GHENT UNIVERSITY

Faculty of Bioscience Engineering

UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Spatially explicit modeling on networks: understanding patterns & describing processes

Gisele Helena Barboni Miranda

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: _____

Gisele Helena Barboni Miranda

Spatially explicit modeling on networks: understanding patterns & describing processes

Doctoral dissertation submitted to the Institute of Mathematics and Computer Sciences - ICMC-USP and to the Faculty of Bioscience Engineering - Ghent University - UGent, in partial fulfillment of the requirements for the degree of the Doctorate Program in Computer Science and Computational Mathematics (ICMC-USP) and Doctor of Applied Biological Sciences (UGent), in accordance with the international academic agreement for Ph.D. double degree signed between ICMC-USP and UGent. FINAL VERSION

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Odemir Martinez Bruno (ICMC-USP, Brazil)

Advisor: Prof. Dr. Bernard De Baets (UGent, Belgium)

Co-advisor: Prof. Dr. ir. Jan Marcel Baetens (UGent, Belgium)

USP – São Carlos
July 2019

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

M672s Miranda, Gisele Helena Barboni
 Spatially explicit modeling on networks:
 understanding patterns & describing processes /
 Gisele Helena Barboni Miranda; orientador Odemir
 Martinez Bruno; coorientador Bernard De Baets. --
 São Carlos, 2019.
 300 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. Network Characterization. 2. Network
 Descriptor. 3. Cellular Automata. 4. Spatio-
 temporal Pattern. 5. Pattern Recognition. I. Bruno,
 Odemir Martinez, orient. II. De Baets, Bernard ,
 coorient. III. Título.

Gisele Helena Barboni Miranda

**Modelagem espacialmente explícita em redes:
compreendendo padrões e descrevendo processos**

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP e à *Faculty of Bioscience Engineering - Ghent University* - UGent, como parte dos requisitos para obtenção do título de Doutora em Ciências - Ciências de Computação e Matemática Computacional (ICMC - USP) e Doutora em Ciências Biológicas Aplicadas, de acordo com o convênio acadêmico internacional para dupla titulação de doutorado assinado entre o ICMC-USP e a UGent. VERSÃO REVISADA

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. Odemir Martinez Bruno (ICMC-USP, Brazil)

Orientador: Prof. Dr. Bernard De Baets (UGent, Belgium)

Coorientador: Prof. Dr. ir. Jan Marcel Baetens (UGent, Belgium)

**USP – São Carlos
Julho de 2019**

To my family, for being the first supporters of my career.

ACKNOWLEDGEMENTS

I cannot be more grateful to the people who have inspired, supported and helped me during this important step of my life. My Ph.D. journey would have been much more difficult without your presence!

I want to express my sincere thanks to my three supervisors. At USP, I want to thank Odemir for his guidance along the whole Ph.D. The first discussions about this thesis proposal, our conversations about science and his support in all the steps of my Ph.D. will be very good memories of this period. I want to thank Bernard and Jan for their guidance. They showed me a new horizon for this thesis and contributed immensely to my Ph.D. proposal. All the three are examples of great researchers for me and I am very proud to have worked with them. In addition, I am very thankful for all their efforts to transform this collaboration in a joint Ph.D.

I am thankful to the Institute of Mathematics and Computer Sciences and the São Carlos Physics Institute for providing the necessary infrastructure for the development of this project. I am also very grateful for being part of University of São Paulo since my Bachelor. I want to thank all the professors that have taught me so much along all these years. I am also thankful to Ghent University (UGent) and to the Faculty of Bioscience Engineering. The opportunity to develop part of my thesis at UGent has enriched a lot my research experience. In addition, I want to thank all the staff members from both USP and UGent that helped us so much with all the administrative procedures required for this joint Ph.D., especially professor Adenilso Simão, Aline Martinez and Marleen Vereecken. Also, special thanks to Livia Ricci, for all her support at the Physics Institute in São Carlos, even when I was not in Brazil.

I am also very grateful to Nathalie Bossuyt and Sciensano for providing the data concerning the incidence of influenza cases in Belgium, and, to the group of Prof. Dr. Bernd Hellingrath at Münster University. During a research visit, I could exchange ideas about epidemic modeling and simulation with his group. Moreover, part of the simulations presented in this thesis were carried out using the High Performance Computing (HPC) infrastructure at UGent.

I want to thank my parents, Renato and Gilda, for being my first inspiration and for always motivating me to go ahead with my dreams. They are my greatest example. I also want to thank my sister, Renata, for being my best friend and a person with whom I can always share my dreams. Also, to all my other relatives for being an inspiration and so supportive along these years. A very special thanks to Sina. During my Ph.D. I had the opportunity to meet this wonderful person with whom I am planning our future steps. Thank you for all your support!

Along the way to the conclusion of this Ph.D. I could meet and work with many colleagues that made all this journey much more pleasurable. I am very thankful to all my colleagues of the Scientific Computing Group (SCG), which inspired me with their work, support and friendship: João, Núbia, Humberto, Mariane, Lucas, Ângela, Marcela, Alex, Rayner, Leonardo, Marcos, Jarbas and the new members that are arriving at our group, welcome! A very special thanks to Jeaneth, with whom I could closely collaborate in some parts of my Ph.D. project with fruitful results. I also want to express all my gratitude to the colleagues I have made at Ghent University both from Kermit and Biomath groups: José, Wouter, Aisling, Christina, Raúl, Andreia, Niels, David, Dailé, Lisellote, Bac, Shuyun, Wenwen, Mengzi, Michael, Kamal and many others. I have very good memories from you all!

Finally, I want to thank the funding support of the following agencies and institutions, that were essential for the development of this work: the Coordination of Improvement of Higher Education Personnel (CAPES), the São Paulo Research Foundation - FAPESP (grants #15/05899-7 and #18/00147-5), and, Ghent University (BOF grant #01SF0516).

*“We can’t solve problems by using the same
kind of thinking we used when we created them.”
(Albert Einstein)*

ABSTRACT

In contrast to established approaches that analyze networks based on their structural properties, networks can also be studied by investigating the patterns that are evolved by a discrete dynamical system built upon them, such as cellular automata (CAs). Combined with networks these tools can be used to map the relationship between the network architecture and its impact on the patterns evolved by the governing spatially discrete dynamical system. This thesis focuses on the investigation of discrete spatially explicit models (SEMs), among which are CAs, for network analysis and characterization. The relationship between network architecture and its dynamic aspects concerning pattern formation is studied. Additionally, this work aims at the development of evolutionary methods that can be employed for extracting features from such patterns and then be used as network descriptors. In order to achieve this goal, methods that integrate the network structure with the SEMs were proposed, implemented and analyzed. The proposed family of network automata is characterized by birth-survival dynamics that results in different categories of spatio-temporal patterns. Such patterns were quantitatively assessed and used to characterize different network topologies and perform classification tasks in the context of pattern recognition. Inspired by the classic Life-like CA, the proposed Life-like Network Automata (LLNA) illustrate how such tasks can be performed in real-world applications. In addition, the rock-paper-scissors (RPS) model, normally implemented on square lattices, was investigated by defining it on networks. The obtained results confirm the potential of the proposed quantitative analysis of the spatio-temporal patterns for network classification. This quantitative analysis was performed for a set of different pattern recognition tasks and for the majority of them, the classification performance improved. In addition, the reliability of LLNA as a general tool for pattern recognition applications was demonstrated in a diverse scope of classification tasks. The applicability of structural network descriptors was also highlighted in the context of shape characterization in computer vision. Through the proposed approach, the link between these network descriptors and the shape properties, such as angle and curvature, was illustrated. Moreover, when chosen adequately, the network descriptors led to a better classification performance for different shape recognition tasks. Regarding the RPS model, we demonstrated that the presence of long-range correlations in some networks directly influence the RPS dynamics. Finally, it was shown how a commuter network can be used to predict influenza outbreaks. All the proposed methods use different aspects of network analysis and contribute to the study of CAs and other SEMs on irregular tessellations, in contrast to the commonly used regular topologies. In addition, new insights were obtained concerning pattern recognition in networks through the use of spatio-temporal patterns as network descriptors.

Keywords: Network Characterization, Network Descriptor, Cellular Automata, Spatio-temporal Patterns, Pattern Recognition.

RESUMO

Em contraste às abordagens clássicas que analisam redes com base em suas propriedades estruturais, as redes também podem ser estudadas investigando-se os padrões desenvolvidos por um sistema dinâmico discreto construído sobre essas redes, como os autômatos celulares (CAs). Combinadas às redes, essas ferramentas podem ser usadas para se mapear a relação entre a arquitetura da rede e seu impacto nos padrões obtidos pelo sistema dinâmico subjacente. Esta tese está focada na investigação de modelos discretos espacialmente explícitos (SEMs), entre os quais os CAs, para análise e caracterização de redes. A relação entre a arquitetura da rede e seu aspecto dinâmico em relação à formação de padrões é investigada. Além disso, este trabalho visa o desenvolvimento de métodos evolutivos que podem ser usados para extrair características de tais padrões para, então, serem usados como descritores de redes. Para atingir este objetivo, métodos que integram a estrutura da rede com os SEMs foram propostos, implementados e analisados. A família de redes-autômatos proposta é caracterizada por uma dinâmica de nascimento-sobrevivência que resulta em diferentes categorias de padrões espaço-temporais. Tais padrões foram avaliados quantitativamente e utilizados para caracterizar diferentes topologias de redes e realizar tarefas de classificação no contexto do reconhecimento de padrões. Inspirados pelo clássico *Life-Like CA*, a rede-autômato proposta, *Life-like (LLNA)*, ilustra como tais tarefas podem ser realizadas em aplicações mais realistas. Além disso, o modelo de *rock-paper-scissors (RPS)*, normalmente implementado em reticulados quadrados, foi investigado usando-se redes como tesselações. Os resultados obtidos confirmam o potencial da análise quantitativa proposta dos padrões espaço-temporais para classificação de redes. Essa análise quantitativa foi realizada para um conjunto de tarefas de reconhecimento de padrões, e, para a maioria dessas tarefas, o desempenho da classificação melhorou. Além disso, a confiabilidade do LLNA como uma ferramenta genérica para reconhecimento de padrões foi demonstrada para várias tarefas de classificação de diferentes escopos. A aplicabilidade de descritores estruturais de redes também foi destacada no contexto de caracterização de formas em visão computacional. Através da abordagem proposta, a ligação entre esses descritores de rede e as propriedades da forma, como ângulo e curvatura, foi ilustrada. Além disso, quando escolhidos adequadamente, os descritores de rede levam a um melhor desempenho de classificação para diferentes tarefas de categorização de formas. No que diz respeito ao modelo RPS, demonstramos que a presença de correlações de longo alcance nas redes afeta diretamente a dinâmica do modelo. Finalmente, foi apresentado como uma rede de transporte pode ser usada para prever surtos de gripe. Todos os métodos propostos utilizam diferentes aspectos da análise de redes e contribuem para o estudo de CAs e outras SEMs em tesselações irregulares, uma vez que estes modelos são geralmente descritos em topologias regulares. Além disso, uma nova metodologia foi proposta em relação ao reconhecimento de padrões em redes através do uso de padrões espaço-temporais como descritores da rede.

Palavras-chave: Caracterização de Rede, Descritor de Rede, Autômato Celular, Padrões Espaço-Temporais, Reconhecimento de Padrões.

ABSTRACT

In tegenstelling tot gevestigde technieken die netwerken analyseren op basis van hun structurele eigenschappen, kunnen deze laatste ook worden bestudeerd door de patronen te onderzoeken die bekomen worden met een discreet dynamisch systeem dat op deze netwerken is gebouwd, zoals cellulaire automaten (CA's). In combinatie met netwerken kunnen deze methodes worden gebruikt om de relatie in kaart te brengen tussen de netwerkachitectuur en hun impact op de patronen die zijn ontwikkeld door het onderliggende ruimtelijk discrete dynamische systeem. Dit proefschrift richt zich op het onderzoek van discrete ruimtelijk expliciete modellen (SEM's), waaronder CA's, voor netwerkanalyse en -karakterisatie. De relatie tussen netwerkachitectuur en de dynamische aspecten ervan met betrekking tot patroonvorming wordt bestudeerd. Bovendien richt dit werk zich op de ontwikkeling van evolutionaire methoden die kunnen worden gebruikt voor het extraheren van kenmerken uit dergelijke patronen, die vervolgens kunnen worden gebruikt als netwerkkarakteristiek. Om dit doel te bereiken werden methoden die de netwerkstructuur integreren met de SEM's voorgesteld, geïmplementeerd en geanalyseerd. De voorgestelde familie van netwerkautomaten wordt gekenmerkt door dynamica van geboorte-overleving die resulteert in verschillende categorieën van spatio-temporele patronen. Dergelijke patronen werden kwantitatief beoordeeld en gebruikt om verschillende netwerktopologieën te karakteriseren en classificatietaken uit te voeren in de context van patroonherkenning. Geïnspireerd door de klassieke Life-Like CA, illustreren de voorgestelde Life-Like Network Automata (LLNA) hoe dergelijke taken kunnen worden uitgevoerd in realistische applicaties. Daarnaast werd het RPS-model (Rock-Paper-Scissors), normaal geïmplementeerd op vierkante roosters, onderzocht door het op netwerken te definiëren. De bekomen resultaten bevestigen het potentieel van de voorgestelde kwantitatieve analyse van de spatio-temporele patronen voor netwerkclassificatie. Deze kwantitatieve analyse werd uitgevoerd voor een set van verschillende patroonherkenningstaken en voor de meeste van deze taken verbeterden de classificatieresultaten. Bovendien werd de betrouwbaarheid van LLNA als algemeen hulpmiddel voor patroonherkennings toepassingen aangetoond in een diverse waaier aan classificatietaken. De toepasbaarheid van structurele netwerkkarakteristieken werd ook benadrukt in de context van vormkarakterisering in computervisie. Met de voorgestelde aanpak werd de link geïllustreerd tussen deze netwerkkarakteristieken en de vormeigenschappen, zoals hoek en kromming. Bovendien leiden de netwerkkarakteristieken, indien adequaat gekozen, tot een betere classificatie voor verschillende taken van vormherkenning. Voor wat betreft het RPS-model hebben we aangetoond dat de aanwezigheid van langeafstandscorrelaties in sommige netwerken de RPS-dynamiek rechtstreeks beïnvloedt. Ten slotte werd getoond hoe een pendelaarsnetwerk kan worden gebruikt om uitbraken van influenza te voorspellen. Alle voorgestelde methoden gebruiken verschillende aspecten van netwerkanalyse en dragen bij tot de studie van CA's en andere SEM's op onregelmatige tessellaties, meestal beschreven voor reguliere topologieën. Daarnaast werden nieuwe inzichten verkregen over patroonherkenning in netwerken door het gebruik van spatio-temporele patronen als netwerkkarakteristieken.

Sleutelwoorden: Netwerkkarakterisatie, Netwerkkarakteristieken, Cellulaire Automaten, Spatio-temporele Patronen, Patroonherkenning.

LIST OF FIGURES

Figure 1 – Thesis organization per topic.	39
Figure 2 – The graph of the Königsberg bridges	44
Figure 3 – Graphical representations of different types of graphs and their corresponding adjacency matrices.	47
Figure 4 – Random networks: (a) Network with $\langle k \rangle = 8$. (b) Degree distribution	48
Figure 5 – Network evolution as a function of the rewiring probability p	50
Figure 6 – Clustering coefficient $\langle cc \rangle$ and average path length $\langle l \rangle$ as a function of the rewiring probability p	50
Figure 7 – Scale-free networks. (a) Presence of <i>hubs</i> . (b) Degree distribution	51
Figure 8 – Feature extraction from the network	58
Figure 9 – The most common neighborhoods of 2-dimensional CAs with radius = 1: (a) <i>von Neumann</i> and (b) <i>Moore</i>	65
Figure 10 – Rule 90: Illustration of a transition function of the ECA.	66
Figure 11 – ECA classification: (a) stable, rule 4; (b) periodic, rule 51; (c) chaotic, rule 30, (d) complex, rule 110.	67
Figure 12 – Rule 777: Illustration of a transition function of a totalistic one-dimensional CA	68
Figure 13 – Patterns obtained through the evolution of one-dimensional CAs for different values of m and different rules	69
Figure 14 – Patterns obtained through the evolution of life-like CAs for six different rules.	71
Figure 15 – All possible ways that a single defect (black circle) can propagate in the evolution space of an ECA (arrows).	75
Figure 16 – Evolution dynamics of a sample network according to Eq. (5.2).	93
Figure 17 – Spatio-temporal patterns for three different network types obtained by applying the neighborhood density NA	94
Figure 18 – Average Shannon entropy as a function of the parameter δ for different network models: random, small-world, scale free and geographical.	95
Figure 19 – Classification accuracy (%) of network models as a function of the parameter δ	96
Figure 20 – Illustration of density intervals defined by x_a, x_b and y_a, y_b	98
Figure 21 – Illustration of the birth-survival transition function for Rule B33/S23.	99
Figure 22 – Spatio-temporal patterns obtained through the evolution of the BSNA for $R = 2$ and for two different tessellations	101
Figure 23 – Spatio-temporal patterns for obtained through the evolution of the BSNA for $R = 3$ and using a regular network as the NA tessellation.	102

Figure 24 – Spatio-temporal patterns for obtained through the evolution of the BSNA for $R = 3$ and using a random network as the NA tessellation.	103
Figure 25 – Average Lyapunov exponent (LE) over 50 simulations of the BSNA on (a) a regular network and on (b) a random network	105
Figure 26 – Average Lempel-Ziv complexity over 50 simulations of the BSNA on (a) a regular network and on (b) a random network	106
Figure 27 – Average word length complexity over 50 simulations of the BSNA on (a) a regular network and on (b) a random network	108
Figure 28 – Average Lyapunov exponent over 50 simulations of the BSNA on a regular network	109
Figure 29 – Average Lyapunov exponent over 50 simulations of the BSNA on a random network	110
Figure 30 – Illustration of density intervals for the generalized birth-survival NA	114
Figure 31 – The rule space as a function of the parameter R	115
Figure 32 – Comparison between the Moore neighborhood for the classic Life-like CA and the intervals of the neighborhood density used by LLNA	117
Figure 33 – Space-time diagrams for different network models	119
Figure 34 – Entropy ($\vec{\mu}_S$), word length ($\vec{\mu}_W$) and Lempel-Ziv complexity ($\vec{\mu}_L$) histograms for different network models.	120
Figure 35 – Accuracy of network classification as a function of the number of time steps and the initial configuration	121
Figure 36 – Histograms of the Shannon entropy distribution for each network model and for different values of the number of nodes	121
Figure 37 – Comparison between BSNA and GenBSNA concerning the classification accuracy of the network models for different values of R	122
Figure 38 – Accuracy distribution obtained by the BSNA for different rule spaces defined by R	123
Figure 39 – Accuracy distribution obtained with the GenBSNA for different rule spaces defined by R	124
Figure 40 – Sensitivity to the initial distribution of states	126
Figure 41 – Regular grid illustrating a sample configuration of the RPS.	131
Figure 42 – Representation of the interactions in the stochastic version of the RPS model	131
Figure 43 – Effect of mobility on biodiversity. Probability of extinction as a function of mobility for different grid sizes.	135
Figure 44 – Random rewiring of the links of a grid, which is controlled by Q , also referred to as quenched randomness	137
Figure 45 – Illustration of the rock-paper-scissors game on an irregular topology.	137
Figure 46 – Probability of extinction as a function of mobility for scale-free networks	139
Figure 47 – Probability of extinction as a function of mobility for regular networks	140

Figure 48 – Average time to first extinction as a function of mobility for scale-free networks	141
Figure 49 – Average time to first extinction as a function of mobility for regular networks	142
Figure 50 – Average cluster size per Monte Carlo generation for scale-free networks with $N = 400$	143
Figure 51 – Average cluster size per Monte Carlo generation for scale-free networks with $N = 1600$	144
Figure 52 – Average cluster size per Monte Carlo generation for regular networks with $N = 1600$	145
Figure 53 – Average number of clusters per Monte Carlo generation for scale-free networks with $N = 1600$	146
Figure 54 – Pattern recognition in networks using the Life-like Network Automata.	152
Figure 55 – Training and test strategies of LLNA used in pattern recognition applications.	155
Figure 56 – Accuracy obtained for all the LLNA rules regarding the correct classification of network models	158
Figure 57 – Synthetic network characterization with LLNA for classifying network models	159
Figure 58 – Synthetic network characterization with LLNA for classifying network models in combination with $\langle k \rangle$	161
Figure 59 – Confusion matrix for rule $B01678/S0457$ using the following feature vector $[\langle \mu_S \rangle, \langle \mu_W \rangle, \langle \mu_L \rangle]$. Each cell corresponds to the percentage (%) of correct classified instances, where filled cells represent non-zero elements.	163
Figure 60 – Canonical analysis for each network model given the distinct values of $\langle k \rangle$ as classes	164
Figure 61 – Synthetic network characterization with LLNA for classifying scale-free models	164
Figure 62 – Comparison between the accuracy obtained with LLNA and the accuracy obtained with structural network measures as feature vector	165
Figure 63 – Characterization of real-world applications with LLNA using the <i>metabolic-dataset</i>	168
Figure 64 – Characterization of real-world applications with LLNA using the <i>social-dataset</i>	171
Figure 65 – Comparison between BSNA and GenBSNA concerning the classification accuracy of the <i>social-dataset</i> for different values of the parameter R	172
Figure 66 – Stomata distributions of the <i>Tradescantia zebrina</i> leaves and the corresponding networks for the three lighting conditions and four δ_T values.	173
Figure 67 – Characterization of real-world applications with LLNA using the <i>stomata-dataset</i>	174
Figure 68 – Fungal growth dynamics shown for the <i>Rhizoctonia solani</i> species given $RH = 65\%$ and $T = 15^\circ$	176
Figure 69 – Average degree, average hierarchical degree of level 2, $\langle k_2 \rangle$, and, average hierarchical degree of level 3, $\langle k_3 \rangle$, for each species	178

Figure 70 – Average degree, average hierarchical degree of level 2, $\langle k_2 \rangle$, and, average hierarchical degree of level 3, $\langle k_3 \rangle$, for each species	179
Figure 71 – Sampling of the <i>fungi-dataset</i> and the corresponding classification tasks . . .	180
Figure 72 – Construction of a network from the shape contour	186
Figure 73 – Comparison between curvature signal (dashed line) and the degree of each node as a function of T_l for different shapes.	188
Figure 74 – Comparison between curvature signal (dashed line) and other structural measures	189
Figure 75 – Interpolation analysis between different shapes.	191
Figure 76 – Average clustering coefficient ($\langle cc \rangle$) as a function of T_l for different geometric shapes.	191
Figure 77 – Examples of images belonging to the three datasets evaluated in this section. (a) <i>geom-shapes</i> , (b) <i>generic-shapes</i> and (c) <i>leaves</i> dataset.	193
Figure 78 – Classification accuracy (%) and standard deviation using the proposed shape descriptor (Φ) and different classifiers	194
Figure 79 – Classification accuracy (%) and standard deviation using the proposed shape descriptor (Φ) for a single T_l	195
Figure 80 – Transformations applied to the <i>leaves</i> dataset dataset	197
Figure 81 – Color-map representation of the degree (k) of the pixels belonging to the shape contour for the <i>leaves</i> dataset and for $T_l = 0.325$	198
Figure 82 – Color-map representation of the clustering coefficient (cc) of the pixels belonging to the shape contour for the <i>leaves</i> dataset and for $T_l = 0.325$. . .	199
Figure 83 – Color-map representation of the betweenness centrality (b) of the pixels belonging to the shape contour for the <i>leaves</i> dataset and for $T_l = 0.325$. . .	200
Figure 84 – Classification accuracy for different degradation levels.	201
Figure 85 – Normalized weekly incidence of influenza-like illness (ILI) cases in Belgium for the seasons 2010–2011 to 2015–2016. The normalized ILI data are shown from week 35 of year i to week 34 of year $i + 1$	207
Figure 86 – Normalized weekly incidence of ILI cases (obs. ILI cases) and the SIR model solution representing the simulated counterpart (sim. ILI cases) for the season 2010–2011.	209
Figure 87 – Proposed approach for real-time prediction of ILI cases.	210
Figure 88 – Weekly ILI incidence per week for every season starting from the onset (week 0) given in Table 21.	211
Figure 89 – Construction of the confidence region	212
Figure 90 – Selection of acceptable model parameters	214
Figure 91 – Predicted numbers of ILI cases (gray) throughout season 2010–2011 for different starting weeks of the simulated period	214

Figure 92 – Simulated (gray) and normalized (black) numbers of ILI cases in Belgium for season 2010–2011	216
Figure 93 – Simulated (gray) and normalized (black) numbers of ILI cases in Belgium for season 2014–2015	217
Figure 94 – MAE, RMSE and correlation between the normalized and the predicted numbers of ILI cases per season for different time lags.	219
Figure 95 – Administrative districts of Belgium and the commuting network	220
Figure 96 – Strength of the districts	220
Figure 97 – Weekly ILI incidence in the Belgian districts for the six analyzed seasons: 2010–2011 up to 2015–2016.	222
Figure 98 – Weekly ILI incidence in the Belgian districts for the six analyzed seasons: 2010–2011 up to 2015–2016.	223
Figure 99 – Simulated and normalized numbers of ILI cases in Brussels for season 2010– 2011 up to 2012–2013 obtained without using commuting data	225
Figure 100–Simulated and normalized numbers of ILI cases in Brussels for season 2013– 2014 up to 2015–2016 obtained without using commuting data	226
Figure 101–Simulated and normalized numbers of ILI cases in Brussels for season 2010– 2011 up to 2012–2013 obtained using commuting data	227
Figure 102–Simulated and normalized numbers of ILI cases in Brussels for season 2013– 2014 up to 2015–2016 obtained using commuting data	228
Figure 103–MAE and RMSE between the normalized and the predicted numbers of ILI cases per season for different time lags	231
Figure 104–Average LE over 50 simulations of the BSNA on a small-world network . . .	263
Figure 105–Average LE over 50 simulations of the BSNA on a scale-free network . . .	264
Figure 106–Average LE over 50 simulations of the BSNA on a geographical network . . .	264
Figure 107–Average LZ over 50 simulations of the BSNA on a small-world network . . .	265
Figure 108–Average LZ over 50 simulations of the BSNA on a scale-free network . . .	265
Figure 109–Average LZ over 50 simulations of the BSNA on a geographical network . . .	266
Figure 110–Average WL over 50 simulations of the BSNA on a small-world network . . .	266
Figure 111–Average WL over 50 simulations of the BSNA on a scale-free network . . .	267
Figure 112–Average WL over 50 simulations of the BSNA on a geographical network . . .	267
Figure 113–Average LE over 50 simulations of the BSNA on a small-world network . . .	268
Figure 114–Average LE over 50 simulations of the BSNA on a scale-free network . . .	269
Figure 115–Average LE over 50 simulations of the BSNA on a geographical network . . .	270
Figure 116–Average LZ over 50 simulations of the BSNA on a regular network	271
Figure 117–Average LZ over 50 simulations of the BSNA on a random network	272
Figure 118–Average LZ over 50 simulations of the BSNA on a small-world network . . .	273
Figure 119–Average LZ over 50 simulations of the BSNA on a scale-free network . . .	274
Figure 120–Average LZ over 50 simulations of the BSNA on a geographical network . . .	275

Figure 121–Average WL over 50 simulations of the BSNA on a regular network	276
Figure 122–Average WL over 50 simulations of the BSNA on a random network	277
Figure 123–Average WL over 50 simulations of the BSNA on a small-world network . .	278
Figure 124–Average WL over 50 simulations of the BSNA on a scale-free network . . .	279
Figure 125–Average WL over 50 simulations of the BSNA on a geographical network .	280
Figure 126–Probability of extinction as a function of mobility for random networks . . .	282
Figure 127–Probability of extinction as a function of mobility for geographic networks .	283
Figure 128–Probability of extinction as a function of mobility for small-world networks	283
Figure 129–Average time to first extinction as a function of mobility for random networks	284
Figure 130–Average time to first extinction as a function of mobility for geographical networks	285
Figure 131–Average time to first extinction as a function of mobility for small-world networks	285
Figure 132–Simulated and normalized numbers of ILI cases in Belgium for season 2011– 2012	288
Figure 133–Simulated and normalized numbers of ILI cases in Belgium for season 2012– 2013	289
Figure 134–Simulated and normalized numbers of ILI cases in Belgium for season 2013– 2014	290
Figure 135–Simulated and normalized numbers of ILI cases in Belgium for season 2015– 2016	291
Figure 136–Simulated and normalized numbers of ILI cases in Antwerpen district for season 2010–2011 up to 2012–2013	293
Figure 137–Simulated and normalized numbers of ILI cases in Antwerpen district for season 2013–2014 up to 2015–2016	294
Figure 138–Simulated and normalized numbers of ILI cases in Gent district for season 2010–2011 up to 2012–2013	295
Figure 139–Simulated and normalized numbers of ILI cases in Gent district for season 2013–2014 up to 2015–2016	296
Figure 140–Simulated and normalized numbers of ILI cases in Antwerpen district for season 2010–2011 up to 2012–2013 using commuting data	297
Figure 141–Simulated and normalized numbers of ILI cases in Antwerpen district for season 2013–2014 up to 2015–2016 using commuting data	298
Figure 142–Simulated and normalized numbers of ILI cases in Gent district for season 2010–2011 up to 2012–2013 using commuting data	299
Figure 143–Simulated and normalized numbers of ILI cases in Gent district for season 2013–2014 up to 2015–2016 using commuting data	300

LIST OF TABLES

Table 1	– Set of rules of the birth-survival NA belonging to the rule space defined by $R = 2, R = 3$ and $R = 4$	100
Table 2	– Generalized birth-survival NA for $R = 2, R = 3$ and $R = 4$	116
Table 3	– Accuracy (%) obtained using different feature vectors for the classification of network models. The networks have $N = 500, 1000, 1500$ and 2000 nodes, and, mean degree varying from $\langle k \rangle = 4, 6, \dots, 16$. The results are shown for the 10 rules obtained in the rule selection step.	159
Table 4	– Performance measures corresponding to the best accuracy as shown in Table 3.	159
Table 5	– Robustness analysis. Classification accuracy (%) given different values of ρ_N , and, also without noise ($\rho_N = 0$)	160
Table 6	– Accuracy (%) obtained using different feature vectors for the classification of network models in combination with their $\langle k \rangle$	161
Table 7	– Performance measures corresponding to the best accuracy as shown in Table 6 for the following network models: “RAN” (random), “SW” (small-world), “SF” (scale-free) and “GEO” (geographical).	162
Table 8	– Accuracy (%) obtained using different feature vectors for the classification of scale-free models. The results are shown for the 10 rules obtained in the rule selection step.	163
Table 9	– Performance measures corresponding to the best accuracy of Table 8.	165
Table 10	– Classification accuracy (%) given different values of ρ_N , and, also without noise ($\rho_N = 0$) using structural network measures	166
Table 11	– Accuracy (%) obtained using different feature vectors for the classification of the metabolic networks. The results are shown for the 10 rules obtained in the rule selection step.	169
Table 12	– Performance measures corresponding to the best accuracy as shown in Table 11	169
Table 13	– Accuracy (%) obtained using different feature vectors for the classification of social networks. The results are shown for the 10 rules obtained in the rule selection step.	170
Table 14	– Performance measures corresponding to the best accuracy as shown in Table 13	170
Table 15	– Accuracy (%) obtained using different feature vectors for the classification of the stomata networks. The results are shown for the 10 rules obtained in the rule selection step.	175
Table 16	– Performance measures corresponding to the best accuracy as shown in Table 15	175

Table 17 – Classification accuracy for the three classifications tasks using structural network descriptors and LLNA.	180
Table 18 – Summary of the performance obtained for each synthetic dataset.	181
Table 19 – Summary of the performance obtained for each real-world application.	181
Table 20 – Classification accuracy (%) given the rotation, scale and noise transformations.	201
Table 21 – Characteristics of the Belgian influenza seasons 2010–2011 to 2015–2016. The onset is the first week at which the epidemic threshold was exceeded. This threshold is provided by ECDC and calculated using the MEM method.	208
Table 22 – Total population in Belgium per year in the period of 2010 to 2016.	208
Table 23 – Estimated model parameters β and γ , mean absolute error (MAE), root mean squared error (RMSE) and Pearson correlation coefficient between the normalized and simulated numbers of ILI cases for seasons 2010–2011 to 2015–2016.	210

LIST OF ABBREVIATIONS AND ACRONYMS

k -NN	k -Nearest Neighbor
AUC	Area Under the Curve
BSNA	Birth-Survival Network Automata
CA	Cellular Automaton
CAs	Cellular Automata
CN	Complex Networks
CV	Computer Vision
ECA	Elementary Cellular Automata
ECDC	European Centre for Disease Prevention and Control
EISN	European Influenza Surveillance Network
GenBSNA	Generalized Birth-Survival Network Automata
GFT	Google Flu Trends
HE	Hurst Exponent
ILI	Influenza-Like Illness
LE	Lyapunov exponent
LLNA	Life-like Network Automata
LZ	Lempel-Ziv complexity
MAE	Mean Absolute Error
ML	Machine Learning
NA	Network Automata
NB	Naive Bayes
ODE	Ordinary Differential Equations
PDE	Partial Differential Equation
PR	Pattern Recognition
RH	Relative Humidity
RMSE	Root Mean Square Error
RPS	Rock-Paper-Scissors
S	Shannon entropy
SEM	Spatially Explicit Model
SGPs	Sentinel General Practitioners
SIR	Susceptible-Infected-Recover

SVM	Support Vector Machine
WHO	World Health Organization

LIST OF SYMBOLS

$\mathcal{G}$ — Graph or Network

$\mathcal{V}$ — Set of vertices of a network

$\mathcal{E}$ — Set of edges of a network

A — Adjacency matrix

$\mathcal{N}$ — Number of network nodes

cc_i — Clustering coefficient of node i

k_i — Degree of node i

$\langle k^n \rangle$ — n -th moment of the degree distribution

$\langle k \rangle$ — Average degree

H — Entropy of the degree distribution

$\langle k^h \rangle$ — Average hierarchical degree

ρ_P — Pearson correlation coefficient

$\langle cc \rangle$ — Average clustering coefficient

$\mathcal{X}$ — Feature vector

$\mathcal{C}$ — Cellular Automata

$\mathcal{T}$ — Tessellation

$\mathbb{S}$ — Set of states

s_0 — Initial configuration of the cellular automata

$\mathcal{N}$ — Neighborhood

ϕ — Transition function

S — Shannon entropy

λ — Lyapunov exponent of the cellular automata

ρ_i — Neighborhood density for node i

δ — Threshold for the neighborhood density

σ — Initial configuration of the network automata

R — Normalization factor of the density interval
 x_a — Initial point of the density interval for the birth condition
 x_b — End point of the density interval for the birth condition
 y_a — Initial point of the density interval for the survival condition
 y_b — End point of the density interval for the survival condition
 n_R — number of rules of the network automata
 r_i — A given rule of the network automata
 X — Set containing the birth conditions
 Y — Set containing the survival conditions
 BX/SY — A rule of the generalized birth-survival network automata
 $\vec{\mu}_S$ — Distribution of the Shannon entropy
 $\vec{\mu}_L$ — Distribution of the Lempel-Ziv complexity
 $\vec{\mu}_W$ — Distribution of the word length
 μ — Reproduction rate
 α — Selection rate
 ε — Exchange rate
 M — Mobility
 $\langle \mu_S \rangle$ — average Shannon entropy
 $\langle \mu_W \rangle$ — average word length
 $\langle \mu_L \rangle$ — average Lempel-Ziv complexity
 ρ_N — Noise rate
 δ_T — Distance threshold
 W — Weighted adjacency matrix
 T_l — Transformation over W
 φ — CNDescriptor feature vector
 C — Curvature
 Φ — Generalized feature vector of CNDescriptor
 β — Transmission rate
 γ — Recovery rate

CONTENTS

1	INTRODUCTION	35
1.1	Contextualization	35
1.2	Objectives	38
1.3	Scope of the thesis	39
I	Theoretical Background	41
2	GRAPHS & NETWORKS	43
2.1	Introduction	43
2.2	Graphs: <i>basics</i>	45
2.3	Network Theory	46
2.3.1	<i>Theoretical models of Complex Networks</i>	46
2.3.1.1	<i>Random networks</i>	47
2.3.1.2	<i>Small-World networks</i>	49
2.3.1.3	<i>Scale-free networks</i>	51
2.3.1.4	<i>Geographical networks</i>	52
2.3.2	<i>Structural characterization</i>	52
2.3.2.1	<i>Connectivity measurements</i>	53
2.3.2.2	<i>Hierarchical degree</i>	53
2.3.2.3	<i>Degree correlations</i>	54
2.3.2.4	<i>Clustering</i>	54
2.3.2.5	<i>Distance measurements</i>	55
2.3.2.6	<i>Centrality indices</i>	55
2.4	Pattern Recognition & Networks	56
2.4.1	<i>Modeling data as networks</i>	56
2.4.2	<i>Feature extraction</i>	58
2.4.3	<i>Learning Strategies</i>	59
2.4.3.1	<i>Supervised learning</i>	59
2.4.3.2	<i>Unsupervised learning</i>	61
2.5	Conclusions	62
3	CELLULAR AUTOMATA & SPATIO-TEMPORAL PATTERNS	63
3.1	Introduction	63

3.2	Mathematical Definition	64
3.3	Cellular Automata Models	66
3.3.1	<i>Elementary Cellular Automata</i>	66
3.3.2	<i>Totalistic cellular automata</i>	68
3.3.3	<i>Two-dimensional cellular automata</i>	69
3.4	Analysis of Spatio-temporal Patterns	71
3.4.1	<i>Hamming distance</i>	71
3.4.2	<i>Entropy-related measures</i>	72
3.4.3	<i>Word count & word entropy</i>	72
3.4.4	<i>Lempel-Ziv complexity</i>	73
3.4.5	<i>Hurst exponent</i>	73
3.4.6	<i>Lyapunov exponent</i>	74
3.5	Conclusions	75
4	STATE OF THE ART ON NETWORK AUTOMATA	77
4.1	Introduction	77
4.2	Cellular Automata on Irregular Tessellations	77
4.3	Cellular Automata on Graphs and Complex Networks	78
4.4	Automata Models in the Study of Scale-free Topologies	83
4.5	Networks & Pattern Recognition Applications	84
4.6	Partially Self-avoiding Deterministic Walks on Networks	85
4.7	Conclusions	87

II Spatially Explicit Modeling on Networks Using Cellular Automata 89

5	NETWORK AUTOMATA: MODELING & ANALYSIS	91
5.1	Introduction	91
5.2	Network Automata Based on Neighborhood Density	92
5.2.1	<i>Modeling the network automata</i>	92
5.2.2	<i>Spatio-temporal patterns</i>	93
5.2.3	<i>Quantitative analyses using the Shannon entropy</i>	94
5.3	Birth-Survival Network Automata (BSNA)	97
5.3.1	<i>Definition</i>	97
5.3.2	<i>Rule space & notation</i>	98
5.3.3	<i>Spatio-temporal patterns</i>	99
5.3.4	<i>Quantitative analyses of spatio-temporal patterns</i>	102
5.4	Conclusion	111

6	GENERALIZED BIRTH-SURVIVAL NETWORK AUTOMATA . . .	113
6.1	Introduction	113
6.2	A Generalized Family of Birth-Survival Network Automata (GenB-SNA)	113
6.2.1	<i>Rule space & notation</i>	115
6.3	Life-like Network Automata (LLNA)	116
6.3.1	<i>Spatio-temporal patterns</i>	118
6.3.2	<i>Analysis of parameters</i>	118
6.3.3	<i>Comparison with BSNA</i>	121
6.3.4	<i>Sensitivity to initial distribution of states</i>	125
6.4	Conclusion	127
7	ROCK-PAPER-SCISSORS: A NETWORK ANALYSIS	129
7.1	Introduction	129
7.2	The Rock-Paper-Scissors Game on grids	130
7.2.1	<i>Mean field analysis of fundamental ecosystems processes</i>	132
7.2.2	<i>Individual-based modeling</i>	133
7.3	From Regular Grids to Irregular Tessellations	135
7.4	RPS on Networks	136
7.4.1	<i>Experimental design</i>	138
7.4.2	<i>Results</i>	138
7.4.2.1	<i>Probability of extinction</i>	138
7.4.2.2	<i>Time to the first extinction</i>	140
7.4.2.3	<i>Patchiness</i>	142
7.4.2.3.1	Average cluster size per time	143
7.4.2.3.2	Average number of clusters per time	146
7.5	Conclusion	147

III Understanding Patterns and Describing Processes on Networks 149

8	USING LLNA FOR PATTERN RECOGNITION APPLICATIONS .	151
8.1	Introduction	151
8.2	Experimental Design	153
8.2.1	<i>Description of the datasets</i>	153
8.2.2	<i>Generation of the network models</i>	154
8.2.3	<i>Training & test strategies</i>	154
8.2.4	<i>Feature extraction & Classification</i>	155
8.2.5	<i>Performance evaluation</i>	156

8.3	Pattern Recognition in Synthetic Networks	157
8.3.1	<i>Classifying network models</i>	157
8.3.1.1	<i>Robustness to noise</i>	159
8.3.2	<i>Classifying network models in combination with mean degree</i>	160
8.3.3	<i>Classifying scale-free models</i>	162
8.3.4	<i>Comparison with structural network measures</i>	164
8.4	Pattern Recognition in Real-World Applications	166
8.4.1	<i>Identifying organisms using metabolic networks</i>	166
8.4.1.1	<i>Context</i>	166
8.4.1.2	<i>Classification</i>	167
8.4.2	<i>Identifying structural patterns in social networks</i>	169
8.4.2.1	<i>Context</i>	169
8.4.2.2	<i>Classification</i>	170
8.4.2.3	<i>Comparison between BSNA and GenBSNA</i>	171
8.4.3	<i>Classifying stomata distribution patterns</i>	172
8.4.3.1	<i>Context</i>	172
8.4.3.2	<i>Classification</i>	173
8.4.4	<i>Identifying fungal growth patterns</i>	175
8.4.4.1	<i>Context</i>	175
8.4.4.2	<i>Classification</i>	176
8.4.4.2.1	<i>Connectivity analysis</i>	176
8.4.4.2.2	<i>Classification using LLNA</i>	177
8.5	Conclusion	181
9	STRUCTURAL NETWORK DESCRIPTORS IN COMPUTER VI- SION	183
9.1	Introduction	183
9.2	Modeling Images as Networks	185
9.3	Shape Properties & Structural Measurements	186
9.3.1	<i>Curvature Analysis</i>	187
9.3.2	<i>Interpolation analysis</i>	190
9.4	Shape Classification	192
9.4.1	<i>Datasets</i>	192
9.4.2	<i>Training and validation strategies</i>	192
9.4.3	<i>Evaluation of classification performance</i>	193
9.4.4	<i>Single threshold analysis</i>	195
9.4.5	<i>Robustness & noise tolerance</i>	196
9.5	Conclusion	200
10	PREDICTION OF INFLUENZA OUTBREAKS	203

10.1	Introduction	203
10.2	An Overview of the Epidemic Seasons in Belgium	206
10.3	The SIR Epidemic Model	207
10.4	Real-time Prediction of ILI Incidence	210
10.4.1	<i>The Confidence Region</i>	211
10.4.2	<i>Prediction of ILI incidence</i>	213
10.5	Accuracy of the Global Analysis	214
10.5.1	<i>Error analysis</i>	218
10.6	Incorporating the Commuting Network	218
10.6.1	<i>An overview of the ILI incidence in the Belgian districts</i>	221
10.6.2	<i>SIR epidemic model with commuting data</i>	221
10.6.2.1	<i>Accuracy of the local analysis</i>	221
10.6.2.2	<i>Error analysis</i>	229
10.7	Conclusion	229
 IV Conclusions & Future Directions		233
11	CONCLUSIONS & PERSPECTIVES	235
11.1	General Conclusions	235
11.1.1	<i>Part I: Theoretical Background</i>	235
11.1.2	<i>Part II: Spatially Explicit Modeling on Networks Using Cellular Automata</i>	236
11.1.3	<i>Part III: Understanding Patterns and Describing Processes on Networks</i>	238
11.2	Future Perspectives	239
11.3	Bibliographical Production	240
 BIBLIOGRAPHY		243
 APPENDIX A BIRTH-SURVIVAL NETWORK AUTOMATA: PARTIALLY-ORDERED SETS & MEASURES		263
 APPENDIX B RPS DYNAMICS ON NETWORKS		281
B.1	Probability of Extinction	282
B.2	Time to extinction	284
 APPENDIX C ILI PREDICTION PLOTS FOR THE BELGIAN EPIDEMIC SEASONS		287
C.1	Prediction of ILI numbers using the global approach	287

C.2	Prediction of ILI numbers using the local approach	292
-----	--	-----

INTRODUCTION

1.1 Contextualization

There are a lot of systems in nature that can be classified as complex. Among the most classic examples are the human nervous system, ant colonies, the financial market, and many others. A complex system consists of a population of individuals interacting with each other, and from this interaction a global dynamics emerges which is the result of the action of each individual (BOCCARA, 2010; MITCHELL, 2006). Complexity is related to the macroscopic behavior of a system whose properties are the result of a non-linear relationship between its constituent elements, e.g., despite the intricate dynamics presented by a single neuron, it is relatively simple when compared to the behavior of the whole nervous system (MITCHELL, 2006). The mapping between individual dynamics and collective behavior is nontrivial and the concept of non-linearity is important in this context, since the decomposition of a complex system in its isolated parts does not allow for the understanding of the system as a whole (BARRAT; BARTHELEMY; VESPIGNANI, 2008).

One of the most common characteristics of complex systems is the emergent behavior resulting from a collective and self-organized dynamics and the absence of a central controller (BOCCARA, 2010). Moreover, the complexity of the overall behavior of such system can be characterized by the pattern being generated, by the information processing that the system performs, and, by the degree of adaptability of these patterns (MITCHELL, 2009). Among the mathematical tools for modeling complex systems, we can mention the methods based on differential equations and individual-based methods. While the first class makes use of population densities, the second class deals directly with spatially distributed agents, representing the elements of the system, as is the case, for example, of Cellular Automata (CAs).

In the last years, the study of complex systems was motivated by the use of networks as modeling tools. The representation of such systems by means of the interactions among

their constituents can provide a more robust analysis of their structure and dynamics, since it makes it possible to study the connectivity of the involved constituents and not only the elements themselves. Also many biological systems have been studied by resorting to network modeling, such as genetic and metabolic networks (JEONG *et al.*, 2000), protein networks (KROGAN *et al.*, 2006), disease spreading networks (PASTOR-SATORRAS; VESPIGNANI, 2005), ecological networks (WILLIAMS *et al.*, 2002), neurological networks (BULLMORE; SPORNS, 2009), and many others (COSTA *et al.*, 2011b). The metabolic networks served as model networks for retrieving important topological characteristics of many real networks, like the scale-free property (JEONG *et al.*, 2000; BARABÁSI; ALBERT, 1999). Such scale-free networks are characterized by many vertices with a few connections and a few vertices that are highly connected, i.e. so-called hubs. The models that describe such networks usually incorporate important characteristics, such as growth and preferential attachment, i.e., new nodes can be added to the network over time and these nodes have a high probability to be connected with the network hubs (BARABÁSI; ALBERT, 1999; DOROGOVTSSEV; MENDES, 2002).

Besides, among the important factors for the increasing interest in the area of complex networks are the evolution of computers, allowing faster data processing, and, the emergence of large databases (BARRAT; BARTHELEMY; VESPIGNANI, 2008; MITCHELL, 2009). With increasing availability and accessibility, significant advances have been made in the analysis of networks with large numbers of vertices. These networks present a non-trivial topology and may present community structures and hierarchies (BOCCALETTI *et al.*, 2006; FORTUNATO, 2010). In this way, statistical tools in combination with graph theory allow for a more robust study of the network topology, previously considered as a random structure, in many cases due to the lack of experimental data. Random networks were largely explored in the 1950s by Erdős & Rényi (ERDÖS; RÉNYI, 1960).

From the works of Watts & Strogatz (WATTS; STROGATZ, 1998) on the behavior of small-world networks, and of Barabási & Albert (BARABÁSI; ALBERT, 1999) about the characteristics of scale-free networks, new methods have contributed to the structural and evolutionary modeling and analysis of networks. The importance of network characterization lies in the understanding of intrinsic patterns and properties that are the result of its topology. Two approaches to network analysis can be distinguished, namely structural and dynamic methods. For the former, an extensive number of measures have been used to distinguish the underlying network models (COSTA *et al.*, 2007). The clustering coefficient, for instance, quantifies the small-world phenomenon on networks (WATTS; STROGATZ, 1998), while distances and path lengths are important in the case of geographical networks for which the spatial position of nodes is relevant. On the other hand, dynamic approaches aim at modeling the physical processes that take place on networks (NEWMAN, 2003; BOCCALETTI *et al.*, 2006) and investigate the properties of the underlying networks through the patterns evolved by spatially explicit dynamical systems that are built upon these networks to mimic the processes at stake. Propagation of epidemics, analysis of network robustness and noise tolerance, synchronization,

among others, are examples of dynamic processes.

Cellular Automata are just one class of spatially explicit discrete models that can be used for modeling complex systems. Initially studied as models of growth and self-reproduction processes, CAs have become an important tool for studying the complexity and formation of spatio-temporal patterns (WOLFRAM, 1984a). Such patterns are the result of local interactions of a cell and its neighborhood (SHALIZI, 2006). Although they are typically built upon regular tessellations, they can as well be built upon networks (TOMASSINI; GIACOBINI; DARABOS, 2005; TOMASSINI, 2006) or irregular tessellations (BAETENS; DE BAETS, 2012). The latter provide a means to represent real systems more realistically. Besides being used for mimicking biological processes (MALLET; PILLIS, 2006; WHITE; REY; SÁNCHEZ, 2007; ULZURRUN *et al.*, 2015), CAs can also give rise to emergent patterns that result from the interactions between its entities. When defined on a network, these patterns can be used to investigate the relationship between the network architecture and its dynamics.

When combined with networks, CAs can be used to map the relationship between network architecture and its dynamic aspects from the perspective of pattern formation. In this way, CAs can now be investigated on arbitrary topologies and not only on regular tessellations. The relation between the changes in the topology and the corresponding impact on the dynamics of CAs is one of the main objectives of this work. We aim at investigating whether or not the spatio-temporal patterns generated by the dynamics of the CAs can be used in the characterization of the network topology, and consequently in the context of pattern recognition.

