attribute directionality, for each original edge, a random number n is chosen according to a uniform probability, being in the interval between zero and one. If $n \leq Ne^{\leftrightarrow}$, the original edge is converted into two edges that have both directions (in and out), otherwise, a single edge is created, and its direction is attributed with equal probability. After that, we considered only the strongly connected component. So, there are no possibilities for the agent to become trapped. Observe that the KN networks did not follow this methodology because KN implies in an intrinsically directed network.

For almost all networks and dynamics, the results of directed and undirected versions were similar, see Figures 66 and 67. Here we included the KN model, that leads to the lowest reconstruction times (T_{90}) for almost all cases. The lowest necessary time is possibly related to the fact that the KN network construction can be understood as a kind of random walk (see Figure 67). This network has some key aspects: (i) k_{in} and k_{out} tend to be equal for every node and (ii) the reciprocity is smaller than the remaining models. By considering T_{90}^C , the lowest values were found when we employed the BA network, as shown in Figure 67(b). While S_{90} obtained for the simulations on directed networks were found to be similar to those obtained for undirected versions, S_L resulted in smaller values for directed than for the undirected cases. See Figures 67(c) and (d). The exceptions of this trend were found to be BA and KN networks.

4.5.5 Use case: language compression

We apply our methodology to real-world networks that are related to language. This comparison consists in computing the compression ratio (R_L) of real texts by using dictionaries obtained from books and network degrees. First, the texts were pre-processed (see Section 2.6.1). We removed the stop-words and punctuation symbols, and lemmatized the remaining words. To be possible to compare the results among different networks, only the words that take part in all networks were considered. The 100 considered books are organized into 20 authors, as follows:

1. **Arthur Conan Doyle:** “Uncle Bernac - A Memory of the Empire”, “The Tragedy of the Korosko”, “The Valley of Fear”, “The War in South Africa”, and “Through the Magic Door”;

Figure 67 – Results of all the employed network models and varying the random walk dynamics for directed networks. These results are presented as parallel coordinates and present the averages and standard deviations of the measurements. Items (a), (b), (c), (d), (e), and (f) represent T_{90} , T_{90}^C , S_{90} , S_L , R_{90} , and R_L .

Source: Elaborated by the author.

2. **Bram Stoker:** Dracula's Guest, The Jewel Of Seven Stars, The Lady of the Shroud, Lair of the White Worm, and The Man;
3. **Charles Darwin:** Coral Reefs, The Expression of the Emotions in Man and Animals, Geological Observations on South America, The Different Forms of Flowers on Plants of the Same Species, and Volcanic Islands;
4. **Charles Dickens:** American Notes, A Tale of Two Cities, Hard Times, Great Expectations, and Barnaby Rudge: A Tale of the Riots of Eighty;
5. **Thomas Hardy:** A Changed Man; and Other Tales, A Pair of Blue Eyes, Far from the Madding Crowd, Jude the Obscure, and The Hand of Ethelberta;
6. **Pelham Grenville Wodehouse:** My Man Jeeves, Tales of St. Austin's, The Adventures of Sally, The Clicking of Cuthbert, and Livro 5: The Man with Two Left Feet;
7. **Alan Poe:** The Works of Edgar Allan Poe (Volumes 1 to 5);
8. **Hector Hugh Munro (Saki):** Beasts and Super Beasts, The Chronicles of Clovis, The Toys of Peace, When William Came, and The Unbearable Bassington;

9. **Herman Melville:** *Moby Dick*, *The Piazza Tales*, *The Confidence-Man: His Masquerade*, *Typee: A Romance of the South Seas*, and *White Jacket; Or, The World on a Man-of-War*;
10. **Zane Grey:** *The Lone Star Ranger: A Romance of the Border*, *Riders of the Purple Sage*, *The Mysterious Rider*, *To the Last Man*, and *The Call of the Canyon*;
11. **Andrew Lang:** *The Blue Fairy Book*, *The Arabian Nights Entertainments*, *The Yellow Fairy Book*, *The Violet Fairy Book*, and *The Pink Fairy Book*;
12. **Richard Harding Davis:** *Real Soldiers of Fortune*, *Notes of a War Correspondent*, *Cinderella*, and *Other Stories*, *The Congo and Coasts of Africa*, and *Soldiers of Fortune*;
13. **Henry James:** *The Turn of the Screw*, *The Portrait of a Lady — Volume 1*, *The Ambassadors*, *The Real Thing and Other Tales*, and *The American*;
14. **B. M. Bower:** *Cabin Fever*, *The Trail of the White Mule*, *Lonesome Land*, *The Lookout Man*, and *The Long Shadow*;
15. **Washington Irving:** *The Sketch-Book of Geoffrey Crayon*, *Knickerbocker's History of New York*, *Complete*, *Chronicle of the Conquest of Granada*, from the mss. of Fray Antonio Agapida, *The Alhambra*, and *Tales of a Traveller*.
16. **Herbert George Wells:** *The War of the Worlds*, *A Short History of the World*, *The First Men in the Moon*, *Tales of Space and Time*, and *The World Set Free*;
17. **Horatio Alger:** *Ragged Dick*, *Or, Street Life in New York with the Boot-Blacks*, *Adrift in New York: Tom and Florence Braving the World*, *Brave and Bold; Or, The Fortunes of Robert Rushton*, *Fame and Fortune; or, The Progress of Richard Hunter*, and *The Errand Boy; Or, How Phil Brent Won Success*;
18. **Jane Austen:** *Pride and Prejudice*, *Emma*, *Sense and Sensibility*, *Persuasion*, and *Mansfield Park*;
19. **Mark Twain:** *Adventures of Huckleberry Finn*, *The Adventures of Tom Sawyer*, *The Prince, and the Pauper*, *A Connecticut Yankee in King Arthur's Court*, and *Roughing It*;
20. **Nathaniel Hawthorne:** *The Scarlet Letter*, "Mosses from an Old Manse, and Other Stories", "The House of the Seven Gables, Twice Told Tales, and The Blithedale Romance.

In order to measure statistics from the books, we considered 100 titles, and for each title, b a set of $1 \leq N \leq 99$ other books are randomly chosen to create the Huffman code.

The words that represent nodes in the networks were also lemmatized. The employed words networks were: the Edinburgh Associative Thesaurus (EAT) (KISS *et al.*, 1973), WordNet (MILLER, 1995), and a co-occurrence network (AMANCIO; OLIVEIRA JR; COSTA, 2012b) created from the 100 books considered. The Huffman dictionary was also created by considering these three networks, and the average and standard deviations were computed by considering all books. In this case, the books were considered being the temporal series.

In order to calculate R_L , the words were considered representing 14-bit symbols. The R_L results are shown in Figure 68. By comparing the results obtained from the network topologies, the highest values of R_L were obtained when the co-occurrence network was employed, followed by EAT. The lowest performance was found when WordNet was employed, where $R_L < 1$. In other words, the book lengths were found to be higher than the original sizes. Furthermore, the compression ratio obtained when we employed statistics of real books as Huffman dictionary was found to be slightly better than the previous network approach when 40 or more books were considered.

Figure 68 – Average compression ratios (R_L) computed by employing word statistics of real books and the average R_L from network topology (degree) as Huffman dictionary. In the case of network topologies, the dashed lines represent the average value of R_L .

Source: Elaborated by the author.

4.5.6 Conclusions

The areas of complex networks and information theory are typically studied independently. Because we believe that both areas share elements and can be studied in a complementary way, we proposed an integrated framework. An example of real world dynamics that is represented by the proposed framework is the spoken/written language. Our proposed approach consists of some random walk dynamics, which obtains a sequence

of symbols (a time series), and is used to reconstruct the network. This time series is represented in a twofold manner: (i) By considering the original symbols and (ii) by considering the compression of the time series. Observe that this compression takes into account the network topology, more specifically, the node degrees.

Here, we considered some distinct network topologies and four different random walk dynamics. Some notable results were found, such as different network topologies led to distinct performances of transmission and compaction. In the majority of the cases, BA networks oscillated between the best and worst performances, depending on the chosen random walk. In the case of directed networks, KN networks led to better performance for many tests. Furthermore, the two variations of geographic networks, Wax and Geo, led to significant distinct results. By considering the compression ratios, dynamics executed on BA networks provide higher values, except for the RWID dynamics. This result is due to the fact that the power-law degree distribution creates a more efficient Huffman coding because the repetition of symbols is more evident than the remaining models.

CONCLUSIONS AND FUTURE WORKS

In this thesis, we presented a two-fold analysis of complex networks, which are: (i) text mining and (ii) information obtained from dynamics. In the first part, we present methodologies that aim at analyzing texts by considering different scales of analysis. These approaches consist of methodologies that grasp information from words to sets of paragraphs. In the remaining of the thesis, we studied dynamical aspects of knowledge and text representations. Furthermore, these studies were also based on some different scales of analysis. The relationship between the aspects involved are presented in Figure 1.

The first methodology consists of comparing global statistics of the network and local measurements. First of all, we created word networks (co-occurrence) and computed many centrality measurements. The first analysis consists of computing some statistics from these measurements, which were used as features for a classification task. In order to create these networks the text is preprocessed, and the stopwords are removed. In the case of local measurements, we considered two versions of the network with or without stopwords and the features were obtained from specific words of the text. As a case example, we considered a corpus with texts divided into two categories, fact and fiction. The highest accuracies were found for classification by using the local measurements. Additionally, by comparing these measurements, the tests performed with networks that kept stopwords were found to provide the best results. So, the pattern of connectivity of these words keeps essential information of the texts and can be employed for their classification. The results were not better than the traditional methods, but can be employed to other analyses as additional features to classify texts in more complicated tasks (e.g., in tasks that analyzes writing style).

In order to represent texts in a larger dimension, we proposed a network in which the connections are created not only according to the immediate co-occurrence. For that, we proposed three network-based approaches, in which we only considered the nouns and verbs. The first approach consists of an evolution of the co-occurrence networks, where

the words (represented by nodes) are connected to a given neighborhood. The second method is based on the connection between all words in a paragraph. We considered the principle that all words of a given paragraph are semantically related in the document. The latter approach consists of a technique that removes nodes with less statistical significance. Observe that all these network models were developed aiming at highlighting the subjects of the text. From the created networks, a methodology based on community detection was employed, and the communities were associated with subjects. To compare the obtained results, we considered a bag-of-words based technique that used a clustering algorithm. The tests were executed in two datasets, having distinct granularity levels. Remarkably, for all tests, we found that at least one of the networked methods provided better results than the bag-of-words approaches. The connectivity information of the presented approaches can be employed to reveal information of the semantic organization of the texts. Furthermore, these pieces of information can be employed in techniques that depend on the characterization of the textual semantics.

Because the networks above mentioned identified characteristics of subjects in texts, as the next scale of analysis, we propose a network in which the nodes represent paragraphs. Here, the network is weighted according to the semantical similarities between paragraphs (the paragraphs are represented by nodes). In order to keep only the most essential information, we used a threshold and removed the lower weighted edges. By considering this paragraph-based approach, we computed many network measurements and characterized their dependency on the text subject and language. Interestingly, the measurement that better characterized real texts is the modularity, which accounts for the characteristics of communities of the networks. This characterization of the network measurements was employed in a case example, in which we analyze the unknown manuscript called Voynich. First, we show that the modularity of this manuscript is compatible with that of real texts. Furthermore, the features that better describe real-language documents were employed in classification and the text was classified with 100% of accuracy as real text. So, our results indicate that the network obtained from Voynich manuscript is compatible with networks of real texts.

As an evolution of the paragraph-based model, another possibility is to consider chunks of text. Here, we proposed a multi-scale method, in which the scale consists of the number of paragraphs that defines the chunk sizes. So, in this approach, the nodes represent sets of Δ adjacent paragraphs. As well as in the paragraph-based model, the nodes are connected according to the semantical similarity, and the lower valued edges are removed, according to a threshold. The long-range connections of this network represent regions in which the text subject repeats, and the short-range connections depict the temporal evolution. The main characteristic of this methodology is to analyze the text regarding the multi-scale analysis, which is provided by Δ parameter and the threshold used to delete the lower weighted edges. We employ this methodology to classify between

real and shuffled documents. Shuffled texts do not contain a coherent story. As a result, our classifications indicated that our network grasps information from the unfolding of the story. We also employ this methodology to analyze the book “Alice’s Adventures in Wonderland”, and it was possible to trace the parallels between the novel and the respective network. Furthermore, we exemplified the multi-scale characteristics by considering two different thresholds. Consequently, each version grasped information of a given scale of the story.

Apart from the distinct methodologies to analyze the texts, in the second part of this thesis, we considered characteristics of the structure and dynamics of knowledge representation. The first study, that considers the macro scale of the network, comprises the comparison between areas of knowledge and their connectivity. More specifically, we compare the areas of theoretical, applied physics, and a control group. By considering two case studies, the nodes were classified as part of an area according to the subject of the journals. The relationships between the papers were modeled through network representations, in which we considered two versions, directed and undirected networks. We used complex network-based measurements that are related to information concerning paths. From this analysis, we found that both areas, applied and theoretical physics, are equally interrelated. So, both areas are complementary because studies produced in one area are directly or indirectly used by the other.

At a lower scale of analysis, we considered the dynamics of knowledge acquisition, which was based on the true self-avoiding random walk dynamics. In this dynamics, the agents represent researchers, and the network nodes are pieces of knowledge that can be discovered. The concept of stochastic flight, also known as jump, was applied to simulate the situation in which a researcher changes his/her research area. For this purpose, we define an attraction field, which is more effective to the closer the researchers. Observe that this is a multi-agent dynamics, and the field of attraction is used to simulate the agent interactions. We evaluate some network topologies that comprise real networks and models, and varied the parameters of the dynamics, representing different strategies. Through the dynamics execution, some interesting characteristics were found. Remarkably, for different topologies, the strategy that provides the best result varies. For instance, by considering the jump probability, the highest values optimized the results in BA network, but the lowest values improved the discovery in WAX, which is a geographical network. Moreover, we analyzed the performance of the dynamics in distinct regions of the network. In general, the discovery performance is better in the core than in the border of the network. However, the dynamics did not follow this trend for BA and WOS for some values of jump probability.

The previously presented dynamics accounted for the relationship among researchers in the task of discovering new knowledge. In the following, we simulated the learning task

for a reasoning engine, which is a lower-scale analysis because it simulates a single agent learning. Here, this engine includes a network whose nodes communicate to their neighbors and transmit some copies of their information. We exemplify the dynamics execution by considering the task of mounting a puzzle, which is also modeled as a network. When the dynamics starts, the parts of the puzzle are stored by the reasoning engine network, and some strategies of initial position were considered, called assigned priority. All in all, the best results were found for assigned priority as the higher values of the eigenvector centrality. We also considered these dynamics by analyzing the effect of limited memory sizes, called buffer. We investigated the performances of the dynamics by considering a fixed size buffer for all nodes of the reasoning engine network. The number of well-succeeded puzzles assembles increases according to higher buffer sizes. Furthermore, we considered buffers with varied sizes, by following the network node degrees of the reasoning engine. This strategy was found to be much more efficient than the fixed buffer size. In this comparison we considered the same average degree for both tests.

The lowest-scale analysis of these dynamics examines transmission of networks. More specifically, we considered a dynamics that obtain information about the network that shall be transmitted. This information is stored as a sequence of organized symbols (time series). To improve the efficiency of the transmission, it is possible to compact the obtained time series. For that, we employed the method called Huffman, which is based on information theory. Observe that the time series transmission, by considering the compression, creates a connection between the areas of complex networks and information theory. The dynamics employed to obtain time series consists of three versions of random walks, which are: standard version, biased according to the degree, biased according to the inverse of degree, and true self-avoiding random walks. Furthermore, we tested these dynamics in distinct network topologies to understand the differences in compression. The majority of the best and worst performances are found when the dynamics are executed on BA network. By analyzing the compression, BA also plays an important role. The highest values of compression ratio were observed for all dynamics, except for random walk biased according to the inverse of degree, which is due to the power-law degree distribution of BA network. For the majority of the cases, the best performance among directed networks was reached by KN. Additionally, we found significant variations between the geographic models.

There are many possibilities for future work. First, by considering all of the proposed network models that represent documents, features of more than one model can be combined to obtain better results in classification. In the case of networks used to detect subjects, in order to improve the outcomes, more information can be added to this representation. The mesoscopic networks can also be created by using sentences, or groups of words to represent small texts and capture a lower level of information.

The comparison between theoretical and applied physics can be extended and used to compare many other areas of study. For example, the contrast between physics and math, or chemistry. Furthermore, bigger datasets with more articles related to physics can be used to validate the findings of this thesis. In the case of future works regarding dynamics, all of them can be executed on larger networks, and other topologies can be employed, including network models and real-world networks. Also, other parameters can be added to the dynamics, including different characteristics to the agents. In the case of network transmission, there is the possibility to test other versions of random walk dynamics and other algorithms of compression.

BIBLIOGRAPHY

AGIRRE, E.; SOROA, A. Personalizing pagerank for word sense disambiguation. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics**. [S.l.], 2009. p. 33–41. Citation on page [76](#).

AGUIRRE, R.; NYERGES, T. An agent-based model of public participation in sustainability management. **Journal of Artificial Societies and Social Simulation**, v. 17, n. 1, p. 7, 2014. Citation on page [37](#).

AKIMUSHKIN, C.; AMANCIO, D. R.; JR, O. N. O. On the role of words in the network structure of texts: Application to authorship attribution. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 495, p. 49–58, 2018. Citation on page [38](#).

ALBERT, R.; BARABÁSI A, L. Statistical mechanics of complex networks. **Rev. Mod. Phys.**, American Physical Society, v. 74, n. 1, p. 47–97, Jan. 2002. Citation on page [38](#).

ALVAREZ-LACALLE, E.; DOROW, B.; ECKMANN, J.-P.; MOSES, E. Hierarchical structures induce long-range dynamical correlations in written texts. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 103, n. 21, p. 7956–7961, 2006. Citation on page [127](#).

AMANCIO, D. R. Authorship recognition via fluctuation analysis of network topology and word intermittency. **J. Stat. Mech. Theor. Exp.**, IOP Publishing, v. 2015, n. 3, p. P03005, 2015. Citation on page [76](#).

_____. Comparing the topological properties of real and artificially generated scientific manuscripts. **Scientometrics**, Springer, v. 105, n. 3, p. 1763–1779, 2015. Citation on page [114](#).

_____. A complex network approach to stylometry. **PLoS ONE**, Public Library of Science, v. 10, n. 8, p. e0136076, 2015. Citation on page [76](#).

_____. Probing the topological properties of complex networks modeling short written texts. **PLoS ONE**, v. 10, n. 2, p. e0118394, 2015. Citation on page [106](#).

AMANCIO, D. R.; ALTMANN, E. G.; OLIVEIRA JR, O. N.; COSTA, L. F. Comparing intermittency and network measurements of words and their dependence on authorship. **New Journal of Physics**, IOP Publishing, v. 13, n. 12, p. 123024, 2011. Citations on pages [52](#), [54](#), [73](#), [76](#), and [203](#).

AMANCIO, D. R.; ALTMANN, E. G.; RYBSKI, D.; OLIVEIRA JR, O. N.; COSTA, L. F. Probing the statistical properties of unknown texts: application to the voynich manuscript. **PLoS ONE**, Public Library of Science, v. 8, n. 7, p. e67310, 2013. Citations on pages [76](#), [81](#), [100](#), [103](#), [112](#), and [113](#).

AMANCIO, D. R.; ALUISIO, S. M.; OLIVEIRA JR, O. N.; COSTA, L. F. Complex networks analysis of language complexity. **EPL**, v. 100, n. 5, p. 58002, 2012. Citation on page 76.

AMANCIO, D. R.; COMIN, C. H.; CASANOVA, D.; TRAVIESO, G.; BRUNO, O. M.; RODRIGUES, F. A.; COSTA, L. F. A systematic comparison of supervised classifiers. **PloS one**, Public Library of Science, v. 9, n. 4, p. e94137, 2014. Citation on page 109.

AMANCIO, D. R.; NUNES, M. G.; OLIVEIRA JR, O. N.; COSTA, L. F. Extractive summarization using complex networks and syntactic dependency. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 391, n. 4, p. 1855–1864, 2012. Citations on pages 50 and 73.

_____. Extractive summarization using complex networks and syntactic dependency. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 391, n. 4, p. 1855–1864, 2012. Citation on page 76.

AMANCIO, D. R.; OLIVEIRA JR, O. N.; COSTA, L. F. Identification of literary movements using complex networks to represent texts. **New Journal of Physics**, IOP Publishing, v. 14, n. 4, p. 043029, 2012. Citations on pages 38, 114, and 203.

_____. Structure-semantics interplay in complex networks and its effects on the predictability of similarity in texts. **Physica A: Statistical Mechanics and its Applications**, v. 391, n. 18, p. 4406 – 4419, 2012. ISSN 0378-4371. Citation on page 168.

_____. Using complex networks to quantify consistency in the use of words. **Journal of Statistical Mechanics: Theory and Experiment**, IOP Publishing, v. 2012, n. 01, p. P01004, 2012. Citation on page 55.

AMANCIO, D. R.; SILVA, F. N.; COSTA, L. F. Concentric network symmetry grasps authors' styles in word adjacency networks. **EPL (Europhysics Letters)**, IOP Publishing, v. 110, n. 6, p. 68001, 2015. Citations on pages 51, 86, and 87.

AMARAL, L. A. N.; SCALA, A.; BARTHELEMY, M.; STANLEY, H. E. Classes of small-world networks. **Proceedings of the national academy of sciences**, National Acad Sciences, v. 97, n. 21, p. 11149–11152, 2000. Citation on page 58.

AMIT, D. J.; PARISI, G.; PELITI, L. Asymptotic behavior of the "true" self-avoiding walk. **Physical Review B**, APS, v. 27, n. 3, p. 1635, 1983. Citation on page 67.

ANGELOVA, R.; WEIKUM, G. Graph-based text classification: learn from your neighbors. In: **ACM. Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval**. [S.l.], 2006. p. 485–492. Citation on page 76.

ARENAS, A.; DANON, L.; DIAZ-GUILERA, A.; GLEISER, P. M.; GUIMERA, R. Community analysis in social networks. **The European Physical Journal B-Condensed Matter and Complex Systems**, Springer, v. 38, n. 2, p. 373–380, 2004. Citation on page 57.

ARNOLD, K.; ZUBERBÜHLER, K. Language evolution: semantic combinations in primate calls. **Nature**, Nature Publishing Group, v. 441, n. 7091, p. 303–303, 2006. Citation on page 67.

ARRUDA, G. F.; BARBIERI, A. L.; RODRÍGUEZ, P. M.; RODRIGUES, F. A.; MORENO, Y.; COSTA, L. F. Role of centrality for the identification of influential spreaders in complex networks. **Physical Review E**, APS, v. 90, n. 3, p. 032812, 2014. Citations on pages 55, 56, and 66.

ARRUDA, H. F.; COMIN, C. H.; COSTA, L. F. Minimal paths between communities induced by geographical networks. **Journal of Statistical Mechanics: Theory and Experiment**, IOP Publishing, v. 2016, n. 2, p. 023403, 2016. Citation on page 54.

_____. How integrated are theoretical and applied physics? **Scientometrics**, Springer, v. 116, p. 1113–1121, 2018. Available: <<https://link.springer.com/article/10.1007/s11192-018-2777-8>>. Citations on pages 134, 136, 137, 138, and 139.

_____. Intelligent complex networks. **arXiv preprint arXiv:1809.09201**, 2018. Citation on page 150.

ARRUDA, H. F.; COSTA, L. F. Text networks (cdt-4). **Preprint on ResearchGate**, p. 1–7, 2018. Citation on page 79.

ARRUDA, H. F.; COSTA, L. F.; AMANCIO, D. R. Topic segmentation via community detection in complex networks. **Chaos: An Interdisciplinary Journal of Nonlinear Science**, AIP Publishing, v. 26, n. 6, p. 063120, 2016. Citation on page 90.

_____. Using complex networks for text classification: Discriminating informative and imaginative documents. **EPL (Europhysics Letters)**, IOP Publishing, v. 113, n. 2, p. 28007, 2016. Citation on page 83.

ARRUDA, H. F.; MARINHO, V. Q.; COSTA, L. F.; AMANCIO, D. R. Paragraph-based representation of texts: A complex networks approach. **Information Processing & Management**, Elsevier, v. 56, p. 479–494, 2019. Citation on page 100.

ARRUDA, H. F.; MARINHO, V. Q.; LIMA, T. S.; AMANCIO, D. R.; COSTA, L. F. An image analysis approach to text analytics based on complex networks. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 510, p. 110–120, 2018. Citation on page 114.

ARRUDA, H. F.; SILVA, F. N.; COMIN, C. H.; AMANCIO, D. R.; COSTA, L. F. Connecting network science and information theory. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 515, p. 641–648, 2019. Citation on page 160.

ARRUDA, H. F.; SILVA, F. N.; COSTA, L. F.; AMANCIO, D. R. Knowledge acquisition: a complex networks approach. **Information Sciences**, Elsevier, v. 421, p. 154–166, 2017. Citation on page 139.

ARRUDA, H. F.; SILVA, F. N.; MARINHO, V. Q.; AMANCIO, D. R.; COSTA, L. F. Representation of texts as complex networks: a mesoscopic approach. **Journal of Complex Networks**, v. 6, n. 1, p. 125–144, 2018. Available: <<http://dx.doi.org/10.1093/comnet/cnx023>>. Citation on page 114.

BARABÁSI, A.-L.; ALBERT, R. Emergence of scaling in random networks. **science**, American Association for the Advancement of Science, v. 286, n. 5439, p. 509–512, 1999. Citation on page 60.

BARRAT, A.; BARTHELEMY, M.; PASTOR-SATORRAS, R.; VESPIGNANI, A. The architecture of complex weighted networks. **Proceedings of the National Academy of Sciences of the United States of America**, National Acad Sciences, v. 101, n. 11, p. 3747–3752, 2004. Citation on page [48](#).

BARRAT, A.; BARTHELEMY, M.; VESPIGNANI, A. **Dynamical Processes on Complex Networks**. New York, NY, USA: Cambridge University Press, 2008. ISBN 9-780-521-879-507. Citations on pages [38](#), [66](#), and [139](#).

BARTHELEMY, M. Spatial networks. **Physics Reports**, Elsevier, v. 499, n. 1, p. 1–101, 2011. Citation on page [62](#).

BASSETT, D. S.; BULLMORE, E. T. Small-world brain networks. **The neuroscientist**, Sage Publications, v. 12, n. 6, p. 512–523, 2006. Citation on page [58](#).

_____. Small-world brain networks revisited. **The Neuroscientist**, SAGE Publications Sage CA: Los Angeles, CA, v. 23, n. 5, p. 499–516, 2017. Citation on page [58](#).

BATISTA, J.; COSTA, L. F. Knowledge acquisition by networks of interacting agents in the presence of observation errors. **Physical Review E**, APS, v. 82, n. 1, p. 016103, 2010. Citations on pages [139](#) and [148](#).