Some works present models that integrate networks and CAs. For instance, Watts (WATTS, 1999) discusses CA on small-world networks for tasks such as the density classification problem and synchronization. Tomassini et al. (TOMASSINI; GIACOBINI; DARABOS, 2005) discuss properties of small-world networks on the global computing capacity of CAs. Marr & Hütt (MARR; HÜTT, 2009; MARR; HÜTT, 2012) study the dynamics of evolving networks using CAs. Their results indicate a strong connection between the entropy measures obtained from the evolved spatio-temporal patterns and the vertex degree distribution of the underlying network. However, the relationship between changes in topology and the corresponding impacts on the patterns evolved by spatially explicit dynamical systems is still little explored. Furthermore, few efforts have been spent to the development of evolutionary methods because this is both computationally challenging and highly multidisciplinary. The present thesis fits in the context of pattern recognition in networks. Despite the breakthrough in the area of structural network analysis, the area of pattern recognition and classification problems has been little explored within the context of networks.

1.2 Objectives

There is a growing interest in the use of networks as modeling tools. They can represent objects in space or individuals in social and collaborative networks. Different systems can be represented by the interactions between their individuals and through the study of the mechanisms involved in such interactions we are able to characterize and compare networks. This flexibility of representation together with the increasing interest in the study of complex systems have motivated several works in the area of complex networks. The research progress in this area over the last two decades has helped to understand the structural and dynamic characteristics of networks and such results paved the way for pattern recognition. Pattern analysis is an intrinsic task of human beings and that is why we perform this task so easily in our daily life. However, automatic pattern recognition still presents many challenges. Pattern recognition in networks aims at the characterization of networks by extracting information regarding the correlation between vertices and their relationship with topology. Such information may lead to the comprehension of network patterns that are intrinsically related to the network model. Therefore, the choice of adequate network descriptors is crucial for this kind of applications.

The focus of the present project is the proposal and the investigation of methods that explore the topology of the networks through the use of discrete spatially explicit models, such as CAs, which have shown to be important tools in the study of complexity and emerging phenomena. These characteristics motivate the exploration of this model in the mapping between the architecture of the network and its dynamic aspects from the perspective of pattern formation. Although governed by simple deterministic rules, CAs can generate very complex spatio-temporal patterns. We aim at investigating the influence of network topology on the formation of spatio-temporal patterns generated by CAs. In this way, they will be explored on different topologies and not only on regular tessellations. Specifically, this work has the following objectives:

- Development of evolutionary methods for extracting features from the patterns that are evolved by spatially discrete dynamical systems defined on networks, which can subsequently be used as network descriptors;
- Investigation of the relationship between network topology and the patterns evolved by the governing spatially discrete dynamical systems;
- Network characterization: study and analysis of pattern recognition applications that involve network modeling through the use of the proposed network descriptors.

1.3 Scope of the thesis

Figure 1 illustrates the structure of this work. We start by exploring CAs as a spatially discrete modeling tool in order to propose methods for both network characterization and the study of the influence of topology on CAs' spatio-temporal patterns. In further steps, we perform a study of network-based applications, which we divided into three groups: (i) implementation and validation of the proposed approaches in pattern recognition applications; (ii) study of biodiversity by investigating the rock-paper-scissors model on irregular tessellations, and, (iii) prediction of influenza-like illness cases in Belgium.

Figure 1 – Thesis organization per topic.

Source: Elaborated by the author.

In the first part of this thesis we introduce the main theoretical background of this work. In Chapter 2 we discuss the main concepts of graph theory and networks, focusing on definitions and methods that will appear throughout this document. We will also discuss pattern recognition in the context of network analysis. In Chapter 3 we present a formal definition of CAs and their main applications. We also introduce some famous CA models, such as the Elementary CAs. In addition, we present different methods that can be used for the analysis of spatio-temporal patterns, such as entropy related measures and the Lyapunov exponent for CAs. Finally, in the last chapter of Part I, Chapter 4, we present the state of the art on cellular automata on irregular tessellations.

In Part II, we introduce the proposed spatially explicit modeling approach on networks using CAs. In Chapter 5, we discuss the concept of Network Automata and we introduce a family of such automata that is defined on irregular topologies and presents a birth-survival dynamics. The proposed network automata make use of the density of alive neighbors in a cell's neighborhood, which is a key concept for modeling a CA on an irregular topology. In Chapter 6,

we present a generalization of the birth-survival network automata and its study on different network topologies. In addition, we introduce the Life-like Network Automata (LLNA), which is a specific case of this general family of birth-survival network automata and is inspired by the Life-like CA. Therefore, we will present a formal definition of this classic CA on irregular topologies together with its corresponding rule space resembling the Moore neighborhood. In the last chapter of this part, Chapter 7, we investigate how the structural properties of a set of networks obtained from different theoretical models may influence the dynamics of the rock-paper-scissors CA. We present a variant of the square grid version of this CA that includes mobility of individuals in addition to selection and reproduction and we extend this version to irregular topologies.

In Part III we will explore pattern recognition applications that make use of data represented as networks. The network characterization is one of the motivations of using machine learning techniques, which can lead to the understanding of the patterns related to the network structure. In Chapter 8 we employ the LLNA in real-world pattern recognition applications and we demonstrate its general applicability. Pattern recognition using structural network measures is discussed in Chapter 9. We propose a generalized approach for shape characterization in computer vision based on network analysis, which accounts for different categories of measurements. In addition, we demonstrate how such measurements are related with the properties of shape boundaries. In Chapter 10, we present an approach for predicting the number of influenza-like illness cases in Belgium in real-time by resorting to a dynamically calibrated compartmental model, which furthermore takes into account the dynamics of historical influenza seasons and the network of commuting data. Finally, in Part IV, we present the general conclusions and the future prospects of this thesis.

Part I

Theoretical Background

GRAPHS & NETWORKS

2.1 Introduction

Although graph drawings have been observed throughout history, like family trees in manuscripts of the Middle Ages and the boards of Mill games in stone carvings in Ancient Egypt, which depicts the graph explicitly (KRUJA *et al.*, 2001), historically, graph theory began with Leonhard Euler in the 18th century with the solution to the famous problem of the Königsberg bridges crossing the Pregel River, where there was an intense traffic of ships and a thriving commerce (NEWMANN; BARABÁSI; WATTS, 2006). This problem consisted in determining whether or not it would be possible to cross the seven bridges by leaving and arriving at the same point, without crossing the same bridge more than once (Fig. 2-(a)). Euler demonstrated analytically that this route was not possible by making use of a graphical representation consisting of a set of points, corresponding to the land strips, and a set of curves or lines corresponding to the bridges, which were connecting the strips (Fig. 2-(b)). Besides solving this problem, Euler had started a new branch of mathematics. Euler's proof was based on a simple observation: vertices with an odd number of connections should be the point of departure or arrival of the route, and, if there are more than two nodes with an odd number of connections, then the path of interest will not exist, as was the case of the Königsberg graph (BARABÁSI, 2002). In other words, Euler showed that the solution of the problem was dependent only on the connectivity of the bridges, i.e., how they are interconnected, but not on the length of the bridges nor on any other geometric features. This was an important finding on how topological concepts can be used to solve real-world problems (ESTRADA, 2012).

After Euler, several mathematicians made important contributions to the field which are related to graph properties (BERGE; MINIEKA, 1976; BOLLOBÁS, 1998). More recently, besides the focus on the study of such properties, other questions have been raised regarding the construction of graphs and how real networks are structured. The area of complex networks (CN)

Figure 2 – The graph of the Königsberg bridges problem. (a) The city and the bridges. (b) The graph representation.

Source: Adapted from [Giuşcă \(2005\)](#).

has emerged from this, strongly influenced by graph theory, concerning the topological properties of the networks, and, by dynamical systems, regarding the evolutionary properties of networks. From the works of Erdős & Rényi ([ERDÖS; RÉNYI, 1960](#)), Milgram ([MILGRAM, 1967](#)), Watts & Strogatz ([WATTS; STROGATZ, 1998](#)), and, Barabási & Albert ([BARABÁSI; ALBERT, 1999](#)), network theory has been established as a multidisciplinary field with many advances in both theoretical and applied research. Besides the long tradition on the use of network analysis in social sciences, the increasing interest in the field is due to physicists that started using methods of statistical mechanics in the analysis of large-scale networks ([BRANDES; ERLEBACH, 2005](#)).

With the scientific and technological advances observed during the last decades resulting in an increase in the power of computers, networks have gained popularity, especially to uncover properties of complex systems. Since then, the corresponding complex networks have shown unprecedented growth in scale and complexity ([NIAZI; HUSSAIN, 2009](#)). On the one hand, there are spatial networks, in which the position of the nodes in Euclidean space plays a fundamental role, such as the Internet, road networks and airline networks. On the other hand, there are networks that only exist in an abstract space such as the World Wide Web and social networks ([GASTNER; NEWMAN, 2006](#)). Apart from that, networks are much more realistic in representing real-world systems as compared to regular tessellations ([COLIZZA; PASTOR-SATORRAS; VESPIGNANI, 2007](#)).

This chapter presents the main theoretical foundations of graph theory and networks. We give particular attention to some definitions and methods that will be used throughout this thesis. This way, we aim to standardize the nomenclature used in this thesis, so the reader can return to it whenever necessary.

2.2 Graphs: *basics*

Formally, a graph $\mathcal{G}$ is defined as a couple $\langle \mathcal{V}, \mathcal{E} \rangle$, in which, $\mathcal{V}$ is a finite and non-empty set of vertices and $\mathcal{E}$ is a set of edges that represent the connections between those vertices, i.e., $\mathcal{E} \subseteq \{(u, v) | u, v \in \mathcal{V}\}$. Using this terminology, we get the following definitions (BOLLOBÁS, 1998; BRANDES; ERLEBACH, 2005):

- **Directed Graph or Digraph:** a graph whose edges present an origin $u \in \mathcal{V}$ and a destination $v \in \mathcal{V}$. The couple (u, v) is used to represent a directed edge.
- **Undirected Graph:** a graph whose edges are not directed. Therefore, the representation of a single edge is given by the pair $\{u, v\}$.
- **Weighted Graph:** a graph that has cost information (weight) associated with its edges. The weight is represented by a function $w : \mathcal{E} \rightarrow \mathbb{R}$, which relates an edge $e \in \mathcal{E}$ to a value $w(e)$. This weight can represent distance, capacity, similarity, etc, depending on the application context. A graph that has no cost associated with its edges is equivalent to a weighted graph with unit value for each edge.
- **Subgraph:** a graph $\mathcal{G}' = \langle \mathcal{V}', \mathcal{E}' \rangle$ is a subgraph of $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ if $\mathcal{V}' \subseteq \mathcal{V}$ and $\mathcal{E}' \subseteq \mathcal{E}$.
- **Multigraph:** when an edge e appears more than once in the set $\mathcal{E}$, a graph is called multigraph and the repeated edges are denoted parallel edges. When there is only one copy of each edge in the set $\mathcal{E}$, the graph is called simple graph.
- **Adjacent vertices:** two vertices are adjacent if they have one edge in common. For directed graphs, u can be adjacent to v , but v may not be adjacent to u .
- **Vertex neighborhood:** the neighborhood of a vertex $v \in \mathcal{V}$ is the set of vertices that are adjacent to v , $\mathcal{N}(v)$.
- **Regular Graph:** a graph whose vertices have the same number of neighbors.
- **Complete Graph:** a graph for which each pair of distinct vertices has an edge among themselves.
- **Path:** a path P is an ordered sequence of edges that connects two vertices of a graph without repetition of any edge, $P = \{(v_1, v_2), (v_2, v_3), \dots, (v_{n-1}, v_n)\}$, being v_1 the origin vertex and v_n the destination vertex.
- **Distance between two vertices:** the distance between two vertices of a graph $d(u, v)$ is the number of edges in the shortest path from u to v taking into account all the possible paths from u to v in $\mathcal{G}$. This distance is also known as the geodesic distance. By definition, the distance from any vertex to itself is 0 and if there is no path from u to v , then $d(u, v) = \infty$.

- **Diameter:** the diameter of a graph is the largest geodesic distance between two vertices of $\mathcal{G}$.
- **Connected Component:** a component or a subgraph $\mathcal{G}'$ of $\mathcal{G}$ is connected if each vertex belonging to $\mathcal{G}'$ can be reached by any other vertex of $\mathcal{G}'$. If $\mathcal{G}$ has only one connected component, then the graph is said to be connected.

Mathematically, a graph is commonly represented by means of an *adjacency matrix* $A = a_{ij}$, which is a square matrix such that

$$a_{ij} = \begin{cases} 1, & \text{if } (i, j) \in \mathcal{E} \\ 0, & \text{if } (i, j) \notin \mathcal{E} \end{cases} \quad (2.1)$$

Figure 3 summarizes this section with the illustration of some types of graphs and their corresponding adjacency matrices. Alternatively, adjacency lists and edge lists can also be used for representing graphs. In the first case, for each vertex i there is an array of vertices that are adjacent to i , therefore the size of the adjacency list is equal to the total number of nodes. In the second case, each pair of vertices belonging to the graph is stored in the edge list.

From now on, the words “graph” and “network” will be used interchangeably in this work. Similar correspondence also holds for “node” and “vertex”, and, “edge” and “link”.

2.3 Network Theory

A network can be any system that can be mathematically represented as a graph. The nodes of this graph correspond to the elements of the system and the links represent the corresponding interactions or relations among these elements (BARRAT; BARTHELEMY; VESPIGNANI, 2008). Such an abstraction can be applied to a huge variety of systems making networks a powerful tool for the study of complex systems. Networks provide a theoretical framework that allows the study of a system through the element-level analysis, the group-level analysis or the whole network analysis. In this section, we introduce the main concepts in network theory starting with some theoretical models that describe the generation of such networks, followed by their structural properties. In addition, we present some background on pattern recognition in networks.

2.3.1 Theoretical models of Complex Networks

Random, small-world and scale-free networks are the main theoretical models proposed to describe complex networks. These models have been well studied over the last two decades.

Figure 3 – Graphical representations of different types of graphs and their corresponding adjacency matrices.

Source: Adapted from [Barrat, Barthelemy and Vespignani \(2008\)](#).

A fourth model will also be introduced in this section, the Waxman random graph, which can also be denoted as a geographic network.

2.3.1.1 Random networks

Initially studied by Erdős & Rényi ([ERDÖS; RÉNYI, 1960](#)) in the 1950s, the random networks are considered as a basic model of a complex network. Using probabilistic methods, these authors studied the properties of such graphs as a function of the increasing number of randomly defined connections. For the construction of a random network, we start from a set V of initially disconnected vertices and a set of edges E . Then, at each iteration, an edge of E is randomly drawn in order to connect a pair of vertices. This process is also called evolution of the network, since it is constructed by the successive addition of edges ([ALBERT; BARABÁSI,](#)

2002).

A second equivalent model for the generation of random networks involves the definition of a connection probability, p , between the network vertices (NEWMAN, 2003). Let N be the total number of vertices of a network. For each vertex $i \in V$, there are $N - 1$ different possibilities of connection with other network vertices. Thus, the maximum number of edges of the network is given by $N(N - 1)/2$ (the division by 2 is justified in non-directed graphs). In this way, considering p and $1 - p$, the probabilities that an edge belongs or not to a network, respectively, and that there are $\binom{N-1}{k}$ different ways of choosing k vertices from $N - 1$, the degree (number of neighbors) distribution of the network nodes follows a Binomial distribution with parameters $\text{Bin}(N - 1, p)$:

$$P(k) = \binom{N-1}{k} p^k (1-p)^{(N-1)-k}. \quad (2.2)$$

For $N \rightarrow \infty$ and $p \ll 1$, given a fixed value of k , the degree distribution tends to a Poisson distribution. The expected value of the degree can be calculated by $\langle k \rangle = (N - 1)p$. Figure 4 shows the visualization of a random network with $\langle k \rangle = 8$ and its corresponding degree distribution. We will also refer to $\langle k \rangle$ as the average degree of the network.

Figure 4 – Random networks: (a) Network with $\langle k \rangle = 8$. (b) Degree distribution

Source: Elaborated by the author.

An interesting point relates to determining the value of p for which a specific property of the network emerges. By means of the results provided by Erdős and Rényi, it is possible to verify that such properties tend to suddenly appear, i.e., there is a critical value, p_c , for which those properties emerge. This value corresponds to the average degree of the network: $\langle k_c \rangle$. If $p < p_c$, the network tends to become fragmented into small components. As p_c increases, the components tend to grow, initially by the connection of isolated nodes and then through the merging of the components themselves. When $p > p_c$, there is only one connected component (giant component),

such that each node can be reached by any other node in the network. The probability p_c is directly related to the percolation phenomenon on random networks (STROGATZ, 2001).

The importance of the work of Erdős and Rényi is reflected in a change of focus in graph theory, which was previously concentrated on the study of regular graphs. In the following years, random networks were used as models for network characterization. However, in the 1990s, two other models of complex networks were proposed that incorporate important characteristics of many real networks. Therefore, such networks have properties that are at the intersection between order and randomness.

2.3.1.2 Small-World networks

The small-world characteristic is derived from Milgram's experiment in the 1960s which aimed at discovering the distance between two persons. Given a network that describes relationships among people, distance represents the geodesic path between two nodes (MILGRAM, 1967). Milgram's experiment was carried out through the exchange of mail letters between people in different regions of the United States. The places of origin and destination were previously chosen so that they were geographically distant. If the sender knew the recipient he would send the letter directly to him, otherwise he would send the letter to someone that he knew and who might have known the recipient. In this way it was possible to calculate the average number of people needed for the letters to reach the recipient. The result showed that, on average, two citizens were connected by approximately 6 different people.

Watts & Strogatz have shown that many real networks present the small-world characteristic (WATTS; STROGATZ, 1998). In such networks, many vertices can be reached by others through a small number of edges, as is the case of the social networks. In addition, these networks also have a rapid spread of information across the network. Another property of small-world networks is a high *clustering coefficient* (cc). This coefficient quantifies the probability that two vertices j and k are connected to each other given that both are connected to vertex i . In other words, the clustering coefficient quantifies the probability that two of a node's friends are also friends themselves, characterizing the presence of loops in a network. The clustering coefficient (WATTS; STROGATZ, 1998) for a node i is defined by Eq. (2.3), where e_i is the number of edges between neighbors of node i , and k_i is the degree of node i .

$$cc_i = \frac{2e_i}{k_i(k_i - 1)}. \quad (2.3)$$

Small-world networks can be classified between random and regular networks regarding their topological organization. A small-world network can be obtained from a regular network with N vertices in which each vertex is connected to the k nearest neighbors. Given a vertex i , one of its edges is randomly chosen from the k initial connections, and it is rewired to another node with probability p , as shown in Fig. 5. When $p = 0$, there are no changes and the network

remains regular. When p approaches 1, all edges are redirected and the network will be closer to a random network. For intermediate values of p , the small-world property becomes apparent.

Figure 5 – Network evolution as a function of the rewiring probability p .

Source: Elaborated by the author.

Regular networks have a high average clustering coefficient ($\langle cc \rangle$) and the average distance (average geodesic path length), $\langle l \rangle$, between any two vertices of the network is large, since the vertices are connected only to their nearest neighbors. On the other hand, for random networks both the average distance and the average clustering coefficient are small. Watts & Strogatz have shown that there is a range of p for which the clustering coefficient is high and still the average distance is small. The coexistence of these two properties is present in many real networks (ALBERT; BARABÁSI, 2002). Figure 6 presents the typical curves of the average clustering coefficient and the average path length for small-world networks.

Figure 6 – Clustering coefficient $\langle cc \rangle$ and average path length $\langle l \rangle$ as a function of the rewiring probability p

Source: Elaborated by the author.

The degree distribution of the small-world network is similar to the corresponding distribution of the random network. In the case of the regular network, since all nodes have

the same degree k , the distribution is a delta function centered in k . As the rewiring probability (p) increases, the distribution tends to expand, but it remains centered at k . Since the rewiring process only changes the degree of one vertex of each edge, each node presents at least $k/2$ edges, then after the rewiring process and for $k > 2$, there are no isolated nodes and the network is composed of a single connected component. As such, the degree distribution has a peak in the mean value k and decays exponentially for large values of k (ALBERT; BARABÁSI, 2002).

2.3.1.3 Scale-free networks

In a famous paper (BARABÁSI; ALBERT, 1999), Barabási & Albert introduced the scale-free model, which incorporates important characteristics of real networks. The degree distribution of these networks follows a power law: $P(k) \sim k^{-\gamma}$, where γ is the scale exponent. Along the last two decades, the scientific community has shown that many real-world networks are scale-free, being characterized by an irregular degree distribution, i.e., only a few vertices are highly connected, which are the so-called *hubs*, whereas the majority of the vertices present few connections. This network is illustrated in Fig 7 together with the degree distribution of a typical scale-free topology.

Figure 7 – Scale-free networks. (a) Presence of *hubs*. (b) Degree distribution

Source: Elaborated by the author.

While the previous models assume a fixed number of network vertices, N , the scale-free model accounts for the network growth. New vertices can be added to the network and the connections between these nodes are established according to the so-called preferential attachment, i.e., a new node is more likely to connect with nodes that present high degrees. Due to these two characteristics, growth and preferential attachment, the scale free model is more realistic when representing real-world networks (BOCCALETTI *et al.*, 2006).

A scale-free network can be generated as follows: from an initial number of isolated nodes, m_0 , at each time step $t = 1, 2, \dots, N - m_0$, a new node j is added to the network, where j has a number of connections $m \ll m_0$. The probability of connecting j to a node i , denoted $P(j \rightarrow i)$, is linearly proportional to the degree of i and is given by:

$$P(j \rightarrow i) = \frac{k_i}{\sum_u k_u}, \quad (2.4)$$

where k_i is the degree of the vertex i and $\sum_u k_u$ is the sum of the degrees of the remaining vertices of the network.

The importance of this network model lies in the fact that the scale-free topology has been found in several real networks, such as metabolic networks, protein-protein networks, the Web, the Internet, collaborative authorship among scientists and many others. In addition to Barabási & Albert's model, Dorogovtsev & Mendes also proposed a model for the generation of scale-free networks (DOROGOVTSSEV; MENDES, 2002). In this model, the construction of the network starts with three vertices and three edges forming a triangle, then a new node i is added to the network at each time step. For that purpose, an edge is randomly selected and i is connected to the two vertices at the ends of the selected edge. Since the nodes connected to a high number of edges are the ones with the highest chance of being selected, the resulting network has a degree distribution similar to a power law.

2.3.1.4 Geographical networks

Although many networks are described in an abstract space, the geographic position of the nodes may be useful in some cases. Such networks are called geographic or spatial. This is the case, for instance, of airport and highway networks and the Internet (at the level of Autonomous Systems (SA), which are "subnetworks" separately managed), among others. Generally, the connections between the vertices of these networks are restricted to the distance or the geographical limitations between them.

One way to generate geographic networks is to distribute n vertices at random in a 2-dimensional space. Then, the edges between these vertices are defined according to a probability $\mathcal{P}$ that decays with the distance between them: $\mathcal{P}(i \rightarrow j) = e^{-\lambda s_{ij}}$, where s_{ij} is the geographic distance between vertices i and j , and λ is the scaling factor (WAXMAN, 1988). This model generates networks with a Poisson degree distribution.

2.3.2 Structural characterization

The structure of a network can be analyzed through the use of measurements which can unveil different topological properties. Important contributions have been made in this area over the last two decades. Statistical methods and new tools were used in the structural analysis

of networks that present a large number of vertices and edges. Many measures can be used to quantify different properties of networks, and, consequently, broaden their understanding. The study of network topology can improve the comprehension of its evolution and dynamics (NEWMAN, 2003; BOCCALETTI *et al.*, 2006). In this section, we present different types of measurements that can be used for the structural characterization of graphs and networks.

2.3.2.1 Connectivity measurements

The degree of a vertex i , denoted as k_i , is the number of connections of this vertex, i.e., the number of neighbors to which i is connected. The vertex degree is then defined by Eq (2.5), where A is the adjacency matrix.

$$k_i = \sum_{j=1}^n a_{ij}. \quad (2.5)$$

Several measures can be derived from k_i . The degree distribution $P(k)$ expresses the fraction of nodes that presents degree equal to k , i.e., $P(k) = P(K = k)$. Consequently, from $P(k)$ it is possible to obtain other statistical moments. The moment of order n is defined as

$$\langle k^n \rangle = \sum_k k^n P(k). \quad (2.6)$$

The first moment is given by the mean degree $\langle k \rangle$, which can also be defined as the average degree of all the nodes of the network, $\langle k \rangle = \frac{1}{N} \sum_{i=1}^n k_i$. The second moment is the variance which accounts for the fluctuations of the network connectivity. Another important measure is the entropy of the degree distribution, H_k , which is related to the network complexity. The entropy is defined as

$$H = - \sum_{k=k_{min}}^{k_{max}} P(k) \log P(k). \quad (2.7)$$

2.3.2.2 Hierarchical degree

The hierarchical degree, k_h , accounts for the connectivity of the node neighbors that are restricted to a hierarchy level h . For instance, the hierarchical degree of level 2 of a given node i , $k_2(i)$, is the sum of the degrees of its neighbors. Therefore, the average hierarchical degree of a network, $\langle k_h \rangle$, where h defines the hierarchy level, is given by:

$$\langle k_h \rangle = \frac{1}{N} \sum_i k_h(i). \quad (2.8)$$

2.3.2.3 Degree correlations

In many real networks there is a degree correlation between the network nodes. The probability that a node with degree k is connected to a node with degree k' depends on k and can be described by the conditional probability $P(k'|k)$. The normalization $\sum_{k'} P(k'|k) = 1$ and the condition $kP(k'|k)p(k) = k'P(k|k')p(k')$ are satisfied for $P(k'|k)$. Due to the fact that many real networks have a sparse degree distribution, and also due to the finite size N of the network, it becomes difficult to estimate $P(k'|k)$. In practice one of the solutions to estimate this probability is the use of the average degree of the nearest neighbors $k_{nn}(i)$, defined by:

$$k_{nn}(i) = \frac{1}{k_i} \sum_{j \in \mathcal{N}_i} k_j = \frac{1}{k_i} \sum_{j=1}^N a_{ij} k_j. \quad (2.9)$$

Let $\mathcal{N}_i$ be the set of neighbors of node i , then $k_{nn}(i)$ can also be expressed in terms of the conditional probability $k_{nn}(k) = \sum_{k'} k' P(k'|k)$. If $k_{nn}(k)$ is an increasing function of k the network is classified as assortative, meaning that nodes with the highest degrees tend to connect with nodes that also have high degree. When the contrary is observed the network is called disassortative (BOCCALETTI *et al.*, 2006). In this case, nodes with the highest degrees tend to connect with nodes with lower degrees.

The Pearson correlation coefficient (Eq. (2.10)) can also be employed in order to quantify the degree correlation, where M corresponds to the total number of edges. The same assortativity analysis can be performed considering the values $\rho_P > 0$ and $\rho_P < 0$. In the case $\rho_P = 0$, there is no degree correlation between the network nodes.

$$\rho_P = \frac{(1/M) \sum_{j>i} k_i k_j a_{ij} - [(1/M) \sum_{j>i} (1/2)(k_i + k_j) a_{ij}]^2}{(1/M) \sum_{j>i} (1/2)(k_i^2 + k_j^2) a_{ij} - [(1/M) \sum_{j>i} (1/2)(k_i + k_j) a_{ij}]^2}. \quad (2.10)$$

2.3.2.4 Clustering

As introduced in Section 2.3.1.2, the clustering coefficient quantifies the probability that two nodes j and k are connected to each other given that they are connected to vertex i . This coefficient is directly related to the small-world characteristic. It is a local measure since it is defined per node. As a global measure, the average clustering coefficient can be used:

$$\langle cc \rangle = \frac{1}{N} \sum_i cc_i. \quad (2.11)$$

Another way to quantify the presence of clusters and cycles is through the use of the transitivity measure: $C = 3N_\Delta/N_3$, where N_Δ is the number of triangles in the network and N_3

is the number of connected triples. The multiplication by 3 is due to the fact that each triangle consists of three connected triples (COSTA *et al.*, 2007).

2.3.2.5 Distance measurements

As defined previously, a path connecting two network nodes, i and j , is a sequence of nodes chosen to go from i to j . The number of edges in this path is called the path length, d_{ij} . The geodesic path is the shortest path between nodes i and j . If there is no path between i and j , then $d_{ij} = \infty$, or d_{ij} can be defined as the largest shortest path of the network increased by one. Yet, another alternative for the latter case is to consider only the largest connected component of the network (COSTA *et al.*, 2007). The average geodesic distance is defined as:

$$\langle l \rangle = \frac{1}{N(N-1)} \sum_{i \neq j} d_{ij}, \quad (2.12)$$

where $1/N(N-1)$ is a normalization factor for the fully connected graph. Another measure related to distance is the efficiency, E , which is defined as $E = \frac{1}{N(N-1)} \sum_{i \neq j} 1/d_{ij}$. The efficiency measures the information traffic. The lower the distance traveled, the higher the network efficiency regarding the information flow.

2.3.2.6 Centrality indices

Centrality indices quantify the intuitive feeling that some nodes or edges are more central than others (KOSCHÜTZKI *et al.*, 2005). The degree of a node, for instance, can be taken as a degree index. In this case, the higher the degree, the more central the node. The choice of the centrality index will depend on the application, since not all the centrality indices are suitable to all applications (KOSCHÜTZKI *et al.*, 2005).

The betweenness is one important centrality measure that quantifies the information load of the network nodes. Let j and k be two non-adjacent nodes of a network, the communication between them depends on the nodes belonging to the paths that link j and k (COSTA *et al.*, 2007). The betweenness of a node i will depend on the number of paths that contains i and that links j and k . Therefore, the betweenness is given by the number of geodesic paths that pass through i and link each pair of nodes of a network (Eq. (2.13)).

$$b_i = \sum_{j \neq k} \frac{n_{jk}(i)}{n_{jk}}, \quad (2.13)$$

where $n_{jk}(i)$ is the number of geodesic paths connecting j and k that pass through i and n_{jk} is the number of geodesic paths connecting j and k . The normalized equation is given by:

$$b_i = \frac{1}{n^2} \sum_{j \neq k} \frac{n_{jk}(i)}{n_{jk}}, \quad (2.14)$$

where n^2 represents the total number of links in a complete graph. The definition of betweenness can also be extended to edges (edge betweenness). In this case the measure is defined as a function of the number of geodesic paths passing through a given edge (BOCCALETTI *et al.*, 2006).

The betweenness calculation involves the use of graph search algorithms. Depending on the size of the network, this procedure can be very costly. Some alternative measures are the maximum flow and measures based on random walks (NEWMAN, 2005). The maximum flow is defined by the number of independent paths between two nodes of the network. Two paths connecting the same pair of nodes i and j are independent if they do not share any edges neither any nodes, except the starting and the arrival nodes. The number of independent paths also corresponds to the number of edges to be removed from the network to disconnect i and j (BOCCALETTI *et al.*, 2006).

All the measures described in this section are defined for non-directed networks. For directed networks, the out-degree k_i^{out} and the in-degree k_i^{in} have to be taken into account with respect to the connections of node i . Many other network measures are described in the literature of graphs and networks. Costa *et al.* give a summary comprising many of them (COSTA *et al.*, 2007).

2.4 Pattern Recognition & Networks

Any discrete system can be represented by the interactions among its elements. Through the study of the mechanisms involving these interactions we can characterize such systems and compare them. This flexibility of representation together with the growing interest in the study of complex systems have motivated several works in the area of complex networks. Consequently, the advances in this area over the last decade helped to understand the structural and the dynamical properties of networks. These results pave the way to the area of pattern recognition in networks. The interactions among the network nodes present a great potential to be explored within this context, since they provide another class of attributes, the relational features. The network topology is a consequence of such interactions and, therefore, provides the basis for the study of the processes that are taking place on the network. Despite the different methods proposed for the analysis of networks, pattern recognition techniques were still little explored.

2.4.1 Modeling data as networks

Different types of data can be modeled as networks, e.g., time series, images, texts, sequence of numbers, metabolic reactions, financial exchanges and many other examples. Creating such a relational dataset can be a challenging task when defining the interactions of the corre-

sponding network. Starting from a set of elements, it is necessary to establish the connections among them. In some cases, the definition of the elements that will be represented by the network nodes is already a task to be optimized, depending on the application context. We would like to illustrate this scenario with three distinct applications.

Computer Vision. Images are 2-dimensional signals defined on a discrete domain. Casanova defines different ways to model networks from images in order to perform texture analysis (BACKES; CASANOVA; BRUNO, 2013; CASANOVA, 2013). In this case, the pixels of the image are defined as the network nodes and the corresponding edges are defined according to some criterion such as color, intensity, adjacency, radius, distance among the pixels, etc. The same approach can be used to shape analysis but now the nodes correspond to pixels of the shape boundaries and the edges are defined by a distance threshold (BACKES; CASANOVA; BRUNO, 2009a; BACKES; BRUNO, 2010; BACKES; CASANOVA; BRUNO, 2014). The authors have shown that the resulting networks are small-world.

Metabolic networks. Metabolic networks describe chemical reactions that are organized into metabolic pathways. In this process, one chemical component is transformed into another through the action of enzymes. These reactions can be represented in directed and weighted networks whose vertices can be metabolites, reactions or enzymes and the edges can represent the flow of the reactants towards the reactions, or, the flow of the enzymes towards the reactions (COSTA *et al.*, 2011b). In the famous work described by Jeong *et al.* (JEONG *et al.*, 2000), the metabolic networks of 43 different organisms were investigated, and the result indicates that these networks present a power law degree distribution. Metabolic networks represent one of the most studied groups of complex networks. To some extent, this is due to the large-scale sequencing projects, which enabled the creation of databases ranging from the genome to the integrated pathways of the connectivity maps of the metabolism for different organisms. In addition, the easy access to these databases supports the study of the dynamics of such networks.

Time series. Regarding the representation of time series as networks, one of the key questions is whether the associated network incorporates some structure of the time series and, consequently, if the processes that generated this series can be characterized using network theory. For example, periodic signals tend to generate more regular networks (LACASA *et al.*, 2008). The characterization of more complex dynamics from experimental time series is a fundamental problem and of importance for different areas. The cycles of the time series can correspond directly to the nodes of the network and the connectivity can be defined as a function of the temporal correlation between the cycles. This representation incorporates the dynamics of the time series into the network topology that can then be quantified through statistical properties (ZHANG; SMALL, 2006).

2.4.2 Feature extraction

The feature extraction step is characterized by the construction of a vector $\vec{\mathcal{X}} = (x_1, x_2, \dots, x_n)$, denoted as feature vector, which contains n descriptors of a network. Assessing network characteristics allows for a network representation by means of a set of features that can be subsequently used in a pattern recognition context (Fig. 8). This characterization can be performed in two ways. First, one can extract structural measures from the network (COSTA *et al.*, 2007; COSTA *et al.*, 2011b; COSTA *et al.*, 2010), such as the degree distribution, the average vertex degree, the clustering coefficient (WATTS; STROGATZ, 1998), as well as other measures cited in Section 2.3.2. In spite of the great variety of measures that can be used to characterize networks, this can also result in a high mutual correlation. Therefore, to avoid the curse of dimensionality (DUDA; HART; STORK, 2012), it is recommended to use techniques such as PCA (Principal Component Analysis) or LDA (Linear Discriminant Analysis), to reduce the dimension of $\vec{\mathcal{X}}$. These measures allow to characterize among other things the size, connectivity and communicability of networks.

Figure 8 – Feature extraction from the network. a) Feature vector obtained through the use of structural measures. b) Feature vector obtained by a spatially explicit model (SEM) defined on the network topology

Source: Elaborated by the author.

The second way to extract features from the network is based on the dynamics evolved by a spatially explicit model (SEM), such as a CA of an IBM (individual-based model) defined on the network topology. In this case the attributes are not inferred directly from the network structure, but rather indirectly by studying the dynamics taking place on it (SILVA; ZHAO, 2012; SILVA; ZHAO, 2016). For instance, self-avoiding deterministic walks (CAMPITELI *et al.*, 2006) have been used for the analysis of both static and dynamic textures in the context of computer vision and pattern recognition (GONCALVES; BRUNO, 2013) and in the classification of complex network models (GONÇALVES; MARTINEZ; BRUNO, 2012). Further in this thesis we will present the LLNA (MIRANDA; MACHICAO; BRUNO, 2016a), inspired by the classic life-like CA and that has been used to solve different pattern recognition problems.

2.4.3 Learning Strategies

Machine Learning (ML) is a research field whose main purpose is the development of methods that are capable of “learning” with accumulated knowledge (DUDA; HART; STORK, 2012). Computer modeling of learning processes is the subject of machine learning and artificial intelligence (MICHALSKI; CARBONELL; MITCHELL, 2013). In general, there are two types of learning, namely, supervised and unsupervised (BISHOP, 2006). In *supervised learning*, the goal is to derive concepts from the data from labeled instances. In this case, all the samples belonging to a dataset present the respective associated classes. This set of labeled instances is also denoted the training set and the learning process consists of a mapping function defined over the training set. Following the construction of this mapping function, the learning process is tested by means of unseen data instances, also called the test set. For labels described as discrete values, the learning process is denoted *classification*, whereas when described as continuous values, *regression* (DUDA; HART; STORK, 2012). For what concerns *unsupervised learning*, there is no prior knowledge about the classes associated with the training instances. Therefore, density functions have to be estimated to describe the analyzed data (BISHOP, 2006). In addition, a *semisupervised learning* approach has been proposed in the literature of machine learning (CHAPELLE; SCHOLKOPF; ZIEN, 2009). Such an approach combines both supervised and unsupervised learning. In this context, a dataset consists of labeled and unlabeled instances and the existing labels of the set of labeled instances are propagated to the unlabeled instances.

2.4.3.1 Supervised learning

Formally, a supervised learning task can be defined as follows (MICHALSKI; CARBONELL; MITCHELL, 2013): consider a training set defined as $\mathcal{X}_{train} = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, where x_i denotes the i -th tuple $x_i = (f_1, \dots, f_d)$, which defines the d -dimensional attributes of the i -th training instance. The y variable accounts for the class labels of the training instances, $y_i \in \mathcal{Y} = L_1, \dots, L_l$. Therefore, the learning process provides the mapping $x \mapsto y$. For the test set, we have $\mathcal{X}_{test} = \{x_{n+1}, \dots, x_{n+u}\}$, for which the labels are unknown. The construction of the mapping $x \mapsto y$ is denoted as the *training phase*, on which the classifier is trained through the use of the training set, and, during the *classification (or validation) phase* the labels of the instances belonging to the test set are predicted by means of the induced classifier (BISHOP, 2006).

In order to avoid any bias in the learning process, the training and the test sets should present no intersections, $\mathcal{X}_{train} \cap \mathcal{X}_{test} = \emptyset$. A well-known training strategy is the k -fold cross validation, on which the dataset is randomly divided into k parts, $k - 1$ parts being used for training and one part used for validation. Then, this process is repeated k times until all the possible configurations have been tested (BISHOP, 2006; DUDA; HART; STORK, 2012). Another common training strategy is *leave-one-out*, which is a specific case of the k -fold cross

validation method with k equals to the number of data points except one for which a prediction is made. Although the leave-one-out strategy can provide a complete performance evaluation of the dataset, it is very costly from the computational point of view. Therefore, it is more suitable for small datasets. The classification performance is given by the number of correctly classified instances of the test set. The average performance can be calculated from all the folds of the cross validation (BISHOP, 2006). Usually, the performance results are shown in a confusion matrix, in which rows and columns account for the predicted and the real class labels $y_i \in \mathcal{L} = \{L_1, \dots, L_l\}$. Therefore, each cell c_{ij} shows the number of instances which were predicted as class i and are from class j . The best performance is achieved when the diagonal cells are the only cells different from 0s.

In the following paragraphs we will briefly present some classification methods that are commonly used for supervised learning:

- **Instance-based learning.** Instance-based learning makes use of a similarity function in order to calculate the distance between the training instances and the new test instances. The most famous algorithm is the k -Nearest Neighbor (k -NN), which simply attributes for a new instance i the class of its k nearest neighbor according to a distance measure (ALTMAN, 1992).
- **Statistical learning theory.** Originally proposed as a two-class classifier, the SVM (Support Vector Machine) is a statistical classifier. SVM builds a hyperplane in an n -dimensional space which is used for classification or regression. The optimal separating hyperplane is the one with the largest distance to the closest training instance of any class. In this way, the solution is derived through the data points that are on the margins (VAPNIK, 1998).
- **Decision trees.** This type of classifiers are tree-like models consisting of vertices and branches that define partitions in order to obtain decision rules. Therefore, each path along the tree corresponds to one part in the space of attributes which optimizes the separation among the classes of interest (QUINLAN, 1987).
- **Artificial Neural Networks (ANNs).** Neural networks are computational methods inspired by biological neural networks. They are represented by means of an interconnected group of neurons that are able to process information through the use of mathematical models. The definition of the network structure, regarding its number of layers and neurons and also its parameters, is a key part in developing neural networks. The back-propagation algorithm is one of the most used learning algorithm and is responsible for the adjustment of the weights of the nodes. In this way, neural networks can be defined as adaptive systems. Consequently, they can be used to find patterns in datasets and, also, for modeling complex relationships between inputs and outputs (HAYKIN, 2008).

2.4.3.2 Unsupervised learning

In the context of unsupervised learning, no prior knowledge is provided during the learning process (JAIN; MURTY; FLYNN, 1999). Therefore, a data set can be formally represented by $\mathcal{X} = \{x_1, x_2, \dots, x_n\}$, n being the total number of instances and x_i , the d -dimensional feature vector of the i -th instance. In contrast to supervised learning, a separation between training and test sets cannot be performed (BISHOP, 2006). As such, information about the data distribution itself should be used during the learning process. This results in the definition of subgroups of the data set for which the instances belonging to the same group or cluster are more similar to each other than instances belonging to other groups. This process is also known as “data clustering” and can be applied to different machine learning and pattern analysis tasks, such as image segmentation and data mining.

Clustering algorithms can be divided in two types: partitional or hierarchical (DUDA; HART; STORK, 2012). Partitional clustering algorithms determine all the clusters at once. The K -means clustering is one of the most well-known examples (MACQUEEN *et al.*, 1967). This clustering method provides a partition of the input instances into K clusters in such a way that each instance will belong to the cluster with the nearest centroid. At each iteration of the algorithm the centroids of each cluster are recalculated until their convergence or until a certain number of iterations. Although widely used, one of the main issues with K -means clustering is its dependence on the initial values of the centroids. Hierarchical clustering algorithms generate successive clusters based on previously established clusters. Such algorithms can be agglomerative or divisive. In the former, each instance of the dataset is considered as one single cluster which is merged in successive larger clusters. In the latter, divisive algorithms begin with the whole set of instances as one big cluster, and then this cluster is successively divided into small clusters (BISHOP, 2006; DUDA; HART; STORK, 2012).

2.5 Conclusions

This chapter presented the main theoretical foundations of graphs and network theory. A formal definition of a graph was introduced as well as different types of graphs. We also introduced the main concepts concerning complex networks. We summarized some well-known network models and described how the structure of a network can be quantitatively characterized. In addition we also discussed some examples of dynamical processes taking place on networks. Finally, we introduced some pattern recognition concepts applied to network analysis. In the next chapter we broaden the theoretical background of this thesis by introducing concepts related to CAs and spatio-temporal patterns.

CELLULAR AUTOMATA & SPATIO-TEMPORAL PATTERNS

3.1 Introduction

Cellular Automata are mathematical models that can be used for modeling many different dynamical systems. Some CAs are characterized by an emergent behavior from simple deterministic rules that can lead to the formation of complex patterns (WOLFRAM, 1985). The first models using CAs are the works of John von Neumann and Stanislaw Ulam about self-replicating machines (automata) (NEUMANN; BURKS *et al.*, 1966). Their aim was to propose a theory of biological development by conceiving a system capable to produce copies of itself (BERTO; TAGLIABUE, 2017). For the representation of such automata, von Neumann adopted Ulam's suggestion for representing the elements of the system in a two-dimensional cellular space, with an infinite number of square cells. Von Neumann's automaton had 29 different states with a complicated dynamics, described by means of a transition function, which was able to perform self-reproduction (BERTO; TAGLIABUE, 2017). This automaton is updated during consecutive time steps. Therefore, it is discrete in both time and space. Due to von Neumann's early death in 1957, he could not complete a paper in which he would present the proposed CA. This task was further carried out by a contemporary who completed and published the unfinished manuscripts (NEUMANN; BURKS *et al.*, 1966). Von Neumann's CA was the first discrete and parallel computational model proposed (CHOPARD, 2012). This CA was shown to be a universal computer, able to emulate a universal Turing machine (BERTO; TAGLIABUE, 2017).

Throughout the subsequent years, Ulam investigated the growth patterns obtained through the spatio-temporal evolution of CAs (SCHRANDT; ULAM, 1967). In the 60s Moore (MOORE, 1962) and Myhill (MYHILL, 1963) proved that not every possible pattern can emerge, except when defined as initial conditions. In the 1970s, the Game of Life, a well-known CA, was introduced by Martin Gardner. This CA presents interesting computational properties (GARDNER,

1970). The Game of Life is defined in a two-dimensional orthogonal grid, in which each cell can be in one of two possible states, dead or alive (0 or 1), and interacts with its eight neighbors. At each time step, new states are defined for every cell according to transition rules that take into account the states of the cells in the neighborhood. The evolution of this CA provides interesting patterns and presents particular properties. The Game of Life became one of the most popular CAs. In addition it is proven to be a universal computer (BERTO; TAGLIABUE, 2017).

Since then, models based on CAs were explored beyond the computational area and some works started using CAs for representing real-world phenomena (MARGOLUS, 1984; TOFFOLI, 1984; VICHNIAC, 1984). In the 1980s, Stephen Wolfram explored one-dimensional CAs by proposing a taxonomy for the behavior of such CAs and consolidating the basis for future research in the area (WOLFRAM, 1984a; WOLFRAM, 2002). The popularity of CAs was also promoted by the evolution of computers. During the eighties and nineties, as computers became more powerful, the CA paradigm started to be appreciated and the number of CA-based models started to increase (SCHIFF, 2011). As a consequence, CAs were highlighted as an alternative to PDE-based models. Although they are not yet accepted as such, the CA paradigm is recognised as one of the milestones in mathematics (PICKOVER, 2009). Moreover, there is still a lack of a unifying framework for designing and analyzing CA-based models which makes that model parameters are often manually tuned (BAETENS, 2012; D'AMBROSIO; SPATARO; IOVINE, 2006). CAs can be seen as an alternative to PDE-based models, since they provide the discrete evolution of many identical components (WOLFRAM, 1984a). A basic approach is to analyze the CA evolution from all their possible initial states (WOLFRAM, 1985). Moreover, due to their flexibility in representing different types of systems, they can be used for modeling non-linear phenomena.

3.2 Mathematical Definition

Formally, CAs consist of a tessellation in a given dimension and its fundamental units are denoted cells. Each cell $i = 1, 2, \dots, n$ is assigned a state s_i at time t . Space, time and states are discrete. The states are updated at each time step according to deterministic rules, also called transition rules. These rules depend on the current state of cell c_i and on the states of its neighborhood. A CA is completely defined by the quintuple (ILACHINSKI, 2001; SCHIFF, 2011): $\mathcal{C} = \langle \mathcal{T}, \mathbb{S}, s_0, \mathcal{N}, \phi \rangle$, where

- $\mathcal{T}$, is an infinite tessellation on an n -dimensional Euclidean space $\mathbb{R}^n$, that consists of cells $c_i, i \in \mathbb{N}$;
- $\mathbb{S}$, a set of m states, such that $\mathbb{S} \in \mathbb{N}$;

- $s_0, \mathcal{T} \rightarrow \mathbb{S}$ assigns to every cell c_i an initial state, i.e., $s(c_i, 0) = s_0(c_i)$. Note that σ_s is the probability of having a state s .
- $\mathcal{N}$, a function that maps each cell c_i to a finite sequence $\mathcal{N}(c_i) = (c_{ij})_{j=1}^{|\mathcal{N}(c_i)|}$ that consists of $|\mathcal{N}(c_i)|$ distinct cells;
- ϕ , an output function, such that $\phi : \mathbb{N} \rightarrow \mathbb{S}$. This function determines the state of cell c_i at time t , i.e., $s(c_i, t)$;

The tessellation $\mathcal{T}$ can be composed of cells of any particular shape, e.g., triangular, square, hexagonal or even irregular. It is assumed that the cells are homogeneous, such that they have the same properties and follow the same transition rules (SHALIZI, 2006). The implementation of CAs requires the definition of boundary conditions, which may be 1) periodic, in which the boundary cells are connected (for instance, for one-dimensional CA, it becomes a circular space); 2) reflexive, in which all the cells of the borders are self-reflected; and 3) fixed, in which the states of the cells at the borders are kept fixed during the CA evolution.

Due to the CA architecture, the state of each cell depends on its local neighborhood and the neighborhood depends on the size of the tessellation and its shape. In the case of two-dimensional CAs, the most common are the Von Neumann and the Moore neighborhoods. The first is defined by the four nearest neighbors in horizontal and vertical directions, and, the second also includes the neighbors in the diagonal directions, as shown in Fig 9.

Figure 9 – The most common neighborhoods of 2-dimensional CAs with radius = 1: (a) *von Neumann* and (b) *Moore*

Source: Elaborated by the author.

According to the CA definition, $\mathcal{N}(c_i)$ is an ordered list of cells $(c_{ij})_{j=1}^{|\mathcal{N}(c_i)|}$. However, there are also families of invariant CAs for which ϕ_i does not depend on the ordering imposed on $\mathcal{N}(c_i)$. Such CAs are called totalistic CAs. Finally, the set of transition functions determines the state of each cell c_i at time t .

3.3 Cellular Automata Models

3.3.1 Elementary Cellular Automata

Elementary Cellular Automata (ECA) represent the simplest class of CAs. Defined in a one-dimensional space, each cell may be in one of two possible states: 0 or 1. Given an arbitrary initial configuration, at each time step, the state of each cell is updated according to local rules that take into account the state of its two closest neighbors and the state of the cell itself. Formally, an ECA is defined by the quintuple: $\mathcal{C} = \langle \mathcal{Z}^1, \mathbb{S} = \{0, 1\}, s_0, \mathcal{N} = (c_{i-r}, c_i, c_{i+r}), \Phi : S^3 \rightarrow S \rangle$, where $r = 1$ is the radius of the neighborhood. Thus, by combining the three cells (center cell and its two neighbors) and the number of states, there is a total of $|\mathbb{S}|_{(c_i)}^{\mathcal{N}} = 2^3 = 8$ possible neighborhood configurations that are taken into account during the ECA evolution. So there is a total of $2^8 = 256$ transition rules.

Stephen Wolfram has introduced a nomenclature for the ECAs, in which each rule is represented by a sequence of 8 bits (WOLFRAM, 1984b; WOLFRAM, 1983). Figure 10 depicts the transition function for rule 90. In the upper part of this figure we can see all the possible neighborhood configurations, where the states of node i and its two neighbors are represented by the binary ordered triples 111, 110, 101, 100, 011, 010, 001, 000. For each triple, we can see the corresponding output where the white cells corresponds to state 0 and the black cells to state 1. Consider for instance, a neighborhood defined by the triple 100, the corresponding output will be 1, which means that the state of cell c_i will be equal to $s(c_i) = 1$ at time $t + 1$. The resulting binary number, 01011010, corresponds to number 90. Therefore, each of the 256 rules that compose the rule set of the ECA can be mapped into their binary counterparts in order to obtain the ECA evolution.

Figure 10 – Rule 90: Illustration of a transition function of the ECA.

Source: Elaborated by the author.

Usually, for some CA families, such as the ECAs, their evolution can be visualized by means of two-dimensional space-time diagrams whose columns represent the CA cells and the rows correspond to the states of each cell at each time step. Some examples can be seen in Fig. 11 in which there are interesting patterns that emerge from the CA evolution. According to these space-time diagrams, Wolfram defined four distinct classes of ECAs: 1) *Stable*, when the initial configurations of the CA cells converge rapidly to a single state which can be 0 or 1, as shown in Fig 11-(a) for rule 4; 2) *Periodic*, when the states of the cells oscillate at each time step, as in Fig 11-(b), for rule 51; 3) *Chaotic*, when the cells present apparently random and non-periodic behavior, with strong dependence on the initial states of each cell, as the case of rule 30, shown in Fig 11-(c), and, finally 4) *Complex*, when the emerging patterns are characterized by having stable structures that can “survive” over a long period of time, although there can also exist structures that interact in a more complex way, resulting in new patterns that can be created, destroyed or propagated. An example can be seen in Fig 11-(d) for rule 110.

Figure 11 – ECA classification: (a) stable, rule 4; (b) periodic, rule 51; (c) chaotic, rule 30, (d) complex, rule 110.

Source: Elaborated by the author.

Figure 11 presents patterns obtained from the same initial configuration, in which only the central cell is in state 1. However, regardless of the initial condition of each cell, the evolution of the CA converges to the same pattern. Although widely used, the classification proposed by Wolfram is qualitative. Along the last decades other methods were proposed to quantitatively

assess the spatio-temporal patterns (LI; PACKARD, 1990; WUENSCHKE; LESSER, 1992; MARTINEZ; ADAMATZKY; ALONSO-SANZ, 2013). ECAs are examples of how complex patterns can emerge from the evolution of simple deterministic rules. Rule 30, for example, is used in a random number generator of the *Mathematica* software¹.

There are other models of one-dimensional CAs. A first class derived from the ECAs is represented by the set of CAs whose neighborhood is composed of 4 elements (the two nearest left neighbors and two nearest right neighbors). The extension to 3 or more states can also be considered. Another important class of CAs are the totalistic ones that take into account the density of neighbors in a certain state independent of the order of these neighbors (WOLFRAM, 1984b).

3.3.2 Totalistic cellular automata

A totalistic CA is a CA for which the corresponding transition functions depend only on the values/states of the cells' neighborhood and not on the spatial arrangement of the neighborhood, as in the case of the ECAs. Similar to the ECAs, the evolution of the one-dimensional totalistic CA can be described by a table that specifies the state of a given cell in the next iteration based on the average values of the cells belonging to its neighborhood and its current state. Given a one-dimensional CA of m states, there are $3m - 2$ possible configurations with regard to the average values of the states of the three nearest neighboring cells, being the left cell, the right cell and the central cell itself. Therefore, there are m^{3m-2} possible totalistic CAs. Each CA can be indexed by a $3m - 2$ -digit code. Consider, for instance, a one-dimensional CA with $m = 3$, i.e., 3 states. The rule (or code) 777 will be represented in the ternary base: 1001210_3 . This rule is illustrated in Fig. 12.

Figure 12 – Rule 777: Illustration of a transition function of a totalistic one-dimensional CA

Source: Elaborated by the author.

In Fig. 13 we can observe different spatio-temporal patterns obtained through the evolution of totalistic CAs. The number of states is mapped to gray-levels. Figs. 13-(a) and (b) depict the patterns of 3-state ($m = 3$) totalistic CAs corresponding to rules 777 and 600. Figs. 13-(c) and (d) depict the patterns of 4-state ($m = 4$) totalistic CAs corresponding to rules 68900 and 12345. The rule numbers are arbitrary and were chosen randomly. It is worth noting that as m increases, the number of possible rules also increases in a non-linear way. In the case of totalistic CAs, the rule space can be smaller than for the non-totalistic counterparts, as the former CAs only take

¹ <http://mathworld.wolfram.com/Rule30.html>.

into account the states of a cell's neighbors and not their spatial position. Still, there is another class of CAs, so-called *outer-totalistic* CAs, which depends on the states of the neighbors in addition to the state of the central cell.

Figure 13 – Patterns obtained through the evolution of one-dimensional CAs for different values of m and different rules. (a) $m = 3$ and rule 777. (b) $m = 3$ and rule 600. (c) $m = 4$ and rule 68900, and, (d) $m = 4$ and rule 12345.

Source: Elaborated by the author.

3.3.3 Two-dimensional cellular automata

Two-dimensional CAs are defined on tessellations of $\mathbb{R}^2$. A well-known two-dimensional class of CAs are the Life-Like CAs, which are inspired by the rules of the Game of Life. They are defined on the same topology and also on the same neighborhood (*Moore* neighborhood). However, the transition rules may vary. Given the transition rules of the Game of Life, we can describe them as follows (DEMONGEOT; GOLÈS; TCHUENTE, 1985):

$$x_i(t+1) = \begin{cases} 1, & \text{if } \begin{cases} x_i(t) = 0 & \text{and } \sum_{j=1}^8 x_j = 3 \\ x_i(t) = 1 & \text{and } 2 \leq \sum_{j=1}^8 x_j \leq 3 \end{cases} \\ 0, & \text{otherwise} \end{cases} \quad (3.1)$$

Equation 3.1 describes the dynamics of the Game of Life CA where a cell is born (B) if it has exactly 3 alive neighbors, and, a cell survives (S), if at least 2 or 3 neighbors are alive. The following notation is also widely used to represent this system: B3/S23, on which B and S are followed by the number(s) which denote how many alive neighbors a node must have to be born or survive.

For the life-like CA, B and S are *strings* containing values from the set $0, 1, \dots, 8$, which means that each combination for birth and survival conditions represent a Life-Like rule. Therefore, there is a total of 2^{18} transition rules. A life-like automaton is completely defined by the quintuple: $\mathcal{C} = \langle \mathcal{Z}^2, \mathbb{S} = \{0, 1\}, s_0, \mathcal{N} = (c_1, \dots, c_9), \Phi : \mathcal{S}^9 \rightarrow \mathcal{S} \rangle$.

A Life-Like CA is denoted “fertile” if there exists a pattern P such that, given a bounding area belonging to its tessellation, there will be at least one alive cell outside this area after a certain time t when this CA is evolved from P . Similarly, using the same criterion, a CA may exhibit “not fertile” patterns. Eppstein (EPPSTEIN, 2010) presents a detailed explanation of growing and decreasing patterns for Life-Like CAs. Some life-like rules, defined generically by B_x/S_y , have been well investigated in relation to their evolution patterns, as such, they present specific names with particular meanings (Fig. 14). However, the vast majority of life-like rules do not have interpretations or analogies with specific patterns. For example, rule B3678/S34678 (*Day & Night*) is stable and presents the property that the alternation between dead and alive states does not modify the behavior of the CA, because the number of living neighbors that leads to birth is the same number that leads to survival (EPPSTEIN, 2010).

Every pattern of rule B1357/S1357 (*Replicator*) is a replicator, which is also considered a “fertile” (exploding) rule. Rule B45678/S2345 (*Walled Cities*) generates patterns similar to cities or regions. Once the “city” has stabilized, it will not grow anymore, but some activity may be present within the city region (“wall”). B357/S238 (*Pseudo-Life*) is a chaotic rule very similar to the Game of Life, although not all the patterns that appear in the last can be seen in *pseudo-life*. One of the reasons is the *glider* instability, which is a very well-known pattern for the Game of Life. Interestingly, rule B1357/S2468 can be used as a pseudorandom number generator, since it has high entropy and high Lyapunov exponent as demonstrated in (MACHICAO; MARCO; BRUNO, 2012). In the work of (JOHNSTON, 2010) a discussion of the standards provided by rule B36/S125 (2x2) is presented. This name is due to the fact that initial patterns formed by blocks of 2x2 living cells will always be delimited to blocks of this size. This rule also has some properties common to elementary CAs and can emulate rule 90 (JOHNSTON, 2010). In Fig. 14, patterns generated by some of these rules can be seen. They represent the patterns obtained by the evolution of the respective CAs after 500 time steps.

Figure 14 – Patterns obtained through the evolution of life-like CAs for six different rules.

Source: Elaborated by the author.

3.4 Analysis of Spatio-temporal Patterns

Although visual assessment has been employed in the analysis of the spatio-temporal patterns resulting from the dynamics of CAs, scientists have also proposed many quantitative approaches to carry out such task (BAETENS; DE BAETS, 2013). These approaches help to go a step further in CA analysis. Some measures try to quantitatively analyze the dynamical properties of CAs regarding their stability and their characteristic behavior (WOLFRAM, 1984b; LANGTON, 1990; ILACHINSKI, 2001). In addition, by considering that each column of a CA corresponds to the evolution of a single node, many tools for time series analysis can also be employed. In this section we introduce some measures for the characterization of CAs' spatio-temporal patterns.

3.4.1 Hamming distance

Given two sequences of the same length, the Hamming distance counts the number of positions that have different symbols, i.e., it gives the number of substitutions that are required to transform one sequence into the other (HAMMING, 1950). Consider, for instance, the following

binary sequences “1001100010” and “1001001010”. The characters in the fifth and the seventh positions of the first sequence should be changed so that this sequence will be equal to the second one. Therefore, the Hamming distance in this case is equal 2. When it comes to the spatio-temporal patterns of CAs, the Hamming distance can be calculated either for each pair of cells in the same diagram or for two distinct diagrams. Note that this measure is dependent on the ordering of the cells in the diagram. As such, special attention should be given for CAs defined on irregular topologies.

3.4.2 Entropy-related measures

The Shannon entropy is a well-known measure derived from information theory. Conceived by the mathematician Claude Shannon in the 1940s as a result of his studies in the field of communication, this entropy measure has several practical applications such as data compression and transmission concerning digital communication tools (SHANNON, 1948). In addition, information theory is relevant in areas such as cryptography and bioinformatics, in which derived measures are used for pattern analysis of gene sequences (MITCHELL, 2009). One way to understand the definition of the Shannon entropy is that it represents the amount of information contained in a message, in which the communication unit can be a sentence, a word or even a single bit. The entropy is defined according to the probabilities of these messages and is not related to the content of the message. It is defined as $S = -\sum_i P_i \log P_i$, where P_i represents the probability that a symbol i appears in a given sequence.

Considering a binary spatio-temporal pattern, for a single cell c_i , the Shannon entropy is defined by $S_i = -(p_i^0 \log_2 p_i^0 + p_i^1 \log_2 p_i^1)$, where p_i^0 is the probability of having zeros in the time series of c_i and p_i^1 is the probability of having ones. This measure quantifies how homogeneous the evolution pattern is and is normalized to the unit interval. Spatio-temporal series containing only zeros or ones have $S_i = 0$, whereas, oscillating patterns have higher Shannon entropy.

3.4.3 Word count & word entropy

A “word” can be arbitrarily defined within the context of a time series and usually is defined as a block of constant states (MARR; HÜTT, 2012). In the case of binary time series for instance, a word can represent a sequence of ones limited by zeros. Consider, for instance, the following sequence $q = (0011101100)$. We can see one word of length three and one word of length two. Alternatively, a word can also be defined as a sequence of zeros limited by ones. Therefore it is possible to compute the distribution of all the word lengths. The maximum word length is the number of evolution steps, which is the case that a cell started and ended the evolution in the same state and did not change it.

In addition to the distribution obtained through the word counting, we can also define the

word entropy, $H_{W_i} = -\sum_{l=1}^L p_i^l \log_2 p_i^l$, which measures how irregular a time series is in relation to the length of the words given a large time scale (MARR; HÜTT, 2012). The term p_i^l represents the probability of finding words of a given length l .

3.4.4 Lempel-Ziv complexity

The Lempel-Ziv complexity (LZ) is derived from the data compression algorithm proposed by Lempel and Ziv (LEMPEL; ZIV, 1976). This algorithm is based on the construction of a dictionary of distinct “words” (blocks of information) from different sizes that can be found in a text. The number of different blocks that a sequence contains is then taken into account for the calculation of the LZ. For constructing the dictionary, the order of appearance of the sub-sequences should be considered.

For example, given the sequence $s = (010101010101010101)$ of length $l = 20$, the first bit on the left defines the first block, then each bit is read until a sequence that is not yet in the dictionary appears. This new block will be then added to the dictionary. The sequence s will provide the following block decomposition: $0|1|01|010|10|101|0101$. Therefore, the total number of blocks is $g = 7$. The Lempel-Ziv complexity is then given by $L = \frac{g \log l}{l} = 1.049$

3.4.5 Hurst exponent

The Hurst Exponent (H_E) is a correlation measure in time series that can be used to detect long-term regularities (HURST, 1951). This measure is widely used to distinguish between random and non-random time series. The Hurst exponent is directly related to the fractal dimension (D) of a time series such that $D = 2 - H_E$, where $0 \leq H_E \leq 1$. Within this context, H_E measures the smoothness of a fractal time series based on the asymptotic behavior of the so-called rescaled range of a process (KALE; BUTAR, 2011). The Hurst exponent can be estimated according to the following equation:

$$H_E = \frac{\log(R/S)}{\log(T)}, \quad (3.2)$$

where R/S corresponds to the value of the rescaled range and T corresponds to the dimension of the data. The rescaled range is a measure of the time series divergence and is defined as the largest difference (range) between the values of a sample previously centered by the mean in relation to the sample standard deviation and proportional to its size, as follows:

$$(R/S)_N = cN^{H_E}, \quad (3.3)$$

where c is a constant that depends on the time series. The value R/S is proportional to N by a power law whose exponent is equal to H_E , therefore, the R/S analysis allows to describe a time

series without a characteristic scale.

Given a time series $F_1, F_2, \dots, F_N$, the first step towards the calculation of R is the subtraction of the sample mean ($\hat{\mu}$) of each value of the series: $S_i = F_i - \hat{\mu}; i = 1, \dots, N$. Consequently, the time series will have zero mean. The next step is to calculate the cumulative deviation (Γ): $\Gamma_i = S_1 + S_i; i = 2, \dots, N$. The range is the difference between the maximum and the minimum value of Γ_i : $R_N = \max(\Gamma_1, \dots, \Gamma_N) - \min(\Gamma_1, \dots, \Gamma_N)$ and R_N will always be a non-negative value. Finally, the rescaled range is given by R_N/S , where S is the standard deviation of the time series. For the estimation of the Hurst exponent, the entire process performed for the calculation of R is repeated for smaller and smaller intervals of the series that are proportional to N : $N/2, N/4, N/8$. As such, for each subdivision of the series (in relation to its size N) it is possible to obtain an average value of R/S . Finally, H_E is given by the angular coefficient of the regression line of $\log(R/S)$ vs $\log(N)$.

The value of H_E varies between 0 and 1. For $H_E = 0.5$, we have an uncorrelated and independent time series. The range $0 < H_E < 0.5$ indicates that a high value in the series will probably be followed by a lower value, which will be, in turn, followed by a higher value and so on. As H_E approaches zero, the oscillation increases. A value in the interval $0.5 < H_E < 1$ indicates that the series is positively self-correlated, i.e., high values tend to be followed by higher values and vice versa.

3.4.6 Lyapunov exponent

For continuous dynamical systems, the Lyapunov exponent measures the average exponential rate of divergence given two close trajectories in the phase space. Therefore, the LE quantifies the sensitivity of the dynamical system to the initial conditions (ECKMANN; RUELLE, 1985; OTT, 2002). The definition of the Lyapunov exponent was extended to CAs by tracking all the possible “defects” resulting from the divergence of two initially close CA configurations (BAETENS; DE BAETS, 2010b; BAETENS; DE BAETS, 2013). The LE for a two-state CA is defined as

$$\lambda = \lim_{t \rightarrow \infty} \frac{1}{t} \log(\varepsilon_t), \quad (3.4)$$

where t corresponds to the number of time steps and ε_t is the number of defects. The calculation of ε_t will be illustrated next.

Given a two-state CA $\mathcal{C}$ with an initial configuration denoted as s_0 . Let us consider its perturbed counterpart, s_0^* , which has only one cell that differs from s_0 , i.e., there is only one cell c_j for which its initial state is different in both initial configurations. In the case of binary CAs, if $s_0(c_j) = 1$, then $s_0^*(c_j) = 0$ and vice versa. The co-evolution of $\mathcal{C}$ from s_0 and s_0^* gives then all the possible defects/perturbations between these configurations at time t for which $s(c_k, t) \neq s^*(c_k, t)$. Fig. 15 illustrates the propagation of such defects from a single cell

for an elementary CA. The black circles correspond to the CA cells with defects and the arrows shows how they are propagated. Due to the neighborhood size of an ECA, a single defect can be propagated for at most 3 cells, as in the example of these figure. Considering this maximum propagation for $t = 3$, then the total number of defects will be $\varepsilon_t = 27$, since one cell can be affected by the same kind of defect through different ways. All the possible ways of defect propagation are then taken into account when calculating ε_t .

Figure 15 – All possible ways that a single defect (black circle) can propagate in the evolution space of an ECA (arrows).

Source: Adapted from [BAETENS and DE BAETS \(2013\)](#).

Therefore, in the case of ECAs, the maximum number of defects at time t is given by 3^t , which shows that ε_t can grow exponentially through time. This way, the Lyapunov exponent can then be defined within the context of CAs. The respective definition complies with the one for continuous dynamical systems ([BAETENS; DE BAETS, 2013](#)). In addition Baetens&DeBaets show that for CAs defined on arbitrary tessellations and that can be represented as graphs, the maximum number of defects at time t is given by C^t , being C the average number of neighbors among the CA cells. It should also be noticed that the definition presented in Eq. 3.4 leads to the largest LE of a CA.

3.5 Conclusions

This chapter presented the main theoretical foundations related to the context of this thesis with regard to the CA theory. A formal definition of CAs was introduced and different types of CAs were presented. Among them, we could highlight the well-known ECA, which is a one-dimensional CA. Other two important definitions were introduced, the totalistic and *outer-totalistic* CAs, for which only the states of neighbors are taken into account irrespective of their spatial position. Finally, we discussed how the spatio-temporal patterns obtained through the CA evolution can be analyzed by means of different quantitative approaches. The next chapter discusses the state of the art on Network Automata by providing a discussion through different

related works on how a CA can be modeled over irregular tessellations like networks and Voronoi diagrams.

STATE OF THE ART ON NETWORK AUTOMATA

4.1 Introduction

Most of the works developed since the popularization of CAs in the 70s use regular tessellations in the Euclidean space. Regular tessellations are homogeneous concerning each cell and its neighborhood. Beyond the ease of implementation, such regular tessellations are also computationally efficient. However, studies discussing CAs on irregular tessellations have been much less explored.

This chapter presents different approaches proposed in related works that lie at the intersection among the areas of graph theory, complex networks, cellular automata and other types of automata. The first section brings a mathematical formalization of CAs on irregular topologies, such as Voronoi tessellation. Subsequently, we introduce some works that relate CAs to classic models of complex networks. In another section, the pattern recognition context in networks will be highlighted. Finally, some methods using partially self-avoiding deterministic walks on networks are discussed.

4.2 Cellular Automata on Irregular Tessellations

CAs defined on irregular tessellations are formally introduced by Baetens&DeBaets ([BAETENS; DE BAETS, 2010b](#); [BAETENS; DE BAETS, 2010a](#); [BAETENS; DE BAETS, 2012](#)). The authors present a *framework* for a mathematical definition of totalistic CAs on irregular tessellations. Such a framework is more general since it is based on the well-known Voronoi tessellation ([OKABE *et al.*, 2009](#)), and can be advantageous for modeling phenomena that occur on surfaces with complex geometries, such as spatial entities in geographic systems.

These entities are usually represented by arbitrary regions of the Euclidean space that share physical, geographic, and biological properties, such as vegetation (BAETENS; DE BAETS, 2012).

Given a CA defined by the quintuple $\mathcal{C} = \langle \mathcal{T}, S, s_0, \mathcal{N}, \Phi \rangle$, its tessellation $\mathcal{T}$ will be represented by the set of polygons that make up the tessellation: $\mathcal{P}_n = \{p_1^n, p_2^n, \dots, p_t^n, \dots\}$. The neighborhood of each cell c_i , $(c_{ij})_{j=1}^{|\mathcal{N}(c_i)|}$ is defined by adjacency, i.e., two cells are neighbors if $p_i^m \cap p_j^m \neq \emptyset$. Similarly to regular tessellations, the neighborhoods of Moore and von Neumann can also be defined for irregular tessellations. The difference between them is that for the von Neumann neighborhood two cells are considered neighbors if their respective polygons' boundaries have more than one point in common.

An important contribution of this work concerns the definition of the transition functions of such irregular CAs. Since the number of neighbors is different for each cell c_i , the sum of the states of the neighboring cells (σ_i) becomes unbounded. To overcome this situation, Baetens & De Baets adopt a threshold θ for this sum: when $\sigma_i > \theta$, the CA's output function is defined by $s(\sigma_i) = s(\theta)$, i.e., when the sum of the states of the neighbors is greater than the threshold, the state of cell c_i is given by θ .

Some questions raised by Baetens & De Baets are related to the dynamical properties of irregular CAs. For instance, how similar they are to the properties of regular CAs and how they can be investigated. In addition, to what extent the dynamical properties of CAs depend on the underlying tessellation? In (BAETENS; DE BAETS, 2010b), a quantitative analysis of the dynamics of CAs is explored using the Lyapunov Exponent (LE) (Section 3.4.6) and the Jacobian. The first is inspired by the study of continuous dynamical systems and provides an analysis of the instability of such systems by means of the trajectories in the phase space given small perturbations in the initial configurations of the cells.

The representation of CAs by graphs is also explored by the authors. For families of totalistic CAs, which are invariant to the ordering of the neighborhood elements, there is an equivalent representation through the use of graphs that can be explored in order to simplify the CA implementation. The graph $G(V, E)$ is composed of the set of vertices $V = \mathcal{T}$ and by the set E of edges such that an edge e_{ij} , defined between cells c_i e c_j , will belong to E , if $c_j \in \mathcal{N}(c_i)$.

4.3 Cellular Automata on Graphs and Complex Networks

In one of the first works that integrates the areas of CAs and CN, Watts (WATTS, 1999) discusses the evolution dynamics of CAs on small-world networks concerning their performance on tasks such as the majority problem and network synchronization. Therefore, the restriction to regular tessellations is relaxed and a Generalized Network-Automata (GNA) is defined. The famous majority problem (also known as the density classification task) (CRUTCHFIELD;

(MITCHELL, 1995; TOMASSINI, 2006) consists of determining the majority state (0 or 1) in an initial configuration. Starting from this configuration, we have the density ρ_0 which represents the fraction of cells in state 1. If $\rho_0 > 0.5$, then at the end of the evolution, all cells must be in state 1 (*on*), otherwise all cells should have a state equal to 0 (*off*). This is a simple problem for a system with a central processor. However, when considering a parallel architecture such as CAs, this task becomes more complex, since the only information that each cell has is the configuration of the states of its neighbors. An analogy can be made to the task of determining which candidate will win an election, given that the only information each voter possesses is the vote of his/her neighbors, as well as his own vote.

The majority problem has been widely explored in literature. The main purpose is to find a rule capable of giving the majority vote. The simple choice of the rule that uses the local majority state among the neighbors of a cell c_i performs very poorly as demonstrated in (DAS; MITCHELL; CRUTCHFIELD, 1994) for CAs in regular topologies. Some solutions make use of genetic algorithms (CRUTCHFIELD; MITCHELL; DAS, 2003). The performance of a rule is typically evaluated by the fraction of correct classifications on $n = 10^4$ different initial configurations defined according to a binomial distribution, and, randomly chosen among the possible 2^N binary strings, N being the number of elements of the CA. This distribution is centered at $\rho_0 = 0.5$. Watts showed that the use of the small-world network as tessellation for the majority problem provides a performance of 0.85, while the highest performance obtained by both evolutionary and exhaustive algorithms was 0.8. The transition rule used by Watts was the local majority which, as discussed earlier, performs very poorly on regular tessellations.

The same approach was used by Watts for the synchronization problem in CAs. Given a random initial configuration of a binary CA, the synchronization task consists of uniform alternation of states by the CA cells after some time, i.e., at time t_i all cells assume state 1, at time t_{i+1} all cells assume state 0, and so on alternately. Watts uses the same rule of the local majority, but in this case, if most of the neighbors of a cell have a state equal to 1 then the cell reverses its state in the next time step. The results for the synchronization task were similar to those obtained for the majority problem. By relaxing the constraint of the regular topology, it is possible to obtain a network automaton with a good performance for the synchronization problem. The performance in this case is evaluated for different initial configurations with density drawn from $[0, 1]$ (TOMASSINI, 2006).

Tomassini *et al.* also investigated the majority and synchronization problems in small-world networks (TOMASSINI; GIACOBINI; DARABOS, 2005; DARABOS; GIACOBINI; TOMASSINI, 2007; DARABOS *et al.*, 2011; DARABOS; MOORE, 2015). The authors used evolutionary algorithms in order to identify the network topology that would provide the best performance for the majority problem. They also used the local majority rule as transition function (the same as used by Watts). In this approach, the structure of a network represents a single individual and the fitness function is calculated by means of the fraction of correct classifications

over $n = 10^4$ different initial configurations, generated using a binomial distribution centered at $\rho_0 = 0.5$. The initial population is obtained starting from a regular network with degree $k = 4$. The structure of this network is gradually changed by the addition of new edges for each node, which happens with low probability. The best network in this set leads to a performance of 0.823. In another experiment, starting with random graphs, the highest performance was 0.821. Besides performance evaluation of the networks, an important contribution of these works relates to the proposal of an evolutionary model of small-world networks in tasks such as the majority problem. The robustness of the topologies evolved through the use of CAs was tested in two different situations. In the first, a failure probability of a rule to generate an incorrect output was defined for each cell while the network structure remained unchanged. In the second case, there are edge failures, i.e., the link between two nodes of the network is removed. The results indicate that the network-automaton offers greater robustness to these failures when compared to regular topologies that perform the same tasks of synchronization and density classification.

Still in the same research branch of Tomassini *et al.*, other works also explore the network automata (NAs). This term is used in (GOG; CHIRA, 2012). In this work the majority problem is addressed through the use of networks and genetic algorithms. In another work (MACÊDO; OLIVEIRA; RIBEIRO, 2013), the dynamics of chaotic ECAs is evaluated in regular and small-world topologies. Smith *et al.* (SMITH *et al.*, 2011) propose a class of NAs such that their topology changes as a function of time, in contrast to the CAs defined on the network. In the latter, only the evolution of the nodes' states is observed. This approach comes from the fact that while the topology may restrict the type of dynamics occurring on the network, in many scenarios, the dynamical processes may influence the evolution of the topology.

Influenced by the advances in the area of complex networks, Marr & Hütt studied known models of CAs on random graphs and small-world networks. In a first work (MARR; HÜTT, 2005), the authors investigated how the topology of the graph can determine and/or restrict the ability of a CA to reveal specific dynamics, for which the authors used the same transition rule in different topologies. This rule was analyzed regarding the spatio-temporal patterns obtained through the evolution of a binary CA modeled on the network structure. Since the number of neighbors of each network node can vary considerably, the transition function uses the density of alive neighbors of node i : ρ_i . In this way, the set of transition rules is given by:

$$x_i(t+1) = \begin{cases} x_i(t), & \text{if } \rho_i \leq \delta, \\ 1 - x_i(t), & \text{if } \rho_i > \delta, \end{cases} \quad (4.1)$$

where $x_i(t)$ is the state of node i at time t and δ is the density threshold above which node i will update its current state. The state of node i itself can be taken into account when calculating ρ_i . In this case, the set of transition rules will be different. The authors also evaluate the effects of this second set of rules.

Firstly, the evolution dynamics given by Eq. (4.1) was analyzed on regular graphs. Using

the same value for the number of nodes n and maintaining the same neighborhood ordering as for the ECAs, the four spatio-temporal patterns defined by Wolfram were obtained by the authors for the set Ω . This can be explained by the fact that for small values of δ , the majority of the nodes will change their current state and an oscillatory pattern will emerge, whereas, for larger values of δ , the stationary pattern becomes more frequent, since the majority of the nodes will not change their current state. For intermediate values of δ , complex and chaotic patterns can be observed. The same analysis of the parameter δ was performed for random and small-world graphs. It was observed that the range of δ for which complex patterns begin to emerge is different for each model. With the transformation of a regular network into a random network the local collective behavior is gradually lost. Beyond the interval in which the small-world characteristic prevails, the long-range correlations are lost, and, as a consequence, the spatio-temporal patterns obtained through the CA evolution become random.

For the quantitative analysis of the spatio-temporal patterns, Marr and Hütt used the Shannon entropy (S) and the word entropy (W). These measures are commonly used to analyze the evolutionary dynamics of CAs. The authors analyzed the trajectory in the plane defined by W and S which shows the word entropy in relation to the Shannon entropy as a function of the parameter δ and used this trajectory as a dynamic signature of a graph. Therefore, this signature represents the complex dynamics of the graph (MARR; HÜTT, 2005).

In another work (MARR; MÜLLER-LINOW; HÜTT, 2007), the same authors use the dynamics defined by Eq. (4.1) on metabolic networks. The results show that this dynamics presents specific properties for this class of networks. In the plane defined by W and S , it is possible to observe that the metabolic networks occupy a low entropy region when compared to equivalent randomized networks, which have the same degree sequence. The randomization process changes important topological properties. Consequently, one thing that can be observed is that the values of the average shortest path and the clustering coefficient decrease. In addition, modular structures of the network are lost.

In a further work (MARR; HÜTT, 2009), the authors extended the methodology discussed earlier to the implementation of CAs on arbitrary topologies. Some restrictions for such modeling are imposed: 1) homogeneity, i.e., the same rule must be applied to all the nodes of the graph; 2) isotropy, ie, the CA rules cannot depend on the ordering of the elements belonging to a node's neighborhood, but depend on the respective density ρ , and, 3) the CA dynamics is given as a function of ρ . In addition the authors present a generalization of the majority rule for binary CAs. The proposed class of CAs is called outer-totalistic because it depends only on the state of cell x_i at time t and on the density of neighboring cells (ρ_i). The presented formalism can be easily extended to the use of more than one density threshold or to a bigger state space. One of the conclusions of the experiments carried out in this article is that the occurrence of complex dynamics varies systematically with the connectivity of regular graphs. The same analysis carried out for the metabolic networks was also performed using the dynamics defined by this outer-

totalistic CA. This analysis showed a decrease of the entropy values. The authors concluded that although the entropy attributes obtained for the majority of the rules do not discriminate between real and randomized topologies, a small set of rules is topologically sensitive. For this set, the entropy values of the real (metabolic) networks are always smaller when compared to the randomized counterparts.

Concluding the approach proposed by Marr and Hütt, in a more recent work, the authors complement their previous results with the investigation of which topological features may be associated with the low entropy values obtained through the spatio-temporal patterns of the NA (MARR; HÜTT, 2012). For that, an approach based on Simulated Annealing was employed. The simulation of the CA evolution is performed using entropy minimization as the target function in order to explore which network properties are highlighted under this dynamics. Starting from a directed random graph with 100 nodes and approximately 200 edges, a CA defined on this topology and governed by the dynamics of Eq. (4.1) is iterated 200 times and the Shannon and word entropies are then calculated. The initial states of each vertex were randomly defined. Subsequently, a randomly chosen edge is rewired. Then, the evolution of the CA, as well as the calculation of entropies are carried out once more. The new graph (obtained after edge rewiring) is maintained if $\Delta E^{(i)} = E^{(i-1)} - E^{(i)} \geq 0$, where $E^{(i-1)}$ is the average entropy for the first configuration of the graph, and, $E^{(i)}$ is the average entropy for the second configuration. Therefore, if the entropy decreases when rewiring the edges, the graph is maintained. Otherwise, if $\Delta E^{(i)} \leq 0$, the new graph is maintained with the following probability: $P_i = \exp(\Delta E^{(i)} / T^{(i)})$, on which $T^{(i)}$ is the temperature defined by $T^{(i)} = T^{(0)}(1 - i/s)$, $T^{(0)}$ being the initial temperature, and s the number of CA evolutions. The entire process is repeated until the end of the simulation. During the simulation, the number of edges is kept fixed.

For the evaluation of the results, the in- and out-degree distributions were calculated for the input and the output graph. No significant differences between the out-degree distributions were observed for the two graphs. Both graphs had binomial distributions. However, the distributions of the in-degree are different for the two graphs. The standard deviation of the in-degree distribution increases linearly with the decrease of the Shannon entropy. In addition, the graph resulting from the simulation of the in-degree distribution approaches a power law. The authors pointed out that although the degree distribution of the graphs resulting from simulation is not comparable to the power law distributions of scale-free networks (in part due to the small number of nodes in the graph), it is worth noting that the increase in the standard deviation is related to the low entropy requirement.

4.4 Automata Models in the Study of Scale-free Topologies

Since many real networks have characteristics of scale-free models, this type of topology and its dynamical properties have caught the attention of the research community. In this section, we describe some works that present different approaches for the study of the dynamics of scale-free networks through the use of automata-based models. In the first work, Aldana ([ALDANA, 2003](#)) presents a study of Boolean networks defined on scale-free topologies. The analysis carried out by the author is inspired by the famous NK model, proposed by Stuart Kauffman in 1969, which describes the dynamics involved in the process of gene regulation and on cell differentiation ([KAUFFMAN, 1969](#)). This model is also known as Random Boolean Networks (RBN). In this model the genome of an organism is represented by a set of N genes that can assume states 1 or 0, which represent the expression of the gene or not, respectively. Kauffman assumes that the genome is a directed network and each gene is controlled by exactly K other genes. In addition, each gene has a probability p of being expressed. An interesting result is that even with these simplifications, it is possible to observe the existence of a transition phase controlled by the parameters K and p . There is a critical value, K_c , below which all disturbances in the initial state of the network do not propagate. For values of K , such that $K > K_c$, the dynamics becomes chaotic. When $K = K_c$ (critical phase) the NK model presents stable properties that are compatible with the gene regulation network of living beings.

Aldana studies the NK model in networks with scale-free topology, performing the same analysis of the phase transitions, but as a function of the γ exponent. For this kind of topology the connectivity (given by the parameter K) does not characterize the network due to the great variability of the nodes' degrees. The author shows that for the scale-free topology there is no need of a low connectivity (K) for the system to remain in the stable phase. This eliminates the necessity of a very restricted adjustment of the parameters K and p in order to achieve stable dynamics. Furthermore, since the phase transition occurs for values of γ in the interval $[2, 2.5]$, there are elements with a much greater degree variation in the stable and critical phases. A final hypothesis of the author based on the obtained results is that the scale-free topology can favor the evolution and a functional adaptation of the network from a biological perspective ([ALDANA, 2003](#)).

In a paper by ([WU; XU; WANG, 2007](#)), the authors use another CA model to study the dynamics of scale-free networks. This model is a two-dimensional CA known as the Greenberg-Hastings model ([GREENBERG; HASTINGS, 1978](#)). This CA is defined on a rectangular grid and each cell can be in one of three possible states: resting (or quiescent), excited or refractory. There is a set of rules that govern the dynamics of the cells at each time step. The authors use this CA as a model of neuronal activation in scale-free networks for the study of the average activity as a function of external stimulus.

In another work (ZHOU; LIPOWSKY, 2005), the Ising model was employed to study the characteristics and the generic properties of scale-free networks. The states of each vertex are described by a Boolean dynamics and the global dynamics of the network is then characterized by the temporal patterns obtained through the network evolution. The results indicate that this dynamics varies with the power law exponent γ , and that there is a critical point for $\gamma = 5/2$. In (BARRAT; WEIGT, 2000), the dynamics of the Ising model was also used in the study of the topological properties of networks, but with the focus on small-world networks.

4.5 Networks & Pattern Recognition Applications

Pattern recognition in complex networks can be employed in the following areas: 1) structural characterization by means of different measures that can be extracted from the network topology and used to compose the feature vector; 2) structural pattern characterization, also named as motifs. These structures are subgraphs that appear in real networks but are statistically little observed in equivalent random networks, i.e., the probability of occurrence of motifs in random networks is very small, and 3) community detection. Communities can be defined as clusters of vertices densely connected with each other.

The objectives of the present thesis can be linked to the first pattern recognition approach, since we propose new network descriptors that are derived from the evolution of a NA. Within this context, related works make use of structural measures to characterize networks and consequently to classify them or their structures. In a survey (COSTA *et al.*, 2007), Costa *et al.* summarize different classes of measurements that can be used for this task. Some of those are presented in Chapter 2. Such measures are used in tasks related to representation, characterization, classification and modeling of complex networks. An application related to the classification of complex network models is also presented in this article. This application is extended in another work (COSTA *et al.*, 2010) by including several steps within the pattern recognition approach. In addition to the characterization of the networks by means of different structural descriptors, the authors also discuss some data reduction methods in this context, such as PCA (Principal Component Analysis) and canonical analysis. The classification of the network models is discussed at several levels: (i) local, which takes into account the clustering of nodes according to local characteristics, allowing a comparison of these clusters with communities; (ii) intermediate, in which the structural properties can be used for the classification of communities, and, (iii) global, in which networks can be grouped according to global characteristics and classified hierarchically.

Silva & Zhao (SILVA; ZHAO, 2015) present a high-level classification method. This class of methods takes into account topological attributes in contrast to the so-called low-level classification which, in turn, makes use of attributes such as physical measures, similarity criteria, distributions, among others. High-level classification methods explore the relationships between

the instances of the dataset incorporating intrinsic patterns that may reflect the topological properties of the networks. Specifically, the authors used the so-called “deterministic tourist walk”, an agent-based dynamical model (LIMA; MARTINEZ; KINOCHI, 2001; STANLEY; BULDYREV, 2001).

Starting from the network built on the input dataset, the tourist walk was employed to capture the complex topological patterns of these networks. The proposed classifier is controlled by a parameter $\lambda \in [0, 1]$, which allows prioritizing the high-level in relation to the low-level classification, i.e., when the spatial complexity of a particular class is high, the classifier will only be based on the proposed high-level classification. On the contrary, λ can be adjusted in order to balance the weight of both classification methods. In this way, the authors used the proposed approach in the context of pattern recognition by combining both types of classification. They observed that the high-level classification is especially useful for more complex datasets.

4.6 Partially Self-avoiding Deterministic Walks on Networks

In addition to CAs, within the context of pattern recognition, the partially self-avoiding deterministic walk, which was introduced in the previous section, can be used (LIMA; MARTINEZ; KINOCHI, 2001; STANLEY; BULDYREV, 2001). In this model, also known as the “tourist walk”, given a d -dimensional space, with N randomly distributed points, the “tourist walk” dynamics is given by the following rule: starting from a site i , move to the nearest site that has not been visited in the last μ time steps. The rule for choosing the nearest site is defined according to the problem at stake, for example, the Euclidean distance between two sites (CAMPITELI *et al.*, 2006). In this way, the parameter μ , which accounts for the system “memory” and becomes a spatial scaling factor, according to the walking rule. Within this context, the tourist walk approaches a CA for which the evolution dynamics is limited to a given topology.

Starting from a city, the tourist performs a trajectory that can be described as having a transient part and an attractor. The transient accounts for the cities visited by the tourist from the initial city until the first city belonging to the attractor. The attractor is defined as the set of cities from which the tourist cannot escape, i.e., the cities that will always be visited by the tourist. The cardinality of this set is defined as the period, ρ , of the attractor. Although simple to implement, the tourist walk may present complex behavior, which can be identified by the joint distribution of the transient time and the attractor period. The corresponding distribution provides the number of walks with transient time $\mathcal{T}$ and attractor period ρ . The joint distribution can then be used in pattern recognition tasks, since it is an intrinsic feature of the tourist walk (CAMPITELI *et al.*, 2006).

The tourist walk has been employed in computer vision for the analysis of static and

dynamic textures (GONCALVES; BRUNO, 2013; GONCALVES *et al.*, 2012) and in the classification of complex network models (GONÇALVES; MARTINEZ; BRUNO, 2012). Modeling the tourist walk on the network topology goes as follows (GONÇALVES; MARTINEZ; BRUNO, 2012): Given a network with a set of vertices $V = \{v_1, \dots, v_n\}$ and a set of edges $W = \{w_{v_i, v_j} | v_i, v_j \in V, w_{v_i, v_j} \in \mathbb{R}\}$, where each component w_{v_i, v_j} represents the weight of the edge that connects vertices i and j , the possible steps that the tourist can choose are completely defined as a function of W . Let $T = \{t_0, t_1, \dots, t_i | t_i \in V\}$ be a list of vertices that keeps the tourist trajectory, the memory M_i is characterized by a subset composed of the last μ vertices visited by the tourist: $M_i = \bigcup_{k=i-\mu}^i t_k$.

At each time step, the tourist moves to the nearest vertex v , which does not belong to the memory M_i . Since more than one vertex may satisfy this condition, the following set represents the possible choices of the tourist: $\zeta = \{v \in V | \arg \min_{v \in \eta(t_i)} w_{v, t_i} | v \notin M_i\}$, in which, $\eta(t_i)$ is the neighborhood of vertex i . Finally, by an iterative process the trajectory is given by the following equation:

$$t_{i+1} = \begin{cases} \zeta_1, & \text{if } n(\zeta) = 1 \\ \phi(\zeta), & \text{otherwise} \end{cases} \quad (4.2)$$

If ζ contains only one vertex, it will be chosen. Otherwise, ϕ is a function that returns the chosen vertex, which can be given randomly or according to any other specific operation (GONÇALVES; MARTINEZ; BRUNO, 2012). After a transient time t , the tourist gets stuck in the ρ attractor. This process is repeated until the attractor is detected, which is defined as follows:

$$End \leftarrow \begin{cases} \zeta = \emptyset \\ \exists \sum_{\rho, it} = 0, 0 \leq \rho, it \leq i \end{cases} \quad (4.3)$$

where $\sum_{\rho, it} = \sum_{j=0}^{\rho-1} |\varphi(t_{it+j}) - \varphi(t_{it+j+\rho})|$ and $\varphi(v)$ is the index of vertex v , e.g., $\varphi(v_1) = 1$, $\varphi(v_2) = 2$. The walk ends if the tourist did not find an attractor and $\rho = 0$, or, if $\sum_{\rho, it} = 0$. For example, $\varphi(v_1) = 1$, $\varphi(v_2) = 2$, and so on. If $\sum_{\rho, it} = 0$, an attractor of period ρ and a transient time $t = it - 1$ is obtained.

The work of Gonçalves *et al.* approaches some of the methods presented in this thesis since it uses a spatial explicit model in order to obtain attributes from spatio-temporal patterns. The authors have successfully employed the proposed methodology in pattern recognition tasks in different applications. In addition, the proposed approach could be easily adapted to the topology of networks, being useful in the study of arbitrary topologies.

4.7 Conclusions

This chapter presented a set of works related to the context of this project. The main relation between the highlighted works is the use of complex networks aiming to model different problems, as well as the characterization and analysis of these networks through methods based on cellular automata theory. In addition, different methods were presented in the context of pattern recognition. The present proposal differs from the presented works as it uses the dynamic evolution of the network automata as an attribute for the characterization of CN. From the space-time pattern of the automaton a descriptor is defined that can be used to classify the network. The objective of the next chapter is to characterize the proposal of this project in an objective way regarding the methodology to be used.

Part II

Spatially Explicit Modeling on Networks Using Cellular Automata

NETWORK AUTOMATA: MODELING & ANALYSIS

5.1 Introduction

The classical CA paradigm is based on regular tessellations. Although the use of irregular tessellations was proposed almost three decades ago ([COUCLELIS, 1985](#)), few studies addressed this topic. Besides, there is a lack of a robust theoretical framework to describe such CAs ([BAETENS; DE BAETS, 2012](#)). In many cases, the use of square or hexagonal tessellations defined on the Euclidean space, for two-dimensional CAs for instance, is due to their simple implementation and computational efficiency. However, the use of a regular spatial domain may limit the applicability of CAs. The constraints of regular topologies within the CA framework have led to questions regarding how such models can be used in real-world problems that present a much more complex geometry. This is the case, for instance, of mass flow models, for which square tessellations can restrict the possible flow directions ([BAETENS; DE BAETS, 2012](#); [D'AMBROSIO; GREGORIO; IOVINE, 2003](#)). Alternative definitions of cellular spaces can overcome some of these limitations ([COUCLELIS, 1985](#)). In addition, there is a variety of well-known tessellations that could be used, such as the Voronoi tessellation and the Delaunay triangulation ([OKABE *et al.*, 2009](#)). More recently, a mathematical formalization of CAs on irregular tessellations was developed by Baetens & De Baets ([BAETENS; DE BAETS, 2012](#)). Their proposal takes into account an enumeration scheme for totalistic and outer-totalistic CAs defined on irregular tessellations.

Using more realistic spatial domains is also one of the main motivations when using network-based models for modeling real-world phenomena. For some families of CAs, like the order-invariant and the outer-totalistic CAs, there exists a graph representation that can be employed in order to simplify their implementation ([BAETENS, 2012](#)). If the CA is order-variant, it relies on the spatial arrangement of the neighborhood besides the states of the neighbors,

therefore, in this case, a graph representation is not possible, since the ordering of the cells is not maintained by the graph structure. If a graph can be generated, then the set of CA cells is represented by the vertices of the graph $\mathcal{V}$ and the CA neighborhood relations are mapped to a set of edges $\mathcal{E}$. The corresponding graph $\mathcal{G}$ is an undirected graph. As important as the construction of the graph is the definition of the transition function of such graph-CA. Recalling that in this thesis we use the words “graph” and “network” interchangeably.

As discussed in Chapter 4, some authors have already proposed models that integrate CAs and graphs and present specific nomenclature for them. In this thesis we will use the term Network Automata (NA) to refer to such combination. This term is in accordance with previous literature and corresponds to a CA that is defined on the network topology. NAs can also be seen as generalization of CAs concerning the tessellation. In this chapter and the subsequent ones, we will introduce three different NAs. The first one is an extension of the NA proposed by Marr&Hütt (MARR; HÜTT, 2005; MARR; HÜTT, 2009) for different theoretical network models. This work is already published (MIRANDA; MACHICAO; BRUNO, 2016b). The second one is based on a “birth-survival” dynamics with a more complex transition function, and, the third one is inspired by the classic Life-like two-dimensional CA (GARDNER, 1970), named as Life-like Network Automata (LLNA). Finally, in the last chapter of this part, we introduce the well-known Rock-Paper-Scissors (RPS) game on networks. The RPS is mainly used as a competition model and has already been well studied on regular topologies. We analyze the effect of the tessellation on the RPS dynamics.

5.2 Network Automata Based on Neighborhood Density

The first NA model that will be discussed in this chapter was introduced in previous works and was used to solve the density classification and synchronization problems on CAs (WATTS, 1999; MARR; HÜTT, 2005; MARR; HÜTT, 2009). The transition function of this NA takes into account the density of the neighbors’ states.

5.2.1 Modeling the network automata

Given a NA described by the quintuple $\mathcal{C} = \langle \mathcal{T}, \mathbb{S}, s_0, \mathcal{N}, \phi \rangle$, we assume the following: (i) each network node corresponds to a NA cell and the tessellation $\mathcal{T}$ is represented by the network. (ii) The set of states $\mathbb{S}$ consists of two elements, where $s_i = 0$ corresponds the “dead” state and $s_i = 1$, the “alive” state, and, $s_i \in \mathbb{N}$. (iii) s_0 is the initial configuration for all cells $c_i \in \mathcal{T}$ i.e., $s(c_i, 0) = s_0(c_i)$. (iv) The neighborhood $\mathcal{N}$ is given by the adjacency matrix A , where $a_{ij} = 1$, if i is connected to j and $a_{ij} = 0$, otherwise. As expected, the degree of each node, k_i , varies, and, therefore, each network has a specific degree distribution. Due to this fact we

adopted the neighborhood density (ρ_i) of node i for a given state $s_k = s$ as follows:

$$\rho_i(s_k) = \frac{1}{k_i} \sum_{j=1}^N a_{ij} \theta_{s_j, s_k}(t), \quad (5.1)$$

where $\theta_{s_j, s_k}(t) = 1$, if $s_j = s_k$, and $\theta_{s_j, s_k}(t) = 0$, otherwise. In addition, a_{ij} defines the neighborhood and θ_{s_j, s_k} accounts for the condition that the nodes have the same state. When the state space is composed of only two elements, ρ_i can be simplified to $\rho_i = \frac{1}{k_i} \sum_{j=1}^N a_{ij} s_j(t)$. Moreover, $\rho_i \in [0, 1]$ and accounts for the proportion of the neighbors of a cell that are alive. Finally, (v) the transition function ϕ incorporates the dynamic of the nodes according to the following equation:

$$x_i(t+1) = \begin{cases} x_i(t), & \text{if } \rho_i \leq \delta, \\ 1 - x_i(t), & \text{if } \rho_i > \delta, \end{cases} \quad (5.2)$$

where δ is a threshold for ρ_i under which the node maintains its state at the next time step, otherwise, the node inverts its state.