BECKWITH, R.; MILLER, G. A. Implementing a lexical network. **International Journal of Lexicography**, Oxford Univ Press, v. 3, n. 4, p. 302–312, 1990. Citation on page [69](#).

BECKWITH, R.; MILLER, G. A.; TENGI, R. Design and implementation of the wordnet lexical database and searching software. 2006. Citation on page [69](#).

BELFIELD, R. **The Six Unsolved Ciphers: Inside the Mysterious Codes that Have Confounded the World's Greatest Cryptographers**. [S.l.]: Ulysses Press, 2007. Citation on page [112](#).

BIRD, S.; KLEIN, E.; LOPER, E. **Natural language processing with Python**. [S.l.]: "O'Reilly Media, Inc.", 2009. Citation on page [69](#).

BLONDEL, V. D.; GUILLAUME, J.-L.; LAMBIOTTE, R.; LEFEBVRE, E. Fast unfolding of communities in large networks. **Journal of statistical mechanics: theory and experiment**, IOP Publishing, v. 2008, n. 10, p. P10008, 2008. Citations on pages [57](#) and [58](#).

BOCCALETTI, S.; LATORA, V.; MORENO, Y.; CHAVEZ, M.; HWANG, D. Complex networks: Structure and dynamics. **Physics reports**, Elsevier, v. 424, n. 4-5, p. 175–308, 2006. ISSN 0370-1573. Citations on pages [47](#) and [57](#).

BOCK, H.-H. Clustering methods: a history of k-means algorithms. In: **Selected contributions in data analysis and classification**. [S.l.]: Springer, 2007. p. 161–172. Citations on pages [95](#) and [206](#).

BONACICH, P. Power and centrality: A family of measures. **American journal of sociology**, JSTOR, p. 1170–1182, 1987. Citation on page [56](#).

BONAVENTURA, M.; NICOSIA, V.; LATORA, V. Characteristic times of biased random walks on complex networks. **Physical Review E**, APS, v. 89, n. 1, p. 012803, 2014. Citation on page [162](#).

BRAMER, M. **Principles of data mining**. [S.l.]: Springer, 2007. Citations on pages [84](#) and [204](#).

BREDE, M.; NEWTH, D. Patterns in syntactic dependency networks from authored and randomised texts. **Complexity International**, v. 12, n. msid23, 2008. Citations on pages [37](#), [64](#), and [73](#).

BREIMAN, L.; FRIEDMAN, J. H.; OLSHEN, R. A.; STONE, C. J. **Classification and Regression Trees**. Belmont, California: Wadsworth International Group, 1984. Citations on pages [84](#) and [205](#).

BRISCOE, T. Introduction to linguistics for natural language processing. **Computer Laboratory, University of Cambridge.–October**, v. 4, 2011. Citation on page [68](#).

BUITINCK, L.; LOUPPE, G.; BLONDEL, M.; PEDREGOSA, F.; MUELLER, A.; GRISEL, O.; NICULAE, V.; PRETTENHOFER, P.; GRAMFORT, A.; GROBLER, J.; LAYTON, R.; VANDERPLAS, J.; JOLY, A.; HOLT, B.; VAROQUAUX, G. API design for machine learning software: experiences from the scikit-learn project. In: **ECML PKDD Workshop: Languages for Data Mining and Machine Learning**. [S.l.: s.n.], 2013. p. 108–122. Citation on page [71](#).

BUNT, H. C.; BOS, J.; PULMAN, S. **Computing meaning**. [S.l.]: Springer, 2014. ISSN 1386-291X. ISBN 978-94-007-7283-0. Citation on page [69](#).

CARRO, A.; TORAL, R.; MIGUEL, M. S. Coupled dynamics of node and link states in complex networks: A model for language competition. **arXiv preprint arXiv:1609.00078**, 2016. Citation on page [139](#).

CHOMSKY, N. *et al.* **New horizons in the study of language and mind**. [S.l.]: Cambridge University Press, 2000. Citation on page [37](#).

CHRISTIANSEN, M. H.; KIRBY, S. **Language evolution**. [S.l.]: Oxford University Press, 2003. Citation on page [67](#).

CLAUSET, A.; NEWMAN, M. E.; MOORE, C. Finding community structure in very large networks. **Physical review E**, APS, v. 70, n. 6, p. 066111, 2004. Citation on page [57](#).

CODLING, E. A.; PLANK, M. J.; BENHAMOU, S. Random walk models in biology. **Journal of the Royal Society Interface**, The Royal Society, v. 5, n. 25, p. 813–834, 2008. Citation on page [66](#).

COKOL, M.; IOSSIFOV, I.; WEINREB, C.; RZHETSKY, A. Emergent behavior of growing knowledge about molecular interactions. **Nature biotechnology**, Nature Publishing Group, v. 23, n. 10, p. 1243–1247, 2005. Citation on page [139](#).

COLEMAN, S.; WEINBERG, E. Radiative corrections as the origin of spontaneous symmetry breaking. **Physical Review D**, APS, v. 7, n. 6, p. 1888, 1973. Citation on page [51](#).

COMIN, C. H. **Estudo da relação estrutura–dinâmica em redes modulares**. PhD Thesis (Tese de Doutorado) — University of São Paulo, 2016. Citation on page [162](#).

CONG, J.; LIU, H. Approaching human language with complex networks. **Physics of life reviews**, Elsevier, v. 11, n. 4, p. 598–618, 2014. Citations on pages [68](#) and [75](#).

CORMEN, T. H.; STEIN, C.; RIVEST, R. L.; LEISERSON, C. E. **Introduction to Algorithms**. 2nd. ed. [S.l.]: McGraw-Hill Higher Education, 2001. Citation on page [161](#).

COSTA, L. F. Learning about knowledge: A complex network approach. **Physical Review E**, APS, v. 74, n. 2, p. 026103, 2006. Citation on page [139](#).

_____. Knitted complex networks. **arXiv/abs: 0711.2736**, 2007. Citation on page [64](#).

COSTA, L. F.; ANDRADE, R. F. S. What are the best concentric descriptors for complex networks? **New Journal of Physics**, IOP Publishing, v. 9, n. 9, p. 311, 2007. Citation on page [49](#).

COSTA, L. F.; CESAR, R. M. **Shape Analysis and Classification: Theory and Practice**. [S.l.]: Taylor & Francis, 2000. (Image Processing Series). ISBN 9780849334931. Citations on pages [117](#) and [209](#).

COSTA, L. F.; OLIVEIRA JR, O. N.; TRAVIESO, G.; RODRIGUES, F. A.; BOAS, P. R. V.; ANTIQUEIRA, L.; VIANA, M. P.; ROCHA, L. E. C. Analyzing and modeling real-world phenomena with complex networks: a survey of applications. **Advances in Physics**, Taylor & Francis, v. 60, n. 3, p. 329–412, 2011. Citation on page [66](#).

COSTA, L. F.; RODRIGUES, F.; TRAVIESO, G.; VILLAS BOAS, P. Characterization of complex networks: A survey of measurements. **Advances in Physics**, Taylor & Francis, v. 56, n. 1, p. 167–242, 2007. ISSN 0001-8732. Citations on pages [48](#), [58](#), and [100](#).

COSTA, L. F.; SPORNS, O.; ANTIQUEIRA, L.; NUNES, M. G. V.; OLIVEIRA JR, O. N. Correlations between structure and random walk dynamics in directed complex networks. **Applied physics letters**, AIP Publishing, v. 91, n. 5, p. 054107, 2007. Citations on pages [68](#) and [162](#).

COSTA, L. F.; TRAVIESO, G. Exploring complex networks through random walks. **Physical Review E**, APS, v. 75, n. 1, p. 016102, 2007. Citation on page [66](#).

COVER, T. M.; THOMAS, J. A. **Elements of information theory**. [S.l.]: John Wiley & Sons, 2012. Citations on pages [204](#), [211](#), and [212](#).

DAI, A. M.; OLAH, C.; LE, Q. V. Document embedding with paragraph vectors. **arXiv preprint arXiv:1507.07998**, 2015. Citations on pages [70](#) and [110](#).

DAS, N. G. Statistical methods, combined edition (volumes i & ii). **Tata McGraw Hill Education Private Limited, PAGES-4**, v. 5, p. 290, 2008. Citation on page [122](#).

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the em algorithm. **Journal of the royal statistical society. Series B (methodological)**, JSTOR, p. 1–38, 1977. Citation on page [130](#).

DICKMANN, I. Conhecimento e linguagem: o problema da anterioridade. **Revista Filosofazer**, v. 14, n. 27, p. 145–150, 2011. Citations on pages [38](#), [41](#), and [133](#).

DIJKSTRA, E. W. A note on two problems in connexion with graphs. **Numerische mathematik**, Springer, v. 1, n. 1, p. 269–271, 1959. Citation on page [53](#).

DOMENICONI, G.; MORO, G.; PASOLINI, R.; SARTORI, C. A study on term weighting for text categorization: a novel supervised variant of tf. idf. In: **Proceedings of the 4th international conference on data management technologies and applications (DATA)**. [S.l.: s.n.], 2015. p. 26–37. Citation on page [70](#).

DUÑABEITIA, J. A.; CHOLIN, J.; CORRAL, J.; PEREA, M.; CARREIRAS, M. Syllabarium: An online application for deriving complete statistics for basque and spanish orthographic syllables. **Behavior research methods**, Springer, v. 42, n. 1, p. 118–125, 2010. Citation on page [213](#).

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**, National Acad Sciences, v. 99, n. 20, p. 12917–12922, 2002. Citation on page [58](#).

EBELING, W.; NEIMAN, A. Long-range correlations between letters and sentences in texts. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 215, n. 3, p. 233–241, 1995. Citation on page [127](#).

EFFORD, N. **Digital Image Processing: A Practical Introduction using Java**. Harlow, England: Addison-Wesley, 2000. Citation on page [116](#).

ERDŐS, P.; RÉNYI, A. On random graphs. **Publicationes Mathematicae Debrecen**, v. 6, p. 290–297, 1959. Citation on page [58](#).

ERKAN, G.; RADEV, D. R. Lexrank: Graph-based lexical centrality as salience in text summarization. **Journal of Artificial Intelligence Research**, v. 22, p. 457–479, 2004. Citation on page [76](#).

ESTRADA, E. **The structure of complex networks: theory and applications**. [S.l.]: Oxford University Press, 2012. Citation on page [53](#).

FELDMAN, R.; SANGER, J. **The text mining handbook: advanced approaches in analyzing unstructured data**. [S.l.]: Cambridge University Press, 2007. Citations on pages [70](#) and [72](#).

FERRER I CANCHO, R. The variation of zipf’s law in human language. **The European Physical Journal B-Condensed Matter and Complex Systems**, Springer, v. 44, n. 2, p. 249–257, 2005. Citation on page [214](#).

FERRER I CANCHO, R.; SOLÉ, R. V. The small world of human language. **Proceedings of the Royal Society of London B: Biological Sciences**, The Royal Society, v. 268, n. 1482, p. 2261–2265, 2001. Citations on pages [37](#), [58](#), and [76](#).

_____. Least effort and the origins of scaling in human language. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 100, n. 3, p. 788–791, 2003. Citations on pages [68](#) and [75](#).

FERRER I CANCHO, R.; SOLÉ, R. V.; KÖHLER, R. Patterns in syntactic dependency networks. **Phys. Rev. E**, American Physical Society, v. 69, p. 051915, May 2004. Citations on pages [37](#), [73](#), [75](#), and [76](#).

FLORY, P. J. **Principles of polymer chemistry**. [S.l.]: Cornell University Press, 1953. Citation on page [66](#).

FORTUNATO, S. Community detection in graphs. **Physics Reports**, v. 486, n. 3–5, p. 75 – 174, 2010. ISSN 0370-1573. Citation on page [93](#).

FORTUNATO, S.; BARTHELEMY, M. Resolution limit in community detection. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 104, n. 1, p. 36–41, 2007. Citation on page [57](#).

FORTUNATO, S.; BERGSTROM, C. T.; BÖRNER, K.; EVANS, J. A.; HELBING, D.; MILOJEVIĆ, S.; PETERSEN, A. M.; RADICCHI, F.; SINATRA, R.; UZZI, B. *et al.* Science of science. **Science**, American Association for the Advancement of Science, v. 359, n. 6379, p. eaao0185, 2018. Citation on page [77](#).

FOSTER, J. G.; RZHETSKY, A.; EVANS, J. A. Tradition and innovation in scientist-research strategies. **American Sociological Review**, SAGE Publications, v. 80, n. 5, p. 875–908, 2015. Citation on page [139](#).

FRANK, E.; HALL, M.; HOLMES, G.; KIRKBY, R.; PFAHRINGER, B.; WITTEN, I. H.; TRIGG, L. Weka-a machine learning workbench for data mining. In: **Data mining and knowledge discovery handbook**. [S.l.]: Springer, 2009. p. 1269–1277. Citation on page [124](#).

FREEMAN, L. C. A Set of Measures of Centrality Based on Betweenness. **Sociometry**, v. 40, p. 35–41, 1977. Citation on page [54](#).

FREEMAN, L. C.; BORGATTI, S. P.; WHITE, D. R. Centrality in valued graphs: A measure of betweenness based on network flow. **Social networks**, Elsevier, v. 13, n. 2, p. 141–154, 1991. Citation on page [54](#).

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. **The elements of statistical learning**. [S.l.]: Springer series in statistics New York, 2001. Citation on page [110](#).

FRUCHTERMAN, T. M.; REINGOLD, E. M. Graph drawing by force-directed placement. **Software: Practice and experience**, Wiley Online Library, v. 21, n. 11, p. 1129–1164, 1991. Citation on page [65](#).

GAO, Y.; LIANG, W.; SHI, Y.; HUANG, Q. Comparison of directed and weighted co-occurrence networks of six languages. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 393, p. 579–589, 2014. Citations on pages [41](#) and [76](#).