Figure 16 illustrates the evolution of a sample network given $\delta = 0.5$ using Eq. (5.2). The first network ($t = 0$) represents the initial configuration where black cells represent the “dead” nodes, while white cells, the “alive” nodes, i.e., $s_i = 0$ and $s_i = 1$, respectively. The subsequent networks depict the configuration at different time steps. When $t = 0$, the majority of the neighbors of cell c_i are alive and its density is equal $\rho_i = 0.67$. Therefore, given $\delta = 0.5$, cell c_i will change its state. However, at the subsequent time steps its density will be always lower than δ and the state of cell c_i remains constant.

Figure 16 – Evolution dynamics of a sample network according to Eq. (5.2).

Source: Adapted from [Miranda, Machicao and Bruno \(2016b\)](#).

5.2.2 Spatio-temporal patterns

It is also possible to represent the evolved pattern of a NA as a spatio-temporal diagram where each column represents a network node and each row represents the states of all nodes at each time step. However, due to the irregular tessellation for what concerns the distribution of number of neighbors across the nodes, it is necessary to define some ordering criterion to

construct the diagram. We chose the node degree k_i for that purpose. Consequently, the left-most nodes (concerning the columns of the diagram) are the ones with the smallest values of k_i , and, the right-most nodes correspond to the ones with the highest values of k_i . Figure 17 illustrates some spatio-temporal patterns obtained for different types of networks: regular (a), small-world (b), and, random (c). All the networks in Fig. 17 have $N = 200$ nodes and they were evolved for 200 time steps.

Figure 17 – Spatio-temporal patterns for three different network types obtained by applying the neighborhood density NA. (a) Regular network. (b) Small-world network. (c) Random network.

Source: Adapted from [Miranda, Machicao and Bruno \(2016b\)](#).

The patterns obtained for the regular network (Fig. 17-(a)) are related to the patterns obtained for the ECAs. Since there are no long-range correlations in this network type, we can observe the maintenance of more structured patterns. On the contrary, in the case of the random network (Fig. 17-(c)), we can see an oscillating pattern and the absence of structures that could lead to chaotic or complex structures, according to the Wolfram classification ([WOLFRAM, 1984b](#)). In the case of the small-world network (Fig. 17-(b)), its spatio-temporal pattern has the characteristics of the two other types.

5.2.3 Quantitative analyses using the Shannon entropy

In order to arrive at a more quantitative analyses of the obtained spatio-temporal patterns, we evaluated the evolution of each node as a binary time series containing zeros and ones. The Shannon entropy was calculated for each node according to the following equation: $S_i = -(p_i^0 \log_2 p_i^0 + p_i^1 \log_2 p_i^1)$. As presented in Section 3.4.2, p_i^0 represents the frequency of zeros in a time series and p_i^1 is the frequency of ones. Given the values of S_i for each node, it is possible to calculate the average Shannon entropy for the entire network. We analyzed the neighborhood density NA in relation to the parameter δ for four different theoretical network models ([MIRANDA; MACHICAO; BRUNO, 2016b](#)) by calculating $\vec{S}$. The result is shown in Fig. 18.

Figure 18 – Average Shannon entropy as a function of the parameter δ for different network models: random, small-world, scale free and geographical.

Source: Adapted from [Miranda, Machicao and Bruno \(2016b\)](#).

We can observe in Fig. 18 that the transition from oscillating patterns (high Shannon entropy) to stable patterns (low Shannon entropy) has a specific behavior for each network model and it occurs in the interval $[0.32, 0.46]$. In previous works, Marr&Hütt ([MARR; HÜTT, 2005](#); [MARR; HÜTT, 2009](#)) have shown that for different intervals of the parameter δ , different spatio-temporal patterns can be obtained for a specific network topology. For low values of δ , oscillating patterns tend to prevail since the majority of the nodes will change their states at each time step. In contrast, for higher values of δ , the majority of the nodes will not change their states, and, consequently, stable patterns will prevail. Finally, for intermediary values of δ , chaotic and complex patterns tend to emerge.

The entropy analysis shows that the topological differences of the theoretical network models can also be reflected by the corresponding spatio-temporal patterns of the evolved NAs. The importance of this result is related to the fact that a specific class of networks may present specific characteristics regarding its evolution dynamics, and, as such, these attributes can be used as network descriptors. In a subsequent analysis, we went on a step further and implemented a pattern recognition application for classifying theoretical network models by using the proposed network descriptor as feature vector. We evaluated the classification accuracy of the network model as a function of the threshold parameter δ . In addition, we evaluated the impact of the initial distribution of the nodes' states in the final accuracy. This initial configuration is described by σ , which represents the percentage of alive neighbors when $t = 0$.

This pattern recognition application first involved the creation of a dataset composed of synthetic networks generated according to the following theoretical models: 1) random, with connection probability of $p = \langle k \rangle / n$; 2) small-world, with rewiring probability $p = 0.1$; 3) scale-

free, with power exponent $\gamma = 1$, and, 4) geographical. This dataset contains 100 networks of each model with mean degree $\langle k \rangle = 8$. Each network model represents a class in this experiment, i.e., the category we want to automatically classify. By evolving each network of the dataset using the neighborhood density NA, we obtained the Shannon entropy distribution for each network, which was then used as network descriptor. We employed the 10-fold cross-validation scheme and the SVM (*Support Vector Machines*) classifier. In this context, accuracy values represent the number of correctly classified networks. Figure 19 presents the results for five different values of σ : 10%, 30%, 50%, 70% and 90%.

Figure 19 – Classification accuracy (%) of network models as a function of the parameter δ , given distinct initial distributions of the cells' states. The σ represents the percentage of alive cells in when $t = 0$.

Source: Adapted from [Miranda, Machicao and Bruno \(2016b\)](#).

We can see from Fig. 19 that in general the highest accuracy values were obtained for $0.3 \leq \delta \leq 0.5$, which also corresponds to the interval for which we observed the transition from oscillating to chaotic patterns in Fig. 18. In the cases where $\sigma = 50\%$ and $\sigma = 70\%$, the accuracy is also high beyond $\delta = 0.7$. The exception was $\sigma = 90\%$, for which the highest accuracy was obtained for $\delta = 0.85$. The highest accuracy obtained for the classification of the network models was 95.56% for $\delta = 0.4$.

The importance of the characterization of theoretical network models relates to the known properties of these models which may be useful when trying to comprehend their spatio-temporal patterns. The first experiment considering four network models as classes (random, small-world, scale-free and geographical) provided a basic classification problem in order to evaluate the proposed network descriptor. In the next section, we introduce a NA family with a more intricate dynamics which also relies on the neighborhood definition provided by Eq. (5.1).

5.3 Birth-Survival Network Automata (BSNA)

5.3.1 Definition

A Birth-Survival network automaton (BSNA) is described by a quintuple of the form $\mathcal{C} = \langle \mathcal{T}, \mathbb{S}, s_0, \mathcal{N}, \phi \rangle$ where:

- $\mathcal{T}$ is the tessellation defined by the network structure;
- $\mathbb{S}$ is a set of states consisting of two values, such that $s_i = 0$ represents the “dead” state and $s_i = 1$ the “alive” state;
- s_0 is the initial configuration of the states for all the cells $c_i \in \mathcal{T}$, i.e., $s(c_i, 0) = s_0(c_i)$;
- $\mathcal{N}$ is the neighborhood function given by the adjacency matrix A ;
- ϕ_i is the transition function defined for each cell c_i as:

$$\phi_i(t+1) = \begin{cases} 1, & \text{if } \begin{cases} \phi_i(t) = 0 \text{ and } \frac{x_a}{R} \leq \rho_i < \frac{(x_b+1)}{R} \\ \phi_i(t) = 1 \text{ and } \frac{y_a}{R} \leq \rho_i < \frac{(y_b+1)}{R} \end{cases} \text{ or} \\ 0, & \text{otherwise} \end{cases} \quad (5.3)$$

In Eq. (5.3), $\phi_i(t+1)$ defines a birth-survival dynamics for cell c_i at time $t+1$, where ρ_i denotes the neighborhood density of c_i and $R \in \mathbb{N}$. If the state of cell c_i at time t is given by $s(c_i, t) = 0$, this means that the cell is dead and can be **born** if ρ_i satisfies a given condition. If $s(c_i, t) = 1$, then the cell is already alive and can **survive** if ρ_i satisfies another condition. Otherwise, if none of the conditions are satisfied, then $s(c_i, t+1) = 0$. Finally, $x_a, x_b \in \mathbb{N}$ and $0 \leq x_a \leq x_b \leq R$. Similarly, $y_a, y_b \in \mathbb{N}$ and $0 \leq y_a \leq y_b \leq R$. Both x_a, x_b and y_a, y_b define intervals within $[0, 1]$. Figure 20 shows the conditions that have to be satisfied in both cases, birth and survival. The interval for birth is more restrictive than the corresponding interval for the survival condition. R is a normalizing factor such that $R \geq 2$. For the example in Fig. 20 we have that $R = 10$, and x_a and y_b are the initial and end points of the interval corresponding to the birth condition, which is defined as $\left[\frac{x_a}{R}, \frac{x_b+1}{R} \right]$. Similar intervals can be described for y_a, y_b .

According to the dynamics defined by $\phi_i(t+1)$, there are two ways for cell c_i to assume the alive state ($s(c_i, t+1) = 1$). In the first case, cell c_i was dead at time t ($s(c_i, t) = 0$) and will be **born** at time $t+1$ if ρ_i is contained in $\left[\frac{x_a}{R}, \frac{x_b+1}{R} \right]$. In the second case, $s(c_i, t) = 1$ and c_i will **survive** if ρ_i is contained in $\left[\frac{y_a}{R}, \frac{y_b+1}{R} \right]$.

Figure 20 – Illustration of density intervals defined by x_a, x_b and y_a, y_b , which are used in the birth-survival transition function (Eq. 5.3). In this example $R = 10$ is the normalization factor.

Source: Elaborated by the author.

5.3.2 Rule space & notation

The transition function of the neighborhood density NA (Eq. (5.2)) is characterized by the parameter δ , a threshold that controls when a given node will update its state. Therefore, δ defines two sub-intervals in the range $[0, 1]$: $[0, \delta]$ and $]\delta, 1]$. In contrast, Eq. 5.3 allows for the definition of more sub-intervals for which the birth or the survival criterion is met (Fig. 20). Therefore, there are different combinations of birth and survival conditions, and, consequently different rules can be contemplated through Eq. (5.3).

Figure 21 illustrates the evolution of a sample network using Eq. (5.3) with $[x_a, x_b] = [3, 3]$, $[y_a, y_b] = [2, 3]$ and $R = 9$. We can see that at t_1 the neighborhood density of cell c_i is $\rho_i = 0.4$, since 2 out of 5 neighbors are alive. Therefore, the birth condition is satisfied, since $0.33 \leq \rho_i < 0.44$, and, at t_2 the state of cell c_i will be $s(c_i, t_2) = 1$. However, at t_2 , when testing the survival condition for cell c_i ($0.22 \leq \rho_i < 0.44$), we can see it is not satisfied, since at t_2 , $\rho_i = 0$, and therefore, $s(c_i, t_3) = 0$. A similar analysis can be performed for the other nodes in Fig. 21.

The transition function exemplified in Fig. 21 can be summarized using the following notation $B33/S23$, where B stands for “birth” and S for “survival” conditions. Such a nomenclature corresponds to a specific region of the density interval, where x_a/R and $(x_b + 1)/R$ are the end points of the interval. The normalization factor R defines the number of subdivisions of the density interval, and, consequently, the rule space of the NA. The rule space is represented by the set of all rules, given R . Therefore, as R increases, the size of the rule space also increases. The total number of rules is given by:

$$n_R = \left[\left(\frac{R(R+1)}{2} \right) + 1 \right]^2. \quad (5.4)$$

Let us consider for instance $R = 2$, then there are $n_R = 16$ possible rules. For $R = 3$, the

Figure 21 – Illustration of the birth-survival transition function for Rule B33/S23.

Rule B33/S23

R=9

 $\begin{cases} \circ & \text{alive} \\ \bullet & \text{dead} \end{cases}$

$$\phi_i(t+1) = \begin{cases} 1, & \text{if } \begin{cases} \phi_i(t) = 0 \text{ and } \frac{3}{9} \leq \rho_i < \frac{(3+1)}{9} \\ \phi_i(t) = 1 \text{ and } \frac{2}{9} \leq \rho_i < \frac{(3+1)}{9} \end{cases} \text{ or} \\ 0, & \text{otherwise} \end{cases}$$

t_1

$$\begin{aligned} \rho_i &= 0.4 \\ s(c_i, t_1) &= 0 \end{aligned}$$

t_2

$$\begin{aligned} \rho_i &= 0 \\ s(c_i, t_2) &= 1 \end{aligned}$$

t_3

$$\begin{aligned} \rho_i &= 0 \\ s(c_i, t_3) &= 0 \end{aligned}$$

Source: Elaborated by the author.

rule space consists of $n_R = 49$ rules, and, so on. Table 1 shows the rule space for $R = 2$, $R = 3$ and $R = 4$. Rule B/S represents empty sets for both birth and survival. But it is possible that only one of the sets is empty, e.g., $B/S00$ and $B00/S$. Note that $B00$ corresponds to the first subdivision of the density interval and therefore it is different from the empty set B . In addition, it is worth noting that rules with the same name in different rule spaces, do not represent the same density intervals. For instance, if $R = 3$, then $B01$ accounts for $2/3$ of the density interval, however, if $R = 4$, then $B01$ accounts for $2/4$ of the density interval. Finally, a set defined for birth or survival conditions represents a single region of the density interval, which can also include the whole density interval. In this case, the condition being tested will always be satisfied.

5.3.3 Spatio-temporal patterns

In order to illustrate the behavior of the BSNA rules, we present some spatio-temporal patterns obtained using Eq. (5.3) and different network models. Figure 22-(a) illustrates the spatio-temporal patterns for all the rules in a rule space defined by $R = 2$ and using regular networks. Figure 22-(b) presents the spatio-temporal patterns for the same set of rules and using random networks. All the networks used to generate the corresponding spatio-temporal diagrams have $N = 400$ nodes and $\langle k \rangle = 8$. In the diagrams, the nodes are ordered by their average degrees, where the left-most columns correspond to nodes with the smallest mean degrees. In addition, Figs. 23 and 24 present similar spatio-temporal patterns for all the rules that can be enumerated

Table 1 – Set of rules of the birth-survival NA belonging to the rule space defined by $R = 2$, $R = 3$ and $R = 4$

R=2	R=3		R=4				
B/S	B/S	B22/S01	B/S	B11/S22	B33/S12	B12/S13	B13/S00
B/S00	B/S00	B22/S12	B/S00	B11/S33	B33/S23	B12/S03	B13/S11
B/S11	B/S11	B22/S02	B/S11	B11/S01	B33/S02	B23/S	B13/S22
B/S01	B/S22	B01/S	B/S22	B11/S12	B33/S13	B23/S00	B13/S33
B00/S	B/S01	B01/S00	B/S33	B11/S23	B33/S03	B23/S11	B13/S01
B00/S00	B/S12	B01/S11	B/S01	B11/S02	B01/S	B23/S22	B13/S12
B00/S11	B/S02	B01/S22	B/S12	B11/S13	B01/S00	B23/S33	B13/S23
B00/S01	B00/S	B01/S01	B/S23	B11/S03	B01/S11	B23/S01	B13/S02
B11/S	B00/S00	B01/S12	B/S02	B22/S	B01/S22	B23/S12	B13/S13
B11/S00	B00/S11	B01/S02	B/S13	B22/S00	B01/S33	B23/S23	B13/S03
B11/S11	B00/S22	B12/S	B/S03	B22/S11	B01/S01	B23/S02	B03/S
B11/S01	B00/S01	B12/S00	B00/S	B22/S22	B01/S12	B23/S13	B03/S00
B01/S	B00/S12	B12/S11	B00/S00	B22/S33	B01/S23	B23/S03	B03/S11
B01/S00	B00/S02	B12/S22	B00/S11	B22/S01	B01/S02	B02/S	B03/S22
B01/S11	B11/S	B12/S01	B00/S22	B22/S12	B01/S13	B02/S00	B03/S33
B01/S01	B11/S00	B12/S12	B00/S33	B22/S23	B01/S03	B02/S11	B03/S01
	B11/S11	B12/S02	B00/S01	B22/S02	B12/S	B02/S22	B03/S12
	B11/S22	B22/S	B00/S12	B22/S13	B12/S00	B02/S33	B03/S23
	B11/S01	B22/S00	B00/S23	B22/S03	B12/S11	B02/S01	B03/S02
	B11/S12	B22/S11	B00/S02	B33/S	B12/S22	B02/S12	B03/S13
	B11/S02	B22/S22	B00/S13	B33/S00	B12/S33	B02/S23	B03/S03
	B22/S	B22/S01	B00/S03	B33/S11	B12/S01	B02/S02	
	B22/S00	B22/S12	B11/S	B33/S22	B12/S12	B02/S13	
	B22/S11	B22/S02	B11/S00	B33/S33	B12/S23	B02/S03	
	B22/S22		B11/S11	B33/S01	B12/S02	B13/S	
16 rules	49 rules		121 rules				

when $R = 3$, using regular and random networks, respectively.

We can observe different categories of spatio-temporal patterns and we will make a description of these patterns based on the categories defined for the ECAs in Section 3.3.1. The first one involves stable patterns for which the nodes keep their states until the end of the simulation. In this category there can be a convergence of all nodes to a single state, as is the case of rules $B11/S$, $B11/S11$, $B11/S01$ and others in Fig. 22-(a), or there is a convergence of the nodes presenting both states, as is the case for rules B/S , $B/S01$, $B00/S$, and others in Fig. 22-(a). In a second category there are oscillating patterns, for which the nodes keep alternating their states, for instance, rule $B01/S00$ in Figs. 22-(a) and (b). In a third category, complex patterns emerge like for rule $B11/S00$ in Fig. 22-(a) and rule $B01/S00$ in Fig. 23. In this category we can observe some similar structures generated by the ECA rules when emerging patterns “survive” for a long time during the evolution of the NA. Finally, the last category consists of rules that lead to apparently random or chaotic behavior, as for instance rules $B00/S11$, $B00/S22$, $B11/S00$ and others in Fig. 24.

Figure 22 – Spatio-temporal patterns obtained through the evolution of the BSNA, for $R = 2$ and for two different tessellations: (a) regular network and (b) random network.

Source: Elaborated by the author.

Figure 23 – Spatio-temporal patterns for obtained through the evolution of the BSNA for $R = 3$ and using a regular network as the NA tessellation.

Source: Elaborated by the author.

5.3.4 Quantitative analyses of spatio-temporal patterns

We can observe that some rules of the BSNA are more restrictive than others concerning the birth and the survival conditions (Eq. (5.3)). For instance, when comparing rules $B12/S00$ and $B11/S01$, a dead cell is more probable to be born in the space defined by $B12/S00$, since the corresponding density interval for the birth condition, $B12$, is wider for this rule. However, similarly, an alive cell has a higher chance of survival in the space defined by rule $B11/S01$. This analysis holds for any value of the normalization parameter R . Therefore we defined an ordering of two birth-survival rules that takes into account the concepts of inclusion and “chance of living”. The first one accounts for the inclusion of a rule r_1 into another rule r_2 . Whereas, the latter, quantifies how probable that a cell c_i will assume state $s(c_i) = 1$ for given a rule r . These definitions are described next.

Figure 24 – Spatio-temporal patterns for obtained through the evolution of the BSNA for $R = 3$ and using a random network as the NA tessellation.

Source: Elaborated by the author.

Definition 5.3.1. Inclusion. Given two rules r_1 and r_2 within the rule space defined by $R = \alpha$, then $r_1 \subseteq r_2$ if, and only if, $B_{r_1} \subseteq B_{r_2}$ and $S_{r_1} \subseteq S_{r_2}$, where B_{r_1} and B_{r_2} are the density intervals for the birth condition of r_1 and r_2 , and, S_{r_1} and S_{r_2} are the density interval for the survival condition of r_1 and r_2 , respectively.

Considering rules $B12/S00$ and $B11/S00$, for instance, we have that $B12/S00 \not\subseteq B11/S00$. However, the following inclusion holds: $B11/S00 \subseteq B12/S00$.

Definition 5.3.2. Chance of living. The “chance of living” of a given rule r is defined as the proportion of the density intervals, within the rule space defined by $R = \alpha$, for which the birth or the survival criteria are satisfied.

Therefore, let us consider the example illustrated in Fig. 20 for which $R = 10$. Regarding the birth condition, it can be satisfied by only 2 out of 10 subdivisions of the density interval.

Whereas, the survival condition is less restrictive since it can be satisfied by 4 out of 10 subdivisions. Therefore, in total, a cell has a chance of 6 out of 20 of being alive in the next time step when considering birth and survival conditions. We will denote the “chance of living” as a percentage, therefore, in the example of Fig. 20 a cell has a “chance of living” of 30%.

The diagrams presented in Figs. 25, 26 and 27 illustrate the inclusion and the “chance of living” definitions. They represent the all the rules given $R = 2$ as partially ordered sets. In the upper part of these diagrams, there is the least restrictive rule, $B01/S01$. This rule will always allow a cell to be alive, since for $R = 2$ the set 01 corresponds to the whole density range $[0, 1]$. Therefore, the “chance of living” of rule $B01/S01$ is 100%. In the second row, we have the rules for which the “chance of living” is 75%, i.e., the conditions for birth and survival together will account for 3 out of 4 sub-intervals of the entire density range. The other rows consist of rules with 50%, 25% and 0% “chance of living”, respectively. Therefore, rule B/S is the most restrictive since a cell will always be in the same state. In addition, the lines connecting the rules illustrate the inclusion relations among them. Finally, for each rule, we calculated a specific measure that is also shown in the diagrams of Figs. 25, 26 and 27.

We carried out a quantitative analysis for the rules defined by $R = 2$. For that purpose, we calculated the average Lyapunov exponent (LE), the average Lempel-Ziv complexity (LZ) and the average word length (WL) for each rule. The results are presented in the diagrams of Figs 25, 26 and 27. Fig. 25 depicts the average LE for two types of networks: regular (Fig. 25-(a)) and random (Fig. 25-(b)). The average LE was calculated after running the BSNA for each network type. The networks have $N = 400$ and $\langle k \rangle = 8$ and they were evolved for $t = 350$ time steps.

As described in Section 3.4.6, the LE measures the sensitivity of a dynamical system to the initial conditions. We repeated the simulation 50 times, by randomly choosing different initially perturbed cells. The final average values of the LE are shown in Fig. 25 for each rule. There is a group of rules for which the LE is $-\infty$ for some initial configurations, but positive for other initial conditions. The small number in parentheses shows the number of simulations, among the total of 50 simulations, for which the LE was not $-\infty$. Therefore, the average LE was calculated considering this number of simulations.

We can observe, for instance, that the rule with the highest LE in Fig. 25-(a) is rule $B11/S00$, which was the only rule that leads to more complex/chaotic patterns when analyzing the spatio-temporal patterns in Fig. 22-(a). In Fig. 25-(b), we can see that rule $B11/S00$ also has a high average LE together with rule $B00/S11$. These results complement the spatio-temporal patterns of Fig. 22-(b), in which these two rules presented a mixture of oscillating and regular patterns. Notice that, the spatio-temporal patterns shown here, as the ones in Fig. 22 represent only one possible evolution for each rule. In addition, by comparing both Figs. 25 and 22, we observe that rules leading to a single state, lead to a Lyapunov exponent $-\infty$, while stable rules converging to two-state oscillating patterns lead to positive small LE, since more defects are propagated during the NA evolution.

Figure 25 – Average Lyapunov exponent (LE) over 50 simulations of the BSNA on (a) a regular network and on (b) a random network. The rules are presented as a Hasse diagram, using the Inclusion relationship (5.3.1). Parameters $N = 400$, $\langle k \rangle = 8$ and $R = 2$.

(a)

(b)

Source: Elaborated by the author.

Figure 26 shows the average LZ for the same network types with the same configurations. Recalling that the Lempel-Ziv algorithm generates a dictionary of distinct "words" of different

Figure 26 – Average Lempel-Ziv complexity over 50 simulations of the BSNA on (a) a regular network and on (b) a random network. The rules are presented as a Hasse diagram, using the Inclusion relationship (5.3.1). Parameters $N = 400$, $\langle k \rangle = 8$ and $R = 2$.

(a)

(b)

Source: Elaborated by the author.

sizes. As the size of the time series increases, the number of words in the dictionary also increases

(Section 3.4.4). Therefore, the higher the number of words, the higher the LZ complexity. This is why the rules that have the highest average LZ are the rules with oscillating patterns, $B00/S00$ and $B01/S00$. However, the average LZ complexity was very similar for rules leading to stable and complex patterns. The fact that the LZ values are equal for many rules in the diagrams of Fig. 26 is related to the procedure for calculating this measure. Therefore, for single-state stable patterns or for two-state stable patterns the same LZ was obtained. Therefore, the chaotic patterns inferred for $B11/S00$ and $B00/S11$ correspond to intermediate values of LZ complexity. This demonstrates that as the number of time steps increases, there is less variation for what concerns the different words that are obtained for the spatio-temporal pattern of these rules when compared to the corresponding patterns of oscillating rules.

Finally, we analyzed the average WL for the rules defined by $R = 2$. The result is illustrated in Fig. 27. We considered a “word” as a sequence of ones bounded by zeros. In this figure we also have a clear distinction among stable, oscillating and complex/chaotic patterns. The average WL was calculated considering 50 simulations and $t = 250$ time steps for the evolution of the NA. As expected, we can see, for instance, that rules leading to oscillating patterns result in an average word length of ~ 1 ($B00/S00$ and $B01/S00$). In contrast, the stable patterns lead to an average word length of ~ 250 , which is the total number of time steps. The patterns with a more complex behavior lead to intermediate values of WL.

We also analyzed the rule space defined by $R = 3$ and the results are showed in Figs. 28 and 29 concerning the Lyapunov exponent for regular and random network topologies, respectively. The corresponding diagrams of these figures can be analyzed together with the spatio-temporal patterns of Figs. 23 and 24, respectively, for a better comparison between the pattern and the obtained measure. The quantitative analyses for the case $R = 2$ can also be applied for the rule space of $R = 3$ and the same three classes of LE can be identified: i) high and positive LE for complex/chaotic patterns, ii) positive LE for oscillating and two-state stable patterns, and, iii) $-\infty$ for single-state stable patterns.

Similar diagrams for the Lyapunov exponent, the Lempel-Ziv complexity and the average word length for the other types of networks approached in this work (small-world, scale-free and geographical) can be found in Appendix A for both $R = 2$ and $R = 3$.

Chance of "living"

100%

75%

50%

25%

0%

Legend:

- WL ≥ 100
- 1 < WL < 100
- WL = 1

Chance of "living"

100%

75%

50%

25%

0%

WL ≥ 100

$1 < WL < 100$

WL = 1

Diagram illustrating a network of nodes representing different states of a system, with arrows indicating the chance of "living" for each state.

Legend:

- Yellow box: WL ≥ 100
- Orange box: $1 < WL < 100$
- Green box: WL = 1

Nodes and their associated WL values:

- Top node: B01/S01 (WL: 250)
- Second row: B00/S01 (WL: 250), B01/S00 (WL: 1), B01/S11 (WL: 250), B11/S01 (WL: 250)
- Third row: B00/S11 (WL: 18.79), B00/S00 (WL: 1.04), B/S01 (WL: 250), B01/S (WL: 250), B11/S11 (WL: 245.04), B11/S00 (WL: 16.88)
- Fourth row: B00/S (WL: 250), B/S00 (WL: 250), B/S11 (WL: 250), B11/S (WL: 250)
- Bottom node: B/S (WL: 250)

Arrows indicate the chance of "living" for each state:

- Top node: 100%
- Second row: 75%
- Third row: 50%
- Fourth row: 25%
- Bottom node: 0%

(b)

Source: Elaborated by the author.

Figure 28 – Average Lyapunov exponent over 50 simulations of the BSNA on a regular network. The rules are presented as a Hasse diagram, using the Inclusion relationship (5.3.1). Parameters $N = 400$, $\langle k \rangle = 8$ and $R = 3$.

Figure 29 – Average Lyapunov exponent over 50 simulations of the BSNA on a random network. The rules are presented as a Hasse diagram, using the Inclusion relationship (5.3.1). Parameters $N = 400$, $\langle k \rangle = 8$ and $R = 3$.

Source: Elaborated by the author.

5.4 Conclusion

In this chapter we presented the concept of NA and we also introduced two different models, that make use of the density neighborhood, ρ_i . This is a key concept for modeling a CA on an irregular topology, such as the network models studied in this thesis. The birth-survival dynamics on irregular topologies is also one of the main concepts proposed here. We investigated the behavior of such NA dynamics on different network topologies and we provided a quantitative analysis of the spatio-temporal patterns. We could mainly distinguish, quantitatively, the following patterns: stable, oscillating and chaotic/complex. In the next chapter, we will introduce a more general NA family with a similar birth-survival dynamics. In addition, we will also introduce the LLNA, mainly inspired by the classic Life-like CA.

GENERALIZED BIRTH-SURVIVAL NETWORK AUTOMATA

6.1 Introduction

In this chapter we will introduce a family of network automata also inspired by the birth-survival dynamics. This family can be defined as a generalization of the BSNA defined in the previous chapter. One of the main differences between both NA is that the rule space is more flexible concerning the birth and survival conditions, allowing more combinations of the subdivisions of the density interval. Besides the formal definitions of the proposed family of NA, we will also provide a formal description of the corresponding rule space and a study of such NA on different network topologies. In the second part of this chapter, we will introduce the Life-like Network Automata (LLNA), which is a specific case of this generalized family of birth-survival NA inspired by the classic Life-Like CA. Therefore, we will present a formal definition of this classic CA on irregular topologies together with the corresponding rule space. The LLNA is already published in ([MIRANDA; MACHICAO; BRUNO, 2016a](#)).

6.2 A Generalized Family of Birth-Survival Network Automata (GenBSNA)

The birth and survival conditions described in Eq. (5.3) are defined by a sub-interval within the range of the neighborhood density. This sub-interval can be represented by the following notation $Bx_a, x_b / Sy_a, y_b$, where x_a, x_b denotes the boundaries of the interval for the birth condition and y_a, y_b for the survival condition. This definition only allows the representation of a single sub-interval, which can also be an empty set or the entire density range. We now present a generalization of the birth-survival representation such that more than one single region

of the neighborhood density range can be defined as conditions for birth and survival. For that purpose let us define the following transition function:

$$\phi_i(t+1) = \begin{cases} 1, & \text{if } \begin{cases} \phi_i(t) = 0 \text{ and } \sum_{j=1}^X h_i^{x_j}(t) > 0 \\ \phi_i(t) = 1 \text{ and } \sum_{j=1}^Y h_i^{y_j}(t) > 0 \end{cases} \quad \text{or} \\ 0, & \text{otherwise,} \end{cases} \quad (6.1)$$

where X and Y are sets such that $X = \{x_1, x_2, \dots, x_n \mid x_j \in \mathbb{N}, 0 \leq x_j < R\}$ and $Y = \{y_1, y_2, \dots, y_m \mid y_j \in \mathbb{N}, 0 \leq y_j < R\}$. Although similar to Eq. (5.3), Eq. (6.1) presents different definitions for the birth and survival conditions. The functions $h_i^{x_j}(t)$ and $h_i^{y_j}(t)$ map if ρ_i is within the density interval according to Eq. (6.2)

$$h_i^{x_j}(t) = \begin{cases} 1, & \text{if } \frac{x_j}{R} \leq \rho_i < \frac{x_j+1}{R} \\ 0, & \text{otherwise} \end{cases} \quad (6.2)$$

and similarly for y_j . Therefore, each element of the sets X and Y represents one subdivision of the density interval, which is defined by R , the normalization factor, such that $R \geq 2$. This way, a single element of X or Y leads to a unique region within the range of the neighborhood density interval. In addition, $\sum_{j=1}^X h_i^{x_j}(t)$ and $\sum_{j=1}^Y h_i^{y_j}(t)$ in Eq. (6.1) are used to verify if at least one element of X and one element of Y satisfies the conditions given by $h_i^{x_j}(t)$ and $h_i^{y_j}(t)$ for the birth and survival conditions, respectively. The density interval is now give by: $\left[\frac{x_j}{R}, \frac{x_j+1}{R} \right)$.

Figure 30 illustrates the subdivisions of the density intervals for the sets $X = \{3, 4, 7\}$, $Y = \{0, 4, 5, 8\}$ and $R = 10$. Therefore, a single element of the sets X and Y corresponds to a unique subdivision of the density interval, i.e., each subdivision of the density interval will be represented by an element from the sets X and Y . In the next section, we discuss the rule space and the rule notation of such generalized birth-survival NA (GenBSNA).

Figure 30 – Illustration of density intervals for the generalized birth-survival NA. The intervals correspond to sets $X = \{3, 4, 7\}$ and $Y = \{0, 4, 5, 8\}$, which are used in the definition of the transition function in Eq. 6.1. In this example $R = 10$ is the normalization factor.

Source: Elaborated by the author.

6.2.1 Rule space & notation

We will use a similar notation for the rule space of the GenBSNA, with B representing the birth condition and S the survival condition. However, now B will be followed by the elements of X and S will be followed by the elements of Y , as BX/SY . Considering, for instance, the example represented in Fig. 30, the respective birth-survival dynamics would be $B347/S0458$. This is same notation scheme used for the classic Life-Like CAs. Therefore, each element of X or Y corresponds to a subdivision of the neighborhood density.

Within this context, let us consider the rule space of such NA family. Given $R = 2$, the X and Y sets are composed of two elements, $\{0, 1\}$. Therefore, the total number of rules is $n_R = 2^{(2+2)}$, which corresponds to the same 16 rules of the BSNA, when $R = 2$. When $R = 3$, the X and Y sets are composed of three elements, $\{0, 1, 2\}$, therefore $n_R = 2^{(3+3)}$, i.e., we have 64 rules for $R = 3$. For $R = 4$, we have 256 rules, and so on. We can observe that from $R = 3$ the size of the rule space is different when compared to the size of the rule space of the BSNA for the same R . Figure 31-(a) presents the sets X and Y for different values of R . Generically, the number of rules of the GenBSNA, given R , is calculated as $n_R = 2^{(2R)}$. Figure 31-(b) shows the exponential growth of the number of rules as a function of R . As R increases, the number of combinations BX/SY that can be represented also increases. However, it becomes difficult to explore the rule space for large values of R .

Figure 31 – The rule space as a function of the parameter R . (a) X and Y sets for different values of parameter R . (b) Number of rules as a function of the parameter R .

Source: Adapted from [Miranda et al. \(2018\)](#).

Table 2 presents the rules for $R = 2$, $R = 3$ and $R = 4$ of the GenBSNA. Similarly, rule B/S represents empty sets for both birth and survival. But it is possible that only one of the sets is empty, e.g., $B/S0$ or $B0/S$. Note that, in the notation of this generalized family of NA, $B0$

denotes the first subdivision of the density interval and it is different from the empty set B . Again, rules with the same nomenclature in different rule spaces (different R), do not represent the same density intervals. In addition, Definitions 5.3.1 and 5.3.2 are also valid for the GenBSNA.

Table 2 – Generalized birth-survival NA for $R = 2$, $R = 3$ and $R = 4$

R=2	R=3		R=4					
B/S	B/S	B01/S	B/S	B1/S	B3/S	B02/S	B12/S	B23/S
B/S0	B/S0	B01/S0	B/S0	B1/S0	B3/S0	B02/S0	B12/S0	B23/S0
B/S1	B/S1	B01/S1	B/S1	B1/S1	B3/S1	B02/S1	B12/S1	B23/S1
B/S01	B/S2	B01/S2	B/S2	B1/S2	B3/S2	B02/S2	B12/S2	B23/S2
B0/S	B/S01	B01/S01	B/S3	B1/S3	B3/S3	B02/S3	B12/S3	B23/S3
B0/S0	B/S02	B01/S02	B/S01	B1/S01	B3/S01	B02/S01	B12/S01	B23/S01
B0/S1	B/S12	B01/S12	B/S02	B1/S02	B3/S02	B02/S02	B12/S02	B23/S02
B0/S01	B/S012	B01/S012	B/S03	B1/S03	B3/S03	B02/S03	B12/S03	B23/S03
B1/S	B0/S	B02/S	B/S12	B1/S12	B3/S12	B02/S12	B12/S12	B23/S12
B1/S0	B0/S0	B02/S0	B/S13	B1/S13	B3/S13	B02/S13	B12/S13	B23/S13
B1/S1	B0/S1	B02/S1	B/S23	B1/S23	B3/S23	B02/S23	B12/S23	B23/S23
B1/S01	B0/S2	B02/S2	B/S012	B1/S012	B3/S012	B02/S012	B12/S012	B23/S012
B01/S	B0/S01	B02/S01	B/S013	B1/S013	B3/S013	B02/S013	B12/S013	B23/S013
B01/S0	B0/S02	B02/S02	B/S023	B1/S023	B3/S023	B02/S023	B12/S023	B23/S023
B01/S1	B0/S12	B02/S12	B/S123	B1/S123	B3/S123	B02/S123	B12/S123	B23/S123
B01/S01	B0/S012	B02/S012	B/S0123	B1/S0123	B3/S0123	B02/S0123	B12/S0123	B23/S0123
	B1/S	B12/S	B0/S	B2/S	B01/S	B03/S	B13/S	B012/S
	B1/S0	B12/S0	B0/S0	B2/S0	B01/S0	B03/S0	B13/S0	B012/S0
	B1/S1	B12/S1	B0/S1	B2/S1	B01/S1	B03/S1	B13/S1	...
	B1/S2	B12/S2	B0/S2	B2/S2	B01/S2	B03/S2	B13/S2	...
	B1/S01	B12/S01	B0/S3	B2/S3	B01/S3	B03/S3	B13/S3	...
	B1/S02	B12/S02	B0/S01	B2/S01	B01/S01	B03/S01	B13/S01	...
	B1/S12	B12/S12	B0/S02	B2/S02	B01/S02	B03/S02	B13/S02	...
	B1/S012	B12/S012	B0/S03	B2/S03	B01/S03	B03/S03	B13/S03	...
	B2/S	B012/S	B0/S12	B2/S12	B01/S12	B03/S12	B13/S12	...
	B2/S0	B012/S0	B0/S13	B2/S13	B01/S13	B03/S13	B13/S13	...
	B2/S1	B012/S1	B0/S23	B2/S23	B01/S23	B03/S23	B13/S23	...
	B2/S2	B012/S2	B0/S012	B2/S012	B01/S012	B03/S012	B13/S012	...
	B2/S01	B012/S01	B0/S013	B2/S013	B01/S013	B03/S013	B13/S013	...
	B2/S02	B012/S02	B0/S023	B2/S023	B01/S023	B03/S023	B13/S023	...
	B2/S12	B012/S12	B0/S123	B2/S123	B01/S123	B03/S123	B13/S123	...
	B2/S012	B012/S012	B0/S0123	B2/S0123	B01/S0123	B03/S0123	B13/S0123	B0123/S0123
16 rules	64 rules		256 rules					

We can conclude that the NA defined by Eq. (5.3) are a specific case of the GenBSNA defined by Eq. (6.1), since the first always defines single regions of the density interval for the birth and survival conditions. Therefore, all the rules of the BSNA can be represented using the generalized definition, however, not all the rules of the latter can be represented by the BSNA.

6.3 Life-like Network Automata (LLNA)

Having presented the GenBSNA based on the birth-survival dynamics, in this section we introduce the Life-like Network Automata (LLNA) (MIRANDA; MACHICAO; BRUNO, 2016a; MACHICAO, 2017), which is a specific case of the GenBSNA and is inspired by the classic Life-like CA. LLNA are the set of NAs for which $R = 9$, since the rule space defined by this value of R allows for a correspondence between the classic CA and the proposed NA. This is illustrated in Fig. 32. The classic Life-like CA uses the Moore neighborhood. In the case of LLNA, the number of neighbors is not bounded by 8, so, we used the neighborhood density (Eq. (5.1)) in order to define the LLNA.

Figure 32 – Comparison between the Moore neighborhood for the classic Life-like CA and the intervals of the neighborhood density used by LLNA. The birth and survival conditions are illustrated for the famous “Life” rule $B3/S23$.

Source: Elaborated by the author.

LLNA also have the same rule space as the classic Life-like, which consists of 262144 rules. Besides the eight neighbors from the Moore neighborhood, the empty set is also considered, and, given the birth and survival conditions, we have 2^{18} possible combinations within the rule space for $R = 9$. Figure 32 depicts the neighborhood of the famous “Life” rule, $B3/S23$, for both Life-like and LLNA. According to this rule, a cell can be born (going from state 0 to 1) if there are 3 neighbors alive, or, it survives (stays in state 1) if 2 or 3 neighbors are alive. Therefore, the sets X and Y may contain up to nine elements $\{0, \dots, 8\}$.

There is an agreement between the number of alive neighbors, given by the Life-like rule, and the density ρ_i of each network node. The first eight intervals are defined as $[\frac{x_j}{9}, \frac{x_{j+1}}{9}]$ and the set X is given by the Life-like rule. The last interval is defined as $[\frac{8}{9}, 1]$, due to boundary conditions of the density range. The same intervals are defined similarly for Y . It is worth noting that the correspondence between the Life-like and the LLNA can be defined only for the rule nomenclature. There is no correspondence between them regarding the rule space.

Finally, since LLNA is part of the GenBSNA, its transition function is given by Eq. (6.1) and Eq. (6.3) can be updated as follows:

$$h_i^{x_j}(t) = \begin{cases} 1, & \text{if } x \neq 8 \text{ and } \frac{x_j}{9} \leq \rho_i < \frac{x_{j+1}}{9} \\ 1, & \text{if } x = 8 \text{ and } \frac{8}{9} \leq \rho_i \leq 1 \\ 0, & \text{otherwise} \end{cases} \quad (6.3)$$

For instance, given rule $B3/S23$, $h_i^3(t) = 1$ must hold for the birth condition and $h_i^2(t) = 1$ or $h_i^3(t) = 1$, for the survival condition.

6.3.1 Spatio-temporal patterns

In this section we will explore the spatio-temporal patterns obtained using LLNA. Figure 33-(a) presents the spatio-temporal diagrams obtained for networks generated according to four theoretical network models: random, small-world, scale-free and geographical. These networks were evolved using rule $B1357/S2468$ according to Eq. (6.1) as an example. All the networks have $N = 500$ nodes and different values of mean degree $\langle k \rangle$, and, they were evolved for $t = 500$ time steps. Initially, at $t = 0$, a state is assigned to each network node drawn from a uniform distribution. Similarly to the BSNA spatio-temporal diagrams, we also used the degree of the nodes to order their time series in the diagram where the left-most column corresponds to nodes with the smallest values of k_i , and, the right-most, to the ones with the largest values of k_i .

There are four patterns that can be observed in these diagrams: stable, oscillating, chaotic and complex. All of these patterns can be observed in Fig. 33-(b), obtained using a scale-free network with $\langle k \rangle = 8$. The colors at the bottom of this figure are related to values of the Shannon entropy. Red cells correspond to the highest entropy values, while the blue cells to the lowest. For the scale free network, hubs give rise to chaotic and complex patterns that can be seen on the right-most side of this diagram. These patterns can also be seen in the diagrams of other network models. As $\langle k \rangle$ increases, random patterns become more frequent. Consequently, the entropy also increases due to the addition of new edges to the network.

We calculated the Shannon entropy (S), the Word Length (WL) and the Lempel-Ziv (LZ) complexity for the spatio-temporal patterns in Fig. 33. We investigated how these measures vary for the different topologies studied in this chapter. Except for the WL, which was calculated for the entire network, the S and the LZ were calculated per network node. For these two measures, we first generated the corresponding distributions ($\bar{\mu}_S$ and $\bar{\mu}_L$) based on the measures of each node, and, the WL distribution ($\bar{\mu}_W$) was obtained directly from the diagram. The corresponding histograms are shown in Fig. 34. The networks used to generate the corresponding histograms have $N = 500$, $\langle k \rangle = 4$ and $t = 350$.

Through the comparison of $\bar{\mu}_S$ among the different networks, we can observe that the scale-free network has a high frequency of nodes with high a entropy. These nodes also have the highest $\langle k \rangle$, as observed in Fig 33. The histograms of the other networks also have a high frequency of nodes with high entropy. This is related to the presence of oscillating spatio-temporal patterns. Regarding $\bar{\mu}_W$, we can observe that the most frequent words are the smallest ones. In addition, the Lempel-Ziv distribution, $\bar{\mu}_L$, also differs among the network models.

6.3.2 Analysis of parameters

In this section we analyze the influence of three parameters on the evolution of LLNA: i) the number of time steps t ; ii) the percentage of the initial alive nodes σ , and, iii) the number of

Figure 33 – Space-time diagrams. (a) Space-time diagrams for different network models: random, small-world, scale-free and geographical. These networks were evolved using rule *B1357/S2468*. The networks have $N = 500$ nodes and four different values of $\langle k \rangle$. The number of evolution steps is given by $t = 500$. (b) Space-time diagram of a scale-free network with $\langle k \rangle = 8$. The colors at the bottom of the diagram correspond to the entropy of each network node. The blue cells correspond to the lowest entropy, while the red cells have the highest entropy.

Source: Adapted from [Miranda, Machicao and Bruno \(2016a\)](#).

nodes N . We evaluated these parameters in a pattern recognition task, similar to the one presented in Section 5.2.3, where we classified theoretical network models by using network descriptors. Each network model corresponds to a class in this experiment: random, small-world, scale-free and geographical. The dataset used for this task contains networks with $N = 500$, $N = 1000$, $N = 1500$ and $N = 2000$ nodes. There are a total 100 network for each combination of N and the four classes, leading to 1600 networks. We used the k -NN classifier to categorize the network models and a 10-fold cross validation scheme and we evaluated the classification performance using the histogram of the Shannon entropy as feature vector.

Figure 35-(a) presents the classification accuracy of the network models as a function

Figure 34 – Entropy ($\vec{\mu}_S$), word length ($\vec{\mu}_W$) and Lempel-Ziv complexity ($\vec{\mu}_L$) histograms for different network models. All networks have $N = 500$ nodes and $\langle k \rangle = 4$, and, the evolution time is given by $t = 350$.

Source: Adapted from [Miranda, Machicao and Bruno \(2016a\)](#).

of the number of evolution steps (t) for three different rules belonging to the LLNA rule space. We observe that the accuracy increases as the number of time steps increases. Although there is also a rapid convergence of the accuracy values for the three analyzed rules, this behavior cannot be assumed for all the rules. In the case of patterns for which the accuracy values do not converge, increasing the number of time steps may provide more details. Figure 35-(b) shows the percentage of initial alive nodes, σ , as a function of accuracy. For this experiment, we used rule $B135678/S03456$. The highest values of accuracy were obtained for the values of σ closest to a uniform distribution of states, *i.e.* $\sigma = 50\%$. We can also notice that the standard deviation, calculated over 100 runs of the cross-validation, starts to increase as the initial distribution of the states becomes uneven.

Finally, we also analyzed the influence of the network size. Figure 36 shows the histograms of the Shannon entropy, $\vec{\mu}_S$, using rule $B01678/S0457$ for the different networks. We can see that, distinctly from t and σ , the size of the network (N) does not influence the network descriptor $\vec{\mu}_S$. The histograms for distinct values of N are very similar given the same network model.

An analysis of the LLNA rule space concerning the classification accuracy will be presented in Chapter 8.

Figure 35 – Accuracy of network classification as a function of (a) the number of time steps (t) for the three highlighted rules, and, (b) the initial distribution of alive nodes (σ) using rule B135678/S03456.

Source: Adapted from [Miranda, Machicao and Bruno \(2016a\)](#).

Figure 36 – Histograms of the Shannon entropy distribution ($\vec{\mu}_S$) for each network model and for different values of the number of nodes $N = 500, 1000, 1500, 2000$. All these networks have $\langle k \rangle = 8$ and $t = 350$ time steps.

Source: Adapted from [Miranda, Machicao and Bruno \(2016a\)](#).

6.3.3 Comparison with BSNA

Within the context of pattern recognition, another parameter that may influence the final accuracy of a classifier is the NA rule. Does a given rule provide a better performance than

another rule? Is a rule a potential solution for a specific classification problem? In order to investigate these questions, we analyzed the classification accuracy of different rule sets, as defined by the parameter R , for both the BSNA and the GenBSNA. For this task we used a similar dataset as presented in the previous section. However, in addition to networks of different sizes, we also included networks with different average degrees, $\langle k \rangle$. We considered seven values of $\langle k \rangle$: 4, 6, 8, 10, 12, 14 and 16. For each network model and for each combination $N/\langle k \rangle$ there are 100 networks, leading to 11200 networks in the dataset.

For each R and for each rule belonging to the corresponding rule space we applied a cross-validation procedure by adopting the network models as classes. We evaluated values of R within the range $\{2, 3, \dots, 9\}$. Then, for each R we selected the rule which led to the highest classification accuracy. Figure 37 presents the results for the best rules and for each value of R for both the BSNA and the GenBSNA. Notice that the GenBSNA given $R = 9$ is the LLNA. We can observe that the GenBSNA performed better for all the values of R .

Figure 37 – Comparison between BSNA and GenBSNA concerning the classification accuracy of the network models for different values of R . The accuracy shown in the bar-plots corresponds to the highest accuracy achieved by a rule within the rule space defined by R .

Source: Elaborated by the author.

Figures 38 and 39 present the respective accuracy distributions for all rules given the different values of R for the BSNA and the GenBSNA, respectively. Each bar-plot of these figures corresponds to the fraction of rules (y-axis) within the rule space for which the cross-validation procedure gave rise to a classification accuracy within a specific range. Note that the accuracy is given as a percentage of correctly classified instances. Through these histograms we have information not only about the best rule but also about the performance of the entire rule sets. We can see for the GenBSNA that as R increases most rules are situated in the accuracy range $[70\%, 100\%]$ (Fig. 39). This is not observed for the BSNA (Fig. 38) for which the majority of the rules achieved accuracy lower than 70%. Therefore, the GenBSNA contain much more rules that provided better classification performance. However, for both NA, as R increases, the probability of obtaining higher accuracy values also increases.

Given that the GenBSNA has a bigger rule space when compared to the BSNA, given the

Figure 38 – Accuracy distribution obtained by the BSNA for different rule spaces defined by R .

Source: Elaborated by the author.

same R , all the rules of the BSNA can be represented in GenBSNA. A large rule space can give

Figure 39 – Accuracy distribution obtained with the GenBSNA for different rule spaces defined by R .

Source: Elaborated by the author.

the impression that many solutions are available, but a huge rule space may not always guarantee a better solution. For instance, from $R = 4$ on, the performance is very similar for the GenBSNA.

We know that the size of the rule space together with the network size have a big impact on the computing time. Therefore, since for each application every single rule must be evaluated regarding its performance, a smaller rule space makes it possible to handle huge datasets.

6.3.4 Sensitivity to initial distribution of states

In addition to the reported parameters which may influence the LLNA evolution, we have also assessed the sensitivity to the initial configuration. Given the same value of σ , let us define different initial configurations, $s_0^1, s_0^2, \dots, s_0^n$. We investigated to what extent the network descriptor is affected by these different initial configurations. Specifically, in this experiment we evaluated the Shannon entropy distribution, $\bar{\mu}_S$, as network descriptor.

For a single network of each model (scale-free, random, small-world, geographic), having $N = 2000$, the LLNA was evolved starting from 1000 different initial configurations. The same rule was used in all the simulations, namely *B01245678/S3456*. Then, we calculated $\bar{\mu}_S$ for each network and for each simulation. Figure 40 presents the average frequency and the corresponding standard deviation of the Shannon entropy for different combinations of parameters. The entropy interval was sampled in twenty bins. Each row of Fig. 40 corresponds to one value of the average degree of the network nodes and each column corresponds to a different value of σ .

We can observe that the variability of the entropy is larger for the networks with $\langle k \rangle = 16$. Although for such a degree the networks have already too many connections, it is still possible to distinguish between the network models. In addition, for the different combinations of parameters shown in Fig. 40, it is only possible to distinguish the network types for higher values of the Shannon entropy. Note that this experiment was performed with rule *B01245678/S3456*.

Figure 40 – Sensitivity to the initial distribution of states. Each row corresponds to the average degree of the networks used in the experiment and each column corresponds to a different value of σ , the initial distribution of alive nodes.

Source: Elaborated by the author.

6.4 Conclusion

In this chapter we have presented a generalized family of NAs that is based on the birth-survival dynamics. This family allows for a bigger rule space. Within this family we have presented the LLNA, which is inspired by the classic Life-like. We have investigated how the number of time steps and the initial configuration influence the spatio-temporal dynamics. In addition, we have evaluated whether the Shannon entropy distribution, obtained through the LLNA dynamics, can be used as a network descriptor.

In the next part of this thesis (Chapter 8) we will go one step further within the context of pattern recognition in networks using both synthetic and real-world datasets. For that purpose we will explore the entire rule space of LLNA and we will investigate how it can be applied to supervised learning tasks.

ROCK-PAPER-SCISSORS: A NETWORK ANALYSIS

7.1 Introduction

One of the most intriguing questions for ecologists concerns how biodiversity is maintained. Scientists have long been investigating the mechanisms which support or limit species coexistence (EBENMAN; JONSSON, 2005; MENDES *et al.*, 2008; LAIRD; SCHAMP, 2015). The loss of biodiversity may be harmful for the functioning of ecosystems and can lead to their extinction. It has been shown that competition is one of the mechanisms that supports biodiversity in ecosystems (LAIRD; SCHAMP, 2006). Previous works pointed out that cyclic dominance among species can promote their coexistence and as such contributes to the maintenance of biodiversity. Many mathematical models proposed in literature incorporate the cyclic dominance and a classic example is the rock-paper-scissors (RPS) game. With some background on evolutionary game theory, the RPS has emerged as a model that has been widely used to describe species diversity (KERR *et al.*, 2002; REICHENBACH; MOBILIA; FREY, 2007; RULQUIN; ARENZON, 2014). In this model three species interact in a non-hierarchical way, i.e., each species has one predator, and, at the same time, it preys on another species. Many food webs display such a behavior, like certain bacterial species (KIRKUP; RILEY, 2004, 2004), coral reefs (BUSS; JACKSON, 1979), and vertebrates (BLEAY; COMENDANT; SINERVO, 2007; SINERVO *et al.*, 2007), amongst others. Besides, some human decision-making processes are guided by similar interactions (TVERSKY, 1969). Variants of the RPS game for more than three species can also be found in literature, supporting the study of richer communities.

However, the cyclic competition is not the only factor that can influence the coexistence of species, and, therefore, it is not the only mechanism to be taken into account when modeling evolutionary systems (KERR *et al.*, 2002; SCHÜTT; CLAUSSEN, 2010; RULQUIN; ARENZON, 2014). Community evenness and the individuals' mobility also play important

roles. The first is related to the distribution of the species, which is often assumed to be uniform despite evidence of the contrary from real-world ecosystems (DALY; BAETENS; DE BAETS, 2015). Secondly, mobility is crucial in many ecosystems and can be driven, for instance, by resource-dependent demographic processes. For instance, it has been found that the coexistence of species is mediated by their dispersal and that there is a critical threshold mobility above which biodiversity is lost (REICHENBACH; MOBILIA; FREY, 2007; REICHENBACH; MOBILIA; FREY, 2008).

In computer simulations, square grids are typically used when simulating the interactions among species, which strongly limits the individuals' degrees of freedom. To a much lesser extent, graphs have been considered (SZABÓ; SZOLNOKI; IZSÁK, 2004; SZOLNOKI; SZABÓ, 2004; SCHÜTT; CLAUSSEN, 2010; LAIRD; SCHAMP, 2015). Yet, other studies have led to the insight that the dynamics of a system can be affected by its underlying topology, making use of network models that are closer to real-world topologies (NEWMANN; BARABÁSI; WATTS, 2006; BOCCALETTI *et al.*, 2006; COSTA *et al.*, 2011b). In this chapter, we investigate how the structural properties of graphs obtained from different network models (random, small-world, scale-free, geographic and regular) influence the dynamics of the RPS game. We start by presenting a variant of the RPS model proposed by Reichenbach *et al.* (REICHENBACH; MOBILIA; FREY, 2007), which includes mobility of individuals in addition to selection and reproduction. Subsequently, we will present some works dealing with the RPS on small-world topologies. Finally, we will present experiments carried out by considering the RPS dynamics on different types of theoretical network models and the obtained results.

7.2 The Rock-Paper-Scissors Game on grids

The RPS game is described by cyclic competition (non-hierarchical) in which species *A* (rock) outcompetes species *B* (scissors), species *B* outcompetes species *C* (paper), and, species *C*, in turn, outcompetes species *A*. Therefore, as described by Laird & Schamp (LAIRD; SCHAMP, 2015), RPS and similar models are frequency dependent. Although competition models can be defined for a variable number of species (ZHANG *et al.*, 2009), we will focus on the three species case in this chapter. When the RPS game is spatially constrained, the regular grid is mostly used. In this case, interactions can only occur locally, since each individual interacts with its nearest neighbors. Let us define a grid containing N available sites, and each of these sites can be occupied by an individual belonging to one of three species, which will be denoted generically as *A*, *B* and *C*, or it can be an empty space, as represented in Fig. 41.

We will describe here a version of the RPS game in which three types of interaction can take place among the individuals: (i) reproduction, (ii) selection (sometimes denoted as competition), and (iii) mobility (exchange). Reproduction can occur when an individual is in an empty site, which is then filled with another individual belonging to the same species.

Figure 41 – Regular grid illustrating a sample configuration of the RPS.

Source: Elaborated by the author.

Figure 42 – Representation of the interactions in the stochastic version of the RPS model. (a) Selection and reproduction occur at rates α and μ , respectively. An empty site is represented by a gray square. (b) All interactions that can take place only locally. Source: Adapted from [Reichenbach, Mobilia and Frey \(2007\)](#).

Reproductions occur at rate μ . The selection process occurs according to the cyclic competition in which every species has one predator and is the predator of another species. The selection process will take place at rate α . Finally, the mobility interactions allow some migrations among the individuals. All individuals can exchange their position with adjacent neighbors at rate ϵ . Therefore, the local interactions of reproduction and selection compete with the migration of

individuals. A schematic representation of the interactions is illustrated in Fig. 42. The three interactions can be characterized according to the following rate equations, where $\emptyset$ denotes an empty site (DALY; BAETENS; DE BAETS, 2015; DALY, 2017):

7.2.1 Mean field analysis of fundamental ecosystems processes

The rate equations described in Eq. (7.1) lead to the following system of ODEs (ordinary differential equations), which accounts for the population densities of each species (A , B and C) through time:

$$\left\{ \begin{array}{l} \frac{dA}{dt} = A[\mu(1-\rho) - \alpha C] \\ \frac{dB}{dt} = B[\mu(1-\rho) - \alpha A] \\ \frac{dC}{dt} = C[\mu(1-\rho) - \alpha B] \end{array} \right. \quad (7.2)$$

We have that $\rho = A + B + C$. System of equations (7.2) holds for a well-mixed population of a large number of individuals (DALY; BAETENS; DE BAETS, 2015; DALY, 2017). While system (7.2) only accounts for the species' densities, it is also possible to incorporate the spatial component using partial differential equation (PDEs). In this way, the population densities can be described as a function of both space and time:

$$\left\{ \begin{array}{l} \frac{\partial A}{\partial t}(\mathbf{r}, t) = \delta \nabla^2 A(\mathbf{r}, t) + \mu A(\mathbf{r}, t)[1 - \rho(\mathbf{r}, t)] - \alpha A(\mathbf{r}, t)C(\mathbf{r}, t) + c_A \xi_A(\mathbf{r}, t), \\ \frac{\partial B}{\partial t}(\mathbf{r}, t) = \delta \nabla^2 B(\mathbf{r}, t) + \mu B(\mathbf{r}, t)[1 - \rho(\mathbf{r}, t)] - \alpha B(\mathbf{r}, t)A(\mathbf{r}, t) + c_B \xi_B(\mathbf{r}, t), \\ \frac{\partial C}{\partial t}(\mathbf{r}, t) = \delta \nabla^2 C(\mathbf{r}, t) + \mu C(\mathbf{r}, t)[1 - \rho(\mathbf{r}, t)] - \alpha C(\mathbf{r}, t)B(\mathbf{r}, t) + c_C \xi_C(\mathbf{r}, t), \end{array} \right. \quad (7.3)$$

where, $\xi(\mathbf{r}, t)$ represents Gaussian white noise with spatio-temporal dependence, ∇^2 is the Laplacian operator, $\mathbf{r} = (r_1, \dots, r_n)$ represents a vector in an n -dimensional space, and, δ is a diffusion coefficient expressed as the area per time. In addition, we have the local density represented by $\rho(\mathbf{r}, t) = A(\mathbf{r}, t) + B(\mathbf{r}, t) + C(\mathbf{r}, t)$

Reichenbach *et al.* (REICHENBACH; MOBILIA; FREY, 2007; REICHENBACH; MOBILIA; FREY, 2008) discuss that this system of equations can give rise to rotating-spiral

waves whose velocity, wavelength and frequency are not influenced by noise in the continuum limit. Therefore, the noise was not taken into account when computing those quantities. As such, the system equations (7.3) was simplified as follows:

$$\frac{\partial z}{\partial t} = M \frac{\partial^2 z}{\partial r^2} + c_1 z - (1 - \iota c_3) |z|^2 z \quad (7.4)$$

where z is a complex variable and the constants c_1 and c_3 depend on the rates α and μ . Eq. (7.4) controls the formation of dynamic spirals, for which it is possible to derive analytic results for the corresponding wavelengths and frequencies. Therefore, the spiral wavelengths, for instance, can be associated with critical mobility values, above which the species co-existence is not possible. Reichenbach *et al.* (REICHENBACH; MOBILIA; FREY, 2007) have shown that for low mobility values, the species diversity can be maintained for a longer time since reproduction and selection occur more frequently among the individuals. On the contrary, when mobility is high, the loss of the spatial homogeneity leads to the loss of biodiversity. Therefore, the authors have shown analytically that it is possible to predict, given the parameter space, whether biodiversity will be lost.

7.2.2 Individual-based modeling

The PDE-based approach provides a theoretical background for understanding the mechanisms that support the species co-existence. However, it does not provide much details on the population dynamics since it takes into account the population as a whole. The system can then be studied from the interactions among the individuals instead of the population level. This individual based approach is focused on tracking changes at the individual level, therefore, they present similarities with CAs, since they are also discrete in both time and space. In this section we present a computational method which was already used to study the RPS model (DALY; BAETENS; DE BAETS, 2015) and was proposed in (REICHENBACH; MOBILIA; FREY, 2007). This method makes use of a stochastic individual-based approach for simulating the RPS dynamics.

We start with the simulation for a regular grid of size $L \times L = N$, as the one shown in Fig. 41, on which the reproduction, selection and mobility can take place according to specified rates μ , α and ϵ (Fig. 42). The computer simulations were implemented according to a modified version of the Gillespie algorithm (GILLESPIE, 1976), which is frequently used for modeling biological and chemical systems. Initially, the sites of the grid are randomly assigned with individuals of the three species and empty sites, in such a way that the final distribution is uniform concerning both the species and the empty sites. Therefore, the initial configuration of the grid is completely even in the example presented here. Then, subsequent steps are carried out according to Algorithm 1.

Algorithm 1 – Gillespie algorithm for simulating the RPS dynamics.

```

1: procedure GILLESPIE( $L, t_{end}$ )
2:   set:  $\mu, \alpha, \varepsilon$ 
3:    $N = L \times L$ 
4:    $t = 0$ 
5:    $r = \mu + \alpha + \varepsilon$ 
6:
7:   while  $t \neq t_{end}$  do
8:     select site  $s_i$                                 ▷ randomly draw a site of the grid
9:     select a neighboring site  $s_j$                   ▷ randomly draw a neighbor site of  $s_i$ 
10:    select interaction:
11:      reproduction, with probability  $\mu/r$ 
12:      selection, with probability  $\alpha/r$ 
13:      exchange, with probability  $\varepsilon/r$ 
14:
15:    if interaction is possible then
16:      update  $s_i$  and/or  $s_j$  accordingly
17:    end if
18:
19:     $t = t + 1$ 
20:  end while
21: end procedure

```

Firstly, the interaction rates must be initialized, as well as the number of time steps for which the algorithm will be evolved. The value of r is a normalization factor for drawing the probability of each interaction. At each time step, a site s_i of the grid is randomly drawn. Then, one of its neighbors is also drawn s_j . We are considering here the von Neumann neighborhood, therefore, one of the four neighbors will be selected. In the next step, an interaction is selected by drawing a random number from the interval $[0, r]$ according to the following probabilities:

- reproduction, given $\frac{\mu}{r}$
- selection, given $\frac{\alpha}{r}$
- exchange, given $\frac{\varepsilon}{r}$

Then, if the selected interaction is possible among the individuals in s_i and s_j , their states are updated accordingly to this interaction. Notice that the interaction is evaluated taking as reference the selected site s_i . The whole process is repeated until the total number of time steps has been reached. Since the cell update is asynchronous, the time steps are aggregated into generations. A single generation represents one Monte Carlo step and corresponds to the number of time steps for which, on average, each individual will be selected once for an interaction event. Since the grid size is given by $N = L \times L$, then one Monte Carlo step corresponds to N time steps.

Using the theory of random walks, Reichenbach *et al.* have found that the area explored by one individual per unit time is given by $M \sim 2\epsilon N^{-1}$, which is referred to as the mobility. Therefore, the corresponding rates of mobility together with reproduction and selection processes can determine whether the species will coexist. Critical mobility values for different grid sizes are presented in Fig. 43. The probability of extinction represents the chance that two species go to extinction after a certain time T which is proportional to the grid size $T(N) \propto N$. Therefore, given a certain number of simulations, the probability of extinction is calculated as the proportion of simulations for which the species went extinct. Figure 43 shows the probability of extinction calculated for 50 simulations as a function of mobility. For this simulation, $\mu = \alpha = 1$. We can observe that as the mobility increases, the probability that the species go to extinction also increases and the coexistence becomes unstable. Above a critical mobility value, species go to extinction for $t < N$.

Figure 43 – Effect of mobility on biodiversity. Probability of extinction as a function of mobility for different grid sizes.

Source: Elaborated by the author.

7.3 From Regular Grids to Irregular Tessellations

Besides the underlying evolutionary processes, the interaction structure on which these processes are taking place can have a huge impact on the competition outcome for what concerns intransitive systems (LAIRD; SCHAMP, 2015). Spatial structures that favor the interaction with nearest neighbors have been used for modeling different coexistence strategies mediated by intransitive relations. When considering real-world examples, such as the biological systems

restricted to two-dimensional substrates (e.g. biofilms), spatial constraints are relevant to the analyzed problem.

In contrast, evolutionary graph theory is more suitable for representing complex interaction structures at the population level. In addition, using graphs for modeling interactions can be more realistic scenarios for many real-world systems (LAIRD; SCHAMP, 2015). Within this context, there are many works addressing small-world topologies for evolving competition dynamics. Between the order of the regular grids and the randomness of stochastic networks (WATTS, 1999), small-world networks have been used to simulate many real-world phenomena.

Szabó *et al.* investigated the RPS dynamics on topologies ranging from regular grids to random graphs (SZABÓ; SZOLNOKI; IZSÁK, 2004; SZOLNOKI; SZABÓ, 2004). The authors discuss how the spatio-temporal distribution of species can be affected by making use of a quenched grid, on which the connections between the sites are redefined but the degree of each site is kept constant, as shown in Fig. 44. They show that for weak quenched randomness (Q), the RPS spatio-temporal patterns remain stable, however as Q approaches a threshold, the amplitude of oscillations in a limit cycle also increases, and, consequently the coexistence of species can be lost. Above such a threshold, the system evolves to monoculture. The authors also analyzed the annealed randomness, i.e., the situation for which the connections of the regular grid are redefined but the degree of the sites is not kept constant, as in Fig. 44. This procedure can be directly compared to the transition of a regular to a random network (see Section 2.3.1.2). For this case, the authors could observe the same phase transitions, although the threshold that leads to species extinction is lower, showing a higher sensitivity to the topological changes.

Other related works also discuss the maintenance of coexistence as a function of structural changes, mobility of individuals or the dispersal rate (BASTOLLA *et al.*, 2005; YING; HUA; WANG, 2007; ZHANG *et al.*, 2009; SCHREIBER; KILLINGBACK, 2013; LAIRD; SCHAMP, 2015). Within this context, we investigated the influence of the topology of the theoretical network models that we have been considering throughout this thesis. For that purpose, we used Algorithm 1 in order to investigate the RPS dynamics on the referred network topologies.

7.4 RPS on Networks

When resorting to networks, the RPS defines cyclic competition among species A , B and C , which now occupy the nodes of a network. We also limit the RPS network analysis to the three species case. In addition, empty spaces are also considered, allowing for the reproduction and the mobility of individuals. In contrast to the grid approach, some network models have long-range correlations, which can lead the species to extinction. Figure 45 presents a sample network on which it is possible to distinguish the three species (red, yellow and blue) and the empty sites (white nodes).

Figure 44 – Random rewiring of the links of a grid, which is controlled by Q , also referred to as quenched randomness. The resulting degree distribution is also uniform, however, depending on the value of Q , the grid can approach a small-world network.

Source: Adapted from Szabó, Szolnoki and Izsák (2004).

Figure 45 – Illustration of the rock-paper-scissors game on an irregular topology.

Source: Elaborated by the author.

We studied the RPS dynamics on different network topologies. For that purpose, we performed extensive simulations and analyses in order to find correlations between the specific structural characteristics of the networks and maintenance of biodiversity in communities with

non-transitive relations.

7.4.1 Experimental design

In order to carry out the proposed study we generated synthetic networks from the following theoretical network models: random, small-world, scale-free, geographical and regular networks. Note that the regular network, similarly to the grid case, is characterized by a uniform degree distribution, however, the spatial arrangement is different. Besides structural differences in the analyzed graphs, we also evaluated the effect of the mean degree $\langle k \rangle$ and the number of nodes N . Therefore we generated networks from the same network model with different combinations of $\langle k \rangle$, $\{4, 6, 8, 10\}$ and N , $\{400, 900, 1600, 10000\}$. The values for N correspond to the same grid sizes considered by Reichenbach *et al.* (REICHENBACH; MOBILIA; FREY, 2007).

In agreement with previous works, the following processes were simulated on these graphs: 1) selection according to the non-hierarchical competition structure, 2) reproduction and 3) migration. The simulations were implemented according to Algorithm 1. Therefore, the interactions are stochastic, meaning that at every iteration one individual and one of its neighbors are randomly chosen, after which only one of the processes occurs, which one is dictated by their respective reaction rates. The number of Monte Carlo generations for each simulation is proportional to N , and, each generation represents the number or time steps that, on average, each individual will be drawn for an interaction event. For all the simulations we adopted $\mu = \alpha = 1$ and mobility varying within the range $[10^{-12}, 10^{-2}]$. The mobility is calculated as a function of the exchange rate ε and N , in the same way for regular grids ($M \sim 2\varepsilon N^{-1}$). In addition, for each combination of network model, $\langle k \rangle$ and N , the simulation was repeated 50 times, with different initial configurations, which were defined according to a uniform distribution concerning the three species and the empty sites.

7.4.2 Results

We investigated how the coexistence of species is affected by tracking the probability of extinction, the time until the first extinction and the patchiness.

7.4.2.1 Probability of extinction

The probability of extinction was calculated taking into account the 50 simulations for each combination of network model, $\langle k \rangle$ and N , therefore, it accounts for the number of simulations in which the species went extinct after a certain time T , such that $T \propto N$. Figure 46 shows the probability of extinction calculated as a function of mobility for scale-free networks. In contrast to the grid case, the probability of extinction is always high irrespective of the average

degree or the network size. Extinctions always occur because of the long range connections even if the mobility is very low. There are only a few exceptions for the networks with lower mean degree, meaning that since there are fewer connections, the time to extinction increases.

Figure 46 – Probability of extinction as a function of mobility for scale-free networks. Each plot corresponds to a different network size: (a) $N = 400$, (b) $N = 900$, (c) $N = 1600$ and (d) $N = 10000$. Within each plot, each curve corresponds to different average degrees, $\langle k \rangle$.

Source: Elaborated by the author.

Figure 47 shows the same plots for regular networks with different parameters. We can observe that the probability of extinction is much smaller for all values of N . This result relates to the absence of long-range connections in this type of network, which presents a high clustering coefficient and many local interactions. In addition, for all values of N , as the average mean degree increases, the probability of extinction also increases, due to a larger number of connections that can favor the interactions.

The same plots for the probability of extinction are presented in Appendix B for random, small-world and geographical networks (Figs. 126, 127 and 128, respectively). We can observe a similar behavior for such networks when comparing to the plots obtained for the scale-free networks. Even when considering very low mobility values, the probability of extinction is very high in almost all the situations. Some exceptions can be seen in the cases where $\langle k \rangle = 4$. For

Figure 47 – Probability of extinction as a function of mobility for regular networks. Each plot corresponds to a different network size: (a) $N = 400$, (b) $N = 900$, (c) $N = 1600$ and (d) $N = 10000$. Within each plot, each curve corresponds to different average degrees, $\langle k \rangle$.

Source: Elaborated by the author.

the geographical network, for instance with $N = 10000$, the observed behavior is similar to the grid case, although the threshold for extinction is much lower for the latter.

7.4.2.2 Time to the first extinction

In addition to the probability of extinction, we also analyzed the average time to the first extinction. Figures. 48 and 49 present the results for scale-free and regular networks. Similarly, section B.2 of Appendix B presents the corresponding plots for the other network models (Figs. 129, 130 and 131). The average time to first extinction is only calculated by taking into account the simulations in which at least one species went extinct.

Recalling that one Monte Carlo (MC) step is the number of time steps for which, on average, each individual will be chosen once to interact, for a network with $N = 400$ nodes, one MC step corresponds to 400 time steps. The analysis of Fig. 48 shows that the species go extinct faster than the time needed to complete one MC step. This is observed for all the values

Figure 48 – Average time to the first extinction as a function of mobility for scale-free networks. Each plot corresponds to a different network size: (a) $N = 400$, (b) $N = 900$, (c) $N = 1600$ and (d) $N = 10000$.

Source: Elaborated by the author.

of N . We can also observe that the lower the degree $\langle k \rangle$, the longer the time to first extinction irrespective of the number of nodes and mobility. However, as N increases, the average time to first extinction also increases. In addition, for higher values of mobility, we can observe the steepest increase of the time to first extinction, specially for larger N . This is due to the fact that as mobility increases, exchange events will happen much faster than competition and selection interactions, and, consequently, it will take longer for species extinction. The conclusions for the scale-free networks also hold for the results obtained for other network models. For small-world networks, for instance, though many short-range connections are still present (high clustering coefficient), the rewiring process from which they are constructed creates enough long-range connections that can lead to a faster extinction.

In the case of regular networks (Fig. 49), the influence of the average degree on the time to first extinction is smaller when compared to the other network models. However, although there is much more variability for the regular case, networks with $\langle k \rangle = 4$ also take a longer time for extinction in a similar way as for the other network models. Note that the time to first

Figure 49 – Average time to the first extinction as a function of mobility for regular networks. Each plot corresponds to a different network size: (a) $N = 400$, (b) $N = 900$, (c) $N = 1600$ and (d) $N = 10000$.

Source: Elaborated by the author.

extinction is calculated only for the simulations in which at least one species went extinct. Since regular networks had the lowest probabilities of extinction, the time to the first extinction for these networks was averaged for a smaller number of simulations, consequently increasing the standard deviation.

7.4.2.3 Patchiness

In order to gain insights into the spatial distribution of species during the RPS evolution, we analyzed the patchiness, which quantifies the number of individuals from the same species located near each other. We calculated the average cluster size per time and the number of clusters per time.

7.4.2.3.1 Average cluster size per time

We consider a cluster as a connected component of a network that is composed of individuals of the same species. The simulation starts with a uniform distribution of species that are randomly assigned to the network sites. Therefore, there exists a large number of small clusters at the beginning of the simulation. If the network topology favors species extinction, the average cluster size will increase after some time steps. However, for the case of the regular topology, this increase is slow.

Figure 50 – Average cluster size per Monte Carlo generation for scale-free networks with $N = 400$. Each plot corresponds to a network with a different $\langle k \rangle$ and each curve corresponds to one mobility value.

Source: Elaborated by the author.

Figure 50 shows the average cluster size as a function of the Monte Carlo generation for scale-free networks with $N = 400$ nodes and with different values of $\langle k \rangle$. Each curve corresponds to one mobility value. We can see that the higher the value of $\langle k \rangle$, the more similar the curves are across the different mobilities. Note that, when the average cluster size is already equal to N , the system is already a mono-culture. Also, since the extinction goes faster, due to the higher connectivity among the individuals, the maximum cluster size is also reached faster.

Figure 51 – Average cluster size per Monte Carlo generation for scale-free networks with $N = 1600$. Each plot corresponds to a network with a different $\langle k \rangle$ and each curve corresponds to one mobility value.

Source: Elaborated by the author.

We also analyzed the effect of N on the average size of the clusters. Figure 51 shows the average cluster size as a function of the Monte Carlo generation for scale-free networks with size $N = 1600$. Comparing Figs. 50 and 51, we can see that the time to mono-culture is proportional to the system size. In addition, the effect of the high mobility is clear. We can observe that for a very high mobility, it takes longer to extinction (see Fig. 48), so more time is required for the system to reach the maximum cluster size.

In conclusion, for the scale-free networks, the lower the values of $\langle k \rangle$ and the lower the number of nodes, the faster the extinction at the highest mobilities. Moreover, increasing $\langle k \rangle$ and/or increasing N will result in a slower increase of the cluster size. We also performed this analysis for scale-free networks with different sizes ($N = 900$ and $N = 10000$). We can see that for $N = 10000$ the effect of high mobility is even more highlighted. In addition, we also performed the same simulations using random, geographic and small-world networks. Similar conclusions presented here for the scale-free networks also hold for these networks.

Here, we also present the analysis for regular networks. Figure 52 shows the average

Figure 52 – Average cluster size per Monte Carlo generation for regular networks with $N = 1600$. Each plot corresponds to a network with a different $\langle k \rangle$ and each curve corresponds to one mobility value.

Source: Elaborated by the author.

cluster size as a function of the Monte Carlo generation for regular networks with sizes $N = 1600$ for different values of $\langle k \rangle$. Recalling that for a fixed value of N the time to extinction is longer for lower $\langle k \rangle$ (Fig. 49), the probability of extinction for this type of networks is very low (Fig. 47) when compared to the other types of network model. Therefore, we can observe that for regular networks the average cluster size increases smoothly. Given the fact that the probability of extinction is very low, the average cluster size increases but does not reach the maximum size (only observed for monocultures). Therefore, the clusters increase in size, which means that the size of the connected components formed by the individuals belonging to the same species will increase, i.e., the individuals tend to be grouped by species. As discussed in Section 7.2.1 for the grid, this corresponds to an increase of the spiral wavelengths.

7.4.2.3.2 Average number of clusters per time

When biodiversity is lost, only a single cluster will remain at the end of the simulation. We illustrate this case for scale-free networks. Figure 53 shows the average number of clusters per Monte Carlo generation for networks of size $N = 1600$. For the first Monte Carlo generations, there is a transient during which different clusters of the same species are joined. This explains the steep decrease of the number of clusters. As the mobility decreases, this transient becomes longer. We can also observe that as $\langle k \rangle$ or N increases, the length of the transient decreases.

Figure 53 – Average number of clusters per Monte Carlo generation for scale-free networks with $N = 1600$. Each plot corresponds to a network with a different $\langle k \rangle$ and each curve corresponds to one mobility value.

Source: Elaborated by the author.

We also calculated the average number of clusters per Monte Carlo generation using random, geographic and small-world networks. Similar results to the scale-free networks were also obtained for these networks. As N increases, the number of clusters tends to converge. Since this type of network favors species coexistence, this result also indicates that the number of individuals per species also tends to converge so that the system remains in equilibrium.

In general, for the different types of networks considered, we observed a considerable influence of their topology on the RPS dynamics, and likewise with the exception of the regular network, the probability of extinction is very high for the different networks and network configurations. This is mainly due to the presence of long-range connections. In addition, for most experiments the time to the first extinction is shorter than the time to complete one MC step. As in the classic grid-based RPS, before a species goes extinct, we observed the same spatial clustering of the species. Initially the species are randomly distributed on the network, but, as the system evolves, the species tend to be arranged in larger clusters.

7.5 Conclusion

In this chapter we presented the results of computer simulations carried out in order to analyze the impact of different network topologies on RPS dynamics. We started by presenting the RPS on regular tessellations. The cyclic evolution is not the only factor that can influence the coexistence of species. The individuals' mobility also plays an important role in species coexistence.

In addition, based on these extensive simulations and corroborating preliminary studies, we conclude that not only the mobility of the species has an important impact on the evolved dynamics, but the connectedness of the species as well. Some spatial arrangements focused on interactions with nearest neighbors have less long-range correlations which can favor biodiversity. In contrast, other kinds of topologies, like the random or scale-free networks did not lead to species diversity, with a short time to first extinction. In the case of the small-world networks, biodiversity will depend on the rewiring probability that generated such a topology. However, rewiring a few links is enough to increase long-range connections, and, consequently, this topology may also favor species extinction. Regular networks have shown to be closer to the grid case. The probability of extinction for regular networks is lower than the grid, even for higher mobility. In general, for random, scale-free, small-world and geographical networks, the lower the degree ($\langle k \rangle$), the longer the time to first extinction, irrespective of the number of nodes and mobility. In addition, the higher the number of nodes, the longer the time to first extinction at high mobility.

In the next part of this thesis, we turn to applications that can be addressed by using the methodology proposed so far in this thesis.

Part III

Understanding Patterns and Describing Processes on Networks

USING LLNA FOR PATTERN RECOGNITION APPLICATIONS

8.1 Introduction

In this part of the thesis we will explore pattern recognition (PR) in the context of network data. Networks are a useful tool for data modeling. In addition, the combination of PR and networks appears as an important alternative for finding, identifying, analyzing, and clustering patterns that are unfeasible to handle with other approaches. Network characterization is one of the goals of using pattern recognition and machine learning techniques, which can lead to an understanding of the intrinsic patterns that are related to the network structure. This characterization can be performed, for instance, through the use of structural network measures (COSTA *et al.*, 2007; BACKES; CASANOVA; BRUNO, 2009a; COSTA *et al.*, 2010). In this chapter, however, we will use the data inferred from a spatially explicit model (SEM) defined on the network structure (GONÇALVES; MARTINEZ; BRUNO, 2012; GONCALVES *et al.*, 2012; MIRANDA; MACHICAO; BRUNO, 2016b; MIRANDA; MACHICAO; BRUNO, 2016a). Figure 54 illustrates the proposed approach.

The datasets used are composed of network data defined within a specific application context. Then, we will apply the LLNA in order to obtain the corresponding spatio-temporal patterns for each network belonging to the dataset. The next step is feature extraction in order to quantify these spatio-temporal patterns using the measures already discussed in Chapter 3, e.g., Shannon entropy, Lempel-Ziv complexity, among others. Finally, the last step consists of assembling a feature vector for each network, which is used in a learning strategy for network classification. One important parameter that should be investigated when using LLNA in PR is the underlying rule. Given a LLNA rule r_i , its performance may be better when compared with the performances of other LLNA rules, given a specific PR application. Therefore, the entire rule space should be investigated.

Figure 54 – Pattern recognition in networks using the Life-like Network Automata.

Source: Elaborated by the author.

We implemented the process described in Fig. 54 for two types of applications: synthetic networks and real-world networks. In the former, we classified theoretical network models. We already performed similar experiments in Chapters 5 and 6. Here we extend this approach for the entire LLNA rule space. In addition, still considering these synthetic networks, we analyzed the influence of the average degree on the classification accuracy. Finally, we also evaluated different classes of scale-free networks, generated according to the Barabási & Albert model and other theoretical models.

In relation to the real-world networks, we performed classification tasks for real-world applications for which the data are represented as networks. We used LLNA for four pattern recognition applications: (i) identifying organisms from the three domains of life, *Archaea*, *Bacteria* and *Eukaryota*, through their metabolic networks. This dataset was first studied by Jeong *et al.* (JEONG *et al.*, 2000); (ii) identifying structural patterns in two online social networks (*Twitter* and *Google+*), using the social interactions obtained from the SNAP database (LESKOVEC; KREVL, 2014); (iii) classifying stomata distributions varying according to different lighting conditions, and, (iv) identifying fungal growth patterns, by focusing on the investigation of the topological properties related to fungal growth dynamics. Since the theoretical network models are well-studied in literature and have known properties, their classification supports the understanding and validation of the LLNA framework. Moreover, classifying real-world networks complements the proposed PR approach and is a strong evidence that LLNA has a general applicability. The majority of the applications presented here are already published in (MIRANDA; MACHICAO; BRUNO, 2016a).

8.2 Experimental Design

In this section we present the design of the experiments performed in this chapter, which includes the description of the datasets and how we generated the theoretical network models, the training and test strategies, the feature extraction and the creation of the feature vectors, the selected classifiers, and, the measures of performance used for assessing the obtained results.

8.2.1 Description of the datasets

We conducted experiments with two distinct types of networks, namely, synthetic networks and real-world networks. The former are organized into three datasets: *synthetic-dataset*, *rule-selection-dataset* and *synthetic-scalefree-dataset*. The latter are the following datasets: *metabolic-dataset*, *social-dataset*, *stomata-dataset* and *fungi-dataset*. Detailed information about these datasets is given next.

- *synthetic-dataset*: composed of networks generated according to the following theoretical models: 1) random, with connection probability of $p = \langle k \rangle / n$; 2) small-world, with rewiring probability $p = 0.1$; 3) scale-free, with linear and non-linear preferential attachments, and, 4) geographical. For each model, we have networks with $\langle k \rangle$: 4, 6, 8, 10, 12, 14, 16; and, N : 500, 1000, 1500 and 2000 nodes. For each of the 28 combinations of $\langle k \rangle$ and N , 100 networks were generated. Therefore, for this dataset, there are in total 11200 networks (2800 of each model);
- *rule-selection-dataset*: composed of networks generated according to the same theoretical models used in *synthetic-dataset* and with the same parameters. However, only networks with $N = 500$ were considered. In addition, only 50 networks for each of the 7 $\langle k \rangle$ values were generated. Therefore, the networks belonging to this dataset are totally different from the samples in the *synthetic-dataset*;
- *synthetic-scalefree-dataset*: composed of scale-free networks generated according to the Barabási & Albert (BARABÁSI; ALBERT, 1999) model and the Dorogovtsev & Mendes (DOROGOVTSSEV; MENDES, 2002) model. For the first model, both linear and non-linear preferential attachments were considered in order to generate the networks: $\alpha = \{0.5, 1.0, 1.5 \text{ and } 2.0\}$. Therefore, there are five classes in this dataset and 100 networks for each class with $\langle k \rangle = 8$ and $N = 1000$ nodes;
- *metabolic-dataset*: composed of 43 metabolic networks describing the metabolic pathways of three types of organisms: *archaea* (6), *bacteria* (32) and *eukaryotes* (5) (JEONG *et al.*, 2000; CCNR/ICENSA, 2018). We divided this dataset into rule selection and classification sets. The first one consists of two randomly selected networks of each organism and was

used to select the set of LLNA rules that provide the best accuracy in distinguishing among the three evolutionary classes. The second is composed of the remaining networks;

- *social-dataset*: composed of 130 networks from the SNAP (*Stanford Network Analysis Project*) platform (LESKOVEC; KREVL, 2014) representing social interactions in Google+ and Twitter (65 of each online social tool). From these 65 networks, 15 were randomly chosen for rule selection and the remaining for classification. The interactions in the referred social networks are the so-called “ego-networks” representing social relationships of a specific user (“ego”);
- *stomata-dataset*: composed of 6 images of leaves from *Tradescantia zebrina* plants containing their stomata distribution patterns. The plants were exposed to three lighting conditions: (i) sunlight (Natural), on which the plants are exposed to the sun light, (ii) 4 hours (L4) of artificial light, on which the plants are exposed to artificial light during 4 hours, and, similarly (iii) 24 hours (L24) of artificial light. There are 6 images for each lighting condition. Two images were chosen randomly for the rule-selection procedure and the other four were used in the classification step;
- *fungi-dataset*: composed of time series of images of two different fungal species, *Rhizoctonia solani* and *Coniophora puteana* (ULZURRUN *et al.*, 2015). The fungal growth on Petri dishes was scanned automatically every hour, with both temperature (15, 20, 25 and 30 °C) and relative humidity (RH) (65, 70, 75 and 80 %) controlled. So, a total of 16 environmental conditions were considered. In total there are four replicates for each of the 16 environmental conditions. Each replicate results in a time series of 269 images.

8.2.2 Generation of the network models

For the generation of the theoretical network models we made use of the *igraph* library for R¹ (CSARDI; NEPUSZ, 2006). The Dorogovtsev & Mendes model and the geographical networks were implemented according to (DOROGOVITSEV; MENDES, 2002) and (COSTA *et al.*, 2007), respectively. The datasets of containing these networks can be downloaded at: <http://scg.ifsc.usp.br/canetworks/>.

8.2.3 Training & test strategies

Figure 55 shows how the training and test procedures were carried out for the PR applications that are tackled in this chapter. The original dataset is divided into two sets. The total number of instances in each set varies for each dataset. The first set is then used for the selection of the best rules. As discussed in Chapter 6, there are different evolution patterns provided by

¹ <https://www.r-project.org>

LLNA rules. Therefore, the entire rule space should be assessed for PR applications. For this task, the selected dataset is then divided into training and test sets. After classification is performed for the rule selection phase, the set of rules that yielded the best classification accuracies are selected. Then, this set of rules is used to calculate the final performance over the second dataset. Hence, the LLNA is only tested for the selected rules. Similarly, during classification phase, the dataset is also divided into training and test sets. In this strategy, the instances used for rule selection are not used to obtain the final classification performances.

We used n -fold cross-validation in all the experiments. The cross-validation procedure was applied 100 times for both the rule-selection dataset and the classification dataset. Therefore, the obtained standard deviation relates to the variation in accuracy for each run of this procedure.

Figure 55 – Training and test strategies of LLNA used in pattern recognition applications.

Source: Elaborated by the author.

When the datasets did not have a uniform distribution of classes, we used a random sub-sample strategy (e.g. *metabolic-dataset*). This procedure consists of choosing at random, for each fold and for each cross-validation, the same number of instances for all the classes. In this case the number of instances available for the least frequent class will be the number of instances to be drawn for the remaining classes. In this way, we guarantee a uniform distribution of classes. In addition, all the available instances can be used in the cross-validation.

8.2.4 Feature extraction & Classification

The Shannon entropy (S), the word length (WL) and the Lempel-Ziv (LZ) complexity were selected as attributes of the spatio-temporal patterns obtained with LLNA. Therefore, the feature vectors are composed of these measures.

The first feature vector, $\vec{\mu}_S$ corresponds to the Shannon entropy distribution which takes values in the interval $[0,1]$. Similarly, to the PR application described in Section 6.3, in order

to construct $\vec{\mu}_S$, the Shannon entropy is first calculated for each network node, and, from these values we generated a histogram by dividing the unit interval into 20 bins. The value of each bin corresponds to one feature. Therefore, $\vec{\mu}_S$ consists of 20 features or attributes that represent the respective entropy frequencies.

In a similar way, we obtained the second feature vector $\vec{\mu}_W$ which is determined by the word length distribution. The maximum WL is bounded by the number of time steps. However, we observed that the frequency of words with length larger than 40 is very low, and, therefore, we considered only words with length smaller than 40. The bins of the histogram have length 2, which also leads to 20 features for $\vec{\mu}_W$. The last feature vector $\vec{\mu}_L$ consists of the LZ distribution which is also divided into 20 bins. The LZ values were normalized by the maximum LZ among all nodes.

In addition to the distributions, we also considered the feature vectors composed by the average values of the same measures as attributes: average Shannon entropy, average WL and average LZ: $\langle \mu_S \rangle, \langle \mu_W \rangle, \langle \mu_L \rangle$. The k-NN (*k-Nearest Neighbors*), with $k = 1$ and SVM (*Support Vector Machines*) classifiers were used in the classification step. Regarding SVM, the value of the c parameter is 1.0 and a polynomial function was used as kernel function.

8.2.5 Performance evaluation

The experiments presented in this chapter were evaluated through the use of a set of performance measures that were calculated for each experiment and for each specific PR application.

Negative (TN), False Positive (FP), type I error, and, False Negative (FN), type II error.

- Accuracy (ACC): the rate of correctly classified instances

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \quad (8.1)$$

- True Positive Ratio (TPR): the proportion of true positives, also known as recall or sensitivity (percentage of positive correctly classified instances for a given class)

$$TPR = \frac{TP}{TP + FN} \quad (8.2)$$

- False Positive Ratio (FPR): the proportion of false positives (percentage of negative incorrectly classified instances for a given class)

$$FPR = \frac{FP}{FP + TN} \quad (8.3)$$

- F1-score: derived from the combination of precision ($\frac{TP}{TP+FP}$) and recall ($\frac{TP}{TP+FN}$):

$$F1 = \frac{2TP}{2TP + FP + FN} \quad (8.4)$$

- Matthews Correlation Coefficient (MCC)

$$\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (8.5)$$

- Area Under a ROC Curve (AUC):

$$AUC_1 = \frac{U_1}{n_1 n_2} \quad (8.6)$$

where n_1, n_2 are the sample size for samples 1 and 2, respectively, and U_1 is the Mann–Whitney U test.

8.3 Pattern Recognition in Synthetic Networks

This section presents three experiments with synthetic networks in order to illustrate the pattern recognition approach of LLNA.

8.3.1 Classifying network models

The classification of network models is implemented here as a basic classification problem which can be further used as a benchmark for other PR applications. If the theoretical network models can be distinguished using LLNA descriptors, it is an indication that such descriptors might also be employed for more complex PR applications that make use of real-world networks.

First, we started by investigating the entire set of LLNA rules for the *rule-selection-dataset*. This way, each LLNA rule was evaluated in relation to the classification performance obtained when such rule was used as the transition function of LLNA. Therefore, each network model corresponds to a class of interest in this experiment: random, small-world, scale-free and geographical. The classification procedure consists of running the LLNA for each of the networks in the *rule-selection-dataset* and extracting the corresponding spatio-temporal features. Finally, the cross-validation procedure is applied.

The histogram of Fig. 56 presents the proportion of rules for which a certain accuracy was obtained. The $k - NN$ classifier and the Shannon entropy distribution ($\vec{\mu}_S$) were used in this experiment. We can observe that for most of the rules the accuracy was greater than 60%. Moreover, for a small set of rules, accuracies greater than 90% were obtained. From this set,

the 10 rules with the highest accuracies were selected, which were then used in subsequent experiments with synthetic networks. We then evaluated if the selected rules would provide a similar performance in other datasets.

Figure 56 – Accuracy obtained for all the LLNA rules regarding the correct classification of network models: random, small-world, scale-free and geographical. The best results were obtained with the highlighted rules.

Source: Adapted from [Miranda, Machicao and Bruno \(2016a\)](#).

We first evaluated the final classification accuracy for the *synthetic-dataset*. The results obtained are shown in Fig. 57-(a) for the four rules that provided the highest accuracies. The referred dataset consists of networks of the four network models, with different values of N and $\langle k \rangle$. High accuracy values were obtained for all feature vectors, except for $[\langle \mu_S \rangle, \langle \mu_W \rangle, \langle \mu_L \rangle]$, containing the average values of each measurement. The maximum accuracy obtained was $99.992 \pm 0.002\%$ for rule $B135678/S03456$ using the combination of the distributions $[\vec{\mu}_S, \vec{\mu}_W, \vec{\mu}_L]$. The distributions also provided a high accuracy when analyzed separately, as the Shannon entropy $\vec{\mu}_S$. Table 3 presents the same accuracy values of Fig. 57-(a) for all the 10 selected rules. In addition, Table 4 presents the performance measurements for the same experiment for the best rule. We can observe that with the best rule, more than 99% was reached, leading to the maximum performance in Table 4.

Figure 57-(b) shows the canonical analysis for the *synthetic-dataset*. We used $\vec{\mu}_S$ as the feature vector and rule $B135678/S03456$ as transition function. The canonical analysis provides a linear combination of the original attributes, maximizing the separation between the classes ([GITTINS, 2012](#)). Therefore, it is a regression analysis. The first and the second canonical variables are the eigenvectors with the highest eigenvalues of the matrix defined by the intra-class variation, and, another matrix that quantifies the inter-class variation. We can observe that there is a clear separation among the four models. This result corroborates the accuracies obtained using the feature vectors composed by the distributions.

Figure 57 – Synthetic network characterization with LLNA. (a) Accuracy (%) and standard deviation resulting from the classification of network models from the *synthetic-dataset*, using five different feature vectors. Results are shown for the best four LLNA rules. The distributions of the Shannon entropy, the word length and the Lempel-Ziv complexity are represented by the vectors $\vec{\mu}_S$, $\vec{\mu}_W$ and $\vec{\mu}_L$, respectively. The vector $[\vec{\mu}_S, \vec{\mu}_W, \vec{\mu}_L]$ consists of the combination of all these distributions, and, $[\langle\mu_S\rangle, \langle\mu_W\rangle, \langle\mu_L\rangle]$ contains the average values of each measure. (b) Canonical analysis performed for the *synthetic-dataset* using rule B135678/S03456 and $[\vec{\mu}_S, \vec{\mu}_W, \vec{\mu}_L]$ as feature vector.

Source: Adapted from Miranda, Machicao and Bruno (2016a).

Table 3 – Accuracy (%) obtained using different feature vectors for the classification of network models. The networks have $N = 500, 1000, 1500$ and 2000 nodes, and, mean degree varying from $\langle k \rangle = 4, 6, \dots, 16$. The results are shown for the 10 rules obtained in the rule selection step.

Rules	$\vec{\mu}_S$	$\vec{\mu}_W$	$\vec{\mu}_L$	$[\vec{\mu}_S, \vec{\mu}_W, \vec{\mu}_L]$	$[\langle\mu_S\rangle, \langle\mu_W\rangle, \langle\mu_L\rangle]$
B01678/S0457	99.37 (± 0.01)	98.87 (± 0.01)	99.55 (± 0.01)	99.946 (± 0.003)	97.34 (± 0.02)
B035678/S023456	98.91 (± 0.02)	99.33 (± 0.02)	98.96 (± 0.02)	99.89 (± 0.01)	84.31 (± 0.03)
B1245678/S3456	97.40 (± 0.03)	96.09 (± 0.01)	97.67 (± 0.02)	99.80 (± 0.01)	95.65 (± 0.01)
B0157/S457	97.19 (± 0.04)	87.74 (± 0.03)	94.60 (± 0.03)	98.67 (± 0.04)	82.60 (± 0.03)
B35678/S03456	98.77 (± 0.02)	99.26 (± 0.03)	99.900 (± 0.004)	99.98 (± 0.01)	68.61 (± 0.05)
B1245678/S03456	96.24 (± 0.04)	96.67 (± 0.03)	95.57 (± 0.02)	99.91 (± 0.01)	95.65 (± 0.01)
B35678/S13456	97.70 (± 0.05)	99.03 (± 0.03)	99.92 (± 0.01)	99.94 (± 0.01)	70.28 (± 0.08)
B01245678/S3456	97.73 (± 0.03)	97.93 (± 0.02)	91.63 (± 0.04)	99.68 (± 0.01)	94.57 (± 0.05)
B035678/S0123456	98.39 (± 0.02)	98.93 (± 0.02)	98.34 (± 0.03)	99.79 (± 0.01)	84.67 (± 0.05)
B135678/S03456	99.47 (± 0.02)	99.912 (± 0.003)	99.76 (± 0.01)	99.992 (± 0.002)	79.30 (± 0.06)

Source: Adapted from Miranda, Machicao and Bruno (2016a).

Table 4 – Performance measures corresponding to the best accuracy as shown in Table 3.

TPR	FPR	F-Measure	MCC	ROC Area	Class
1	0	1	1	1	Random
1	0	1	1	1	Small-world
1	0	1	1	1	Scale-free
1	0	1	1	1	Geographical
1 $\pm$ 0	0 $\pm$ 0	1 $\pm$ 0	1 $\pm$ 0	1 $\pm$ 0	Avg

Source: Adapted from Miranda, Machicao and Bruno (2016a).