GARLASCHELLI, D.; LOFFREDO, M. I. Patterns of link reciprocity in directed networks. **Physical review letters**, APS, v. 93, n. 26, p. 268701, 2004. Citation on page [165](#).

GAZIS, D. C. Influence of technology on science: a comment on some experiences at ibm research. **Research Policy**, Elsevier, v. 8, n. 3, p. 244–259, 1979. Citation on page [38](#).

GERSTEIN, G. L.; MANDELROT, B. Random walk models for the spike activity of a single neuron. **Biophysical journal**, The Biophysical Society, v. 4, n. 1 Pt 1, p. 41, 1964. Citation on page [66](#).

- GHANBARNEJAD, F.; GERLACH, M.; MIOTTO, J. M.; ALTMANN, E. G. Extracting information from s-curves of language change. **Journal of The Royal Society Interface**, The Royal Society, v. 11, n. 101, p. 20141044, 2014. Citations on pages [37](#) and [38](#).
- GIRVAN, M.; NEWMAN, M. E. Community structure in social and biological networks. **Proceedings of the national academy of sciences**, National Acad Sciences, v. 99, n. 12, p. 7821–7826, 2002. Citations on pages [55](#) and [61](#).
- GOEL, R.; SONI, S.; GOYAL, N.; PAPARRIZOS, J.; WALLACH, H.; DIAZ, F.; EISENSTEIN, J. The social dynamics of language change in online networks. **arXiv preprint arXiv:1609.02075**, 2016. Citation on page [139](#).
- GROSS, D.; MILLER, K. J. Adjectives in wordnet. **International Journal of Lexicography**, Oxford Univ Press, v. 3, n. 4, p. 265–277, 1990. Citation on page [69](#).
- GUARDIOLA, X.; DÍAZ-GUILERA, A.; LLAS, M.; PÉREZ, C. J. Synchronization, diversity, and topology of networks of integrate and fire oscillators. **Phys. Rev. E**, v. 62, p. 5565–5570, 2000. Citations on pages [66](#) and [139](#).
- GUIMERA, R.; MOSSA, S.; TURTSCHI, A.; AMARAL, L. N. The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 102, n. 22, p. 7794–7799, 2005. Citation on page [57](#).
- GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. **Journal of machine learning research**, v. 3, n. Mar, p. 1157–1182, 2003. Citation on page [203](#).
- GUYON, I.; WESTON, J.; BARNHILL, S.; VAPNIK, V. Gene selection for cancer classification using support vector machines. **Machine Learning**, v. 46, p. 389–422, 2002. Citation on page [130](#).
- _____. Gene selection for cancer classification using support vector machines. **Machine learning**, Springer, v. 46, n. 1-3, p. 389–422, 2002. Citation on page [204](#).
- HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. The weka data mining software: an update. **ACM SIGKDD explorations newsletter**, ACM, v. 11, n. 1, p. 10–18, 2009. Citations on pages [109](#) and [206](#).
- HAN, J. **Data Mining: Concepts and Techniques**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2005. ISBN 1558609016. Citation on page [71](#).
- HARARY, F. **Graph theory**. [S.l.]: Addison-Wesley Publishing Company, 1969. (Addison-Wesley series in mathematics). ISBN 0201410338, 9780201410334. Citation on page [53](#).
- _____. **Graph Theory**. [S.l.]: Addison Wesley, 1969. (Addison-Wesley Series in Mathematics). Citation on page [53](#).
- HART, M. **Project gutenber**. [S.l.]: Project Gutenberg, 1971. Citation on page [80](#).
- HASTIE, T.; TIBSHIRANI, R. Classification by pairwise coupling. In: **Advances in neural information processing systems**. [S.l.: s.n.], 1998. p. 507–513. Citation on page [206](#).

HAUER, B.; KONDRAK, G. Decoding anagrammed texts written in an unknown language and script. **Transactions of the Association for Computational Linguistics**, v. 4, p. 75–86, 2016. Citation on page [113](#).

HAVLIN, S.; BEN-AVRAHAM, D. New approach to self-avoiding walks as a critical phenomenon. **Journal of Physics A: Mathematical and General**, IOP Publishing, v. 15, n. 6, p. L321, 1982. Citation on page [67](#).

HEIJDEN, F. V. D.; DUIN, R. P. W.; RIDDER, D.; TAX, D. M. J. **Classification, parameter estimation and state estimation - an engineering approach using Matlab**. Chichester: John Wiley & Sons, 2004. Citation on page [145](#).

HOLME, P. Congestion and centrality in traffic flow on complex networks. **Advances in Complex Systems**, World Scientific, v. 6, n. 02, p. 163–176, 2003. Citation on page [54](#).

_____. Detecting degree symmetries in networks. **Physical Review E**, APS, v. 74, n. 3, p. 036107, 2006. Citation on page [51](#).

HUANG, A. Similarity measures for text document clustering. In: **Proceedings of the sixth new zealand computer science research student conference (NZC-SRSC2008)**, Christchurch, New Zealand. [S.l.: s.n.], 2008. p. 49–56. Citation on page [71](#).

HUBERT, L.; ARABIE, P. Comparing partitions. **Journal of classification**, Springer, v. 2, n. 1, p. 193–218, 1985. Citation on page [126](#).

IBE, O. C. **Elements of random walk and diffusion processes**. [S.l.]: John Wiley & Sons, 2013. Citation on page [67](#).

JIN, W.; SRIHARI, R. K. Graph-based text representation and knowledge discovery. In: **ACM. Proceedings of the 2007 ACM symposium on Applied computing**. [S.l.], 2007. p. 807–811. Citation on page [76](#).

JOHN, G. H.; LANGLEY, P. Estimating continuous distributions in bayesian classifiers. In: **Eleventh Conference on Uncertainty in Artificial Intelligence**. San Mateo: Morgan Kaufmann, 1995. p. 338–345. Citations on pages [84](#), [204](#), and [205](#).

JOLLIFFE, I. **Principal component analysis**. [S.l.]: Wiley Online Library, 2002. Citation on page [116](#).

JR, E. A. C.; LOPES, A. A.; AMANCIO, D. R. Word sense disambiguation: A complex network approach. **Information Sciences**, Elsevier, v. 442, p. 103–113, 2018. Citation on page [38](#).

KAISER, M.; HILGETAG, C. C. Edge vulnerability in neural and metabolic networks. **Biological cybernetics**, Springer, v. 90, n. 5, p. 311–317, 2004. Citation on page [52](#).

KANTARDZIC, M. **Data mining: concepts, models, methods, and algorithms**. [S.l.]: John Wiley & Sons, 2011. Citations on pages [84](#) and [205](#).

KAREIVA, P.; SHIGESADA, N. Analyzing insect movement as a correlated random walk. **Oecologia**, Springer, v. 56, n. 2-3, p. 234–238, 1983. Citation on page [66](#).

KEERTHI, S. S.; SHEVADE, S. K.; BHATTACHARYYA, C.; MURTHY, K. R. K. Improvements to platt's smo algorithm for svm classifier design. **Neural computation**, MIT Press, v. 13, n. 3, p. 637–649, 2001. Citation on page [206](#).

KIM, Y.; PARK, S.; YOOK, S.-H. Network exploration using true self-avoiding walks. **Physical Review E**, APS, v. 94, n. 4, p. 042309, 2016. Citations on pages [67](#), [140](#), and [162](#).

KISS, G. R.; ARMSTRONG, C.; MILROY, R.; PIPER, J. An associative thesaurus of english and its computer analysis. **The computer and literary studies**, p. 153–165, 1973. Citation on page [168](#).

KOPONEN, I. T.; NOUSIAINEN, M. Concept networks in learning: finding key concepts in learners' representations of the interlinked structure of scientific knowledge. **Journal of Complex Networks**, v. 2, n. 2, p. 187, 2014. Citation on page [148](#).

KUCERA, H.; FRANCIS, W. **A Standard Corpus of Present-Day Edited American English, for use with Digital Computers (Revised and amplified from 1967 version)**. [S.l.]: Providence, RI: Brown University Press, 1979. Citation on page [79](#).

LAM, P. True self-avoiding walk on critical percolation clusters and lattice animal. **Zeitschrift für Physik B Condensed Matter**, Springer, v. 57, n. 4, p. 301–305, 1984. Citation on page [67](#).

LANCICHINETTI, A.; FORTUNATO, S. Consensus clustering in complex networks. **Scientific reports**, Nature Publishing Group, v. 2, p. 336, 2012. Citation on page [38](#).

LANCICHINETTI, A.; FORTUNATO, S.; RADICCHI, F. Benchmark graphs for testing community detection algorithms. **Physical review E**, APS, v. 78, n. 4, p. 046110, 2008. Citations on pages [14](#), [61](#), and [62](#).

LANCICHINETTI, A.; SIRER, M. I.; WANG, J. X.; ACUNA, D.; KÖRDING, K.; AMARAL, L. A. N. High-reproducibility and high-accuracy method for automated topic classification. **Physical Review X**, APS, v. 5, n. 1, p. 011007, 2015. Citation on page [73](#).

LANGVILLE, A. N.; MEYER, C. D. **Google's PageRank and beyond: the science of search engine rankings**. [S.l.]: Princeton University Press, 2011. Citation on page [56](#).

LAO, N.; MITCHELL, T.; COHEN, W. W. Random walk inference and learning in a large scale knowledge base. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the Conference on Empirical Methods in Natural Language Processing**. [S.l.], 2011. p. 529–539. Citation on page [66](#).

LARSON, R.; ODONI, A. Shortest paths between all pairs of nodes. **Urban Operations Research**, 1981. Citation on page [53](#).

LE, Q.; MIKOLOV, T. Distributed representations of sentences and documents. In: **International Conference on Machine Learning**. [S.l.: s.n.], 2014. p. 1188–1196. Citation on page [110](#).

LEBANON, G.; MAO, Y.; DILLON, J. V. The locally weighted bag of words framework for document representation. **Journal of Machine Learning Research**, v. 8, p. 2405–2441, 2007. Citation on page [37](#).

LEE, W. H. How to identify emerging research fields using scientometrics: An example in the field of information security. **Scientometrics**, Springer, v. 76, n. 3, p. 503–525, 2008. Citation on page [135](#).

LESKOVEC, J.; RAJARAMAN, A.; ULLMAN, J. D. **Mining of massive datasets**. [S.l.]: Cambridge University Press, 2014. Citation on page [70](#).

LEYDESDORFF, L. Betweenness centrality as an indicator of the interdisciplinarity of scientific journals. **Journal of the Association for Information Science and Technology**, Wiley Online Library, v. 58, n. 9, p. 1303–1319, 2007. Citation on page [136](#).

LIU, H.; CONG, J. Language clustering with word co-occurrence networks based on parallel texts. **Chinese Science Bulletin**, Springer, v. 58, n. 10, p. 1139–1144, 2013. Citation on page [76](#).

LIU, H.; XU, C. Can syntactic networks indicate morphological complexity of a language? **EPL**, v. 93, n. 2, p. 28005, 2011. Citation on page [76](#).

LLOYD, S. Least squares quantization in pcm. **IEEE transactions on information theory**, IEEE, v. 28, n. 2, p. 129–137, 1982. Citation on page [130](#).

LONDHE, N.; GOPALAKRISHNAN, V.; ZHANG, A.; NGO, H. Q.; SRIHARI, R. Matching titles with cross title web-search enrichment and community detection. **Proc. VLDB Endow.**, v. 7, n. 12, p. 1167–1178, 2014. Citation on page [57](#).

LONGUET-HIGGINS, H. The symmetry groups of non-rigid molecules. **Molecular Physics**, Taylor & Francis, v. 6, n. 5, p. 445–460, 1963. Citation on page [51](#).

LOPER, E.; BIRD, S. Nltk: The natural language toolkit. In: **Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1**. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002. (ETMTNLP '02), p. 63–70. Citation on page [201](#).

LORDANA, O.; SALLANA, J. M.; SIMOA, P.; GONZALEZ-PRIETOB, D. Robustness of the air transport network. **Transportation Research Part E: Logistics and Transportation Review**, v. 68, p. 155–163, Aug. 2014. ISSN 1366-5545. Citation on page [38](#).

MALECHA, G.; SMITH, I. Maximum entropy part-of-speech tagging in nltk. **unpublished course-related report**, 2010. Citation on page [68](#).

MALIOUTOV, I.; BARZILAY, R. Minimum cut model for spoken lecture segmentation. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics**. [S.l.], 2006. p. 25–32. Citation on page [73](#).

MANN, H. B.; WHITNEY, D. R. On a test of whether one of two random variables is stochastically larger than the other. **Ann. Math. Stat.**, JSTOR, p. 50–60, 1947. Citation on page 207.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. **Introduction to Information Retrieval**. New York, NY, USA: Cambridge University Press, 2008. ISBN 0521865719, 9780521865715. Citation on page 71.

MANNING, C. D.; SCHÜTZE, H. **Foundations of Statistical Natural Language Processing**. Cambridge, MA, USA: MIT Press, 1999. ISBN 0-262-13360-1. Citations on pages 69 and 70.

MARINHO, V. Q.; ARRUDA, H. F.; LIMA, T. S.; COSTA, L. F.; AMANCIO, D. R. On the “calligraphy” of books. In: **TextGraphs**. [S.l.]: Association for Computational Linguistics, 2017. p. 1–10. Citation on page 114.

MARINHO, V. Q.; HIRST, G.; AMANCIO, D. R. Authorship attribution via network motifs identification. In: IEEE. **2016 5th Brazilian Conference on Intelligent Systems (BRACIS)**. [S.l.], 2016. p. 355–360. Citation on page 38.

MARTIN, T.; BALL, B.; KARRER, B.; NEWMAN, M. Coauthorship and citation patterns in the physical review. **Physical Review E**, APS, v. 88, n. 1, p. 012814, 2013. Citation on page 38.

MARTINEZ-ROMO, J.; ARAUJO, L.; BORGE-HOLTHOEFER, J.; ARENAS, A.; CAPITÁN, J. A.; CUESTA, J. A. Disentangling categorical relationships through a graph of co-occurrences. **Physical Review E**, APS, v. 84, n. 4, p. 046108, 2011. Citations on pages 58, 73, and 90.

MASUCCI, A.; RODGERS, G. Network properties of written human language. **Physical Review E**, APS, v. 74, n. 2, p. 026102, 2006. Citations on pages 48, 49, 68, and 73.

MASUCCI, A. P.; KALAMPOKIS, A.; EGUÍLUZ, V. M.; HERNÁNDEZ-GARCIA, E. Wikipedia information flow analysis reveals the scale-free architecture of the semantic space. **PLoS ONE**, Public Library of Science, v. 6, n. 2, p. e17333, 02 2011. Citation on page 76.

MASUCCI, A. P.; RODGERS, G. J. Differences between normal and shuffled texts: Structural properties of weighted networks. **Advances in Complex Systems**, v. 12, n. 01, p. 113–129, 2009. Citation on page 76.

MATOS, F. Documentação de estudos em lingüística teórica e aplicada. **Schielo Brazil**, v. 25, n. 2, 2009. ISSN 0102-4450. Citation on page 69.

MEECHAM, M.; REES-MILLER, J. **Language in social contexts**. 4th. ed. [S.l.]: Contemporary Linguistics, 2001. ISBN 0-312-24738-9. Citation on page 37.

MEHRI, A.; DAROONEH, A. H.; SHARIATI, A. The complex networks approach for authorship attribution of books. **Physica A**, v. 391, n. 7, p. 2429 – 2437, 2012. Citation on page 76.

MEŠTROVIĆ, A.; MARTINČIĆ-IPŠIĆ, S.; MATEŠIĆ, M. Postupak automatskoga slogovanja temeljem načela najvećega pristupa i statistika slogova za hrvatski jezik. **Govor**, Odjel za fonetiku Hrvatskoga filološkoga društva, v. 32, n. 1, p. 3–34, 2015. Citation on page [213](#).

METZLER, D.; CROFT, W. B. Combining the language model and inference network approaches to retrieval. **Information processing & management**, Elsevier, v. 40, n. 5, p. 735–750, 2004. Citation on page [76](#).

MEYER, M.; DEBACKERE, K.; GLÄNZEL, W. Can applied science be ‘good science’? exploring the relationship between patent citations and citation impact in nanoscience. **Scientometrics**, Springer, v. 85, n. 2, p. 527–539, 2010. Citation on page [38](#).

MIHALCEA, R. Random walks on text structures. In: SPRINGER. **International Conference on Intelligent Text Processing and Computational Linguistics**. [S.l.], 2006. p. 249–262. Citation on page [66](#).

MIHALCEA, R.; CORLEY, C.; STRAPPARAVA, C. Corpus-based and knowledge-based measures of text semantic similarity. In: **AAAI**. [S.l.: s.n.], 2006. v. 6, p. 775–780. Citation on page [73](#).

MIHALCEA, R.; RADEV, D. **Graph-based natural language processing and information retrieval**. [S.l.]: Cambridge University Press, 2011. Citations on pages [56](#) and [73](#).

MIHALCEA, R.; TARAU, P.; FIGA, E. Pagerank on semantic networks, with application to word sense disambiguation. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 20th international conference on Computational Linguistics**. [S.l.], 2004. p. 1126. Citations on pages [38](#) and [56](#).

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. Efficient estimation of word representations in vector space. **arXiv preprint arXiv:1301.3781**, 2013. Citation on page [70](#).

MIKOLOV, T.; SUTSKEVER, I.; CHEN, K.; CORRADO, G. S.; DEAN, J. Distributed representations of words and phrases and their compositionality. In: **Advances in neural information processing systems**. [S.l.: s.n.], 2013. p. 3111–3119. Citation on page [110](#).

MILLÁN, V. M. L.; CHOLVI, V.; LÓPEZ, L.; ANTA, A. F. A model of self-avoiding random walks for searching complex networks. **Networks**, Wiley Online Library, v. 60, n. 2, p. 71–85, 2012. Citation on page [67](#).

MILLER, G. A. Nouns in wordnet: a lexical inheritance system. **International journal of Lexicography**, Oxford Univ Press, v. 3, n. 4, p. 245–264, 1990. Citation on page [69](#).

_____. Wordnet: a lexical database for english. **Communications of the ACM**, ACM, v. 38, n. 11, p. 39–41, 1995. Citations on pages [69](#) and [168](#).

MILLER, G. A.; BECKWITH, R.; FELLBAUM, C.; GROSS, D.; MILLER, K. J. Introduction to wordnet: An on-line lexical database*. **International journal of lexicography**, Oxford Univ Press, v. 3, n. 4, p. 235–244, 1990. Citation on page [69](#).

- MIRZASOLEIMAN, B.; BABAEI, M.; JALILI, M.; SAFARI, M. Cascaded failures in weighted networks. **Physical Review E**, APS, v. 84, n. 4, p. 046114, 2011. Citation on page 54.
- MISLOVE, A.; MARCON, M.; GUMMADI, K. P.; DRUSCHEL, P.; BHATTACHARJEE, B. Measurement and analysis of online social networks. In: ACM. **Proceedings of the 7th ACM SIGCOMM conference on Internet measurement**. [S.l.], 2007. p. 29–42. Citation on page 60.
- MONTEMURRO, M. A.; ZANETTE, D. H. Keywords and co-occurrence patterns in the voynich manuscript: An information-theoretic analysis. **PLoS ONE**, Public Library of Science, v. 8, n. 6, p. e66344, 2013. Citations on pages 76, 81, 112, and 113.
- MOON, T. K. The expectation-maximization algorithm. **IEEE Signal processing magazine**, IEEE, v. 13, n. 6, p. 47–60, 1996. Citations on pages 95 and 206.
- MORENO, Y.; NEKOVEE, M.; PACHECO, A. F. Dynamics of rumor spreading in complex networks. **Physical Review E**, APS, v. 69, n. 6, p. 066130, 2004. Citations on pages 66 and 139.
- MOTTER, A. E.; LAI, Y.-C. Cascade-based attacks on complex networks. **Physical Review E**, APS, v. 66, n. 6, p. 065102, 2002. Citations on pages 54, 66, and 139.
- MOTTER, A. E.; MYERS, S. A.; ANGHEL, M.; NISHIKAWA, T. Spontaneous synchrony in power-grid networks. **Nat Phys**, Nature Publishing Group, v. 9, n. 3, p. 191–197, Mar. 2013. ISSN 1745-2473. Citation on page 38.
- NAVIGLI, R. Word sense disambiguation: A survey. **ACM Computing Surveys (CSUR)**, ACM, v. 41, n. 2, p. 10, 2009. Citation on page 48.
- NES, J. G.-G.; LATORA, V. Entropy rate of diffusion processes on complex networks. **Phys. Rev. E**, American Physical Society, v. 78, p. 065102, Dec 2008. Available: <<https://link.aps.org/doi/10.1103/PhysRevE.78.065102>>. Citation on page 67.
- NEWMAN, M. The Structure and Function of Complex Networks. **SIAM review**, JSTOR, v. 45, n. 2, p. 167–256, 2003. ISSN 0036-1445. Citation on page 38.
- NEWMAN, M. E. Assortative mixing in networks. **Physical review letters**, APS, v. 89, n. 20, p. 208701, 2002. Citation on page 53.
- _____. A measure of betweenness centrality based on random walks. **Social networks**, Elsevier, v. 27, n. 1, p. 39–54, 2005. Citation on page 54.
- _____. Modularity and community structure in networks. **Proceedings of the national academy of sciences**, National Acad Sciences, v. 103, n. 23, p. 8577–8582, 2006. Citation on page 57.
- NOH, J. D.; RIEGER, H. Random walks on complex networks. **Physical review letters**, APS, v. 92, n. 11, p. 118701, 2004. Citation on page 66.
- NORTON, H. K.; EMERSON, D. J.; HUANG, H.; KIM, J.; TITUS, K. R.; GU, S.; BASSETT, D. S.; PHILLIPS-CREMINS, J. E. Detecting hierarchical genome folding with network modularity. **Nature methods**, Nature Publishing Group, 2018. Citation on page 57.

PALLA, G.; DERÉNYI, I.; FARKAS, I.; VICSEK, T. Uncovering the overlapping community structure of complex networks in nature and society. **Nature**, Nature Publishing Group, v. 435, n. 7043, p. 814–818, 2005. Citation on page 57.

PAN, R. K.; KASKI, K.; FORTUNATO, S. World citation and collaboration networks: uncovering the role of geography in science. **Scientific reports**, Nature Publishing Group, v. 2, p. 902, 2012. Citation on page 38.

PARESCHI, L.; TOSCANI, G. Wealth distribution and collective knowledge: a boltzmann approach. **Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences**, The Royal Society, v. 372, n. 2028, 2014. Citation on page 148.

PATIL, K.; BRAZDIL, P. Text summarization: Using centrality in the pathfinder network. **Int. J. Comput. Sci. Inform. Syst.**, v. 2, p. 18–32, 2007. Citation on page 56.

PEARSON, K. The problem of the random walk. **Nature**, v. 72, p. 342, 1905. Citation on page 66.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011. Citation on page 71.

PENROSE, M. **Random geometric graphs**. [S.l.]: Oxford University Press, 2003. Citations on pages 62 and 63.

PIANTADOSI, S. T. Zipf's word frequency law in natural language: A critical review and future directions. **Psychonomic bulletin & review**, Springer, p. 1–19, 2014. Citations on pages 37 and 68.

PINKER, S.; BLOOM, P. Natural language and natural selection. **Behavioral and brain sciences**, Cambridge Univ Press, v. 13, n. 04, p. 707–727, 1990. Citation on page 37.

PLATT, J. C. Fast training of support vector machines using sequential minimal optimization. **Advances in Kernel Methods - Support Vector Learning**, 1998. Citation on page 206.

PONCELA, J.; GÓMEZ-GARDEÑES, J.; TRAULSEN, A.; MORENO, Y. Evolutionary game dynamics in a growing structured population. **New Journal of Physics**, IOP Publishing, v. 11, n. 8, p. 083031, 2009. Citation on page 38.

PONS, P.; LATAPY, M. Computing communities in large networks using random walks. **J. Graph Algorithms Appl.**, v. 10, n. 2, p. 191–218, 2006. Citations on pages 58 and 66.

RATNAPARKHI, A. A maximum entropy model for part-of-speech tagging. In: PHILADELPHIA, USA. **Proceedings of the conference on empirical methods in natural language processing**. [S.l.], 1996. v. 1, p. 133–142. Citation on page 68.

RAVASZ, E.; SOMERA, A. L.; MONGRU, D. A.; OLTVAI, Z. N.; BARABÁSI, A.-L. Hierarchical organization of modularity in metabolic networks. **science**, American Association for the Advancement of Science, v. 297, n. 5586, p. 1551–1555, 2002. Citation on page 52.

RAYLEIGH, L. The problem of the random walk. **Nature**, v. 72, n. 1866, p. 318, 1905. Citation on page 66.

REDDY, S.; KNIGHT, K. What we know about the voynich manuscript. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 5th ACL-HLT workshop on language technology for cultural heritage, social sciences, and humanities**. [S.l.], 2011. p. 78–86. Citations on pages 81 and 112.

REKA, A.; JEONG, H.; BARABASI, A.-L. Diameter of the world-wide web. **Nature**, v. 401, n. 6749, p. 130–131, 1999. Citation on page 60.

RODRIGUES, E. P.; BARBOSA, M. S.; COSTA, L. F. Self-referred approach to lacunarity. **Phys. Rev. E**, American Physical Society, v. 72, p. 016707, Jul 2005. Citations on pages 117 and 210.

RODRIGUES, F. A.; COSTA, L. F. Comparison of the interactomic networks of different species in terms of accessibility. **Molecular BioSystems**, Royal Society of Chemistry, v. 6, n. 1, p. 234–240, 2009. Citation on page 66.

_____. A structure–dynamic approach to cortical organization: Number of paths and accessibility. **Journal of Neuroscience Methods**, Elsevier, v. 183, n. 1, p. 57–62, 2009. Citation on page 50.

ROKACH, L.; MAIMON, O. **Data mining with decision trees: theory and applications**. [S.l.]: World scientific, 2014. Citation on page 205.

ROSATO, V.; ISSACHAROFF, L.; MELONI, S.; CALIGIORE, D.; TIRITICCO, F. Is the topology of the internet network really fit to sustain its function? **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 387, n. 7, p. 1689–1704, 2008. Citation on page 52.

ROSSI, L.; TORSELLO, A.; HANCOCK, E. R.; WILSON, R. C. Characterizing graph symmetries through quantum jensen-shannon divergence. **Physical Review E**, APS, v. 88, n. 3, p. 032806, 2013. Citation on page 51.

ROSVALL, M.; BERGSTROM, C. T. Maps of random walks on complex networks reveal community structure. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 105, n. 4, p. 1118–1123, 2008. Citation on page 160.

ROXIN, A.; RIECKE, H.; SOLLA, S. A. Self-Sustained Activity in a Small-World Network of Excitable Neurons. **Phys. Rev. Lett.**, v. 92, p. 198101, 2004. Citation on page 66.