8.3.1.1 Robustness to noise

We evaluated the robustness of LLNA regarding noise tolerance. We performed modifications in the network topology by adding and removing edges according to a noise ρ_N . The

higher the values of ρ_N , the more structural changes are performed on the network topology. Let us consider, for instance, $\rho_N = 20\%$, then 10% of edges will be randomly added to the network structure, while 10% of the existing edges will be removed. Table 5 shows the classification rate of the four network models when considering different values of ρ_N and using SVM classifier. The feature vector consisting of the distributions, $[\vec{\mu}_S, \vec{\mu}_W, \vec{\mu}_L]$, provided the best accuracy. In general, we observe that the increase of ρ_N does not influence the obtained accuracy.

Table 5 – Robustness analysis. Classification accuracy (%) given different values of ρ_N , and, also without noise ($\rho_N = 0$), which corresponds to the results in Table 3.

Rules	Without Noise	$\rho_N = 10\%$	$\rho_N = 20\%$	$\rho_N = 30\%$
B01678/S0457	99.946 (± 0.003)	99.85 (± 0.01)	99.70 (± 0.01)	99.84 (± 0.01)
B035678/S023456	99.89 (± 0.01)	99.90 (± 0.01)	99.79 (± 0.01)	99.95 (± 0.01)
B1245678/S3456	99.80 (± 0.01)	99.79 (± 0.02)	99.51 (± 0.02)	99.73 (± 0.01)
B0157/S457	98.67 (± 0.04)	99.44 (± 0.02)	99.73 (± 0.02)	99.85 (± 0.01)
B35678/S03456	99.98 (± 0.01)	99.96 (± 0.01)	99.91 (± 0.01)	99.990 (± 0.003)
B1245678/S03456	99.91 (± 0.01)	99.80 (± 0.01)	99.61 (± 0.02)	99.68 (± 0.02)
B35678/S13456	99.94 (± 0.01)	99.94 (± 0.01)	99.93 (± 0.01)	99.97 (± 0.01)
B01245678/S3456	99.68 (± 0.01)	99.54 (± 0.01)	99.48 (± 0.02)	99.62 (± 0.02)
B035678/S0123456	99.79 (± 0.01)	99.86 (± 0.01)	99.84 (± 0.01)	99.93 (± 0.01)
B135678/S03456	99.992 (± 0.002)	99.983 (± 0.004)	99.97 (± 0.01)	99.95 (± 0.01)

Source: Adapted from [Miranda, Machicao and Bruno \(2016a\)](#).

8.3.2 Classifying network models in combination with mean degree

We also evaluated the influence of the network average degree, $\langle k \rangle$, on the network descriptors, constructed on the basis of spatio-temporal patterns. In Section 6.3.1 we observed that different evolution patterns are obtained for the same network model given different values of $\langle k \rangle$. Therefore, we investigated in the experiment introduced in this section whether the spatio-temporal pattern of a given network model with specific $\langle k \rangle$ is unique. For this task, we considered the combinations of $\langle k \rangle$ and the network models as classes. Therefore, there are 28 classes in this experiment resulting from the combination of seven values of $\langle k \rangle$ and four network models. This experiment was carried out for the *synthetic-dataset*. The accuracy results are shown in Fig. 58 for the same 10 rules of the previous experiment.

The four rules highlighted in Fig. 58 are the ones that provided the highest accuracies among the ten selected rules. The maximum accuracy ($90.76 \pm 0.07\%$) was obtained when using the combination of the distributions as feature vector and rule B01678/S0457. When comparing the three distributions separately, we observe that $\vec{\mu}_S$ provided the highest accuracies. This shows that it is possible to distinguish the evolution pattern not only for the network models, but also for networks with distinct values of $\langle k \rangle$. Yet, the feature vector composed of the average measures did not obtain a good performance for the classification of the network models. Table 6 presents the accuracy concerning the classification of the network models in combination with the mean

degree $\langle k \rangle$, which were adopted as classes, for all the 10 selected rules and Table 7 shows the corresponding performance measurements for the best rule, *B01678/S0457*.

Figure 58 – Synthetic network characterization with LLNA. Accuracy (%) and standard deviation resulting from the classification of network models in combination with $\langle k \rangle$, using five different feature vectors. Results are shown for the four best LLNA rules.

Source: Adapted from Miranda, Machicao and Bruno (2016a).

Table 6 – Accuracy (%) obtained using different feature vectors for the classification of network models in combination with their $\langle k \rangle$. The results are shown for the 10 rules obtained in the rule selection step.

Rules	$\bar{\mu}_S$	$\bar{\mu}_W$	$\bar{\mu}_L$	$[\bar{\mu}_S, \bar{\mu}_W, \bar{\mu}_L]$	$[\langle \mu_S \rangle, \langle \mu_W \rangle, \langle \mu_L \rangle]$
B01678/S0457	83.35 (± 0.09)	77.8 (± 0.3)	80.87 (± 0.07)	90.76 (± 0.07)	67.0 (± 0.4)
B035678/S023456	84.97 (± 0.06)	82.8 (± 0.2)	82.91 (± 0.07)	90.12 (± 0.07)	60.6 (± 0.5)
B1245678/S3456	80.26 (± 0.08)	73.9 (± 0.1)	80.66 (± 0.05)	86.76 (± 0.08)	54.4 (± 0.4)
B0157/S457	84.33 (± 0.07)	68.7 (± 0.1)	76.8 (± 0.2)	89.34 (± 0.09)	49.1 (± 0.2)
B35678/S03456	77.9 (± 0.1)	77.5 (± 0.1)	75.3 (± 0.1)	86.76 (± 0.09)	62.3 (± 0.5)
B1245678/S03456	80.23 (± 0.08)	72.90 (± 0.08)	80.68 (± 0.06)	86.41 (± 0.07)	54.3 (± 0.4)
B35678/S13456	76.74 (± 0.09)	79.3 (± 0.1)	74.75 (± 0.08)	87.2 (± 0.1)	51.7 (± 0.4)
B01245678/S3456	77.80 (± 0.08)	75.10 (± 0.08)	79.65 (± 0.05)	86.37 (± 0.08)	45.7 (± 0.3)
B035678/S0123456	82.8 (± 0.1)	83.48 (± 0.09)	81.75 (± 0.07)	89.39 (± 0.08)	59.1 (± 0.6)
B135678/S03456	83.82 (± 0.09)	80.8 (± 0.2)	83.48 (± 0.05)	88.93 (± 0.07)	60.9 (± 0.3)

Source: Adapted from Miranda, Machicao and Bruno (2016a).

In addition, Fig. 60 presents the canonical analysis for all network models and the seven values of $\langle k \rangle$ rule *B01678/S0457* and $\bar{\mu}_S$ as feature vector. We observe that not all the distinct values of $\langle k \rangle$ can be discriminated in the plots of Fig. 60. Considering, for instance, the scale-free model, we can distinguish networks with mean degree 4, 6 and 8. However, the other classes are closer to each other. A similar behavior is observed for the geographical model. Moreover, random and small-world networks also have a set of $\langle k \rangle$ values that are also distinguishable from the others. As the average degree increases, the network topology will become highly connected. Consequently, the discrimination between the classes declines. These results are corroborated by measurements derived from the confusion matrix (Table 7). For instance, the values of the area under the curve (AUC) decreases for larger values of $\langle k \rangle$ and for all the network models. In Fig. 59, we can observe the confusion matrix for one realization of the cross-validation regarding rule *B01678/S0457*.

Table 7 – Performance measures corresponding to the best accuracy as shown in Table 6 for the following network models: “RAN” (random), “SW” (small-world), “SF” (scale-free) and “GEO” (geographical).

TPR	FPR	F-Measure	MCC	ROC Area	Class
1	0	0.996	0.996	1	RAN- $\langle k \rangle = 4$
1	0	1	1	1	RAN- $\langle k \rangle = 6$
0.983	0	0.991	0.991	1	RAN- $\langle k \rangle = 8$
0.963	0.002	0.949	0.948	0.998	RAN- $\langle k \rangle = 10$
0.868	0.006	0.858	0.853	0.993	RAN- $\langle k \rangle = 12$
0.743	0.01	0.737	0.727	0.987	RAN- $\langle k \rangle = 14$
0.81	0.005	0.83	0.824	0.995	RAN- $\langle k \rangle = 16$
1	0	1	1	1	SW- $\langle k \rangle = 4$
1	0	1	1	1	SW- $\langle k \rangle = 6$
0.988	0.002	0.971	0.97	0.999	SW- $\langle k \rangle = 8$
1	0	1	1	1	SW- $\langle k \rangle = 10$
0.95	0.001	0.966	0.965	0.999	SW- $\langle k \rangle = 12$
0.968	0	0.979	0.978	0.998	SW- $\langle k \rangle = 14$
0.993	0.001	0.983	0.982	0.999	SW- $\langle k \rangle = 16$
1	0	1	1	1	SF- $\langle k \rangle = 4$
0.985	0.001	0.981	0.981	0.999	SF- $\langle k \rangle = 6$
0.97	0.001	0.977	0.977	0.999	SF- $\langle k \rangle = 8$
0.94	0.002	0.94	0.938	0.998	SF- $\langle k \rangle = 10$
0.893	0.004	0.898	0.894	0.994	SF- $\langle k \rangle = 12$
0.705	0.018	0.647	0.635	0.985	SF- $\langle k \rangle = 14$
0.553	0.011	0.601	0.589	0.984	SF- $\langle k \rangle = 16$
0.993	0	0.996	0.996	1	GEO- $\langle k \rangle = 4$
1	0	1	1	1	GEO- $\langle k \rangle = 6$
0.975	0.001	0.976	0.975	0.999	GEO- $\langle k \rangle = 8$
0.883	0.004	0.881	0.877	0.994	GEO- $\langle k \rangle = 10$
0.75	0.009	0.75	0.741	0.988	GEO- $\langle k \rangle = 12$
0.715	0.012	0.698	0.686	0.985	GEO- $\langle k \rangle = 14$
0.815	0.005	0.836	0.83	0.995	GEO- $\langle k \rangle = 16$
0.9 $\pm$ 0.1	0.003 $\pm$ 0.004	0.9 $\pm$ 0.1	0.9 $\pm$ 0.1	0.996 $\pm$ 0.004	Avg $\pm$ sd

Source: Adapted from [Miranda, Machicao and Bruno \(2016a\)](#).

8.3.3 Classifying scale-free models

In the third experiment with synthetic networks, we evaluated LLNA for classifying different scale-free models. We generated scale-free networks with both linear and non-linear preferential attachment: $\alpha = 0.5, 1.0, 1.5, 2.0$ (Section 2.3.1.3). In addition, we also considered another set of scale-free networks generated according to the method proposed by Dorogovtsev & Mendes ([DOROGOVTSSEV; MENDES, 2002](#)). The *synthetic-scalefree-dataset* was used in this experiment. Similarly to the previous experiments, the performance of each feature vector is shown in Fig. 61-(a).

The feature vector composed of the combination of all the distributions also led to the highest accuracies for the *synthetic-scalefree-dataset*. The maximum accuracy was $98.3 \pm 0.2\%$ using rule *B0157/S457*. The performance of the Shannon entropy and the LZ distributions should also be mentioned. However, there is much variability in regard to the performances

Figure 60 – Canonical analysis for each network model given the distinct values of $\langle k \rangle$ as classes. We used rule $B01678/S0457$ and $\vec{\mu}_S$ as feature vector.

Source: Adapted from [Miranda, Machicao and Bruno \(2016a\)](#).

Figure 61 – Synthetic network characterization with LLNA. a) Accuracy (%) and standard deviation resulting from the classification of scale-free models, using five different feature vectors. Results are shown for the best four LLNA rules. b) The canonical analysis for rule $B0157/S457$ and $[\vec{\mu}_S, \vec{\mu}_W, \vec{\mu}_L]$ as feature vector.

Source: Adapted from [Miranda, Machicao and Bruno \(2016a\)](#).

8.3.4 Comparison with structural network measures

We also compared the classification accuracy obtained with LLNA rules to the one reached using structural network measures. Figure 62 and Table 10 present this comparison

Table 9 – Performance measures corresponding to the best accuracy of Table 8.

TPR	FPR	F-Measure	MCC	ROC Area	Class
0.98	0	0.99	0.987	0.994	$BA\alpha = 1.0$
1	0.003	0.995	0.994	0.999	DM
0.95	0.003	0.969	0.962	0.991	$BA\alpha = 2.0$
1	0.003	0.995	0.994	0.999	$BA\alpha = 0.5$
0.99	0.013	0.971	0.963	0.992	$BA\alpha = 1.5$
1.0 ± 0.0	0.004 ± 0.003	1.0 ± 0.0	1.0 ± 0.0	0.995 ± 0.003	Avg $\pm$ sd

Source: Adapted from [Miranda, Machicao and Bruno \(2016a\)](#).

for the three experiments with the synthetic networks. The column “All” corresponds to the feature vector composed by the combination of all the structural measures. In the case of the classification of the network models (Fig. 62-(a)), both approaches provided similar results reaching to 100% of accuracy. This is due to the fact that the networks used in this experiment were generated from classical theoretical models that have known properties which can usually be characterized by structural measures.

Figure 62 – Comparison between the accuracy obtained with LLNA and the accuracy obtained with structural network measures as feature vector: average degree ($\langle k \rangle$), average hierarchical degree of level 2 ($\langle k_2 \rangle$), average hierarchical degree of level 3 ($\langle k_3 \rangle$), average clustering coefficient ($\langle cc \rangle$), average path length ($\langle l \rangle$), and degree correlation (ρ_P). The LLNA accuracy is highlighted in yellow.

Source: Adapted from [Miranda, Machicao and Bruno \(2016a\)](#).

The performance of LLNA is better illustrated in the other experiments. Concerning the classification of the network models together with $\langle k \rangle$, there was an improvement in accuracy when using LLNA. The maximum accuracy obtained using all the structural measures as feature vector ($[\langle k_2 \rangle, \langle k_3 \rangle, \langle cc \rangle, \langle l \rangle, \rho_P]$) was $65.2 \pm 0.2\%$. Consequently, the improvement in accuracy using LLNA was $25.5 \pm 0.3\%$. Since we aimed at classifying the network model in combination with $\langle k \rangle$, we did not include $\langle k \rangle$ in the feature vector in this case. In the last experiment, concerning the classification of the scale-free models, LLNA once more provided a higher accuracy when compared to the accuracy obtained with the combination of structural measures ($96.20\% \pm 0.03$), yielding an improvement of $2.08 \pm 0.25\%$. Table 10 presents the results of the robustness analysis

for the structural measures.

Table 10 – Classification accuracy (%) given different values of ρ_N , and, also without noise ($\rho_N = 0$) using structural network measures for the different experiments carried out in this section. A comparison with LLNA is also shown.

Measure	Network model				Network model & $\langle k \rangle$	Scale-free models
	without noise	$\rho_N = 10\%$	$\rho_N = 20\%$	$\rho_N = 30\%$		
$\langle k \rangle$	26.2 (± 0.4)	26.0 (± 0.5)	26.1 (± 0.4)	26.2 (± 0.4)	—	—
$\langle k^2 \rangle$	43.4 (± 0.4)	43.5 (± 0.5)	43.5 (± 0.5)	44.2 (± 0.5)	49.1 (± 0.1)	90.6 (± 0.1)
$\langle k^3 \rangle$	44.3 (± 0.2)	44.0 (± 0.3)	43.5 (± 0.4)	43.1 (± 0.4)	51.83 (± 0.08)	86.7 (± 0.2)
$\langle cc \rangle$	76.11 (± 0.02)	75.64 (± 0.02)	75.92 (± 0.02)	75.79 (± 0.02)	28.8 (± 0.3)	76.3 (± 0.2)
$\langle l \rangle$	50.46 (± 0.05)	43.82 (± 0.04)	42.7 (± 0.2)	41.6 (± 0.3)	28.3 (± 0.3)	69.6 (± 0.2)
ρ_P	75.2 (± 0.2)	83.72 (± 0.03)	85.87 (± 0.03)	85.90 (± 0.03)	22.4 (± 0.2)	88.4 (± 0.3)
combined	100.0 (± 0.0)	100.0 (± 0.0)	100.0 (± 0.0)	100.0 (± 0.0)	65.2 (± 0.2)	96.20 (± 0.04)
<i>B135678/S03456</i>						<i>B01678/S0457</i>
LLNA	99.992 (± 0.002)	99.983 (± 0.004)	99.97 (± 0.01)	99.95 (± 0.01)	90.76 (± 0.07)	<i>B0157/S457</i> 98.3 (± 0.2)

Source: Adapted from [Miranda, Machicao and Bruno \(2016a\)](#).

8.4 Pattern Recognition in Real-World Applications

In this section we present how LLNA can be applied to real-world pattern recognition problems. For this task, we introduce four different classification problems.

8.4.1 Identifying organisms using metabolic networks

8.4.1.1 Context

Metabolic networks are among the examples of complex networks that helped to unveil the scale-free characteristic of many real-world systems. The model proposed by Barabási & Albert, for instance, incorporates growth and preferential attachment ([BARABÁSI; ALBERT, 1999](#)), which were not considered in previous models. Metabolic networks describe the chemical reactions of the metabolic pathways that rule the transformations between chemical compounds through the action of enzymes ([COSTA *et al.*, 2011b](#)). One of the most common ways of representing metabolic networks is through a directed graph on which the substrates are connected to each other and the links represent the metabolic reactions.

Starting from the work of Jeong *et al.* ([JEONG *et al.*, 2000](#)), one of the first works that addressed the scale-free characteristic of metabolic networks, they have been extensively studied within the context of the area of complex networks ([BARABASI; OLTVAI, 2004](#); [KANEHISA *et al.*, 2004](#)). In this work, the authors pointed out the power law degree distribution of metabolic networks of 43 organisms belonging to the three domains of life: archaea, bacteria and eukaryote. They found that the network diameter of all the analyzed organisms is very similar. This was explained by the fact that as the complexity of the organisms grows, some nodes get more

connections, so the relative diameter is kept constant (COSTA *et al.*, 2011b). These network are then more tolerant to random failures.

In spite of the similarity in relation to the topology of the metabolic networks of organisms from different evolutionary classes, the study of such networks has also been approached as a tool for the extraction of phylogenetic relationships (HEYMANS; SINGH, 2003; BANERJEE, 2012; FILHO; MACHICAO; BRUNO, 2018). In this way, it is possible to find patterns related to a specific evolutionary class. The aim of using LLNA is related to this approach. We aim at the characterization of the metabolic networks of distinct organisms grouped by evolutionary classes.

8.4.1.2 Classification

We investigated whether LLNA could be employed to distinguish the three classes of organisms. For this task, we use the same dataset which was previously constructed by Jeong *et al.* (JEONG *et al.*, 2000) and is publicly available (CCNR/ICENSA, 2018). This dataset was referred to as *metabolic-dataset* and contains 43 metabolic networks that describe metabolic pathways of three types of organisms: archaea, bacteria and eukaryote. The original database was yielded on the basis of the metabolic reactions found in the WIT database (OVERBEEK; AL., 2000). The corresponding networks were generated taking into account the educt-educt complexes and associated enzymes as representations of nodes and edges respectively.

The first plot of Fig. 63-(a) shows the distribution of the Shannon entropy ($\bar{\mu}_S$) for three networks belonging to each of the three classes of interest. These distributions were obtained from the spatio-temporal patterns evolved by LLNA on the respective metabolic network. Distinct distributions for the three classes are observed. For instance, for the “Eukaryote” class, there is a high frequency of low entropy values, corresponding to the presence of more stable and/or oscillating patterns in the respective spatio-temporal diagram. The separation of the “Eukaryote” class is also clear in Fig. 63-(b), which shows the canonical analysis for the *metabolic-dataset*. The vector $\bar{\mu}_S$ was also used to generate the canonical variables.

The LLNA performance regarding the identification of the metabolic networks is presented in Fig. 63-(c). Specifically for the *metabolic-dataset* we used a re-sampling strategy during the cross-validation, since the number of networks per class is not uniform. This procedure is explained in Section 8.2.3. Therefore, we guarantee a uniform distribution of classes, and we also allow that all the networks in the most frequent classes to be used during training and testing. The feature vector made up by the distribution of the Shannon entropy $\bar{\mu}_S$ provided the highest accuracy for the *metabolic-dataset*, $87 \pm 13\%$, when using rule B05/S13568. Note that another rule selection procedure was carried out for the *metabolic-dataset* in order to select the rules that perform better in the analyzed context. The 10 best rules are presented in Table 11. In addition, the performance measures of this experiment are presented in Table 12. As can be

observed in this table, the network descriptors provided by LLNA can completely distinguish the “Eukaryote” class from the others. The F-measure, the MCC and the AUC achieved 1.0 for this class. When comparing the accuracy of the best LLNA rule to the accuracy obtained using structural measures as feature vector (Fig. 63-(d)), we observe that LLNA provided an increase in the classification accuracy of $23 \pm 23\%$. The best accuracy using structural measures was $64 \pm 10\%$ for the clustering coefficient feature.

Figure 63 – Characterization of real-world applications with LLNA. (a) Shannon entropy distribution, $\bar{\mu}_S$, for three networks belonging to each of the three classes of interest (archaea, bacteria and eukaryotes). These distribution were obtained using rule B05/S13568. (b) The canonical analysis for the *metabolic-dataset*. (c) Accuracy (%) and standard deviation resulting from the classification of metabolic networks for the best four rules and for all feature vectors. (d) Accuracy and standard deviation obtained by using structural network measures as feature vectors. These results were compared to the best LLNA rule (yellow).

Source: Adapted from [Miranda, Machicao and Bruno \(2016a\)](#).

Jeong *et al.* (JEONG *et al.*, 2000) showed that even organisms belonging to different evolutionary classes have metabolic networks with similar a topology. All of them have power-law degree distributions, characterizing them as scale-free networks. In addition, we have shown here that PR algorithms can go a step further in terms of analyzing the network topology as they are able to find subtleties that can be used to distinguish networks within the same topological group, allowing the characterization of sub-categories of networks.

Table 11 – Accuracy (%) obtained using different feature vectors for the classification of the metabolic networks. The results are shown for the 10 rules obtained in the rule selection step.

Rules	$\bar{\mu}_S$	$\bar{\mu}_W$	$\bar{\mu}_L$	$[\bar{\mu}_S, \bar{\mu}_W, \bar{\mu}_L]$	$[\langle\mu_S\rangle, \langle\mu_W\rangle, \langle\mu_L\rangle]$
B05-S13568	87 (± 13)	61 (± 16)	52 (± 18)	77 (± 15)	63 (± 14)
B0346-S03	62 (± 15)	62 (± 7)	52 (± 16)	68 (± 15)	53 (± 12)
B02468-S7	64 (± 16)	33 (± 0)	61 (± 14)	67 (± 15)	54 (± 16)
B0578-S1356	74 (± 15)	60 (± 11)	57 (± 14)	62 (± 14)	65 (± 14)
B01237-S13	68 (± 14)	33 (± 0)	55 (± 16)	62 (± 16)	39 (± 16)
B01247-S2347	61 (± 16)	56 (± 10)	32 (± 15)	57 (± 18)	35 (± 15)
B3468-S12346	39 (± 16)	49 (± 12)	69 (± 16)	54 (± 15)	28 (± 16)
B2345678-S367	27 (± 15)	67 (± 2)	38 (± 17)	48 (± 17)	49 (± 16)
B23468-S45	40 (± 16)	37 (± 14)	42 (± 14)	44 (± 20)	33 (± 18)
B01678-S013578	40 (± 16)	31 (± 16)	33 (± 16)	37 (± 18)	71 (± 16)

Source: Adapted from [Miranda, Machicao and Bruno \(2016a\)](#).

Table 12 – Performance measures corresponding to the best accuracy as shown in Table 11

TPR	FPR	F-Measure	MCC	ROC-AUC	Class
1.0	0.000	1.0	1.0	1.0	<i>Eukaryote</i>
0.667	0.000	0.800	0.756	0.833	<i>Bacterium</i>
1.0	0.167	0.857	0.791	0.917	<i>Archae</i>
0.9 $\pm$ 0.1	0.06 $\pm$ 0.07	0.9 $\pm$ 0.1	0.8 $\pm$ 0.1	0.92 $\pm$ 0.06	Avg $\pm$ sd

Source: Adapted from [Miranda, Machicao and Bruno \(2016a\)](#).

8.4.2 Identifying structural patterns in social networks

8.4.2.1 Context

Social networks have been studied for many decades using different theoretical approaches. They can be characterized as complex systems and have been used to illustrate several properties of complex networks. Social networks can represent academic collaborations, citation networks, personal relations, friendship and communication networks, and so on ([BARABÁSI *et al.*, 2002](#); [GIRVAN](#); [NEWMAN, 2002](#); [NEKOVEE *et al.*, 2007](#)). In the application described in this section, we focus on online social networks, which provide the means for social interactions. Moreover, supported by the technological advances, they can store a huge amount of data. Some of them present tools for sharing and grouping people in communities of specific topics.

Different softwares for constructing social networks can bias the way people connect to each other, yielding in this way, specific structures in the network. We aim to investigate the topology of such networks from two platforms, Google+ and Twitter. We evaluated the LLNA framework regarding the correct identification of the software tool used to create the social network. For this task, we used networks from the SNAP database ([LESKOVEC; KREVL, 2014](#); [MCAULEY; LESKOVEC, 2012](#)) in order to distinguish networks from Google+ and Twitter. In this context, LLNA was used to analyze different structural properties of both types of networks, which correspond to the classes of this application.

8.4.2.2 Classification

We analyzed the performance of LLNA in distinguishing between the two software tools, Google+ and Twitter using the *social-dataset*. Figures 64-(a) and (b) illustrate the differences regarding the spatio-temporal dynamics of each social network. Figure 64-(a) shows the distribution of the LZ complexity, $\vec{\mu}_L$ for two networks of the *social-dataset*. We can observe that $\vec{\mu}_L$ has a different behavior for both classes. While the Twitter distribution is concentrated between 0.6 and 0.9, the Google+ distribution has values distributed across the entire LZ range with a few peaks at the left. The separation between both classes is clear in Fig. 64-(b), which presents the canonical analysis for the *social-dataset*.

Figure 64-(c) presents the accuracies obtained using LLNA and its different feature vectors. The best accuracy was obtained using the distribution of $\vec{\mu}_L$ as feature vector, $92 \pm 1\%$, and, rule B0167/S248 as LLNA transition function. In addition, the feature vector $\vec{\mu}_S$ also had a good performance for the same rule. The rule-selection procedure was also performed for the *social-dataset*. We can observe that the selected rules in this application are different from the ones selected for the other PR applications presented in this chapter (e.g. Table 3 and 11). Tables 13 and 14 present the accuracy values for all the 10 selected rules and the performance measures, respectively.

Table 13 – Accuracy (%) obtained using different feature vectors for the classification of social networks. The results are shown for the 10 rules obtained in the rule selection step.

Rules	$\vec{\mu}_S$	$\vec{\mu}_W$	$\vec{\mu}_L$	$[\vec{\mu}_S, \vec{\mu}_W, \vec{\mu}_L]$	$[\langle \mu_S \rangle, \langle \mu_W \rangle, \langle \mu_L \rangle]$
B0167-S248	92 (± 1)	88 (± 1)	92 (± 1)	89 (± 1)	83 (± 2)
B05-S148	76 (± 1)	82 (± 2)	79 (± 2)	89 (± 1)	81 (± 1)
B06-S148	79 (± 2)	79 (± 2)	75 (± 2)	78 (± 2)	81 (± 2)
B256-S3467	89 (± 1)	78 (± 2)	87 (± 1)	80 (± 1)	80 (± 1)
B05-S1458	84 (± 1)	80 (± 1)	81 (± 2)	82 (± 1)	80 (± 2)
B1236-S245	82 (± 2)	81 (± 1)	84 (± 1)	86 (± 1)	79 (± 1)
B456-S01237	80 (± 2)	65 (± 2)	78 (± 2)	82 (± 2)	79 (± 1)
B01267-S035678	78 (± 2)	75 (± 2)	69 (± 2)	70 (± 2)	79 (± 2)
B012368-S01245	85 (± 1)	82 (± 1)	76 (± 1)	78 (± 1)	79 (± 2)
B08-S1467	84 (± 1)	78 (± 3)	78 (± 2)	75 (± 1)	79 (± 2)

Source: Adapted from Miranda, Machicao and Bruno (2016a).

Table 14 – Performance measures corresponding to the best accuracy as shown in Table 13

TPR	FPR	F-Measure	MCC	ROC-AUC	Class
0.900	0.040	0.928	0.862	0.930	Google+
0.960	0.100	0.932	0.862	0.930	Twitter
0.93 $\pm$ 0.03	0.070 $\pm$ 0.03	0.930 $\pm$ 0.002	0.8 $\pm$ 0.0	0.9 $\pm$ 0.0	Avg $\pm$ sd

Source: Adapted from Miranda, Machicao and Bruno (2016a).

When comparing to the classification performance obtained when using structural measures as feature vectors (Fig. 64-(d)), LLNA also had a better performance. The maximum

Figure 64 – Characterization of real-world applications with LLNA. (a) LZ distribution, $\vec{\mu}_L$, for two networks belonging to the classes of interest (Twitter and Google+). These distribution were obtained using rule B0167/S248. (b) The canonical analysis for the *social-dataset*. (c) Accuracy (%) and standard deviation resulting from the classification of social networks for the best four rules and for all feature vectors. (d) Accuracy and standard deviation obtained by using structural network measures as feature vectors. These results were compared to the best LLNA rule (yellow).

Source: Adapted from [Miranda, Machicao and Bruno \(2016a\)](#).

accuracy obtained using structural measures was $88 \pm 2\%$. Therefore, there was an increase of $4 \pm 3\%$ in the classification rate for the *social-dataset*. The two-class problem of distinguishing Twitter and Google+ shows that the tools provided by each of them may influence the way people connect to each other, resulting in structural differences between both social networks. Still, preferential connection of nodes and presence of hubs may exist in both of them.

8.4.2.3 Comparison between BSNA and GenBSNA

We performed the same comparison as carried out in Section 6.3.3 for the *social-dataset*. We compared the classification accuracies for both families of NAs, the BSNA and the GenBSNA. Recall that the LLNA belongs to the GenBSNA family and has $R = 9$. Fig. 65 presents the results for the best rules and for each R value for both BSNA and GenBSNA. Again, as in the case of the classification of the network models (Fig. 37), we observe that as R increases, the classification accuracy tends to increase. However, for the *social-dataset*, the difference between the accuracy provided by both NAs is larger than in the case of the classification of the network models.

This example also illustrates the trade-off between the size of the rule space and the

Figure 65 – Comparison between BSNA and GenBSNA concerning the classification accuracy of the *social-dataset* for different values of the parameter R . The accuracy shown in the barplots corresponds to the highest accuracy achieved by a rule within the rule space defined by R .

Source: Elaborated by the author.

classification accuracy in PR applications. Although a huge rule space may not always guarantee a better solution, it is important that every single rule is evaluated regarding its performance. In addition, the computing time per dataset should also be considered when choosing an NA for a specific application.

8.4.3 Classifying stomata distribution patterns

8.4.3.1 Context

The stomata distribution in the leaves of plants represents the so-called phenotypic plasticity, which is the plant's ability to adapt to different environmental conditions, such as light, temperature, availability of nutrients, among others. We employed LLNA for the characterization of the stomata distribution patterns of the species *Tradescantia zebrina*. This plant was exposed to different lighting conditions. For this purpose we used an existing image dataset ([FLORINDO et al., 2015](#)), which consists of six images for each light condition: sunlight (natural), 4 hours (L4) and 24 hours (L24) of artificial light. For representing the stomata using a network, each stoma was segmented from the leaf image and its centroids were assessed. For each image, a network was constructed, such that the segmented stomata correspond to the network nodes and the links between them are given according to a distance threshold, δ_T . As δ_T increases, more connections will be established between the centroids, and, therefore, the density of the network will be higher. This can be observed in Figure 66. In the first row, three microscopic images of the three distinct classes are shown. The next rows present the corresponding networks given different values of δ_T . This approach for representing stomata images as networks was adapted from ([BACKES; CASANOVA; BRUNO, 2009a](#)). The main characteristic of this method is the concatenation of the network descriptors obtained at each value of δ_T . We used 16 threshold values with δ_T starting with 0.25 and incremented by 0.0625 until reaching a final threshold of

$\delta_T = 1$.

Figure 66 – Stomata distributions of the *Tradescantia zebrina* leaves and the corresponding networks for the three light conditions and four δ_T values. Each stoma in the microscopic image is represented as a network node and δ_T is a distance threshold for the connection of the nodes.

Source: Elaborated by the author.

8.4.3.2 Classification

Fig. 67 presents the results by applying LLNA to the *stomata-dataset*, which contains networks of three analyzed classes. We obtained the LLNA descriptors for the networks at each threshold δ_T . The bar-plot of Fig. 67-(a) shows the average values of the Shannon entropy $\langle \mu_S \rangle$

at each threshold δ_T for the different light conditions. The separation among the three classes is also highlighted in the canonical analysis shown in Fig. 67-(b). We can see that the class “L24” is linearly separable from the others.

Figure 67-(c) presents the classification results for the *stomata-dataset*. The highest accuracy achieved was $90 \pm 6\%$, using rule B12345/S04568 and the feature vector composed by the average values of the spatio-temporal measures, $[\langle \mu_S \rangle, \langle \mu_W \rangle, \langle \mu_L \rangle]$. Tables 15 and 16 present the classification accuracy for the ten selected rules and the classification performance, respectively. We can observe that the obtained standard deviation is higher for this dataset. This is due to the smaller number of instances for each class. Moreover, there is also a higher heterogeneity regarding the accuracy obtained for the analyzed feature vectors given the same rule. Finally, when compared to structural measures (Fig. 67-(d)), we can observe that the best classification accuracy was $83 \pm 4\%$. Therefore, LLNA provided an increase of $7 \pm 9\%$ in the classification rate.

Figure 67 – Characterization of real-world applications with LLNA. (a) Average Shannon entropy $\langle \mu \rangle$ for different threshold values δ_T , which correspond to the construction of the stomata network. Each plot corresponds to one of the three classes of interest of the *stomata-dataset*. These measure were obtained using rule B12345/S04568. (b) Canonical analysis for the *stomata-dataset*. (c) Accuracy (%) and standard deviation resulting from the classification of the stomata networks for the best four rules and for all feature vectors. (d) Accuracy and standard deviation obtained using structural network measures as feature vectors. These results were compared to the best LLNA rule (yellow).

Source: Adapted from [Miranda, Machicao and Bruno \(2016a\)](#).

The plant plasticity for different lighting conditions is reflected by the network of connections between the stomata centroids. In this case, the proposed method could represent

Table 15 – Accuracy (%) obtained using different feature vectors for the classification of the stomata networks. The results are shown for the 10 rules obtained in the rule selection step.

Rules	$\bar{\mu}_S$	$\bar{\mu}_W$	$\bar{\mu}_L$	$[\bar{\mu}_S, \bar{\mu}_W, \bar{\mu}_L]$	$[\langle\mu_S\rangle, \langle\mu_W\rangle, \langle\mu_L\rangle]$
B12345-S04568	87 (± 7)	58 (± 11)	76 (± 8)	88 (± 7)	90 (± 6)
B01346-S0257	37 (± 9)	90 (± 6)	65 (± 8)	83 (± 6)	54 (± 9)
B3567-S0468	64 (± 11)	44 (± 10)	88 (± 8)	78 (± 9)	69 (± 14)
B0568-S1578	88 (± 7)	56 (± 10)	47 (± 9)	71 (± 7)	61 (± 10)
B125678-S012345	68 (± 13)	54 (± 9)	87 (± 7)	66 (± 7)	84 (± 9)
B34-S024568	88 (± 7)	40 (± 7)	84 (± 10)	85 (± 7)	68 (± 9)
B35678-S012578	60 (± 9)	53 (± 10)	75 (± 9)	68 (± 10)	86 (± 8)
B02345-S01456	70 (± 9)	84 (± 8)	75 (± 8)	81 (± 9)	65 (± 8)
B056-S0348	66 (± 6)	85 (± 9)	66 (± 10)	67 (± 7)	63 (± 10)
B245-S0234568	77 (± 8)	34 (± 7)	87 (± 7)	78 (± 9)	63 (± 9)

Source: Adapted from [Miranda, Machicao and Bruno \(2016a\)](#).

Table 16 – Performance measures corresponding to the best accuracy as shown in Table 15

TPR	FPR	F-Measure	MCC	ROC-AUC	Class
1.000	0.125	0.889	0.837	0.938	Natural
0.750	0.000	0.857	0.816	0.875	L4
1.000	0.000	1.000	1.000	1.000	L24
0.9 $\pm$ 0.1	0.04 $\pm$ 0.06	0.9 $\pm$ 0.1	0.9 $\pm$ 0.1	0.94 $\pm$ 0.04	Avg $\pm$ sd

Source: Adapted from [Miranda, Machicao and Bruno \(2016a\)](#).

the specific characteristics of the three classes of the *stomata-dataset*.

8.4.4 Identifying fungal growth patterns

8.4.4.1 Context

Fungal mycelia are structures that can be directly compared to networks. They are responsible for an efficient transport of nutrients that can be carried out over large distances ([BODDY *et al.*, 1999](#)). Different fungal species have optimal growth conditions that are important for different industrial processes that involve fungi, e.g., production of cheese, antibiotics, biofuel, among others. Two of the most important environmental conditions for the growth dynamics are temperature (T) and relative humidity (RH). In the present application, we aim at the study of fungal growth dynamics by focusing on the structure of networks representing the fungal mycelia. For this purpose, we are using a framework developed by [Ulzurrun *et al.* \(2015\)](#). This framework consists of an automated image analysis method in order to obtain segmented mycelia and, consequently, the network structure. This method addresses many disadvantages of related methods in the analysis of fungal growth dynamics. Therefore, we focus on the investigation of the topological properties related to fungal growth dynamics.

The image analysis framework ([ULZURRUN *et al.*, 2015](#)) uses a line-detection algorithm in order to obtain binary ridge maps of the growing fungi, and then, these maps are converted

into graphs, whose nodes are represented by the growing tips and junctions of the fungus hyphae, and, the edges are represented by the hyphal segments connecting them. The dataset used in this work consists of time series representing the fungal growth, where for each time step, a graph is obtained that represents the topological characteristics of each fungus and for each environmental condition. The *fungi-dataset* is described in detail in Section 8.2.1. This dataset consists of time series of images of two different fungal species, *Rhizoctonia solani* (RS) and *Coniophora puteana* (CP), which are classified as filamentous fungi and are two economically important species. There are a total of 16 combinations of environmental conditions that were analyzed considering temperature (15, 20, 25 and 30 °C) and RH (65, 70, 75 and 80 %). Fig. 68 depicts the fungal growth through a set of three images belonging to a time series of the RS species, given RH= 65% and T=15° for different time frames.

Figure 68 – Fungal growth dynamics shown for the *Rhizoctonia solani* species given RH= 65% and T=15°. The images depicts the following time steps: (a) 46, (b) 78, and, (C) 211.

Source: Adapted from [Ulzurrun et al. \(2015\)](#), [De LIGNE et al. \(2019\)](#).

8.4.4.2 Classification

8.4.4.2.1 Connectivity analysis

As a first step, we investigated whether a set of network measures would be able to characterize the growth dynamics of both species for each environmental condition. This approach is inspired by our previous paper ([MIRANDA; MACHICAO; BRUNO, 2018](#)) in which

we evaluated a set of network measures in the pattern recognition context. Figures 69 and 70 present the average degree, the average hierarchical degree of level 2 ($\langle k_2 \rangle$) and the average hierarchical degree of level 3 ($\langle k_3 \rangle$) calculated for each time step and for each species, CP and RS. We can observe that $\langle k_3 \rangle$ is the most discriminating attribute between both species. This was observed for all environmental conditions, although for some of them, e.g. RH= 65% and T= 30°C, only for some time steps there was a clear separation between both species.

8.4.4.2.2 Classification using LLNA

We also evaluated the classification accuracy using LLNA in three scenarios:

1. Species identification irrespective of the environmental conditions;
2. Species identification given temperature and irrespective of RH;
3. Species identification given RH and irrespective of temperature;

Therefore, we defined three different classification tasks. Firstly, we performed a sampling of the time series for each fungi and for each environmental condition. This sampling was performed by choosing a subset of images corresponding to an equally spaced time interval. Since there are 260 images available per time series, we selected a time interval $\delta_t = 10$, starting from the 10th image of the series. In the end, we used 26 images per fungal species and per environmental condition, and, all the four replicates of the selected images. This sampling scheme is represented in Fig. 71-(a). Then, for each image we obtained the corresponding morphological graph by applying the procedure previously described.

In Fig. 71-(b), we present the three classification tasks of interest. For the first classification task we aimed at a correct identification of the fungi species, independently of the environmental condition, i.e., we were interested in investigating whether it is possible to identify the selected species using network descriptors. In the second task, we aimed at the species identification given temperature. In this scenario we have 8 classes in total, which were obtained by combining each species with each studied temperature. Finally, for the third task we investigated the species identification given RH. Similarly to the second classification task, there are also 8 classes in this context.

Regarding the LLNA attributes we used the Shannon entropy distribution, $\bar{\mu}_S$, obtained for the i -th network in the time series. Specifically for the *fungi-dataset* we first calculated the descriptor μ_{S_i} for each of the 26 samples of one time series and then concatenated them in order to obtain a final descriptor to represent it. Concerning the structural network descriptors we concatenated in one feature vector the average degree, the average hierarchical degree of level 2 ($\langle k_2 \rangle$) and the average hierarchical degree of level 3 ($\langle k_3 \rangle$) in the same way we concatenated the LLNA descriptors for the 26 selected graphs per time series and per replicate.

Figure 69 – Average degree, average hierarchical degree of level 2, $\langle k_2 \rangle$, and, average hierarchical degree of level 3, $\langle k_3 \rangle$, calculated for each time step of the time series and for each species using the following environmental conditions: RH= {65%, 70%} and T= {15, 20, 25, 30°C}.

Source: Elaborated by the author.

The classification accuracy is shown in Table 17. We can observe that for all tasks the classification accuracy was highest when using LLNA. We used cross-validation as the training

Figure 70 – Average degree, average hierarchical degree of level 2, $\langle k_2 \rangle$, and, average hierarchical degree of level 3, $\langle k_3 \rangle$, calculated for each time step of the time series and for each species using the following environmental conditions: RH= {75%,80%} and T= {15,20,25,30°C}.

Source: Elaborated by the author.

strategy in all experiments. Identifying the species independently of the environmental conditions was the task for which we obtained the highest accuracy. This was expected given the fact that

Figure 71 – Sampling of the *fungi-dataset* and the corresponding classification tasks. a) Sampling images from time series. This sampling was performed per fungal species and per environmental conditions. b) Classification tasks concerning the three different scenarios for species identification, given temperature (T) and relative humidity (RH).

Source: Elaborated by the author.

combining the environmental conditions with species identification increases the complexity of the task.

Table 17 – Classification accuracy for the three classifications tasks using structural network descriptors and LLNA.

Classification Task	Structural Descriptors	LLNA
1	71.37 ($\pm$ 2.92)	88.12 ($\pm$ 3.79)
2	11.69 ($\pm$ 2.43)	40.75 ($\pm$ 2.27)
3	25.75 ($\pm$ 2.32)	39.19 ($\pm$ 3.64)

The obtained results are also in accordance with the results obtained by De Ligne *et al.* (De LIGNE *et al.*, 2019). The authors analyzed the combined effect of temperature and relative humidity in the growth dynamics of the two species, CP and RS. The same 16 combinations of temperature and RH were analyzed in this work. The authors found that for some combinations of the environmental conditions the differences between the two species were more pronounced than for others.

Finally, Tables 18 and 19 summarize the performance of all the classification tasks presented in this chapter for both the synthetic datasets and the real-world applications, respectively. The tables also show the configurations of each experiment in terms of the number of networks used for rule selection and classification, and, the classes of each experiment.

Table 18 – Summary of the performance obtained for each synthetic dataset.

		Rule Selection	Classification	LLNA performance		Performance – structural measures
Datasets	Classes	# networks	# networks	Accuracy(%)	Selected rule	Accuracy(%)
synthetic	Random, Small-world, Scale-free, Geographical	1400	11200	99.99 ± 0.002	B135678/ S03456	100 ± 0
	{Random, Small-world, Scale-free, Geographical} * {4,6,8,10,12,14,16}			90.76 ± 0.07	B01678/ S0457	65.2 ± 0.2
scale-free	$\alpha = 0.5, 1.0, 1.5, 2.0$	250	500	98.3 ± 0.2	B0157/ S457	96.2 ± 0.04

Table 19 – Summary of the performance obtained for each real-world application.

		Rule Selection	Selection	LLNA performance		Performance - structural measures
Datasets	Classes	# networks	# networks	Accuracy(%)	Selected rule	Accuracy(%)
metabolic	Archaea, Bacterium and Eukaryote	6	37	87 ± 13	B05/ S13568	64 ± 10
social	Twitter, Google+	30	100	92 ± 1	B0167/ S248	88 ± 2
stomata	Natural, 4h, 24h	6	12	90 ± 6	B12345/ S04568	83 ± 4
fungi	Species: CP, RS	$104^\dagger$	$104^\dagger$	88 ± 4	B2478/ S148	71 ± 3
		† per replicate (26 networks * 2 replicates * 2 species)				

8.5 Conclusion

In this section we have presented real-world applications using LLNA, within the PR context, for which the datasets are represented as networks. We have demonstrated the importance of the rule selection step, since for different datasets the performance of the rules may differ. Therefore, the selected rules must be validated for each application. This issue can be explored in future studies by using optimization algorithms in order to reduce the time taken for the training phase. The proposed method outperformed a method using structural measures for the classification of both synthetic and real-world networks, showing that LLNA is a good choice for PR in networks. Therefore, potentially any PR application whose data is represented as a network can consider LLNA.

STRUCTURAL NETWORK DESCRIPTORS IN COMPUTER VISION

9.1 Introduction

In this chapter, we introduce a method for shape characterization in computer vision which is based on the network defined over the pixels on the shape contour. In addition, we present some PR tasks that make use of the proposed approach. This method was already published in (MIRANDA; MACHICAO; BRUNO, 2018).

Shape boundaries are one of the main attributes that are commonly used in Computer Vision (CV) for the characterization of objects. PR applications based on shape analysis cover different areas, such as medical imaging (RANGAYYAN; MUDIGONDA; DESAUTELS, 2000; HEIMANN; MEINZER, 2009; SELLE *et al.*, 2002; TSAI *et al.*, 2003), neuroscience (GERIG *et al.*, 2001; TORRES; FALCAO; COSTA, 2004; FAN *et al.*, 2007), agriculture (AITKENHEAD *et al.*, 2003; NETO *et al.*, 2006; BROSNAN; SUN, 2002; COSTA *et al.*, 2011a; HEMMING; RATH, 2001; PLOTZE *et al.*, 2005), remote sensing (BENZ *et al.*, 2004; BLASCHKE, 2010), among many others. Methods proposed in literature are based on classical approaches, such as, Fourier descriptors (OSOWSKI *et al.*, 2002; ZHANG; LU *et al.*, 2001), wavelets (DAVATZIKOS; TAO; SHEN, 2003; SOARES *et al.*, 2006), fractal dimension (PLOTZE *et al.*, 2005; BACKES; BRUNO, 2010; BACKES; FLORINDO; BRUNO, 2012; BRUNO *et al.*, 2008), curvature scale space (CSS) (MOKHTARIAN; BOBER, 2013), to mention but a few. More recently, another class of methods has drawn attention for performing shape classification. These methods are based on the structural properties of shape boundaries and are strongly influenced by graph theory and network analysis. (BACKES; BRUNO, 2010; BACKES; CASANOVA; BRUNO, 2009a; BACKES; BRUNO, 2013; WU; LU; DENG, 2015; SUI *et al.*, 2016; SCABINI *et al.*, 2017). Therefore, they take into account the topological properties of the network defined over the shape contour, instead of just considering the shape as a chain of connected points. In addition, such

methods are invariant under rotation and can also be scale invariant (BACKES; CASANOVA; BRUNO, 2009a).

Backes *et al.* proposed a method, named CNDescriptor, for shape analysis that is based on the connectivity among pixels of the shape contour. In a paper published in 2009 (BACKES; CASANOVA; BRUNO, 2009a), the authors present this one-parameter model, in which each pixel corresponds to a vertex of a graph and the connections are defined according to that parameter, which represents a distance threshold T . Given a T value, there exists one configuration of this graph and a set of structural measures can be obtained from it. Therefore, the shape can now be represented by a feature vector corresponding to network descriptors that can be used to unveil the topological properties of the shape boundary. In the cited paper, the authors also explore some properties of the graphs that are obtained with proposed method. For instance, for some T values, the obtained graphs have a high clustering coefficient, which characterizes them as small-world. Moreover, in addition, the CNDescriptor is invariant under rotation and scale and has shown to be robust regarding the presence of noise. The authors also evaluated the CNDescriptor in the context of image degradation by removing parts of the contour. Promising results were obtained and an extension of this approach for texture analysis was proposed (BACKES; CASANOVA; BRUNO, 2013).

However, although the CNDescriptor has been well suited for shape analysis tasks, it only accounts for degree-related measures. However, a much more robust set of measures could improve the classification performance regarding shape analysis applications. In a survey of 2007, Costa *et al.* emphasize how the measures extracted from the network structure can be used to describe the network topology, and then, be used for network characterization (COSTA *et al.*, 2007). The authors list many measures, such as connectivity related measures (mean degree and degree distributions and correlations), distances and path lengths as well as hierarchical and spectral measurements, among other categories. In addition, the classical pattern recognition approach using such measures is also addressed for network analysis (COSTA *et al.*, 2010).

In this chapter we present an extension of the work of Backes *et al.* (BACKES; CASANOVA; BRUNO, 2009a) by proposing a generalized approach for shape characterization in computer vision. This approach is based on network analysis and accounts for different categories of measures. In addition, we demonstrate how these measures are related to the shape properties by using an interpolation method between regular shapes. This procedure allowed us to analyze how the internal angles of shapes influence the final network descriptor. Moreover, we also compared the curvature signal of such shapes with structural network measures. We observed, for instance, that pixels with high curvature have a high clustering coefficient, however, these pixels are not the ones with the highest degrees. Regarding the PR context, we evaluated the performance of the proposed approach in three shape recognition applications: (i) classification of geometric shapes, (ii) classification of generic shape contours, e.g., fishes, animals, tools, among others, and, (iii) classification of plant species using leaf contours. This last application makes use of

the same dataset yielded by Backes *et al.*, allowing us to compare the obtained results with the CNDescriptor.