ROZENFELD, A. F.; COHEN, R.; BEN-AVRAHAM, D.; HAVLIN, S. Scale-free networks on lattices. **Physical Review Letters**, APS, v. 89, n. 21, p. 218701, 2002. Citation on page 62.

RZHETSKY, A.; FOSTER, J. G.; FOSTER, I. T.; EVANS, J. A. Choosing experiments to accelerate collective discovery. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 112, n. 47, p. 14569–14574, 2015. Citation on page 139.

SALLOUM, S. A.; AL-EMRAN, M.; MONEM, A. A.; SHAALAN, K. A survey of text mining in social media: facebook and twitter perspectives. **Adv. Sci. Technol. Eng. Syst. J.**, v. 2, n. 1, p. 127–133, 2017. Citation on page [114](#).

SALOMON, D.; MOTTA, G. **Handbook of data compression**. [S.l.]: Springer Science & Business Media, 2010. Citation on page [212](#).

SALTON, G.; SINGHAL, A.; MITRA, M.; BUCKLEY, C. Automatic text structuring and summarization. **Information Processing & Management**, Elsevier, v. 33, n. 2, p. 193–207, 1997. Citations on pages [76](#) and [100](#).

SAUSSURE, F. D. **Course in general linguistics**. [S.l.]: Columbia University Press, 2011. Citation on page [68](#).

SCHULT, D. A.; SWART, P. Exploring network structure, dynamics, and function using networkx. In: **Proceedings of the 7th Python in Science Conferences (SciPy 2008)**. [S.l.: s.n.], 2008. v. 2008, p. 11–16. Citation on page [65](#).

SEGARRA, S.; EISEN, M.; RIBEIRO, A. Authorship attribution using function words adjacency networks. In: IEEE. **Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on**. [S.l.], 2013. p. 5563–5567. Citation on page [84](#).

SHANNON, C. E. A mathematical theory of communication. **Bell system technical journal**, Wiley Online Library, v. 27, n. 3, p. 379–423, 1948. Citation on page [211](#).

SHENG, L.; LI, C. English and chinese languages as weighted complex networks. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 388, n. 12, p. 2561–2570, 2009. Citations on pages [48](#), [49](#), [68](#), and [73](#).

SHIBATA, N.; KAJIKAWA, Y.; SAKATA, I. Extracting the commercialization gap between science and technology-case study of a solar cell. **Technological Forecasting and Social Change**, Elsevier, v. 77, n. 7, p. 1147–1155, 2010. Citation on page [38](#).

SIDIROPOULOS, A.; KATSAROS, D.; MANOLOPOULOS, Y. Generalized hirsch h-index for disclosing latent facts in citation networks. **Scientometrics**, Akadémiai Kiadó, co-published with Springer Science+ Business Media BV, Formerly Kluwer Academic Publishers BV, v. 72, n. 2, p. 253–280, 2007. Citation on page [38](#).

SIGANOS, G.; FALOUTSOS, M.; FALOUTSOS, P.; FALOUTSOS, C. Power laws and the as-level internet topology. **IEEE/ACM Transactions on Networking (TON)**, IEEE Press, v. 11, n. 4, p. 514–524, 2003. Citation on page [60](#).

SIGMAN, M.; CECCHI, G. A. Global organization of the wordnet lexicon. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 99, n. 3, p. 1742–1747, 2002. Citation on page [69](#).

SILVA, F. N. **Dimensão e simetria em redes complexas: uma abordagem multi-escala**. Phd Thesis (Tese de Doutorado) — IFSC-USP, 2015. Citations on pages [14](#), [16](#), [17](#), [59](#), [60](#), [61](#), [62](#), [63](#), [64](#), [113](#), [119](#), [121](#), [122](#), and [129](#).

- SILVA, F. N.; AMANCIO, D. R.; BARDOSOVA, M.; COSTA, L. F.; OLIVEIRA JR, O. N. Using network science and text analytics to produce surveys in a scientific topic. **Journal of Informetrics**, Elsevier, v. 10, n. 2, p. 487–502, 2016. Citations on pages [57](#), [58](#), [65](#), [66](#), [77](#), [116](#), and [134](#).
- SILVA, F. N.; COMIN, C. H.; PERON, T. K. D.; RODRIGUES, F. A.; YE, C.; WILSON, R. C.; HANCOCK, E. R.; COSTA, L. F. Concentric network symmetry. **Information Sciences**, Elsevier, v. 333, p. 61–80, 2016. Citations on pages [51](#) and [66](#).
- SILVA, F. N.; TRAVENCOLO, B. A.; VIANA, M. P.; COSTA, L. F. Identifying the borders of mathematical knowledge. **Journal of Physics A: Mathematical and Theoretical**, IOP Publishing, v. 43, n. 32, p. 325202, 2010. Citation on page [139](#).
- SILVA, F. N.; VIANA, M. P.; TRAVENÇOLO, B. A. N.; COSTA, L. F. Investigating relationships within and between category networks in wikipedia. **Journal of informetrics**, Elsevier, v. 5, n. 3, p. 431–438, 2011. Citation on page [134](#).
- SILVA, T. C.; AMANCIO, D. R. Word sense disambiguation via high order of learning in complex networks. **EPL (Europhysics Letters)**, IOP Publishing, v. 98, n. 5, p. 58001, 2012. Citation on page [77](#).
- SIMON, H. Ha simon, *biometrika* 42, 425 (1955). **Biometrika**, v. 42, p. 425, 1955. Citation on page [214](#).
- SKLANSKY, J. Finding the convex hull of a simple polygon. **Pattern Recogn. Lett.**, Elsevier Science Inc., New York, NY, USA, v. 1, n. 2, p. 79–83, Dec. 1982. ISSN 0167-8655. Citations on pages [117](#) and [209](#).
- SKYUM, S. A simple algorithm for computing the smallest enclosing circle. **Information Processing Letters**, v. 37, n. 3, p. 121–125, 1991. Citations on pages [117](#) and [209](#).
- SMOLA, A. J.; SCHÖLKOPF, B. A tutorial on support vector regression. **Statistics and computing**, Springer, v. 14, n. 3, p. 199–222, 2004. Citation on page [206](#).
- SPORNS, O. Graph theory methods for the analysis of neural connectivity patterns. In: **Neuroscience databases**. [S.l.]: Springer, 2003. p. 171–185. Citation on page [52](#).
- SPORNS, O.; ZWI, J. D. The small world of the cerebral cortex. **Neuroinformatics**, Springer, v. 2, n. 2, p. 145–162, 2004. Citation on page [58](#).
- STEELS, L. Language as a complex adaptive system. In: SPRINGER. **International Conference on Parallel Problem Solving from Nature**. [S.l.], 2000. p. 17–26. Citation on page [37](#).
- STELLA, A.; QUEIROZ, S. de; DUXBURY, P.; STINCHCOMBE, R. Series study of the one-dimensional 'true' self-avoiding walk. **Journal of Physics A: Mathematical and General**, IOP Publishing, v. 17, n. 9, p. 1903, 1984. Citation on page [67](#).
- STROGATZ, S. H. Exploring complex networks. **Nature**, Nature Publishing Group, v. 410, n. 6825, p. 268–276, Mar. 2001. ISSN 0028-0836. Citations on pages [37](#) and [58](#).
- TAMASSIA, R. **Handbook of graph drawing and visualization**. [S.l.]: CRC press, 2013. Citation on page [65](#).

THULASIRAMAN, K.; SWAMY, M. N. **Graphs: theory and algorithms**. [S.l.]: John Wiley & Sons, 2011. Citation on page [53](#).

TOHALINO, J. V.; AMANCIO, D. R. Extractive multi-document summarization using multilayer networks. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 503, p. 526–539, 2018. Citations on pages [38](#) and [50](#).

TRAVENTÇOLO, B. A. N.; COSTA, L. d. F. Accessibility in complex networks. **Physics Letters A**, Elsevier, v. 373, n. 1, p. 89–95, 2008. Citations on pages [49](#), [50](#), [66](#), and [141](#).

TUSSEN, R.; BUTER, R.; LEEUWEN, T. V. Technological relevance of science: An assessment of citation linkages between patents and research papers. **Scientometrics**, Akadémiai Kiadó, co-published with Springer Science+ Business Media BV, Formerly Kluwer Academic Publishers BV, v. 47, n. 2, p. 389–412, 2000. Citation on page [38](#).

VÉRONIS, J. Hyperlex: lexical cartography for information retrieval. **Computer Speech & Language**, v. 18, n. 3, p. 223 – 252, 2004. ISSN 0885-2308. Citation on page [90](#).

VIANA, M. P.; BATISTA, J. L.; COSTA, L. F. Effective number of accessed nodes in complex networks. **Physical Review E**, APS, v. 85, n. 3, p. 036105, 2012. Citation on page [50](#).

WANG, Y.; XU, W. Leveraging deep learning with lda-based text analytics to detect automobile insurance fraud. **Decision Support Systems**, Elsevier, v. 105, p. 87–95, 2018. Citation on page [114](#).

WARD JR, J. H. Hierarchical grouping to optimize an objective function. **Journal of the American statistical association**, Taylor & Francis Group, v. 58, n. 301, p. 236–244, 1963. Citation on page [216](#).

WASSERMAN, S.; FAUST, K. **Social network analysis: Methods and applications**. [S.l.]: Cambridge university press, 1994. Citation on page [58](#).

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of ‘small-world’ networks. **nature**, Nature Publishing Group, v. 393, n. 6684, p. 440–442, 1998. Citations on pages [52](#) and [59](#).

WAXMAN, B. M. Routing of Multipoint Connections. **IEEE Journal on Selected Areas in Communications**, v. 6, n. 9, p. 1617–1622, 1988. Citation on page [62](#).

WEST, D. B. **Introduction to graph theory**. [S.l.]: Prentice hall Upper Saddle River, 2001. Citation on page [53](#).

WITTEN, I. H.; FRANK, E.; HALL, M. A.; PAL, C. J. **The WEKA Workbench. Online Appendix for “Data Mining: Practical Machine Learning Tools and Techniques”**. [S.l.]: Morgan Kaufmann, 2016. Citation on page [109](#).

WU, J.; ZHANG, G.; REN, Y. A balanced modularity maximization link prediction model in social networks. **Information Processing & Management**, Elsevier, v. 53, n. 1, p. 295–307, 2017. Citation on page [76](#).

YAN, E.; DING, Y. Applying centrality measures to impact analysis: A coauthorship network analysis. **Journal of the Association for Information Science and Technology**, Wiley Online Library, v. 60, n. 10, p. 2107–2118, 2009. Citation on page [136](#).

YU, D.; WANG, W.; ZHANG, S.; ZHANG, W.; LIU, R. Hybrid self-optimized clustering model based on citation links and textual features to detect research topics. **PLoS ONE**, Public Library of Science, v. 12, n. 10, p. e0187164, 2017. Citation on page 77.

ZHAO, W.; MAO, J.; LU, K. Ranking themes on co-word networks: Exploring the relationships among different metrics. **Information Processing & Management**, Elsevier, v. 54, n. 2, p. 203–218, 2018. Citation on page 76.

ZLATIĆ, V.; GABRIELLI, A.; CALDARELLI, G. Topologically biased random walk and community finding in networks. **Phys. Rev. E**, American Physical Society, v. 82, p. 066109, Dec 2010. Available: <<https://link.aps.org/doi/10.1103/PhysRevE.82.066109>>. Citation on page 67.

ZNADBERGEN, R. **Text Analysis - Transcription of the Text**. 2018. Available: <<http://www.voyrich.nu/transcr.html>>. Citation on page 81.

LIST OF PUBLICATIONS

In this Appendix, we show the entire list of published papers and preprints produced during the period in which this study was developed. First, we present the list of papers that take part of this thesis, in which the articles are ordered according to the thesis organization, as follows:

1. ARRUDA, H. F.; COSTA, L. F.; AMANCIO, D. R. Using complex networks for text classification: Discriminating informative and imaginative documents. **EPL(Europhysics Letters)**, IOP Publishing, v.113, n.2, p.28007, 2016.
2. ARRUDA, H. F.; COSTA, L. F.; AMANCIO, D. R. Topic segmentation via community detection in complex networks. **Chaos: An Interdisciplinary Journal of Nonlinear Science**, AIP Publishing, v.26, n.6, p.063120, 2016.
3. ARRUDA, H. F.; MARINHO, V. Q.; COSTA, L. F.; AMANCIO, D. R. Paragraph-based representation of texts: A complex networks approach. **Information Processing & Management**, Elsevier, v. 56, p. 479–494, 2019.
4. ARRUDA, H. F. de; SILVA, F. N.; MARINHO, V. Q.; AMANCIO, D. R.; COSTA, L. da F. Representation of texts as complex networks: a mesoscopic approach. **Journal of Complex Networks**, v. 6, n. 1, p. 125–144, 2018.
5. MARINHO, V. Q.; ARRUDA, H. F.; LIMA, T. S.; COSTA, L. F.; AMANCIO, D. R. On the “calligraphy” of books. In: **TextGraphs**. [S.l.]: Association for Computational Linguistics, 2017.
6. ARRUDA, H. F.; MARINHO, V. Q.; LIMA, T. S.; AMANCIO, D. R.; COSTA, L. da F. An image analysis approach to text analytics based on complex networks. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 510, p. 110–120, 2018.