9.2 Modeling Images as Networks

The CNDescriptor is based on the dynamic evolution of a network built from the contour points (BACKES; CASANOVA; BRUNO, 2009a) and controlled by the distance threshold T . Let S be the contour of an image, such that $S = \{s_1, s_2, \dots, s_N\}$, where $s_i = [x_i, y_i]$ are the coordinates of point i , represented by integer values. Given a graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$, each point i of the contour corresponds to a node in $\mathcal{G}$, and, therefore, $S = V$. The set $\mathcal{E}$ is non-directed and defines the corresponding weight, given by the Euclidean distance, for each pair of nodes, as follows:

$$d(s_i, s_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}. \quad (9.1)$$

The adjacency matrix W of this weighted network with N nodes is an $N \times N$ matrix, such that $w_{ij} = W([w_i, w_j]) = d(s_i, s_j)$, which is then normalized to the interval $[0, 1]$: $W = \frac{W}{\max_{w_{ij} \in W}}$.

Initially, each node is connected to all the others, therefore $\mathcal{G}$ is a regular graph. Then, a threshold transformation, T_l , is applied for obtaining a new set of edges $\mathcal{E}'$. This set consists of edges whose weights are *smaller than* T_l . When T_l is small, the number of edges is also small and the resulting network will contain many isolated connected components. On the contrary, when T_l is large, the network will be almost fully connected. For intermediate values of T_l , other properties are observed, such as the small-world property, characterized by the presence of many connected triples. This property was demonstrated for different shapes (BACKES; CASANOVA; BRUNO, 2009a). Consequently, T_l controls the connectivity of the network. Additionally, the CNDescriptor can also be implemented by taking into account the connections that are *greater than* T_l . Therefore, in this case, two points will be connected if the distance between them is greater than T_l .

The shape descriptor is obtained through the transformations defined by T_l . Connectivity measures are extracted from the network topology for every T_l . The evolution of the network as a function of T_l provides distinct attributes that are then combined in a feature vector. T_l is formally defined by $\delta_{T_l}(W)$ operation as follows:

$$A_{T_l} = \delta_{T_l}(W) = \forall_{W \in W} \begin{cases} a_{ij} = 0, & \text{if } w_{ij} \geq T_l \\ a_{ij} = 1, & \text{if } w_{ij} < T_l \end{cases}, \quad (9.2)$$

where A_{T_l} is an unweighted and thresholded matrix. The series of $\delta_{T_l}(W)$ transformations is then used to perform shape characterization where T_l is step by step incremented by T_{inc} . Figure 72 illustrates this transformation. In the left part of this figure, we have the networks generated by

using the *smaller than* comparison. We observe that as T_l increases more connections are created. Similarly, in the right part of the figure, we can see that the networks generated by using the *greater than* comparison, denoted as $\Delta_{T_l}(W)$. In this case, as T_l increases, less connections are established.

Figure 72 – Construction of a network from the shape contour. In the left-side the connections are established for different threshold values by using the *smaller than* comparison, δ_{T_l} . The right-side illustrates the comparison *greater than*, Δ_{T_l} .

Source: Adapted from [Miranda, Machicao and Bruno \(2018\)](#).

The measures calculated at each T_l value are: the *average degree*, denoted here as k_μ and the *max degree*, (k_κ). The final feature vector, φ , is given by

$$\varphi = [k_\mu(T_0), k_\kappa(T_0), k_\mu(T_1), k_\kappa(T_1), \dots, k_\mu(T_Q), k_\kappa(T_Q)], \quad (9.3)$$

where k_μ and k_κ are the *average degree* and *maximum degree* for each $\delta_{T_l}(W)$.

In addition to average and maximum degrees, we propose a more generalized set of measures that can be calculated from the network constructed from the contour points. This was motivated by the fact that different sets of measures are capable of describing distinct structural properties of shapes, and, thus improve the classification performance in PR applications.

9.3 Shape Properties & Structural Measurements

We studied the relationship between the properties of shape boundaries and structural measures. This relationship is incorporated in the dynamic evolution of the underlying network given $\delta_{T_l}(W)$, and, such network is directly influenced by the shape contour. In addition to degree-related measures, we went a step further and analyzed a more general set of structural measures comprising clustering, distance and centrality measures. Therefore, our approach generalizes the method proposed by Backes *et al.* regarding the features that can be obtained

from the network evolution. We will denote this generalized feature vector as Φ , in contrast to the CNDescriptor φ (Eq. (9.3)).

We started by investigating regular geometric shapes, since their properties, like angle and curvature, are easier to assess. For instance, when considering a network built from a shape contour, we observe that the degree of a node also depends on the value of T_l . Consequently, if the internal angles of a shape are small, then the distances between the contour pixels are also small, increasing the value of the average degree. In the next sections we illustrate how structural measures are related to the curvature signal. In addition, we perform an interpolation analysis for a set of shapes.

9.3.1 Curvature Analysis

Curvature is an important tool for image analysis and can be used, for instance, in shape recognition. As discussed in the famous work of Attneave (ATTNEAVE, 1954), the human visual system is able to recognize a given shape only by taking into account the points that have high curvature. We calculated the curvature of shape contours of a set of shapes depicted in Fig. 73. In the method used for this task, the curvature signal is estimated from the Discrete Fourier Transform (DFT) (CESAR; COSTA, 1996). Firstly, the shape contour is represented as a two-dimensional signal, that is linearized by taking the x and y coordinates. Then, the DFT of x and y and their corresponding spectra are used in order to estimate the curvature as follows: starting with a regular curve, C , parameterized as $C(t) = (x(t), y(t))$, where t is defined over an interval Γ of the real line $\mathcal{R}$, we have that the curvature of a single point is given by

$$k(t_0) = \frac{\dot{x}\ddot{y} - \ddot{x}\dot{y}}{[(\dot{x})^2 + (\dot{y})^2]^{3/2}}, \quad (9.4)$$

where the first and the second derivatives with respect to t are evaluated at $t = t_0$. The Fourier Transform is used to derive expressions of $\dot{x}$, $\ddot{x}$, $\dot{y}$ and $\ddot{y}$ (CESAR; COSTA, 1996).

Figure 73 shows the comparison between the curvature signal C (dashed line) and the degree of the nodes. Each row of this figure corresponds to a given shape and each column to different values of T_l , which are in increasing order from left to right. Recalling that each pixel belonging to the shape contour corresponds to one node in the underlying network, and, the measures accounted in the plots of Fig. 73 were calculated for each node (x -axis). Therefore, the x -axis of each plot corresponds to the pixels of the shape contour. In addition, the measures were normalized between 0 and 1.

The starting point for the calculation of the curvature is illustrated in Fig. 73 as a green dot for each shape. Mainly for regular shapes, the aliasing effect is more pronounced and reflects very small oscillations in the obtained curvature, like the hexagon. In spite of that, points with high curvature can be easily identified in the plots. As expected, these points are closer to the

Figure 73 – Comparison between curvature signal (dashed line) and the degree of each node as a function of T_l for different shapes.

Source: Adapted from [Miranda, Machicao and Bruno \(2018\)](#).

corners of the analyzed shapes. Consider, for instance, the triangular shape then, there exist three equally spaced peaks, corresponding to the three corners. When analyzing the curvature together with the degree of each node, we observe that points that have the highest curvature also have the lowest degrees. In the case of regular shapes, this is due to the fact that the points in the corners are distant from the points located on the sides of the shape (between the corners). Consequently, they have less neighbors. However, as T_l increases the degree distribution tends to be more uniform. When T_l is very small, the points in the corner are the ones with the highest

degrees since the network presents only a few connections. In the case of the star shapes, we observe an oscillating behavior due to the alternations between the internal angles. Finally, in the last two rows we see the comparison between curvature and degree for irregular shapes. The general assumptions made for the regular shapes are also valid here.

We performed a similar comparison for other structural measures and the corresponding results are shown in Fig. 74 for the triangular shape. The first row of this figure accounts for the clustering coefficient (cc). In this case, the pixels with the highest curvature are also the pictures with the highest cc . This is due to the fact that the pixels belonging to the corners are nearer to each other, and, as such, the connections established in this region are usually connected triples, increasing the cc value. The connected triples appear less frequently in the sides of the triangle. In general, as T_l increases, the cc also increases, what reflects the density of the network connections along the network evolution. Within this context, if a pixel i is connected to a pixel $i - 1$, given a radius r , then pixel $i + 1$ will also be connected to pixel $i - 1$. For all the analyzed shapes the cc was always greater than 0.5.

Figure 74 – Comparison between curvature signal (dashed line) and other structural measures: clustering coefficient (cc), betweenness (b), hierarchical degree of level 2 (k_2) and hierarchical degree of level 3 (k_3) (from top to bottom), as a function of the threshold parameter T_l for the triangular shape.

Source: Adapted from [Miranda, Machicao and Bruno \(2018\)](#).

The betweenness centrality b was also explored here (second row of Fig. 74). We can observe that similarly to the degree, the betweenness of the nodes increases as the curvature decreases, meaning that pixels with high curvature are not on the shortest paths. This is due to the fact that, as T_l increases, the network gets highly connected and new pathways are created among the pixels, which not always include the pixels in the corners. Finally, the last two rows of Fig. 74 correspond to the hierarchical degree of levels two and three, respectively. Again, we

observe that the general behavior of both measures follows the same trend obtained for the node degree (Fig. 73).

In addition, Figs. 73 and 74 illustrate the dynamical evolution of the underlying networks. Depending on the values of T_l , the network will be sparsely, highly connected or will present intermediate behavior. The plots presented in these figures were generated by adopting the *smaller than* comparison, although the same analysis can be carried out for the *greater than* comparison.

9.3.2 Interpolation analysis

We performed an interpolation process that involving the generation of a series of shapes that represent intermediate steps to fill the space between two geometric shapes. Consider, for instance, Fig. 75-(a) that shows an interpolation between a circle and a square. Let us denote the first and the last shapes of the series as the reference shapes. We then calculated for each of these shapes the corresponding average clustering coefficient $\langle cc \rangle$ and the average path length $\langle l \rangle$ which are shown in the two plots that are below the image series. These measures were calculated for different values of T_l . In addition, each color corresponds to one of the reference or interpolated shapes. It is possible to observe that along the interpolation, internal angles begin appearing. Consequently, the curvature for some pixels and the $\langle cc \rangle$ start to increase, whereas the average path length decreases. This behavior is observed in small-world networks (WATTS; STROGATZ, 1998; COSTA *et al.*, 2007). For the circular shape (light gray curve), as expected, the $\langle cc \rangle$ is constant, irrespective of T_l . Regarding the square (dark blue curve), the $\langle cc \rangle$ and the $\langle k \rangle$ values increase with T_l , since more connections are established.

In a similar way, Fig. 75-(b) and (c) show the linear interpolation between a circle and a star, and, a circle and an irregular shape. The same general analysis performed for the first interpolation also applies in these cases. However, the $\langle cc \rangle$ for the last two comparisons is higher when $T_l > 0.55$, due to the presence of acute angles. The $\langle cc \rangle$ values also increase with the irregularity of the shapes. Meanwhile, there were no significant differences concerning the path length for the three analyzed interpolations. The $\langle l \rangle$ value is always high for thresholds $T_l < 0.1$ and present a fast decay. This relates to the fact that for small T_l the network is not yet a single connected component, as such, the connection between two nodes may not exist or is very large. We calculated $\langle l \rangle$ by considering the missing paths as having distance one value of magnitude higher than the largest possible geodesic path.

Figure 76 illustrates the relationship between the internal angles of a shape and the clustering coefficient for regular shapes. From left to right the angles become more acute and the clustering coefficient tends to increase. When a shape has more acute angles then the distance between the pixels decreases and the number of connections increases.

Figure 75 – Linear interpolation between (a) a circle and a square, (b) a circle and a star, and (C) a circle and an irregular shape. The plots on the left show the clustering coefficient as a function of T_l , and, the plots on the right show the average path length also as a function of T_l .

Source: Adapted from [Miranda, Machicao and Bruno \(2018\)](#).

Figure 76 – Average clustering coefficient ($\langle cc \rangle$) as a function of T_l for different geometric shapes.

Source: Adapted from [Miranda, Machicao and Bruno \(2018\)](#).

9.4 Shape Classification

We evaluated the structural network measures as feature vectors for shape classification in three distinct PR applications. This was performed by taking into account the variations of T_l . Additionally, we also evaluated the performance of the proposed feature vector in the presence of noise concerning the shape structure, and, when applying rotation and scale transformations. We start by describing the datasets that were used in each application, then we present the training strategies and the classification performance obtained in each experiment.

9.4.1 Datasets

Three datasets composed of images containing shape contours were used in the experiments. The first dataset, denoted as *geom-shapes* dataset, consists of 250 geometric shapes belonging to 10 classes: circle, triangle, square, pentagon, hexagon, star, arrow, *pacman*, heart and moon. These shapes were manually designed by different people. Shapes of the same class contain variations among themselves which were incorporated due to the free-hand design. Therefore, these shapes incorporate variations regarding structure and size. Some samples of images belonging to this dataset are shown in Fig. 77-(a).

The second dataset consists of 99 generic shapes from 9 different classes containing structural variations among instances that are in the same class, and, it will be referred to as *generic-shapes* dataset. This dataset has been used as a benchmark in many shape recognition applications (SHARVIT *et al.*, 1998; SEBASTIAN; KLEIN; KIMIA, 2004). Figure 77-(b) shows some examples of this dataset.

The last dataset used in our experiments was collected by Backes *et al.* (BACKES; CASANOVA; BRUNO, 2009a). This dataset contains real-world leaf contours from different plants and will be referred to as *leaves* dataset. It consists of 600 images of leaf contours from 30 distinct classes of plants, being 20 images per class. The selected plants comprise trees of the Brazilian vegetation and its complete description is available in (BACKES; CASANOVA; BRUNO, 2009b). Figure 77-(c) shows some examples belonging to the *leaves* dataset. The three datasets have the same number of instances per class.

9.4.2 Training and validation strategies

We made use of the n -fold cross-validation strategy in order to evaluate the performance of the shape classification tasks for the three datasets. All the tasks consisted in correctly identifying the classes of each shape. The cross-validation procedure was carried out 100 times per dataset in order to obtain the corresponding standard deviation, since the instances are randomly assigned to each fold at each run. In addition, we performed the classification tasks

Figure 77 – Examples of images belonging to the three datasets evaluated in this section. (a) *geom-shapes*, (b) *generic-shapes* and (c) *leaves* dataset.

Source: Adapted from [Miranda, Machicao and Bruno \(2018\)](#).

using three classifiers: k -NN (k -Nearest Neighbors) for $k = 1$, an optimized version of SVM (Support Vector Machines) ([PLATT, 1998](#); [KEERTHI et al., 2001](#)) (with c parameter equals to 1.0 and a polynomial as kernel function) and NB (Naive Bayes) ([MARON, 1961](#)).

Similarly to other PR applications presented in Chapter 8, the classification performance is given by the number of correctly classified instances in relation to the total number of instances of a dataset, which is also referred to as accuracy. Finally, we selected the following measures to made up the feature vector: $\langle k \rangle$ (mean degree), $\langle k_h \rangle$ (average hierarchical degree of levels 2 and 3), $\langle cc \rangle$ (clustering coefficient), $\langle l \rangle$ (average path length), ρ (degree correlation) and $\langle b \rangle$ (average betweenness).

9.4.3 Evaluation of classification performance

We evaluated the classification performance of both CNDescriptor and the proposed generalized descriptor for the three shape datasets: *geom-shapes*, *generic-shapes* and *leaves*. The average accuracy (%) and the corresponding standard deviation are shown in the bar-plots of Fig. 78 on which each row corresponds to one dataset and each column to one classifier. The different colors (purple and green) correspond to the different types of comparisons allowed during the construction of the network: *smaller than* and *greater than* T_l . Whereas the different shadings, light and dark, correspond to the CNDescriptor, φ , and to the generalized descriptor, Φ , respectively. Within each plot, the bar groups on the x -axis represent the number of T_l values used during the network evolution, which correspond to the sampling of the range $[0, 1]$. Then, for each value of T_l , the selected network measures are calculated and concatenated in the feature vector.

Consequently, the more T_l values are used, the bigger the feature vector. As an example, let us consider the first bar group of the upper left plot (*geom-shapes* dataset & k-NN classifier), 13 values equally spaced in the range $[0, 1]$ were used. The total number of descriptors in the feature vector is then given by the number of T_l values times the number of network measures. More formally, the proposed feature vector is represented as follows: $\Phi = [\langle k \rangle(T_0), \langle k^2 \rangle(T_0), \langle k^3 \rangle(T_0), \langle cc \rangle(T_0), \langle l \rangle(T_0), \rho(T_0), \langle b \rangle(T_0), \dots, \langle k \rangle(T_{n_T}), \langle k^2 \rangle(T_{n_T}), \langle k^3 \rangle(T_{n_T}), \langle cc \rangle(T_{n_T}), \langle l \rangle(T_{n_T}), \rho(T_{n_T}), \langle b \rangle(T_{n_T})]$.

Figure 78 – Classification accuracy (%) and standard deviation the three datasets: *geom-shapes*, *generic-shapes* and *leaves* and for different classifiers. Both the CNDescriptor (φ) and the proposed descriptor (Φ) were used, as well as the different comparisons allowed during the construction of the network: *smaller than* and *greater than* T_l .

Source: Adapted from [Miranda, Machicao and Bruno \(2018\)](#).

Through the analysis of Fig. 78 we observe the classification performance as a function of the number of descriptors. The performance tends to decrease for small T_l values, although for some datasets this effect is more pronounced than others, as is the case for the *geom-shapes* dataset. Regarding the *generic-shapes* dataset, the performance was more uniform across all the values evaluated as the number of descriptors. When analyzing the two comparisons, *smaller than* and *greater than* T_l , no significant differences concerning the final performance were observed and a very similar accuracy was obtained, except in a few cases (e.g, *Naive Bayes* vs *geom-shapes* dataset).

When comparing Φ and φ for the *geom-shapes* dataset, we observe that using more measures improved the classification performance. In this case, the feature vector Φ leads to a better performance than φ . But, this improvement was not observed for the other two classifiers. However, the difference in performance between the two feature vectors is small for the other

classifiers, so that they might be used interchangeably in many cases. A similar analysis can be applied to the results obtained for the *generic-shapes* dataset, although for k -NN the improvement in the classification accuracy for the feature vector Φ is more evident for a small number of T_l values (# of descriptors). This result leads to the conclusion that as more measures are used in the feature vector, the sampling of the T_l interval can be reduced.

Finally, for the last dataset (*leaves* dataset) the classification performance was considerably improved through the use of Φ , especially for the SVM classifier. This improvement can be observed even for a small number of descriptors. Therefore, the more the variability regarding the network measures obtained at each T_l , the better performance can be achieved. This results also show that Φ can be computationally more efficient, since less threshold values can be used, reducing the dimension of the feature vector. It is important to highlight that the *leaves* dataset represents real-world shapes, and therefore, is a more challenging task concerning both the number of classes (30) and the number of images per class (20). Whereas, the *geom-shapes* dataset was analyzed mainly as a benchmark dataset. This is one of the reasons for achieving a high accuracy for this dataset using both Φ and φ as feature vectors.

9.4.4 Single threshold analysis

An important aspect of the proposed method for shape recognition is the network evolution, which allows the subsequent calculation of measures for each T_l . When T_l is small, the network will be poorly connected. In contrast, a high T_l will result in a highly connected network. Consequently, the concatenation of the measures at each T_l allows the improvement in shape classification tasks. As such, the most extreme thresholds have a smaller contribution in the classification performance, and the intermediate values of T_l play a more important role. We performed an experiment in order to evaluate this hypothesis. For this task, we analyzed the classification accuracy by using only the measures from a single threshold as feature vectors. The obtained results are shown in Fig. 79.

Figure 79 – Classification accuracy (%) and standard deviation for the three datasets: *geom-shapes*, *generic-shapes* and *leaves*, using the proposed shape descriptor (Φ) for a single T_l .

Source: Adapted from [Miranda, Machicao and Bruno \(2018\)](#).

We can observe that, as expected, the extreme thresholds lead to less accurate results. In addition, for all datasets, the performance achieved by the Φ descriptor is much higher than the performance obtained with φ . Such a result was already observed in the previous experiment for a small number of T_l values, corroborating the fact that increasing the variability of the feature vector, in relation to the measures extracted from the network, increases the classification performance. Nevertheless, it is worth noting that the concatenation of measures for more values of T_l is always better than using only a single threshold.

9.4.5 Robustness & noise tolerance

The CNDescriptor has shown to be invariant to scale and rotation transformations and robust to noise and image degradation (BACKES; CASANOVA; BRUNO, 2009a). We carried out an experiment in order to evaluate the same properties for the proposed feature vector, Φ . We analyzed how the classification accuracy is influenced by these transformations. This experiment was performed using the *leaves-rotated* dataset.

We started by evaluating the rotation and scale transformations. The *leaves-rotated* dataset consists of the same images of the *leaves* dataset rotated by the 6 different angles: 7° , 35° , 104° , 132° , 201° and 298° . Similarly, the *leaves-scaled* dataset consists of the same original dataset for which the images were scaled according to the following factors: 200%, 175%, 150% and 125%. The other properties evaluated were the noise tolerance and the robustness to degradation. The *leaves-noise* dataset was created by adding noise on the original images. This noise is uniformly generated and limited in the range $[-n \dots n]$, where n corresponds to the intensity of the noise. Since the shape is a two-dimensional contour, the noise is added on both x and y coordinates. Four different noise values were considered. Regarding the robustness to degradation, we evaluated the *leaves-degraded* dataset, consisting of 17 different levels of degradation which can be from two types, namely continuous and random. In the former, only adjacent pixels are removed from the shape boundary, whereas, in the latter, the pixels are randomly removed. In both cases, the higher the degradation level, then more pixels will be removed. Figure 80 summarizes all the types of transformations discussed in this section and that were applied to the *leaves* dataset. All the datasets described here were also constructed in the work of Backes *et al.* (BACKES; CASANOVA; BRUNO, 2009a)

We start the robustness analysis of the proposed network descriptor by comparing the distributions of three measures per pixel for all the analyzed transformations. Figures 81, 82 and 83 present the degree (k), the clustering coefficient (cc) and the betweenness centrality (b) distributions per pixel, respectively, and using the *smaller than* comparison for $T_l = 0.325$. The distributions are shown per row making use of a color-map scheme. By analyzing the degree distribution (Fig. 81), we observe that the pixels located near the leaf center are highly connected. It is also possible to observe that for the degraded contours, the missing parts directly influence

Figure 80 – Transformations applied to the *leaves* dataset: rotation, continuous degradation, random degradation, addition of noise and scale.

Source: Adapted from [Miranda, Machicao and Bruno \(2018\)](#).

the degree of the pixels.

On the contrary, the clustering coefficient (Fig. 82) is higher for the pixels located near the leaf tips. At $T_l = 0.325$, the distance is smaller between the pixels of the tips, what creates more connected triples in these regions, and, consequently increases the clustering coefficient. In spite of that, it worth noting that the clustering coefficient is already high (above 0.5) for all the pixels. For the pixels located near the leaf center, the cc is high, but the degree is also high. This may be explained by the fact that the pixels in this regions have a high number of connections which are not always linked as connected triples. A similar behavior is observed even for the degraded contours.

Regarding the betweenness color-map (Fig. 83), its behavior resembles the degree color-map, since the betweenness is also higher for the pixels located near the leaf center. These pixels are part of many pathways of the contour network. As such they are central for the connections of the leaf tips. Consequently, they have a high betweenness centrality.

Additionally, we carried out an experiment in order to evaluate the classification performance given the different versions of the plant leaves dataset. For this task, each transformation was evaluated separately. Again, the classification accuracy relates to the number of correctly classified instances across the 30 classes of leaves. T_l was sampled in 13 equally spaced intervals.

Figure 81 – Color-map representation of the degree (k) of the pixels belonging to the shape contour for the *leaves* dataset and for $T_l = 0.325$.

Degree

Source: Adapted from [Miranda, Machicao and Bruno \(2018\)](#).

The results are presented in Table 20.

The network obtained at each T_l will be the same before and after rotation. Therefore, the network descriptor will not change since this transformation does not affect the adjacency matrix of the network. The variations observed in Table 20 concerning the classification performance for the rotation transformation are related to the random sampling of the cross-validation folds. Besides, other linear transformations such as scale, do not change the network descriptor. We observe, for instance, that the accuracy obtained for rotation and scale are similar to the accuracy for the original dataset. Specifically for the scale transformation, the performance was improved for all scaling factors. This can be explained by the fact that when performing the interpolation

Figure 82 – Color-map representation of the clustering coefficient (cc) of the pixels belonging to the shape contour for the *leaves* dataset and for $T_l = 0.325$.

Source: Adapted from [Miranda, Machicao and Bruno \(2018\)](#).

process together with scaling, a better discrimination of the classes was obtained. Concerning the robustness to noise, in general, we can observe a slight decrease in the classification accuracy as the noise level increases.

Finally, Fig. 84 shows the classification accuracy (%) for the degradation levels concerning the *leaves-degraded* dataset. As expected, as the amount of degradation increases, the classification task becomes more complex and the performance decreases, since parts of the shape are removed.

Figure 83 – Color-map representation of the betweenness centrality (b) of the pixels belonging to the shape contour for the *leaves* dataset and for $T_l = 0.325$.

Betweenness

Source: Adapted from [Miranda, Machicao and Bruno \(2018\)](#).

9.5 Conclusion

We have discussed in this chapter an approach for shape recognition in the context of PR applications. We extended the method proposed by Backes *et al.* ([BACKES; CASANOVA; BRUNO, 2009a](#)) by proposing a more general set of measures that can be calculated from the network defined over the shape contour. We also illustrated through the use of regular and irregular shapes how these structural measures are related to the curvature signal of the shape contours and other shape properties. Therefore, we went a step further in the study of the relationship between the shape architecture and the network structure. Additionally, we

Table 20 – Classification accuracy (%) given the rotation, scale and noise transformations.

	Scale Factor	k-NN	SVM	NB
Scale	1.25	86.55 ± 0.51	82.52 ± 0.52	73.64 ± 0.54
	1.50	87.10 ± 0.44	83.44 ± 0.56	74.50 ± 0.54
	1.75	86.93 ± 0.49	83.37 ± 0.61	75.18 ± 0.46
	2.00	87.67 ± 0.53	83.92 ± 0.58	74.64 ± 0.54
	Degree	k-NN	SVM	NB
Rotation	7°	82.84 ± 0.61	80.65 ± 0.58	72.06 ± 0.41
	35°	82.06 ± 0.60	77.90 ± 0.53	72.07 ± 0.48
	104°	80.40 ± 0.56	79.12 ± 0.66	72.58 ± 0.55
	132°	85.37 ± 0.53	78.59 ± 0.61	72.25 ± 0.48
	201°	81.21 ± 0.54	76.80 ± 0.55	71.96 ± 0.51
	298°	81.85 ± 0.59	77.04 ± 0.59	72.28 ± 0.49
	Noise level	k-NN	SVM	NB
Noise	1	77.08 ± 0.59	76.25 ± 0.68	68.30 ± 0.56
	2	75.82 ± 0.68	78.43 ± 0.47	68.35 ± 0.56
	3	75.42 ± 0.59	77.36 ± 0.59	67.00 ± 0.67
	4	76.54 ± 0.58	76.84 ± 0.64	67.87 ± 0.61
Original		k-NN	SVM	NB
		84.81 ± 0.51	79.99 ± 0.62	72.22 ± 0.52

Source: Adapted from [Miranda, Machicao and Bruno \(2018\)](#).

Figure 84 – Classification accuracy for different degradation levels.

Source: Adapted from [Miranda, Machicao and Bruno \(2018\)](#).

showed how the sampling of the distance threshold T_l influences the performance of the network descriptor.

In Backes *et al.* ([BACKES; CASANOVA; BRUNO, 2009a](#)), the authors showed that the CNDescriptor outperformed the classic algorithms for shape analysis, such as, Fourier descriptors ([OSOWSKI *et al.*, 2002](#)), Zernike moments ([ZHENJIANG, 2000](#)) and Multiscale fractal dimension ([TORRES; FALCAO; COSTA, 2004](#); [PLOTZE *et al.*, 2005](#)). We showed that for the *geom-shapes* and the *generic-shapes* datasets the proposed shape descriptor improves the

classification performance when compared to the CNDescriptor, although the increase in the classification accuracy was dependent on the classifier. The highest increase was observed for the SVM classifier. Additionally, the NB classifier had a slightly better performance for these two datasets when using CNDescriptor. Regarding the *leaves* dataset, the classification performance was greatly improved for all classifiers when using the proposed feature vector Φ , even when considering a smaller number of descriptors. In addition, for the *generic-shapes* dataset, the proposed approach also outperforms the method proposed in (SÁ JUNIOR; BACKES, 2015), which introduces a shape descriptor that is constructed based on the statistics of straight-line segments.

Finally, we have shown that when using a set of measures which accounts for different structural properties of networks, the number of distance thresholds T_l can be smaller, making the method computationally more efficient. Therefore, there exists a trade off between the number of measures and the number of thresholds T_l when creating the feature vector.

PREDICTION OF INFLUENZA OUTBREAKS

10.1 Introduction

As described by the World Health Organization (WHO) (WHO, 2016), seasonal influenza is a public health concern for the human population. It is also one of the main causes of death worldwide. Annually, influenza epidemics will result in 3 to 5 million cases, which can be classified as severe, and about 250,000 to 500,000 deaths (WHO, 2016; Influenza - Sciensano, 2017). Three types of viruses, referred to as A, B and C, may lead to seasonal influenza. Some of them have subtypes, e.g., the H1N1 and H3N2 viruses are subtypes of influenza A. Besides physical contact, influenza spreads easily by airborne transmission in crowded areas when an infected person coughs or sneezes (WHO, 2016). People of different ages are susceptible to influenza, but children, senior people, pregnant women, as well as healthcare workers are exposed to the highest health risk (MAK *et al.*, 2008; NOKLEBY; NICOLL, 2010). The seasonal influenza is more frequent during the winter in countries with a temperate climate. However, the outbreaks are irregular in tropical regions and can occur during the whole year. The assessment of the epidemic intensity, its duration and peak, as well as the corresponding time to the onset are the main purpose of the monitoring campaigns. The importance of accurately defining the onset is related to the management of adequate public health policies and infrastructure, which includes the implementation of public health protocols and alerts for hospitals (NSOESIE *et al.*, 2014).

One of the recommendations of the European Centre for Disease Prevention and Control (ECDC) is to vaccinate the population (AHMED *et al.*, 2013; ECDC, 2009; RODRIGUEZ-BLANCO *et al.*, 2012). ECDC also provides technical assistance in monitoring epidemics. Some studies explore the financial impact of vaccination in reducing influenza-related costs (RYAN *et al.*, 2006; UHART *et al.*, 2016), while others focus on the relationship between influenza incidence and the prevailing climatological conditions (BALLESTER *et al.*, 2011; BALLESTER

et al., 2016; PIN *et al.*, 2012; REICHERT *et al.*, 2004). In the case of Belgium, the Unit of Epidemiology and Public Health and the National Influenza Centre (NIC) of Sciensano are institutions responsible for the influenza surveillance system (Influenza - Sciensano, 2017). There exists in this country a network of Sentinel General Practitioners (SGPs), whose main purpose is the early detection of an influenza epidemic, as well as, the determination of its intensity and duration. Furthermore, characterizing the different viruses is also one of its goals, because the virus type determines the vaccine strains that are needed for the next season. The SGPs network encompasses about 1.5% of all the Belgian GPs (General Practitioners). Such network is representative of the GPs in terms of age, sex and geographical spread. Weekly, the GPs provide reports concerning the number of patients presenting an influenza-like illness (ILI). Among the general criteria for ILI, we can cite high fever, respiratory and systemic symptoms ILI (Influenza - Sciensano, 2017).

Although the seasonality is one of the main characteristics of influenza epidemics in temperate regions, the definition of its onset is very challenging problem. This is mainly due to the differences in meteorological conditions from year to year, the different virus types and the human population dynamics. It is possible to observe a high variability of the onset, peak and duration across consecutive epidemic seasons. For instance, the epidemic of 2014–2015 in Belgium was classified as intense and of long duration. In addition, the predominant virus type was Influenza A(H3N2), as was the case in Europe and North America. A higher number of hospitalizations, which were related to laboratory-confirmed influenza, was observed for this epidemic season. For the same season, there was also an excess all-cause mortality among elder people (more than 65 years old) in most of the reporting countries, according to the European system for monitoring excess mortality for public health action (ECDC Summary, 2015; WHO Review, 2015). This scenario indicates the importance of public health tools for the real-time monitoring and forecasting of influenza related cases.

Distinct modelling approaches have been proposed in literature concerning the influenza forecasting. Time series, agent-based, metapopulation and compartmental models are among the proposed approaches (BRAUER, 2008; KANE *et al.*, 2014; SOEBIYANTO; ADIMI; KIANG, 2010). Recent reviews (NSOESIE *et al.*, 2014; CHRETIEN *et al.*, 2014) highlight some important attributes that have to be considered when implementing these epidemiological models, such as the scope and the limitations of the proposed approach and a clear description of the global evaluation measures, since it will ease the comparison of different approaches. Another challenging task is to measure the performance of real-time approaches (NSOESIE *et al.*, 2014), since different data sources are used. Within this context, digital data sources have emerged along the last decade, such as the Google Flu Trends (GFT) (DUGAS *et al.*, 2013; GFT, 2017; NSOESIE; MARARTHE; BROWNSTEIN, 2013; SHAMAN; KARSPECK, 2012), Wikipedia logs (GENEROUS *et al.*, 2014; MCIVER; BROWNSTEIN, 2014), Twitter (PAUL; DREDZE; BRONIATOWSKI, 2014), and others (GINSBERG *et al.*, 2009). In a recent study using GFT data, the authors conclude that such data can be used for predicting the week with the highest ILI

when these data allow to infer trends, however it is not possible to forecast the highest number of ILI cases (NSOESIE; MARARTHE; BROWNSTEIN, 2013). The GFT tool is available for researches but Google stopped publishing real-time estimates on-line since 2015.

Real-time prediction methods, in contrast to retrospective approaches, are based on a dynamic calibration of the model parameters as new data become available. Shaman *et al.* (SHAMAN; KARSPECK, 2012) propose an approach in order to identify the epidemic onset which makes use of a data assimilation method (the ensemble adjustment Kalman filter). The proposed approach provides a posterior estimate of the parameters by combining data observations with an ensemble of simulations. The authors show that it is possible to predict the weekly incidence of ILI through the use of a retrospective approach. Nsoesie *et al.* (NSOESIE; MARARTHE; BROWNSTEIN, 2013) proposes a forecasting approach for influenza which relies on individual-based modeling and an optimization procedure for finding the optimal parameter set using GFT data. Recently, a framework for early detection of ILI was proposed (WON *et al.*, 2017). The authors used different data sources, as the GFT, the European Influenza Surveillance Network (EISN) and Saúde 24 (Serviço Nacional de Saúde, 2017). It was shown that by combining data from different sources the model accuracy was improved concerning the onset prediction of the epidemics. The predictive potential of the GFT data was also shown in another study (DUGAS *et al.*, 2013) which made use of generalized linear auto-regressive moving average models.

Apart from classical epidemiological models, a variety of methods have been discussed in literature which allow to go from coarse to more fine grain models. Within this context, many modeling approaches have resorted to networks, which were demonstrated to be an excellent tool for modeling real-world phenomena. Networks have been used to analyze from epidemiological processes (PASTOR-SATORRAS; VESPIGNANI, 2001; PASTOR-SATORRAS; VESPIGNANI, 2005; ARRUDA; RODRIGUES; MORENO, 2018) to computer viruses (MURRAY, 1988; KEPHART; WHITE; CHESS, 1993) and other spreading phenomena like information flow in social networks (WENG; MENCZER; AHN, 2013) and economic crisis (GARAS *et al.*, 2010). When defined on networks, each node may represent individuals in a contact network, for instance, or the physical boundaries between different regions of a country. Therefore, different epidemiological levels can be modeled in this context.

In this chapter, we present an approach for the real-time prediction of ILI cases on a weekly basis of an ongoing epidemic up to its end. The proposed approach can be described in two steps. The first accounts for the construction of a confidence region which is defined based on the ILI data of previous epidemic seasons. In this way, the confidence region incorporates the variability of the ILI incidence among the historical data records of influenza, as well as, the time of occurrence of a specific ILI number. We employed the SIR epidemic model (KERMACK; MCKENDRICK, 1927) for ILI prediction from a given week on. In the second step, a dynamic calibration of the SIR model is performed, on which an optimization algorithm is used in order

to select the parameter set that minimizes the difference between the simulated and the observed numbers of ILI cases. The proposed approach was evaluated in the context of the Belgian SGP data for the influenza seasons 2010–2011 up to 2015–2016.

The proposed method for ILI prediction was first applied at the level of the Belgium population, which will be denoted as global analysis. Then, further the proposed method was extended to the Belgian *arrondissements* (districts), which are administrative regions of Belgium. For this task we incorporated into the model the network of commuting data, represented by the people traveling on a daily basis from one *arrondissement* to another. Therefore, we could incorporate data from another demographic level. We start by presenting the data of the epidemic seasons used in this work and a brief introduction about the SIR model. Then, we introduce the proposed approach for real-time prediction of ILI cases. We, then, show how it was applied to the data of the whole country and later we present the local approach which incorporates the commuting data.

The global approach for the real-time prediction of ILI cases has been accepted for publication in the *Epidemics* journal ([MIRANDA et al., 2019](#)).

10.2 An Overview of the Epidemic Seasons in Belgium

In Fig. 85 we can see the numbers of ILI cases in Belgium for seasons 2010–2011 to 2015–2016. The data used in this work was provided by Sciensano, an institution that supports health-policy-making in Belgium through research ([Sciensano, 2017](#)). The observed ILI values were normalized with respect to the catchment population of the 43 Belgian *arrondissements*. The catchment population is the average number of patients that visits the SGP network per district. The normalization is performed by dividing the number of ILI cases of each district by the corresponding catchment population, and, then this number is multiplied by the total population of the district. In this way we can extrapolate the ILI cases of the catchment population of each district to the district's total population. The resulting data will be referred to as the normalized data.

In this work, we selected week 35 of the year as the first week of a new season. Therefore, a season corresponds to the period from week 35 of year a to week 34 of year $a + 1$. Through the analysis of Fig. 85, we can also observe the seasonality of ILI incidence during the last six seasons. The ILI cases are observed mostly in the same period of the year. However the variability among the epidemic seasons is still very high concerning the peak of the ILI curve, as well as the duration of the season and the week in which the highest weekly incidence is reported. Among the studied years, seasons 2012–2013 and 2014–2015 presented the highest ILI incidence. They were classified by Sciensano as high intensity seasons with an ILI consultation rate of about 1000 per 100,000 inhabitants ([Influenza - Sciensano, 2017](#)). Their duration was

also longer when compared to the other epidemic seasons. Although with a lower ILI peak than seasons 2012–2013 and 2014–2015, season 2015–2016 may also be considered as a high intensity epidemic.

Figure 85 – Normalized weekly incidence of influenza-like illness (ILI) cases in Belgium for the seasons 2010–2011 to 2015–2016. The normalized ILI data are shown from week 35 of year i to week 34 of year $i + 1$

Source: Adapted from [Miranda et al. \(2019\)](#).

Except for their onset, seasons 2010–2011 and 2011–2012 present very similar behavior and they were characterized as moderate epidemics. Season 2013–2014 is unique among the analyzed seasons. It was classified as mild and of short duration. Table 21 summarizes the main characteristics of the six epidemic seasons in Belgium in terms of their onset, duration, the week of the highest ILI incidence, the consultation rate (the number of consultations per 100,000 inhabitants), the epidemic threshold calculated by ECDC using the MEM (Moving Epidemic Method) ([VEGA et al., 2013](#); [VEGA et al., 2015](#)) method and their intensity. The epidemic threshold accounts for the weekly ILI cases per 100,000 inhabitants. Finally, Table 22 presents the total population per year in Belgium

10.3 The SIR Epidemic Model

In epidemiology, the compartmental models are classical tools used for describing the dynamics of many infectious disease ([KERMACK; MCKENDRICK, 1927](#)). In such models, it is assumed that the individuals of a population can be in one of three compartments, namely Susceptible (S) individuals, that are able to contract the infection; Infected (I) individuals, that are able to propagate the infection, and Recovered or Removed (R) individuals, who have recovered

Table 21 – Characteristics of the Belgian influenza seasons 2010–2011 to 2015–2016. The onset is the first week at which the epidemic threshold was exceeded. This threshold is provided by ECDC and calculated using the MEM method.

Season	Onset (week/ year)	Peak ILI (week/ year)	Duration (weeks)	Consult. Rate (consult/10 ⁵ inhabitants)	Epidemic Threshold (ILI/10 ⁵ inhabitants)	Epidemic Intensity
2010–2011	50/2010	01-04/2011	12	500	148	moderate
2011–2012	05/2012	07/2012	6	531	138	moderate
2012–2013	52/2012	06/2013	12	1000	139	high
2013–2014	06/2014	09/2014	6	311	141	mild
2014–2015	02/2015	06/2015	10	979	140	high
2015–2016	04/2016	09/2016	10	734	143.9	moderate

Source: Adapted from [Miranda *et al.* \(2019\)](#).

Table 22 – Total population in Belgium per year in the period of 2010 to 2016.

Pop. on January 1st	Flanders region	Brussels-Capital region	Walloon region	Belgium
2010	6.251.983	1.089.538	3.498.384	10.839.905
2011	6.306.638	1.119.088	3.525.540	10.951.266
2012	6.350.765	1.138.854	3.546.329	11.035.948
2013	6.381.859	1.154.635	3.563.060	11.099.554
2014	6.410.705	1.163.486	3.576.325	11.150.516
2015	6.444.127	1.175.173	3.589.744	11.209.044
2016	6.477.804	1.187.890	3.602.216	11.267.910

from the infection or died. There exist variations of this model in literature ([ALLEN, 1994](#); [ANDERSON](#); [ROBERT, 1991](#); [HETHCOTE, 2000](#)). The total population is usually fixed in these type of models and, as such, migration and birth processes are neglected. However since epidemic processes usually evolve much faster than birth and death events, this is not an issue for such models ([HETHCOTE, 2000](#)). The SIR model is described by the following system of ordinary differential equations ([KERMACK](#); [MCKENDRICK, 1927](#)):

$$\frac{dS}{dt} = -\frac{\beta IS}{N}, \quad (10.1)$$

$$\frac{dI}{dt} = \frac{\beta IS}{N} - \gamma I, \quad (10.2)$$

$$\frac{dR}{dt} = \gamma I, \quad (10.3)$$

where β is the transmission rate, γ the recovery rate and N the size of the total population. Since the population is constant, then $S(t) + I(t) + R(t) = N$. The basic reproduction number, denoted R_0 , is the expected number of new infections derived from a single infection and is defined as $\frac{\beta}{\gamma}$. If R_0 is larger than 1, then the number of infected people will grow exponentially.

Besides influenza, other infectious disease, such as whooping cough ([LAVINE](#); [KING](#); [BJØRNSTAD, 2011](#)) and measles ([GRENFELL, 1992](#); [GRENFELL](#); [BJØRNSTAD](#); [FINKENSTÄDT, 2002](#)) have also been modeled through SIR and other compartmental models. In order to show that the SIR model can describe the dynamics influenza epidemics, we calibrated the β

and γ parameters in Eqs. (1)–(3) for each of the six Belgian seasons separately using all available ILI data. This procedure was implemented by means of the Nelder-Mead method (NELDER; MEAD, 1965) in order to minimize the sum of squared differences between the normalized and the simulated weekly ILI cases. Fig. 86 presents the normalized and simulated number of ILI cases for season 2010–2011. As can be seen in Table 21, the onset of this season occurred in week 50 of 2010. We can observe that both curves of ILI cases are in close agreement and the peak is accurately described.

Figure 86 – Normalized weekly incidence of ILI cases (obs. ILI cases) and the SIR model solution representing the simulated counterpart (sim. ILI cases) for the season 2010–2011.

Source: Adapted from Miranda *et al.* (2019).

In Table 23 we can observe the calibrated parameters for every season, the mean absolute error (MAE), the root mean squared error (RMSE) and the Pearson correlation, which were used to quantify the discrepancies between the normalized and simulated ILI numbers. From this table we can conclude once more that there is much variability among the seasons regarding the model parameters (β and γ). However, the SIR model accurately described the dynamics of the six epidemic seasons. In addition, the MAE and RMSE are relatively small when compared to the total Belgian population and the Pearson correlation coefficients are very high, which indicates a pronounced linear relationship between the normalized and simulated ILI numbers.

As already introduced in Table 21, the onsets of the epidemic seasons correspond to the first week in which the epidemic threshold was exceeded. When the start of the period used in the simulation of the SIR model approaches the true onset, a more accurate description of the epidemic behavior is obtained. Therefore, the epidemic onset is one of the requisites in order to reliably simulate the dynamics of the ILI cases. In addition, by shifting the outbreak curves such that week zero corresponds to the true onset provides a way for mutually comparing the influenza seasons. Fig. 88 presents all the six Belgian epidemic seasons shifted to their respective onsets. We can still observe the high variability among the studied seasons even when comparing them by their onsets.

Table 23 – Estimated model parameters β and γ , mean absolute error (MAE), root mean squared error (RMSE) and Pearson correlation coefficient between the normalized and simulated numbers of ILI cases for seasons 2010–2011 to 2015–2016.

Season	$\hat{\beta}$	$\hat{\gamma}$	MAE	RMSE	Corr.
2010–2011	4.635539	4.280132	2091.08	3035.25	0.985
2011–2012	8.376870	7.657043	1825.54	2722.60	0.989
2012–2013	4.231897	3.722846	2072.34	3448.29	0.994
2013–2014	8.607093	8.152276	860.92	1252.99	0.993
2014–2015	6.068778	5.288085	2609.47	3982.24	0.994
2015–2016	4.714837	4.207067	2676.26	4479.31	0.986

Source: Adapted from [Miranda *et al.* \(2019\)](#).

10.4 Real-time Prediction of ILI Incidence

In this section we introduce a proposed approach for real-time prediction of ILI incidence. We start with the construction of a confidence region from previous epidemic seasons, then we discuss how the SIR model parameters should be recalibrated in order to include the ILI data that become available on a weekly basis. In a final step we are able to obtain predictions for an ongoing epidemic. All these processes are illustrated in Fig. 87, which summarizes the proposed approach and that will be described in the next subsections.

Figure 87 – Proposed approach for real-time prediction of ILI cases.

Source: Adapted from [Miranda *et al.* \(2019\)](#).

10.4.1 The Confidence Region

Due to the periodic behavior of the epidemic seasons, time series are often used in the analysis of ILI incidence (KANE *et al.*, 2014; SOEBIYANTO; ADIMI; KIANG, 2010; HELD; PAUL, 2012; VIBOUD *et al.*, 2003). In this work, we treated each epidemic season as a time series of ILI incidence, which was shifted in such a way that week 0 corresponds to the onset, as shown in Fig. 88. In this way, it is possible to compare different seasons on a common basis, as well as to model previous epidemics resulting in more accurate predictions for upcoming seasons. Within the context of this work, the so-called confidence region is a single construct that captures the variability among the studied seasons.

Figure 88 – Weekly ILI incidence per week for every season starting from the onset (week 0) given in Table 21.

Source: Adapted from Miranda *et al.* (2019).

In order to generate the confidence region, firstly we account for the variability among the ILI numbers, denoted by n , at a given week across the seasons. For this task, we calculate the average weekly ILI cases, which is denoted by μ_n with the corresponding standard deviation, σ_n . Then, $\mu_n + \sigma_n$ is used to define the boundaries of the so-called confidence region along the vertical axis (Figs. 89-(a) and (b)). In a similar way, we also take into account the variability in time, i.e., the week of occurrence of a given n value. For that purpose, the interval between the minimum and the maximum normalized ILI number among the analyzed seasons was sampled in order to obtain equally spaced ILI incidence values along the vertical axis. For every season, linear interpolation was used to retrieve the time of occurrence, denoted by t , of each ILI value. Consequently, the resulting weeks were averaged per ILI incidence across the seasons. Therefore, we could obtain μ_t and σ_t , after which $\mu_t + \sigma_t$ was used to define the boundaries of the confidence region along the horizontal axis (Figs. 89-(c) and (d)).

Figure 89 – Construction of the confidence region. (a) Average weekly incidence of ILI cases, μ_n , and the corresponding standard deviation (sd), σ_n . (b) Boundary of the confidence region along the vertical axis defined by $\mu_n + \sigma_n$. (c) Average time of occurrence per ILI incidence value, μ_t , and the corresponding standard deviation (sd), σ_t . (d) Boundary of the confidence region along the horizontal axis defined by $\mu_t + \sigma_t$. (e) Union of the envelopes defined in (b) and (d). (f) Final confidence region.

Source: Adapted from [Miranda et al. \(2019\)](#).

Finally, the two regions are merged which jointly accounts for the variability among the ILI numbers in a certain week and the time of occurrence of a certain number of ILI number (Fig. 89-(e)). This was achieved by taking the union of the envelopes shown in Figs. 89-(b) and (d). The final confidence is shown in Fig. 89-(f). Both σ_n and σ_t can be scaled in order to account for less or more variability in the magnitude of the ILI numbers and the time at which these number were observed.

10.4.2 Prediction of ILI incidence

The constructed confidence region can be used together with the calibrated SIR model for predicting the ILI numbers for a specific season from week i on, given data up to week $i - 1$. Let us define the onset of season j as w_0^j , then in week w_0^j the data of season j up to the onset week are available, so the reported numbers of ILI cases from some week w_{-m}^j up to week w_0^j can be used in order to predict the ILI numbers from week w_1^j on. In this way, only data from weeks w_{-m}^j up to w_0^j are used to calibrate the parameters of Eqs. (1)–(3) using the Nelder-Mead optimization algorithm ((NELDER; MEAD, 1965)). Due to the heuristic nature of this optimization method, initial guesses should be provided for the β and γ so that the search can be initialized. However, the initial guesses may affect the final values of the calibrated parameters, and, consequently, the predicted ILI numbers. In order to overcome this issue, the confidence region is used to select meaningful parameter sets, which occurs as follows: given each initial values of β and γ , and, the ILI data up to week i , it is verified if the predicted ILI number for week $i + 1$ (which is obtained through the use of the calibrated model parameters) falls within the confidence region. If this condition is satisfied the corresponding parameter set is added to the set of acceptable parameters.