7. ARRUDA, H. F.; COMIN, C. H.; COSTA, L. F. How integrated are theoretical and applied physics? **Scientometrics**, Springer, v. 116, p. 1113–1121, 2018.
8. ARRUDA, H. F.; SILVA, F. N.; COSTA, L. F.; AMANCIO, D. R. Knowledge acquisition: a complex networks approach. **Information Sciences**, Elsevier, v. 421, p. 154–166, 2017.
9. ARRUDA, H. F.; COMIN, C. H.; COSTA, L. F. Intelligent complex networks. **arXiv preprint arXiv:1809.09201**, 2018.
10. ARRUDA, H. F.; SILVA, F. N.; COMIN, C. H.; AMANCIO, D. R.; COSTA, L. F. Connecting network science and information theory. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 515, p. 641–648, 2019.
11. ARRUDA, H. F.; COSTA, L. F. Text networks (cdt-4). **Preprint on ResearchGate**, p. 1–7, 2018.

Apart from the papers listed above, the following related articles were also developed during the Ph.D. period:

- LIMA, T. S.; ARRUDA, H. F.; SILVA, F. N.; COMIN, C. H.; AMANCIO, D. R.; COSTA, L. F. The Dynamics of Knowledge Acquisition via Self-Learning in Complex Networks. **Chaos: An Interdisciplinary Journal of Nonlinear Science**, AIP Publishing, v.28, n.8, p.083106, 2018.
- GEWERS, F. L.; FERREIRA, G. R.; Arruda, H. F.; SILVA, F. N.; COMIN, C. H.; AMANCIO, D. R.; COSTA, L. F. Principal Component Analysis: A Natural Approach to Data Exploration. **arXiv preprint arXiv:1804.02502**. 2018.

We also presented a poster whose abstract is published as follows:

- ARRUDA, H. F de; SILVA, F. N.; COMIN, C. H.; AMANCIO, D. R.; COSTA, L. da F. Apresentação de Conteúdos Didáticos: Auxílio ao Planejamento Utilizando Redes Complexas. In: **Congresso de Graduação**, 2017.

LIST OF STOP WORDS

List of the employed *stopwords*, which were obtained from NLTK: the Natural Language Toolkit (LOPER; BIRD, 2002).

i, me, my, myself, we, our, ours, ourselves, you, your, yours, yourself, yourselves, he, him, his, himself, she, her, hers, herself, it, its, itself, they, them, their, theirs, themselves, what, which, who, whom, this, that, these, those, am, is, are, was, were, be, been, being, have, has, had, having, do, does, did, doing, a, an, the, and, but, if, or, because, as, until, while, of, at, by, for, with, about, against, between, into, through, during, before, after, above, below, to, from, up, down, in, out, on, off, over, under, again, further, then, once, here, there, when, where, why, how, all, any, both, each, few, more, most, other, some, such, no, nor, not, only, own, same, so, than, too, very, s, t, can, will, just, don, should, now, d, ll, m, o, re, ve, y, ain, aren, couldn, didn, doesn, hadn, hasn, haven, isn, ma, mightn, mustn, needn, shan, shouldn, wasn, weren, won, and wouldn.

MACHINE LEARNING

In this appendix, we present the employed data mining techniques, which were mainly used for the task of classifying different networks. In the case of classification through complex networks, a possibility is to use the centrality measurements as features ([AMANCIO *et al.*, 2011](#); [AMANCIO; OLIVEIRA JR; COSTA, 2012a](#)). For instance, the average degree of the network nodes or other the other centralities. Observe that the vector that storages the measured characteristics is called *vector of characteristics*. Furthermore, the network-based features can include information different from the traditional measurements. So, there is the possibility to use the models and methods proposed in this thesis together with the traditional measurements.

C.1 Standardization

In order to give the same importance for attributes that are in various scales of magnitude, before attribute selection and classification tasks, data can be standardized. This process re-scales the distribution to have zero-mean and unit-variance and is computed as follows

$$X' = \frac{X - \langle X \rangle}{\sigma(X)}, \quad (\text{C.1})$$

where X is the original vector of features, and $\langle X \rangle$ and $\sigma(X)$ are average and standard deviations of X , respectively.

C.2 Attribute selection

There are many possibilities of attribute selection methods ([GUYON; ELISSEEFF, 2003](#)). In this section, we present the method employed in this thesis.

C.2.1 Information gain

Information Gain (BRAMER, 2007), IG, is based on the information theory. More details about information theory can be seen in Appendix E. IG is based on the concept of mutual information (COVER; THOMAS, 2012), which is defined from Shannon Entropy, $H(X)$, where X is a discrete distribution (see equation E.1). In order to compute information gain, $I(X, Y)$ is calculated (see equation E.4), in which X is the class of attributes, and Y is the vector of attributes. This method is computed for all of the analyzed classes. This method does not take into consideration that two or more distinct attributes can be used together in a classification. In other words, information gain quantifies the capability of each attribute separately.

C.2.2 SVM attribute selection

SVM attribute selection (GUYON *et al.*, 2002b) uses SVM classifier (see Section C.3.4) to evaluate which are the best attributes. In this case, the linear versions of SVM are employed, which consists of linear discriminant classifiers. In order to compute in problems of more than two classes, SVM attribute selection uses the *one-vs-all* solution to rank the attributes.

C.3 Classification

In order to test and compare the characteristics obtained from complex networks, we use some classifiers. These methods consist of two steps, the training, and test. In training, a set of data from a training set is employed to fit the classifier parameters according to their respective classes. Moreover, in the test, by considering the already trained classifier, the classification is executed, in which a data from the test set is classified as part of one of the classes. All the classifiers employed to this thesis are not sophisticated because we aim at testing the network-based features and not finding the best classifier.

C.3.1 Naive Bayes

With basis on Bayes theorem, the Naive Bayes (JOHN; LANGLEY, 1995) classifier was proposed. This classifier consists of creating a separated probability distribution for each of the characteristics. One possibility is to compute the average, $\langle X_c \rangle$, and standard deviation, σ_c , of all the measurements, X_c , for the classes separately. So, it is possible to define a Gaussian distribution, as follows

$$P(x = v|c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} \exp^{-\frac{(v-\langle X_c \rangle)^2}{2\sigma_c^2}}, \quad (\text{C.2})$$

where v is a given value used to calculate the probability. From this probability, the sample is classified as

$$\hat{y} = \arg \max_c P(c) \prod_{x \in X} P(x|c), \quad (\text{C.3})$$

where $P(c)$ is the probability of the class c happens. Variations of this method were proposed in the literature. One possibility is to consider a kernel density estimation (JOHN; LANGLEY, 1995) instead of a Gaussian distribution.

C.3.2 KNN

The method named as K nearest neighbors (KNN) (KANTARDZIC, 2011) consists of an approach that considers a given sample as part of the same class as their nearest neighbors. More specifically, for a given sample of the test set, this method consists of finding, in the training set, the K most adjacent samples. So, the sample is classified as the dominant class of these samples. For that, many measurements of distance can be employed. Commonly, the Euclidean distance is used.

C.3.3 CART

There are many algorithms of classification based on decision trees (ROKACH; MAIMON, 2014). One of them is the method of classification and regression trees (CART) (BREIMAN *et al.*, 1984). One advantage of using tree-based approaches is that the tree can also be used to better understand the data. This method is non-parametric and is based on a binary tree, which is generated by using “cuts” from information gain values, in the training step. In the classification step, the tree is crossed from the root to the leaves, in which the paths are defined by conditional questions for each of the nodes. According to the answer, the next step is determined.

Form the training data, the tree is built, where t_p is the parent node, and t_l and t_r are children of left and right, respectively, X is a matrix $m \times n$, in which m is the number of features and n is the number of samples, and $x_i \in X$, where $0 < i \leq n$. The classes with respect to the samples are stored in the vector Y , with n positions, and K defines the number of classes. For each node of the tree, a cut is done, which defines a neighbor to move. This cut is defined by a question, as follows: If the value is higher or equal to a given value, walk to t_r , otherwise to t_l . In order to define these questions, an impurity function $i(t)$ is defined, and its variation $\Delta i(t)$ is computed as

$$\Delta i(t) = i(t_p) - P_l i(t_l) - P_r i(t_r), \quad (\text{C.4})$$

where P_l and P_r are the probabilities to go to the left or to the right node, respectively. By considering the maximization of equation C.4, it is possible to predict which is the best question $x_j \leq x_j^R$, raying j from 1 to n , in which x_j^R is the best value of cut for x_j . One possibility of impurity function is to consider the information gain.

C.3.4 SVM

Support vector machine (SVM) (SMOLA; SCHÖLKOPF, 2004) consists of a classifier that considers the classes separated by a hyperplane. By considering the training set, a hyperplane is computed to provide the highest possible separation between the data. This information is then used to classify the test set. There are different possibilities to implement this classifier (SMOLA; SCHÖLKOPF, 2004). Here, we considered the Sequential Minimal Optimization (SMO) (PLATT, 1998; KEERTHI *et al.*, 2001), which reduces the necessary amount of used memory to be linear in the training step. Furthermore, SVM was developed to consider only two classes, but to classify samples that are organized into more than two classes, the strategy of pairwise classification is employed (HASTIE; TIBSHIRANI, 1998). Observe that, in this thesis, we use the implementation provided by Weka software (HALL *et al.*, 2009).

C.4 Data clustering

As well as the classifiers, the clustering algorithms separate the data into groups. However, they are based only on the information of the data. In other words, there is no step of training. Many different algorithms have been proposed in the literature. In the present thesis, we employ k-means and expectation maximization.

C.4.1 K-means

For a given set of characteristics, with N distinct samples, the algorithm of K -means clusters the data into K distinct groups (BOCK, 2007), in which $K < N$. This method consists of a minimization of the Euclidean distance between the samples in X and the centroids of the groups, respectively, which can be found by minimizing the equation

$$J = \sum_{i=1}^k \sum_{x \in s_i} \|x - c_i\|^2, \quad (\text{C.5})$$

where S is the set of groups $\{S_1, S_2, \dots, S_K\}$, x is the vector of characteristics of a given sample (that takes part in s_i), and c_i is the centroid of the group s_i (average value). Thus a method can be implemented through Algorithm 2.

C.4.2 Expectation maximization

Expectation maximization (EM) (MOON, 1996) is a method based on finding the maximum likelihood to estimate the parameters of a statistical model that uses a Gaussian mixture. The likelihood consists of the probability of some parameters of the method. Taking X as the set of N samples and θ as the vector of parameters to be optimized, this method maximizes the log-likelihood, $\log P(x; \theta)$, of $x \in X$. This method aims at finding

Algorithm 2 – K-means algorithm.

```

1: Data:  $K$  is the number of groups and  $X$  is a matrix with  $N$  samples;
2: Chose  $K$  centroids,  $\{c_1, c_2, \dots, c_K\}$ , as a random choice;
3: while centroids are updated do
4:   for  $i$  do  $i \leq N$ 
5:     Update the centroid that is nearest to the sample  $x_i$ ;
6:   end for
7:   for  $i$  do  $i \leq K$ 
8:     Update the value of the centroid  $c_i$  with the respective average values;
9:   end for
10: end while
11: Result:  $k$  groups are defined for the input data.

```

the k groups of samples, which are randomly selected to start the method. Furthermore, an initial estimative, θ , of the parameter vector is created. From this vector, for each iteration, the log-likelihood is computed. As the next step, the log-likelihood function of the probability distributions is maximized, and this step is repeated until the convergence of the method.

C.5 Quantification of relevance of the characteristics

The relevance of the characteristics can be quantified according to a method that follows. Being $F = \{f_1, f_2, \dots\}$ a set of Φ distinct attributes, the set F_c of all possible combinations of attributes is computed. In other words, F_c has 2^Φ features. For a given classifier, the accuracies of all combinations of features in F_c are computed. The elements of F_c are ordered in decreasing order according to their respective accuracies. For each attribute $f_i \in F$ a function is associated

$$O(f_i, k) = \sum_{j=1}^k o(f_i, j) \quad (\text{C.6})$$

where $o(f_i, j) = 1$ if the feature f_i takes part in $F_c(j)$, otherwise $o(f_i, j) = 0$. This function quantifies the frequency of the characteristics in the better classifications. In order to quantify the relevance of one characteristics, f_i , the following equation is defined

$$R(f_i) = \sum_{k=1}^{2^{\Phi-1}} O(f_i, k). \quad (\text{C.7})$$

One advantage of using this method is that the inter-relationship between the characteristics is considered (MANN; WHITNEY, 1947).

IMAGE ANALYSIS

In this appendix, we briefly present the methods presented in Section 3.5.3 of this thesis to extract image-based features. The features computed here are defined for binary images.

- **Area:** This measurement is defined as the number of pixels of the image;
- **Perimeter:** This measurement is defined as the arc length of the external border of an object of the image;
- **Euler number:** for a given image, Euler number is defined from the number of holes, which is computed as:

$$e = N_o - N_h, \quad (\text{D.1})$$

where N_o and N_h are the total number of objects and holes in the image, respectively;

- **Minimum enclosing circle:** The minimum enclosing circle comprises the smallest circle that includes all of the pixels of the object (SKYUM, 1991). From this method we considered x and y coordinates of the center of the circle, and the radius size;
- **Convex hull:** The minimum Convex hull comprises the smallest convex polygon that includes all of the pixels of the object (SKLANSKY, 1982). Here, we calculate the area C_a , perimeter C_p , and residual area C_r of the polygon. Observe that residual area is computed as follows

$$C_r = C_a - A. \quad (\text{D.2})$$

- **Elongation:** In order to quantify how stretched an object is, this measurement considers the two variances of the PCA of the object. More specifically, elongation (COSTA; CESAR, 2000) is defined as

$$El = \frac{\lambda_1}{\lambda_2}, \quad (\text{D.3})$$

where λ_1 and λ_2 are the variances of the PCA axes.

- **Lacunarity:** Lacunarity quantifies the variations of gaps in an image. The lower values of this measurement are found when the holes (gaps) are more uniform. In this thesis, we considered the version proposed by [Rodrigues, Barbosa and Costa \(2005\)](#), which uses a circular window of radius r to compute the measurement (instead of a squared window);
- **Fourier:** The Fourier transform consists in representing a signal (here it is an image) into the frequency domain. By considering this representation, in the magnitude Fourier space, we measured characteristics from non-overlapping rings with fixed width. The following measurements were computed for each ring: entropy, average, and standard deviation.

INFORMATION THEORY

Information theory was initially proposed by Claude E. Shannon (SHANNON, 1948) and comprises the compression and communication of information (COVER; THOMAS, 2012). This theory is critical to the development of our technology regarding digital media. In this appendix, we describe the information theory concepts employed in this thesis.

E.1 Entropy

One of the key points of the information theory is the measurement of entropy, which quantifies how organized a discrete distribution is. This measurement is defined as follows

$$H(x) = - \sum_{x \in X} p(x) \log p(x), \quad (\text{E.1})$$

where X is a discrete distribution and $p(x)$ is the probability function of an symbol x . Observe that the maximum and minimum values of entropy are obtained for X being a uniform distribution and with all elements having the same value, respectively. This concept can be extended to be associated with two discrete distributions, X and Y , and is called joint entropy $H(X, Y)$, defined as

$$H(X, Y) = - \sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(x, y), \quad (\text{E.2})$$

where X and Y are discrete distributions and $p(x, y)$ is the probability of x and y of occurring simultaneously.

Another possibility is to calculate the conditional entropy, $H(X|Y)$, which accounts for the entropy of a given discrete distribution, X , when the entropy of Y is known. This measurement is computed as follows

$$H(X|Y) = H(X, Y) - H(Y). \quad (\text{E.3})$$

When $H(X) \neq H(Y)$ then $H(X|Y) \neq H(Y|X)$.

E.2 Information gain

In order to account for a random variable by comparing with another random variable, it is possible to consider the information gain, which is defined as

$$I(X, Y) = H(Y) - H(Y|X). \quad (\text{E.4})$$

This measurement can also be normalized to result in values between 0 and 1, as follows

$$NMI(X, Y) = \frac{2I(X, Y)}{H(X) + H(Y)}. \quad (\text{E.5})$$

In this case, the logarithmic functions are calculated in base-2. Observe that *NMI* is employed to evaluate the quality of methods of cluster.

E.3 Huffman

To be stored and transmitted, messages can be encoded in sets of organized bits. For instance, a text can be stored as a sequence of symbols, which can be represented by characters of 8 bits. However, this sequence can be compressed if by considering varied sizes of the symbols. Such type of approach has been proposed in literature, which is based on information theory (SALOMON; MOTTA, 2010). One possibility of lossless compression algorithm is the Huffman code (COVER; THOMAS, 2012). The method represents the symbols with varied sizes. A dictionary is created, and the shorter and longer bit sequences are associated with more frequent and more rarely symbols, respectively. A binary tree is defined, whose leaves represent the symbols. Each edge of this tree is associated with a bit, usually left and right children are relative to the bits 0 and 1, respectively. The code of each symbol is associated with one leaf of the tree. More specifically, the respective code is obtained by concatenating all edge values (0 or 1) from the root to the leaf, which creates a binary code. This structure consists of a dictionary that is used to encode the original sequence of symbols. Observe that this dictionary is also used to decode the message. Interestingly, by considering entropy computed in a log with base-2, taking X as the sequence of symbols and R the rate of bits per symbol, then $H(X) \leq R < H(X) + 1$.

LANGUAGE COMPARISON: SYLLABLES-BASED APPROACH

By considering the framework proposed in the section 4.5, we analyze characteristics of language in terms of the level of compactness. First of all, the level of compactness of a language shall consider an scale of the analysis. Here we present a network-based methodology employed to analyze the relationship among syllables in a given language. By considering real characteristics of the languages, we also create stochastic versions, which are compact in terms of use of syllables. Furthermore, we compare the characterizations to find a relationship between real and the respective versions of artificial languages.

F.1 Syllable dataset

We use a syllable dataset of syllables obtained by the group of the professor Zoran Levnajić. This dataset consists in sets of words, syllables, and the respective frequencies, for 10 different languages, as follows: Basque, Croatian, Dutch, English, French, German, Latin, Russian, Slovenian, and Spanish. Observe that in these lexicons, words in all grammatical forms were considered. The data of Basque dataset is obtained from SYLLABARIUM database (DUÑABEITIA *et al.*, 2010). The Croatian lexicon was obtained by Meštrović, Martinčić-Ipšić and Matešić (2015), which comprises two lexicons RJEČNIK and SOBiR. Dutch, English and, German datasets were gathered from the from the Linguistic Data Consortium (CELEX2) database¹. The *online Lexique 3* database² was employed to created the French lexicon. In the case of Latin dataset, some books were chosen from *The Latin Library*³, as follows: De Oratore, books I to III, by Marcus Tullius Cicero, Commentarii de Bello Gallico books I to VIII, by Gaius Julius Caesar, Metamor-

¹ <<https://catalog.ldc.upenn.edu/ldc96l14>>

² <<http://www.lexique.org/>>

³ <<http://www.thelatinlibrary.com/>>

phoses books I to XV, by Publius Ovidius Naso, *De Monarchia*, by Dante Alighieri, *Liber de Regno Sicilie*, by Falcandus, and the Holy Bible (books Exodus, Genesis, Numerix, and Levitixus). In the case of Russian lexicon, the data was obtained from the Grammatical dictionary of Russian of Andrey Anatolyevich Zaliznyak. The Slovenian lexicon was gathered from the *Slovene morphological lexicon*⁴. The Spanish dataset was taken from *The Corpus of Contemporary Spanish*⁵. All the lexicons were normalized by considering the characteristics of Zipf's law. Because all the languages have similar histograms, by considering log-log distributions, through a visual inspection, the words with lower frequency were removed.

F.2 Language model

In order to create artificial distributions of words that represent characteristics of languages, we employed Simon's model (SIMON, 1955; FERRER I CANCHO, 2005). This model consists of creating a distribution of words that follow the Zipf's law, which is present in natural languages. Such a model is based on the repetitions of the most frequent words and starts with a single word. For each iteration, by considering a given probability ψ , a new word is added to the text, and with probability $1 - \psi$ a word that already takes part in the text is added. Observe that the probability of adding this new word follows the frequency of words in the text. In order to control the exponent of Zipf's law (β_S), the following equation is considered to find ψ

$$\beta_S = 1 + \frac{1}{1 - \psi}. \quad (\text{F.1})$$

Zipf's law represents this distribution, which follows the probability $P(f) \sim f^{-\beta_S}$, where f is the frequency of words and $\beta_S > 0$. We employ this methodology to create a text with the same number of words as the original, n_w dataset. For that, we the process of creating words finishes when the vocabulary size of the artificial language is $n_w + 1$, and the letter is removed. So the resultant artificial vocabulary has n words.

We employ the information of word size, regarding the number of syllables, obtained from the lexicon dataset (see Section F.1). In other words, these artificial languages are created by considering the original word sizes. The most frequent words have the lowest number of syllables. Starting from the small word of a given artificial version the real syllables are randomly assigned. More specifically, all the syllables of the dataset are stored in a shuffled list L . For each word the first x syllables of L are selected to compose the word, where x is the size of the word. If these syllables create a sequence equal to an already assigned word, the entire list is reshuffled and a new set of syllables is selected. With respect to each real language, we created a set of 30 artificial languages.

⁴ <<http://eng.slovenscina.eu/sloleks>>

⁵ <http://spanishfn.org/tools/cea/english>

F.2.1 Network model

We created syllables-based networks by considering the ten different employed languages. For each language, the network consists in nodes, that represent syllables, and the connections that are created when the pair of syllables co-occur at least in one word. The order of these pairs of syllables also defines the direction of the edges. For each edge, we considered the weight as the sum of all respective word frequencies. In other words, we took into consideration the frequency of co-occurrence in terms of the language. These networks are created for all of the considered languages and their respective artificial version. More specifically, we create 30 artificial-based networks for each language.

F.2.2 Results

First of all, we created one network for each of the considered real languages. Also, for each language, we created their respective 30 networks of the artificial version. By considering this networks, we computed several complex network measurements. For each network and for each measurement X (computed from network nodes), we considered average, $\langle X \rangle$, and standard deviation, $\sigma(X)$. The measurements are listed as follows:

- Degree: k ;
- Degree (in): k_{in} ;
- Degree (out): k_{out} ;
- Strength: s ;
- Strength (in): s_{in} ;
- Strength (out): s_{out} ;
- Clustering coefficient: cc ;
- Neighborhood: n ;
- Neighborhood (in): n_{in} ;
- Neighborhood (out): n_{out} ;
- Accessibility: α^2 and α^3 ;
- Generalized accessibility: α^∞ ;
- Backbone symmetry: Sb^2 , Sb^3 , and Sb^4 ;
- Merged symmetry: Sm^2 , Sm^3 , and Sm^4 ;

- Eccentricity: Ecc ;
- Eccentricity (in): Ecc_{in} ;
- Eccentricity (out): Ecc_{out} ;
- Closeness centrality: C ;
- Closeness centrality (in): C_{in} ;
- Closeness centrality (out): C_{out} ;
- Eigenvector centrality (directed): EC^{dir} ;
- Eigenvector centrality (undirected): EC^{un} ;
- Betweenness centrality (directed): B^{dir} ;
- Betweenness centrality (undirected): B^{un} ;
- Edge betweenness centrality (directed): EB^{dir} ;
- Edge betweenness centrality (undirected): EB^{un} ;
- Modularity: Q .

Observe that modularity, Q , is a global measurement. As a consequence, statistics were not computed from Q . We plot a PCA projection by considering all networks and all of these measurements, as shown in Figure 69(a). Observe that two main groups are formed, which represents networks created from real and artificial languages. Interestingly, we found that the standard deviations of the measurements were less determinant to separate the groups than the averages (see Figures 69(b) and (c)).

In order to better understand the data presented in PCA projection shown in Figure 69(a), we computed the distances between samples, see Figure 70. In the case of the samples obtained from artificial languages, for each group, we considered the average position. These distances were also employed to calculate a dendrogram, which was computed by considering Ward's method (WARD JR, 1963). This method consists of a minimum variance method, which is computed from a recursive function that is optimized. Two predominant groups are formed for real and artificial languages, except for the artificial version of English that is part of the group of real languages. Additionally, inside of these groups, there are also remarkable subgroups. For both real and artificial cases, two of the Germanic languages, Dutch and German, are found to be near, which are followed by the respective versions of French. Both versions of English are also near to this group. Interestingly, English is a Germanic language that incorporated many cognate words from French (Romance language). The remaining real languages are also grouped, in which Basque is

Figure 69 – PCA projection computed from network measurements obtained from real and artificial languages (30 artificial samples for each real network). The circles and squares represent artificial and real languages, respectively. More specifically, item (a) represents the PCA of all considered measurements, items (b) and (c) represent the PCA only with the standard deviation of the measurements and with the remaining measurements (without the standard deviation), respectively.

Source: Elaborated by the author.

fund to be the farthest language. By considering the remaining artificial versions, Basque and Latin is the nearest pair of languages. Furthermore, the Slavic languages are grouped (Russian, Slovenian, and Croatian). Conversely, Spanish, which is a Romance language, also takes part in this group.

We also considered the relationships of languages by computing the cosine distances and considering all computed measurements. Figure 71 shows the heatmap of distances and the respective dendrogram. In this case, other characteristics were observed. First, the group of real languages is well defined. In contrast with the previous results (shown in Figure 70) this group is formed only by real languages. Also considering the real case, Basque and English were found to be the languages with the farthest distances when compared to the other real languages. Additionally, the artificial version of Basque is nearest to the majority of the real languages, which includes the real case of Basque.

Figure 70 – The heatmap represents the Euclidean distances between networks created from all (real and artificial) networks, by considering the two first dimensions of PCA. The character star represents the artificial languages. The dendrogram and the matrix order are computed from the heatmap matrix by considering Ward’s method.

Source: Elaborated by the author.

Remarkably, the pairs of French and English are near for both artificial and real languages. From this result, we believe that the influence of one language in another can be reflected in their structures of syllables.

In the following, we analyzed the individual characteristics of each measurement. For that, we compare the network measurements obtained from artificial with real languages, as shown in Figure 72. This analysis consists of computing the *z-score*, also known as the standard score, which is computed as follows

$$z(X) = \frac{x - \langle X^{(A)} \rangle}{\sigma(X^{(A)})}, \quad (\text{F.2})$$

where A is a given language, x is the measurement value obtained from the real language-based network, $X^{(A)}$ is the set of measurements obtained from A , $\langle X^{(A)} \rangle$ is the average of $X^{(A)}$, and $\sigma(X^{(A)})$ represents the standard deviation of $X^{(A)}$. This measurement accounts for the distance of a given measurement between both versions concerning the standard

Figure 71 – The heatmap represents the cosine distances between networks created from all (real and artificial networks) by considering all employed network measurements. The character star represents the artificial languages. The dendrogram and the matrix order are computed from the heatmap matrix by considering Ward's method.

Source: Elaborated by the author.

deviation. In general, the values obtained from real languages tend to be far from the respective artificial-based measurements. In contrast with the results presented in Figures 69(b) and (c), the measurements that are better described by the artificial model are not only the standard deviations. Furthermore, the measurements that discriminate the most between the two groups are based on network degree. So, the optimization regarding the word frequency distributions also influenced the pattern of syllables organization. The node degrees have a strong relationship with almost all remaining measurements. By comparing among the languages, in few cases, the computed z-score have a high variation.

Figure 72 – For all the considered measurements, the module of z-score computed from the measurement obtained from the real and the group of artificial language networks. The measures are ordered in increasing order, according to the average z-scores of all languages.

Source: Elaborated by the author.