The selection of the final parameter set is illustrated in Fig. 90. Fig. 90-(a) shows all the ILI curves for season 2014–2015 obtained from 30 different initial guesses assigned randomly for β (between 0 and 12.5) and γ (between 0 and β). Each curve is the result of a single run of the optimization algorithm. In Fig. 90-(b), we can see those curves that were selected through the use of the confidence region. Therefore, only the curves for which the prediction for the subsequent week in relation to the current week (red dot) falls within the confidence region are shown. Notice that the confidence region for season 2014–2015 is constructed based on the ILI data of the remaining seasons. Finally, in order to obtain the final curve for ILI prediction, the values of β and γ for the curves shown in Fig. 90-(b) are averaged and the resulting ILI incidence curve for $\hat{\beta}$ and $\hat{\gamma}$ is shown in Fig. 90-(c).

Figure 91 illustrates the predicted ILI numbers for season 2010–2011 when more ILI data become available for this season. The curves shown in each plot are the result after 300 runs of the optimization algorithm. The shaded area depicts the confidence region constructed on the basis of the other seasons. In Fig. 91-(a), we can observe that the simulated period starts at the true onset of season 2010–2011 (Table 21) and only ILI data from weeks 0 and -1 were used for model calibration. As more ILI data become available, these can then be used for model recalibration, as the case for weeks 2, 4 and 6.

Figure 90 – Selection of acceptable model parameters. (a) ILI incidence curves obtained for 30 runs of the optimization algorithm starting from 30 initial guesses of β and γ . (b) ILI incidence curves whose prediction for week $i + 1$, after current week (curr. week), red dot, falls within the confidence region. (c) ILI incidence curve generated using the average of the parameter values ($\hat{\beta}$ and $\hat{\gamma}$) that give rise to ILI incidence curves falling within the confidence region.

Source: Adapted from [Miranda *et al.* \(2019\)](#).

Figure 91 – Predicted numbers of ILI cases (gray) throughout season 2010–2011 for different starting weeks of the simulated period, together with the normalized ILI data (black). The epidemic onset and the start of the simulated period, the current week (curr. week), are shown in blue and red, respectively.

Source: Adapted from [Miranda *et al.* \(2019\)](#).

10.5 Accuracy of the Global Analysis

In order to evaluate the accuracy of the proposed approach concerning the prediction of the ILI numbers, we adopted a cross-validation scheme among the seasons 2010–2011 to 2015–2016. In this context, each season was considered once as the “current” season (the validation season), while the remaining seasons were used for the construction of the confidence region and will be denoted as the calibration set. We carried out 300 runs of the optimization algorithm in order to obtain the calibrated parameter set for all the epidemic seasons analyzed in this work. The parameter sets leading to ILI curves which were not enclosed by the confidence region, as described in the previous section, were excluded. We adopted the average population size in Belgium across the years 2010 up to 2016 (Table 22) as the value of N in Eqs. (1)–(3). In this

section, we will denote as simulated period the time lapse between the first week from which ILI data are used and the last week for which the ILI number is predicted.

Figure 92 presents the predicted numbers of ILI cases for season 2010–2011 for different starting weeks of the simulated period. The corresponding confidence region was generated using data from season 2011–2012 up to 2015–2016. Fig. 92-(a) shows the normalized and predicted ILI numbers when the start of the simulated period is initially set to -1, i.e., one week before the onset. The first plot of this figure, therefore, represents the case where only ILI data up to the epidemic onset (week 0) are available and the corresponding predictions are obtained from week 1 on. The subsequent plots show the predictions when ILI data for a next weeks are available. Consequently, for each plot depicted in Fig. 92-(a) the model has to be recalibrated, so that the newly data can be incorporated to the model. Similar results are presented in Figs. 92-(b) and (c), however for different starting weeks of the simulated period: -2 and -3, respectively. Depending on the behavior of the normalized data the starting week of the simulation may have a significant influence on the final agreement between normalized and predicted ILI values. This can happen mainly when there are oscillations regarding the ILI number at the start of the epidemic period. In the case of season 2010–2011, using data up to week 3, the best agreement was obtained when the simulated period started one week before the epidemic onset (Fig. 92-(a)). For predictions beyond the third week, the start of the simulation period did not impact the results in a significant manner.

Similarly, Fig. 93 presents the simulation outputs for season 2014–2015, which was one of the seasons described as of high intensity (Fig. 85). Together with 2012–2013, both seasons have a great influence on the confidence region shape in relation to its upper boundary. In the case of season 2014–2015, the number of weeks used before the onset had a greater influence on the finals results when compared to the previous example of season 2010–2011, mainly for the first weeks.

Moreover, for some weeks of season 2014–2015 there was a low number of parameter sets leading to curves falling within the confidence region. In the worst case, when starting three weeks before the onset, no predictions were obtained for ILI data up to week 4 (Fig. 93-(c)), and the success rate is less than 3% when using data up to weeks 2 and 3. However, using data up to five or more weeks, more than 96% of parameter sets were selected through the confidence region checking. This is mainly due to the fact that season 2014–2015 presents a very high peak (see Fig. 88). Therefore, during the generation of the confidence region (with the remaining seasons) such high peak will not be in the calibration set, and, consequently, many predictions for the analyzed season may not be covered by the confidence region. Although, season 2012–2013 (which has a similar behavior as season 2014–2015) was used to construct the confidence region for season 2014–2015, its peak happened a few weeks later in time and there is much more variability among both seasons after the peak. Therefore, this is why the descending part of the confidence region could cover predictions for all the weeks. The corresponding plots for the

Figure 92 – Simulated (gray) and normalized (black) numbers of ILI cases in Belgium for season 2010–2011. The blue dot corresponds to the epidemic onset and the red dot to the current week (curr. week) up to which the normalized ILI data were used for model calibration. Simulated period starting (a) the week before the onset (b) 2 weeks before the onset and (c) 3 weeks before the onset.

Source: Adapted from [Miranda et al. \(2019\)](#).

Figure 93 – Simulated (gray) and normalized (black) numbers of ILI cases in Belgium for season 2014–2015. The blue dot corresponds to the epidemic onset and the red dot to the current week (curr. week) up to which the normalized ILI data were used for model calibration. Simulated period starting (a) the week before the onset (b) 2 weeks before the onset and (c) 3 weeks before the onset.

Source: Adapted from [Miranda et al. \(2019\)](#).

other four epidemic seasons can be found in section C.1 of Appendix C.

10.5.1 Error analysis

We calculated the mean absolute error (MAE) and the root mean square error (RMSE) between the normalized and the simulated numbers of ILI cases for all the six seasons. The corresponding values are summarized in Fig. 94. In addition, we also calculated the cross-correlation between both time series considering different time lags. Each row in Fig. 94 corresponds to an epidemic season and each column to a different starting week. The dots in the plots correspond to the average error over 10 predicted ILI incidence values (week 0-9). The corresponding standard deviation is also presented in the figure.

For the majority of the seasons and starting weeks the smallest errors were obtained for a zero lag. This indicates that there is no shift between the time series of the normalized and the predicted values. Moreover, the highest correlation was also observed for zero lag. Among all the six seasons, the largest error values were obtained for season 2014–2015, corroborating the results presented in Fig. 93. One of the reasons is that this season defines the upper boundary of the confidence region. The best agreement was obtained for season 2013–2014. In general, the start of the simulated period did not have a great influence on the error indices and correlation coefficients.

10.6 Incorporating the Commuting Network

As presented in the previous sections, the proposed approach for ILI prediction was applied to the entire Belgian population. In a more fine-grained analysis, our next step consisted of extending this approach to each Belgian district by considering the local ILI incidence as this complies with the granularity of the available data. As explained in Section 10.2, the observed ILI values were normalized with respect to the catchment population of the 43 Belgian districts.

The contribution of infected individuals coming from other districts was accounted for in order to simulate the epidemic spread in a specific district. The proposed local analysis makes use of a spatial component, given by the commuting network. Spatial components have a huge impact on influenza and many other communicable diseases (MURRAY, 2001). In our case, the spatial component describes the spread of the epidemics across a network whose nodes are the Belgian districts. In Belgium, the network of general practitioners registers ILI cases, typically at the district level (ECDC, 2009; Sciensano, 2017). In addition, commuting data are also publicly available at this level. Such statistics include people traveling from one district to another on a daily basis.

Figure 95-(a) shows the Belgian districts. These districts are administrative regions

Figure 94 – MAE (mean absolute error), RMSE (root mean square error) and correlation between the normalized and the predicted numbers of ILI cases per season (rows) and per starting week of the simulated period (columns) for different time lags.

Source: Adapted from [Miranda et al. \(2019\)](#).

between the municipalities and the provinces. The district of Brussels, for instance, includes the city of Brussels and 18 more municipalities. There are in total 43 districts. A network was created in which each district is represented by a node and the weight of the edges corresponds to the number of people commuting between each district (Fig. 95-(b)). Therefore, this network is a complete weighted graph. In addition, Fig. 96 presents, in descending order, the strength of each district, which is calculated as the sum of people living in the other districts and commuting to the

reference district. As expected, the district with the highest strength is Brussels, whose capital is the city of Brussels. The districts of Antwerp and Ghent are at the second and third position, respectively. The commuting data used in the work described in this section was obtained from the Belgian census of 2011.

Figure 95 – Administrative districts of Belgium and the commuting network. (a) The 43 Belgian districts are administrative regions. (b) Network generated using the Belgian commuting data. Each node represents one district and the links are given by the number of people commuting between the districts.

Source: Adapted from Heyden (2006).

Figure 96 – Strength of the districts, being the sum of the number people belonging to all the others districts and commuting to the district of focus.

Source: Elaborated by the author.

10.6.1 An overview of the ILI incidence in the Belgian districts

Figures 97 and 98 show the ILI incidence per Belgian district for the six epidemic seasons presented in Section 10.2 (2010–2011 up to 2015–2016). The seasons are shown as a single time series comprising the six seasons. Considering districts with a higher population density, such as Brussel (Brussels), Antwerpen, Hasselt and Turnhout, we observe well-defined epidemic peaks. For Gent, we also observe these peaks, although they are more pronounced for seasons 3, 5 and 6. This is also the case for Brugge, Charleroi, Leuven, Luik and other districts. In addition, for some districts no such peaks can be seen. This can be due to the population size, characterizing a small district, but also due to under-reporting cases. Recalling that each district consists of a set of municipalities, the data shown in Figs. 97 and 98 reflect the ILI incidence at the level of this set. For some districts (Aarlen, Aat, Bastenaken and Borgworm) most data were missing, so it was impossible to generate predictions for these districts. The ILI data were not reported for the district of Virton, which explains why there are only 42 plots.

10.6.2 SIR epidemic model with commuting data

Besides constructing the confidence region for each district separately, we also accounted for the influence that each district has on the others. Consider a district d_i with S_i susceptible individuals. People belonging to S_i can get infected through contact with individuals already infected, I_i , that live in the same district, d_i , but also by people living in another district d_j , that travel to district d_i , which will be denoted as I_j . Equation (10.4) describes this dynamics. The sum in this equation accounts for the influence of the infected individuals in other districts and C_j is the proportion of commuting people from district d_j . G corresponds to the set of all districts for which the sum in this equation is defined except the analyzed district i .

$$\frac{dS_i}{dt} = -\frac{\beta S_i I_i}{P_i} - \sum_{j \in G \setminus \{i\}} \beta S_i I_j \frac{C_j}{P_j}, \quad (10.4)$$

$$\frac{dI_i}{dt} = \frac{\beta S_i I_i}{P_i} + \sum_{j \in G \setminus \{i\}} \beta S_i I_j \frac{C_j}{P_j} - \gamma I_i, \quad (10.5)$$

$$\frac{dR_i}{dt} = \gamma I_i. \quad (10.6)$$

10.6.2.1 Accuracy of the local analysis

We evaluated the accuracy of the proposed local analysis concerning the prediction of the ILI incidence by adopting the same cross-validation scheme used for the global analysis for seasons 2010–2011 to 2015–2016 (Section 10.5). Therefore, given a district d_i with six epidemic seasons, each of them is considered once as the “current” season (the validation season), and

Figure 97 – Weekly ILI incidence in the Belgian districts for the six analyzed seasons: 2010–2011 up to 2015–2016.

Source: Elaborated by the author.

the remaining seasons (calibration set) are used for the construction of the confidence region.

Figure 98 – Weekly ILI incidence in the Belgian districts for the six analyzed seasons: 2010–2011 up to 2015–2016.

Source: Elaborated by the author.

We studied two approaches for the local analysis. The first one is the approach proposed for the

whole Belgian data (global analysis). Therefore, for each district and for each epidemic season, we first generated the confidence region and then we obtained a prediction. In a second approach, the confidence region is constructed in the same way, but the ILI prediction is performed by incorporating the commuting network (Fig. 95-(a)).

Each simulation is the result of 300 runs of the optimization algorithm in order to obtain the calibrated parameter set for all the epidemic seasons. As discussed before, the parameter sets leading to ILI curves that are not enclosed by the confidence region are excluded. In this local analysis the corresponding average population size of each district i , across the years 2010 up to 2016, was considered as the total population in Eqs. (1)–(3).

Figures 99 and 100 present the predicted ILI numbers for season 2010–2011 up to 2015–2016 for the district of Brussels and for a simulated period starting one week before the onset. Similarly to the previous plots showing the results of the simulated ILI numbers (Figs. 92 and 93), the blue dot corresponds to the onset of the epidemic. In this local analysis, we adopted the same onset week as the one reported in Table 21. The red dot corresponds to the current week up to which the normalized ILI data is used to calibrate the model. The sequence of plots shows the behavior of the predicted curves as the data from new weeks are available. Following the general behavior for the data of the whole country, seasons 2012–2013 and 2014–2015 have the highest peak, and, hence define the upper part of the confidence region.

One issue that can be observed at the local level is that the onset of an epidemic season does not always correspond to the global onset. The epidemic can start in one district and reach other districts later. We observe, for instance, that for season 2010–2011, the ILI values at the beginning of the simulated period are oscillating, which implies that the first predicted curve obtained for this season is a decreasing curve. However, from week 1, this behavior is changed. Still, we adopted the global onset in order to have a common basis for comparing the predictions. In addition, for highly populated districts, such as Brussels, the local and the global dynamics are very similar. Other seasons of the Brussels district also have such an oscillating behavior of the ILI values when approaching the epidemic peak (as season 2015–2016). This may result in a high prediction error. For week 0 of this season no predictions were obtained. However, as more ILI data become available the prediction error will decrease.

Similarly, Figs. 101 and 102 present the predicted numbers of ILI cases in the district of Brussels for seasons 2010–2011 up to 2015–2016 when using the commuting network. The simulation was performed for the same period, which starts one week before the onset. In general, we observe a similar behavior when compared to Figs. 99 and 100. With this approach we could obtain predictions for all the weeks and for all seasons. Seasons 2010–2011 and 2011–2012, for instance, present a very similar behavior in spite of their onset. Seasons with a high peak usually have a steep increase in the ILI number, which can be an issue for the predictions obtained for the first weeks of the simulated period, as the case of seasons 2012–2013 and 2014–2015.

In Section C.2 of Appendix C, we present the corresponding plots showing the simulated

Figure 99 – Simulated (gray) and normalized (black) numbers of ILI cases in *Brussel* (Brussels) district for season 2010–2011 up to 2012–2013 obtained without using commuting data. The blue dot corresponds to the epidemic onset and the red dot to the current week (curr. week) up to which the normalized ILI data were used for model calibration. The simulated period starts one week before the onset.

Source: Elaborated by the author.

and normalized ILI predictions for two other districts, Antwerpen and Gent. We can observe that Antwerpen district has a general behavior as Brussels, since it also has a high population

Figure 100 – Simulated (gray) and normalized (black) numbers of ILI cases in the *Brussel* (Brussels) district for season 2013–2014 up to 2015–2016 obtained without using commuting data. The blue dot corresponds to the epidemic onset and the red dot to the current week (curr. week) up to which the normalized ILI data were used for model calibration. The simulated period starts one week before the onset.

Source: Elaborated by the author.

density. The obtained predictions for Gent illustrate a different scenario, mainly for seasons 2010–2011 and 2011–2012. These seasons can be characterized as low intensity and present

Figure 101 – Simulated (gray) and normalized (black) numbers of ILI cases in the *Brussel* (Brussels) district for season 2010–2011 up to 2012–2013 obtained using commuting data (Eq. 10.4). The blue dot corresponds to the epidemic onset and the red dot to the current week (curr. week) up to which the normalized ILI data were used for model calibration. The simulated period starts one week before the onset.

Source: Elaborated by the author.

many oscillations in the start of the simulated period. Antwerpen and Brussels presented a larger confidence region for all the seasons when compared to the ones obtained for Gent. This is

Figure 102 – Simulated (gray) and normalized (black) numbers of ILI cases in *Brussel* (Brussels) district for season 2013–2014 up to 2015–2016 obtained using commuting data (Eq. 10.4). The blue dot corresponds to the epidemic onset and the red dot to the current week (curr. week) up to which the normalized ILI data were used for model calibration. The simulated period starts one week before the onset.

Source: Elaborated by the author.

mainly due to the higher variability of the seasons for these two districts.

10.6.2.2 Error analysis

We also performed an error analysis for the proposed local approach. We calculated the mean absolute error and the root mean square error considering both the simulations with and without the commuting network. Fig. 103 shows the corresponding error values for the districts of Brussels, Antwerpen and Gent. The dots in the plots correspond to the average error over 10 predicted ILI incidence values (week 0-9) for 11 different time lags. The errors correspond to the predictions shown in the previous section. The simulated period starts one week before the onset.

It is expected that the smallest errors would be obtained for a zero time lag, which indicates that the ILI predictions correspond to the normalized ILI numbers. This is achieved for the majority of the cases in Fig. 103. The largest errors were obtained for season 2014–2015, irrespective of the district, as in the global analysis. The smallest errors were mostly obtained for seasons 2010–2011 and 2011–2012.

Specifically, when comparing both approaches (with and without the commuting data) for the Brussels district, we observed three possible outcomes: (i) in the first one, similar results were obtained for both approaches, as is the case for seasons 2010-2011 and 2015-2016; (ii) secondly, the best performance was achieved for each of the districts when using ILI data only, which is the case of seasons 2012-2013 and 2013-2014, and, (iii) finally, the use of both ILI and commuting data improved the accuracy of the predictions, such as in the case of seasons 2011-2012 and 2014-2015. The same three outcomes were observed for the other two districts although not for the same seasons. In Antwerpen, for instance, using the commuting data improved the results for season 2013-2014. Regarding Gent, we can observe a great improvement in the ILI prediction for season 2010-2011, but in general, a similar behavior for both approaches was observed for the other seasons.

Regarding the approach that incorporates the commuting network, it should be stressed that the confidence region was calibrated using the nationwide defined onset. However, for the different districts the epidemic onset may change. Therefore, this approach could be improved by resorting to an automated way of defining the local onset.

10.7 Conclusion

In Belgium there is a national weekly reporting system of ILI cases, which relies on a network of General Practitioners. By means of weekly reports, it is possible to capture such variations in order to arrive at real-time predictions. These weekly reports are at the basis of this study and allowed us to propose an approach for ILI prediction at both the global (whole country) and the local (districts) levels.

The applicability of the proposed approach has been demonstrated for six epidemic

seasons from 2010 to 2016. We observed variability among these epidemic seasons concerning the onset, the peak and the duration. The proposed approach is based on the construction of a confidence region that accounts for the variability among epidemics seasons. As ILI data of more epidemic seasons become available, this confidence region becomes more representative. One issue is that seasons having a high peak define the upper boundary of the confidence region. In this case, the numbers of ILI cases can be underestimated, since the analyzed season is not taken into account when defining the confidence region. In addition, seasons presenting a steep increase of the ILI incidence after the onset can also be difficult to predict. These problems can be partly overcome by scaling the confidence region. In this way, it is possible to incorporate more variability in the data. In addition, the dynamic calibration of the SIR model, by incorporating new ILI data every week, allows to tune the dynamics of the ongoing season.

In a second part of this chapter we incorporated the commuting network. Then, individuals from a given district can be infected by contact with individuals belonging to the same district or through contact with commuting individuals. We observed that for some seasons and analyzed districts, this decreased the prediction error. However, the situation in which the local epidemic onset is different from the global one can be an issue, and, therefore, in future analysis, the proposed approach could also account for an automated way for defining the local onset for each district.

Figure 103 – MAE (mean absolute error) and RMSE (root mean square error) between the normalized and the predicted numbers of ILI cases in the districts of Brussels, Gent and Antwerpen per season for different time lags.

Source: Elaborated by the author.

Part IV

Conclusions & Future Directions

CONCLUSIONS & PERSPECTIVES

11.1 General Conclusions

11.1.1 *Part I: Theoretical Background*

In the first part of this thesis, we summarized the main theoretical background that was the basis for developing and describing the methods presented in the subsequent chapters. In Chapter 2 we discussed graph theory and networks. We started by presenting some historical remarks on how graphs have emerged as modeling tools. Then, a mathematical definition of graphs was presented in order to standardize the nomenclature used in this thesis. The link between graphs and networks was then introduced. Subsequently, we presented the main theoretical models proposed along the last two decades which have been used to model many real-world systems, as the scale-free and the small-world networks. In addition to these theoretical models, statistical and other related measures that are typically used for characterizing such networks were introduced. Connectivity related measures, distances and path lengths, clustering coefficients and many others are important attributes of the network topology. Finally, in the last section of this chapter, pattern recognition on networks was discussed. This topic appears in several chapters of this thesis. Within this context, we discussed how data can be modeled using networks, how features can be extracted from the network topology, and, some learning strategies.

In the subsequent chapter (Chapter 3) CA theory was introduced. CAs are spatially explicit models and they were explored in many chapters of this thesis. We presented some historical background on how CAs have become an established tool for modeling and analyzing real-world phenomena. We also presented a mathematical formalization of CAs and we discussed the main tessellations and neighborhoods used to define such models. In addition, the most well-known CA models were introduced, such as the Elementary Cellular Automata, the totalistic

3-state CAs and the two-dimensional Life-Like CA. In order to complement this chapter, we discussed how the spatio-temporal patterns obtained through the evolution of CAs can be quantitatively analyzed. Within this context, some measures were introduced such as entropy-related measures, the Lempel-Ziv complexity among others. In addition, an extended version of the Lyapunov exponent was introduced.

In Chapter 4 of Part I we discussed the state of the art on NAs. We started by presenting definitions of CAs on irregular tessellations like the framework proposed by Baetens & De Baets (BAETENS; DE BAETS, 2010b; BAETENS; DE BAETS, 2010a; BAETENS; DE BAETS, 2012). We then investigated models concerning CAs on graphs and networks and we described pioneering works within this context (WATTS, 1999; TOMASSINI; GIACOBINI; DARABOS, 2005; DARABOS; GIACOBINI; TOMASSINI, 2007; DARABOS *et al.*, 2011). Especially the works of Marr & Hütt (MARR; HÜTT, 2005; MARR; MÜLLER-LINOW; HÜTT, 2007; MARR; HÜTT, 2009; MARR; HÜTT, 2012) were a source of inspiration for the proposed GenBSNA and the LLNA concerning the modeling of the irregular neighborhood. Moreover, in this chapter we described other related automata models in the study of scale-free topologies. Finally, once more networks together with pattern recognition applications were discussed, now paying attention to how spatially explicit models can be used in this context.

11.1.2 Part II: Spatially Explicit Modeling on Networks Using Cellular Automata

In the second part of this thesis we discussed how the classical CA paradigm can be extended to irregular tessellations focusing on networks. The CA paradigm normally uses regular tessellations such as the square lattices. As discussed in this part, regular topologies may constrain the modeling of real-world phenomena that usually have a much more complex geometry. In order to implement the transition from regular to irregular tessellations, a key point that was taken into account is the definition of the CA neighborhood. In Chapter 5, we introduced the concept of NAs, i.e., how a CA can be built on a network. Then we presented a family of NAs inspired by a birth-survival dynamics (BSNA). This NA family has a set of rules that defines when a node is born or survives, as such leading to spatio-temporal dynamics that is intrinsically linked to the network topology. Characterizing such spatio-temporal patterns was a key point in this chapter. Understanding the behavior of such NAs is a first step for their further use in PR applications. Through the use of quantitative measures, such as the Lyapunov exponent, we could identify and classify the obtained patterns. We could see, for instance, that even for some topologies with long-range correlations, such as the random network, some rules can give rise to all the types of behavior, e.g., stable, oscillating and chaotic. In addition, we defined the membership relation between the rules belonging to the BSNA rule space and we proposed how to define the chance of “living” for each rule. Therefore, besides defining a new family of NAs,

we proposed how they can be quantitatively classified in terms of their spatio-temporal patterns.

In Chapter 6, we continued the investigation of the properties of the BSNA family. We proposed a general definition of the BSNA for which the rule space is less restrictive, since it allows for more combinations of birth and survival conditions to be represented through the density intervals. This is the case, for instance, of the LLNA, that is inspired by the classic Life-Like CA. We showed how the LLNA neighborhood resembles the Moore neighborhood by mapping its 8-nearest neighbors to corresponding intervals of neighborhood density. In addition, we moved a step further in the PR context by introducing the definition of a network descriptor and illustrating how it can be constructed with the spatio-temporal measures. We used this network descriptor in order to discriminate networks originating from different theoretical network models. Besides the NA rule, we also evaluated the influence of other parameters on the spatio-temporal patterns and consequently on the network descriptors, such as the initial distribution of alive nodes, the number of evolution steps and the size of the network concerning the number of nodes. We concluded that the network size does not influence the final network descriptor, though, the two other parameters may impact the final classification accuracy. We also provided a comparison between the BSNA and its generalized version (GenBSNA). Finally, we analyzed the sensitivity of the proposed network descriptor to the initial configuration of alive nodes (σ). We showed that, for a fixed value of σ , and for distinct starting configurations of the network nodes, the network descriptor is preserved and can, therefore, be used to differentiate the network models.

In the last chapter of this part (Chapter 7), we moved to a classical model that has been described by different approaches in literature (including CAs), the RPS game. Strongly influenced by evolutionary game theory, the RPS has been widely used to describe species diversity. We first discussed the lattice-based RPS including mobility events besides competition and reproduction interactions. Then, we discussed how the corresponding interactions take place when defined on networks. Again, we could evaluate the effect of topology on a specific dynamics. Through extensive simulations, we showed that for some network models, which present long-range correlations, the time to the first species extinction is very short, even for low mobility values, which is not observed, for instance, for regular networks. In addition to the network type, we also evaluated the influence of the network connectivity and size. We observed that as the connectivity increases, the time to the first extinction decreases, except for regular networks which have a very low probability of extinction. However, as the size of the network increases the time to the first extinction also increases.

11.1.3 *Part III: Understanding Patterns and Describing Processes on Networks*

In Part III, we went from a theoretical to a more applied level, showing how the proposed approaches can be applied to real-world problems. In Chapter 8, we presented a formalization of the validation strategies used in the PR context for network characterization. We presented how the network descriptors and the feature vector were obtained. We first extensively evaluated the LLNA regarding the correct identification of theoretical network models. We assessed its accuracy concerning different classification tasks using synthetic datasets. We highlighted that the characterization of theoretical network models is a basic classification problem that allowed to evaluate the proposed LLNA descriptor, since the properties of such networks are already known. Later, we considered a set of tasks characterizing real-world applications involving network data. We also discussed the importance of the rule selection. A given rule may not lead to the same performance for different datasets. When comparing the performance of LLNA with the one achieved using structural network measures as descriptors, we observed a significant improvement of the correct classification rate for different applications. By applying LLNA in PR problems with a different scope, we demonstrated that it is a general tool and can be tuned to different applications.

In Chapter 9, we moved to the context of Computer Vision applications using network descriptors. We once more evaluated how network descriptors can be employed to characterize networks. For the examples in this chapter, we only used the structural network measures discussed in Chapter 2. The main application is shape characterization. We extended an established approach (BACKES; CASANOVA; BRUNO, 2009a) by investigating a more generalized set of network measures and the corresponding classification accuracy for shape classification. We went one step further by investigating the relationship between shape architecture and network topology, i.e., how the network measures vary according to specific shape properties, such as angle and curvature. For the different datasets analyzed in this chapter, the proposed network descriptors could outperform the ones using connectivity measures. Based on the results obtained in this chapter, we can see once more how networks can be used to analyze different applications.

Finally, in Chapter 10 we presented an approach for prediction of influenza-like illness incidence, which can be used as a decision-support tool for medical centers and other health-care systems. The initial approach based only on the confidence region, that accounts for the previous epidemic seasons, was extended in order to incorporate the commuting network. In this way, it was possible to account for the influence of the commuting infected individuals. The implementation of the proposed approach is straightforward, since it is based on an established modeling framework. Moreover, it can be extended to other countries with seasonal influenza outbreaks. Once there are data available from these countries, the confidence region can be constructed in order to characterize the historical trends of such data. Shifting the seasons to their

epidemic onset allowed for the characterization of an epidemic despite the temporal displacement of the epidemic start for the different seasons.

11.2 Future Perspectives

In this section some future perspectives that rely on the contributions of the methods proposed in this thesis are presented. The area of pattern recognition in networks has been little explored. This scenario paves the way for the proposal, implementation and analysis of methods that can be used for network analysis and classification. Moreover, the increasing interest in networks as modeling tools shows that they are suitable to represent different real-world phenomena.

A first extension would be the use of optimization methods for the rule selection process of the BSNA and its generalized version. Considering the LLNA family, for instance, there are a total of 2^{18} possible rules, i.e., 262144 rules. For some NAs, the rule space is too vast to be explored computationally. In addition, for networks with a large number of nodes and edges the rule selection step might become an issue that is difficult to handle. When analyzing the performance in Chapters 6 and 8, it was shown that as the value of the parameter R increases, the classification accuracy also increases until its convergence.

Therefore, there is a need to explore the rule space defined by different values of R . The use of genetic algorithms is then proposed for the rule selection step, when evaluating a NA in the context of a classification task. Inspired by the natural selection process, these algorithms yield solutions to optimization problems (MITCHELL, 1998). In the context of NAs, each rule corresponds to a possible solution (an individual) for the classification task. In addition, concerning the rule selection process, we ask whether there is a set of rules that can be eliminated from the training set since they never provide a good classification accuracy. A starting point for such an investigation would be the obtained results for the synthetic and the real-world datasets analyzed in this work (Chapter 8). For this task, the “chance of living” of each rule (Definition 5.3.2) should be compared with its corresponding classification accuracy for each dataset.

Considering the applications in computer vision, a future research insight focus on the application of LLNA for image classification tasks, such as shape and texture analysis. Given the discrete nature of digital images, different approaches for image analysis rely on the connectivity among the pixels. Such graphs can be explored through the use of LLNA and other NAs introduced in this thesis. This is the case, for instance, for the classification of stomata distribution patterns that make use of images from the plant leaf (Chapter 8).

In relation to the proposed local approach for predicting the number of ILI cases at the level of the districts, one of the main issues that may influence the final performance of the

proposed approach is the definition of the onset of the epidemic. This local approach relies on the onset defined at the national level. The threshold, however, may not correspond to the local threshold in a given district. Therefore, for the ILI prediction at the level of districts, future investigation of the local thresholds should be carried out. Another future prospect is to resort to a more fine grained level of analysis through the construction of a network of contacts among the individuals. This approach can be addressed through the use of geo-location tools in combination with census data. Nowadays, there exist different tools that can be used to retrieve such data (BROECK *et al.*, 2011). Finally, extending the proposed approach to other countries is another future possibility, since it allows a broader evaluation of the proposed approach.

11.3 Bibliographical Production

Articles in scientific journals

- **Miranda, G. H. B.**, Bossuyt, N., Baetens, J. M., Bruno, O. M., De Baets, B. Real-time modeling of influenza outbreaks in Belgium. *Epidemics*, 2019.
- **Miranda, G. H. B.**, Machicao, J., & Bruno, O. M. (2018). An optimized shape descriptor based on structural properties of networks. *Digital Signal Processing*, 82, 216-229.
- Machicao, J., Corrêa Jr, E. A., **Miranda, G. H.**, Amancio, D. R., & Bruno, O. M. (2018). Authorship attribution based on Life-Like Network Automata. *PloS one*, 13(3), e0193703.
- **Miranda, G. H. B.**, Machicao, J., & Bruno, O. M. (2016). Exploring Spatio-temporal Dynamics of Cellular Automata for Pattern Recognition in Networks. *Scientific Reports*, 6, 37329.
- **Miranda, G. H.B.**, Machicao, J., & Bruno, O. M. (2016). Network Analysis Using Spatio-Temporal Patterns. In *Journal of Physics: Conference Series* (Vol. 738, No. 1, p. 012011). IOP Publishing.

Articles in preparation

- **Miranda, G. H. B.**, Baetens, J. M., Bruno, O. M., De Baets, B. A family of birth-survival network automata. *Chaos*, 2019.

Articles published in conference proceedings

- **Miranda, G. H. B.**, Machicao, J., Baetens, J. M., Bruno, O. M., De Baets, B. A Family of Network Automata Based on Neighborhood Density. In: 24th International Workshop

on Cellular Automata and Discrete Complex Systems (AUTOMATA 2018), 2018, Ghent. Automata 2018 - Exploratory papers.

Abstracts published in conference proceedings

- **Miranda, G. H. B.**, Baetens, J. M., Bruno, O. M., De Baets, B. (2018). Towards the prediction of influenza-like illness cases in Belgium. In *11th European Conference on Mathematical and Theoretical Biology*, 2018, Lisbon, Portugal.
- **Miranda, G. H. B.**, Baetens, J. M., Bruno, O. M., De Baets, B. (2017). Interference between connectedness and species dynamics in in silico ecosystems governed by non-hierarchical competition. *Proceedings of CCS'17: Conference on Complex Systems 2017.*, Cancun, Mexico.
- **Miranda, G. H. B.**, Bruno, O. M. Pattern Recognition in Complex Networks through the use of Cellular Automata. In *8th International Conference on Discrete Models of Complex Systems (Summer Solstice 2016).*, Aveiro, Portugal.

BIBLIOGRAPHY

AHMED, F.; LINDLEY, M. C.; ALLRED, N.; WEINBAUM, C. M.; GROHSKOPF, L. Effect of influenza vaccination of healthcare personnel on morbidity and mortality among patients: systematic review and grading of evidence. **Clinical Infectious Diseases**, Oxford University Press, v. 58, n. 1, p. 50–57, 2013. Citation on page 203.

AITKENHEAD, M.; DALGETTY, I.; MULLINS, C.; MCDONALD, A. J. S.; STRACHAN, N. J. C. Weed and crop discrimination using image analysis and artificial intelligence methods. **Computers and electronics in Agriculture**, Elsevier, v. 39, n. 3, p. 157–171, 2003. Citation on page 183.

ALBERT, R.; BARABÁSI, A.-L. Statistical mechanics of complex networks. **Reviews of modern physics**, APS, v. 74, n. 1, p. 47, 2002. Citations on pages 48, 50, and 51.

ALDANA, M. Boolean dynamics of networks with scale-free topology. **Physica D: Nonlinear Phenomena**, Elsevier, v. 185, n. 1, p. 45–66, 2003. Citation on page 83.

ALLEN, L. J. Some discrete-time SI, SIR, and SIS epidemic models. **Mathematical Biosciences**, Elsevier, v. 124, n. 1, p. 83–105, 1994. Citation on page 208.

ALTMAN, N. S. An introduction to kernel and nearest-neighbor nonparametric regression. **The American Statistician**, Taylor & Francis Group, v. 46, n. 3, p. 175–185, 1992. Citation on page 60.

ANDERSON, R. M.; ROBERT, M. **Infectious Diseases of Humans: Dynamics and Control**. [S.l.]: Oxford University Press, 1991. Citation on page 208.

ARRUDA, G. F. de; RODRIGUES, F. A.; MORENO, Y. Fundamentals of spreading processes in single and multilayer complex networks. **Physics Reports**, Elsevier, 2018. Citation on page 205.

ATTNEAVE, F. Some informational aspects of visual perception. **Psychological review**, American Psychological Association, v. 61, n. 3, p. 183, 1954. Citation on page 187.

BACKES, A. R.; BRUNO, O. M. Shape classification using complex network and multi-scale fractal dimension. **Pattern Recognition Letters**, Elsevier, v. 31, n. 1, p. 44–51, 2010. Citations on pages 57 and 183.

_____. Polygonal approximation of digital planar curves through vertex betweenness. **Information Sciences**, Elsevier, v. 222, p. 795–804, 2013. Citation on page 183.

BACKES, A. R.; CASANOVA, D.; BRUNO, O. M. A complex network-based approach for boundary shape analysis. **Pattern Recognition**, Elsevier, v. 42, n. 1, p. 54–67, 2009. Citations on pages 57, 151, 172, 183, 184, 185, 192, 196, 200, 201, and 238.

_____. **Leaves dataset**. 2009. Available: <<http://scg.ifsc.usp.br/dataset/ShapeCN.php>>. Accessed: 08/08/2018. Citation on page 192.

_____. Texture analysis and classification: A complex network-based approach. **Information Sciences**, Elsevier, v. 219, p. 168–180, 2013. Citations on pages [57](#) and [184](#).

_____. Contour polygonal approximation using the shortest path in networks. **International Journal of Modern Physics C**, World Scientific, v. 25, n. 02, p. 1350090, 2014. Citation on page [57](#).

BACKES, A. R.; FLORINDO, J. B.; BRUNO, O. M. Shape analysis using fractal dimension: A curvature based approach. **Chaos: An Interdisciplinary Journal of Nonlinear Science**, AIP, v. 22, n. 4, p. 043103, 2012. Citation on page [183](#).

BAETENS, J. **Discrete spatio-temporal modelling paradigms for environmental processes: SWOT analysis, application portfolio and Lyapunovian stability analysis**. Phd Thesis (PhD Thesis) — Ghent University, 2012. Citations on pages [64](#) and [91](#).

BAETENS, J. M.; DE BAETS, B. Continuous cellular automata on irregular tessellations: mimicking steady-state heat flow. In: IEEE. **Nature and Biologically Inspired Computing (NaBIC), 2010 Second World Congress on**. [S.l.], 2010. p. 53–59. Citations on pages [77](#) and [236](#).

_____. Phenomenological study of irregular cellular automata based on lyapunov exponents and jacobians. **Chaos: An Interdisciplinary Journal of Nonlinear Science**, AIP Publishing, v. 20, n. 3, p. 033112, 2010. Citations on pages [74](#), [77](#), [78](#), and [236](#).

_____. Cellular automata on irregular tessellations. **Dynamical Systems**, Taylor & Francis, v. 27, n. 4, p. 411–430, 2012. Citations on pages [37](#), [77](#), [78](#), [91](#), and [236](#).

BAETENS, J. M.; DE BAETS, B. A lyapunov view on the stability of two-state cellular automata. In: **Irreducibility and Computational Equivalence**. [S.l.]: Springer, 2013. p. 25–33. Citations on pages [71](#), [74](#), and [75](#).

BALLESTER, J.; ROBINE, J.-M.; HERRMANN, F. R.; RODÓ, X. Long-term projections and acclimatization scenarios of temperature-related mortality in Europe. **Nature Communications**, Nature Publishing Group, v. 2, p. 358, 2011. Citations on pages [203](#) and [204](#).

BALLESTER, J.; RODÓ, X.; ROBINE, J.-M.; HERRMANN, F. R. European seasonal mortality and influenza incidence due to winter temperature variability. **Nature Climate Change**, Nature Research, v. 6, n. 10, p. 927–930, 2016. Citations on pages [203](#) and [204](#).

BANERJEE, A. Structural distance and evolutionary relationship of networks. **Biosystems**, Elsevier, v. 107, n. 3, p. 186–196, 2012. Citation on page [167](#).

BARABÁSI, A.-L. **Linked: The New Science of Networks**. [S.l.]: Perseus, Cambridge, MA, 2002. Citation on page [43](#).

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. Citations on pages [36](#), [44](#), [51](#), [153](#), and [166](#).

BARABÁSI, A.-L.; JEONG, H.; NÉDA, Z.; RAVASZ, E.; SCHUBERT, A.; VICSEK, T. Evolution of the social network of scientific collaborations. **Physica A: Statistical mechanics and its applications**, Elsevier, v. 311, n. 3-4, p. 590–614, 2002. Citation on page [169](#).

BARABASI, A.-L.; OLTVAI, Z. N. Network biology: understanding the cell's functional organization. **Nature reviews genetics**, Nature Publishing Group, v. 5, n. 2, p. 101, 2004. Citation on page [166](#).

BARRAT, A.; BARTHELEMY, M.; VESPIGNANI, A. **Dynamical processes on complex networks**. [S.l.]: Cambridge university press, 2008. Citations on pages [35](#), [36](#), [46](#), and [47](#).

BARRAT, A.; WEIGT, M. On the properties of small-world network models. **The European Physical Journal B-Condensed Matter and Complex Systems**, Springer, v. 13, n. 3, p. 547–560, 2000. Citation on page [84](#).

BASTOLLA, U.; LÄSSIG, M.; MANRUBIA, S. C.; VALLERIANI, A. Biodiversity in model ecosystems, i: coexistence conditions for competing species. **Journal of Theoretical Biology**, Elsevier, v. 235, n. 4, p. 521–530, 2005. Citation on page [136](#).

BENZ, U. C.; HOFMANN, P.; WILLHAUCK, G.; LINGENFELDER, I.; HEYNEN, M. Multi-resolution, object-oriented fuzzy analysis of remote sensing data for gis-ready information. **ISPRS Journal of photogrammetry and remote sensing**, Elsevier, v. 58, n. 3, p. 239–258, 2004. Citation on page [183](#).

BERGE, C.; MINIEKA, E. **Graphs and Hypergraphs**. [S.l.]: North-Holland Publishing Company, 1976. Citation on page [43](#).

BERTO, F.; TAGLIABUE, J. Cellular automata. In: ZALTA, E. N. (Ed.). **The Stanford Encyclopedia of Philosophy**. Fall 2017. [S.l.]: Metaphysics Research Lab, Stanford University, 2017. Citations on pages [63](#) and [64](#).

BISHOP, C. **Pattern Recognition and Machine Learning (Information Science and Statistics), chapter 3, pages 138–147**. [S.l.]: Springer-Verlag New York, Inc, 2006. Citations on pages [59](#), [60](#), and [61](#).

BLASCHKE, T. Object based image analysis for remote sensing. **ISPRS journal of photogrammetry and remote sensing**, Elsevier, v. 65, n. 1, p. 2–16, 2010. Citation on page [183](#).

BLEAY, C.; COMENDANT, T.; SINERVO, B. An experimental test of frequency-dependent selection on male mating strategy in the field. **Proceedings of the Royal Society of London B: Biological Sciences**, The Royal Society, v. 274, n. 1621, p. 2019–2025, 2007. Citation on page [129](#).

BOCCALETTI, S.; LATORA, V.; MORENO, Y.; CHAVEZ, M.; HWANG, D.-U. Complex networks: Structure and dynamics. **Physics reports**, Elsevier, v. 424, n. 4, p. 175–308, 2006. Citations on pages [36](#), [51](#), [53](#), [54](#), [56](#), and [130](#).

BOCCARA, N. **Modeling complex systems**. [S.l.]: Springer Science & Business Media, 2010. Citation on page [35](#).

BODDY, L.; WELLS, J. M.; CULSHAW, C.; DONNELLY, D. P. Fractal analysis in studies of mycelium in soil. **Geoderma**, Elsevier, v. 88, n. 3-4, p. 301–328, 1999. Citation on page [175](#).

BOLLOBÁS, B. **Modern graph theory**. [S.l.]: Springer Science & Business Media, 1998. Citations on pages [43](#) and [45](#).

BRANDES, U.; ERLEBACH, T. **Network analysis: methodological foundations**. [S.l.]: Springer Science & Business Media, 2005. Citations on pages [44](#) and [45](#).

BRAUER, F. Compartmental models in epidemiology. In: _____. **Mathematical Epidemiology**. [S.l.]: Springer Berlin Heidelberg, 2008. p. 19–79. Citation on page 204.

BROECK, W. Van den; GIOANNINI, C.; GONÇALVES, B.; QUAGGIOTTO, M.; COLIZZA, V.; VESPIGNANI, A. The gleamviz computational tool, a publicly available software to explore realistic epidemic spreading scenarios at the global scale. **BMC infectious diseases**, BioMed Central, v. 11, n. 1, p. 37, 2011. Citation on page 240.

BROSNAN, T.; SUN, D.-W. Inspection and grading of agricultural and food products by computer vision systems—a review. **Computers and electronics in agriculture**, Elsevier, v. 36, n. 2, p. 193–213, 2002. Citation on page 183.

BRUNO, O. M.; PLOTZE, R. de O.; FALVO, M.; CASTRO, M. de. Fractal dimension applied to plant identification. **Information Sciences**, Elsevier, v. 178, n. 12, p. 2722–2733, 2008. Citation on page 183.

BULLMORE, E.; SPORNS, O. Complex brain networks: graph theoretical analysis of structural and functional systems. **Nature Reviews Neuroscience**, Nature Publishing Group, v. 10, n. 3, p. 186–198, 2009. Citation on page 36.

BUSS, L.; JACKSON, J. Competitive networks: nontransitive competitive relationships in cryptic coral reef environments. **The American Naturalist**, University of Chicago Press, v. 113, n. 2, p. 223–234, 1979. Citation on page 129.

CAMPITELI, M. G.; BATISTA, P. D.; KINOUCI, O.; MARTINEZ, A. S. Deterministic walks as an algorithm of pattern recognition. **Physical Review E**, APS, v. 74, n. 2, p. 026703, 2006. Citations on pages 58 and 85.

CASANOVA, D. **Redes complexas em visão computacional com aplicações em bioinformática**. Phd Thesis (PhD Thesis) — Universidade de São Paulo, 2013. Citation on page 57.

CCNR/ICENSA. **Metabolic Dataset. CCNR/ICeNSA: Interdisciplinary Center for Network Sciences & Applications**. 2018. Available: <<http://www3.nd.edu/~networks/resources.htm>>. Accessed: 08/08/2018. Citations on pages 153 and 167.

CESAR, R. M.; COSTA, L. D. F. Towards effective planar shape representation with multiscale digital curvature analysis based on signal processing techniques. **Pattern Recognition**, Elsevier, v. 29, n. 9, p. 1559–1569, 1996. Citation on page 187.

CHAPELLE, O.; SCHOLKOPF, B.; ZIEN, A. Semi-supervised learning (chappelle, o. et al., eds.; 2006)[book reviews]. **IEEE Transactions on Neural Networks**, IEEE, v. 20, n. 3, p. 542–542, 2009. Citation on page 59.

CHOPARD, B. Cellular automata modeling of physical systems. **Computational Complexity: Theory, Techniques, and Applications**, Springer, p. 407–433, 2012. Citation on page 63.

CHRETIEN, J.-P.; GEORGE, D.; SHAMAN, J.; CHITALE, R. A.; MCKENZIE, F. E. Influenza forecasting in human populations: a scoping review. **PloS One**, Public Library of Science, v. 9, n. 4, p. e94130, 2014. Citation on page 204.

COLIZZA, V.; PASTOR-SATORRAS, R.; VESPIGNANI, A. Reaction–diffusion processes and metapopulation models in heterogeneous networks. **Nature Physics**, Nature Publishing Group, v. 3, n. 4, p. 276, 2007. Citation on page 44.

COSTA, C.; ANTONUCCI, F.; PALLOTTINO, F.; AGUZZI, J.; SUN, D.-W.; MENESATTI, P. Shape analysis of agricultural products: a review of recent research advances and potential application to computer vision. **Food and Bioprocess Technology**, Springer, v. 4, n. 5, p. 673–692, 2011. Citation on page 183.

COSTA, L. d. F.; JR, O. N. O.; TRAVIESO, G.; RODRIGUES, F. A.; BOAS, P. R. V.; ANTQUEIRA, 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. Citations on pages 36, 57, 58, 130, 166, and 167.

COSTA, L. d. F.; RODRIGUES, F. A.; TRAVIESO, G.; BOAS, P. R. V. Characterization of complex networks: A survey of measurements. **Advances in physics**, Taylor & Francis, v. 56, n. 1, p. 167–242, 2007. Citations on pages 36, 55, 56, 58, 84, 151, 154, 184, and 190.

COSTA, L. da F.; BOAS, P. V.; SILVA, F.; RODRIGUES, F. A pattern recognition approach to complex networks. **Journal of Statistical Mechanics: Theory and Experiment**, IOP Publishing, v. 2010, n. 11, p. P11015, 2010. Citations on pages 58, 84, 151, and 184.

COUCLELIS, H. Cellular worlds: a framework for modeling micro—macro dynamics. **Environment and planning A**, SAGE Publications Sage UK: London, England, v. 17, n. 5, p. 585–596, 1985. Citation on page 91.

CRUTCHFIELD, J. P.; MITCHELL, M. The evolution of emergent computation. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 92, n. 23, p. 10742–10746, 1995. Citation on page 79.

CRUTCHFIELD, J. P.; MITCHELL, M.; DAS, R. Evolutionary design of collective computation in cellular automata. **Evolutionary dynamics—exploring the interplay of selection, neutrality, accident, and function**, p. 361–411, 2003. Citation on page 79.

CSARDI, G.; NEPUSZ, T. The igraph software package for complex network research. **Inter-Journal**, Complex Systems, p. 1695, 2006. Available: <<http://igraph.org>>. Citation on page 154.

DALY, A. **Putting ecological theories to the test : individual-based simulations of synthetic microbial community dynamics**. XXXIV, 278 p. Phd Thesis (PhD Thesis) — Ghent University, 2017. Citation on page 132.

DALY, A. J.; BAETENS, J. M.; DE BAETS, B. The impact of initial evenness on biodiversity maintenance for a four-species in silico bacterial community. **Journal of theoretical biology**, Elsevier, v. 387, p. 189–205, 2015. Citations on pages 130, 132, and 133.

D'AMBROSIO, D.; GREGORIO, S. D.; IOVINE, G. Simulating debris flows through a hexagonal cellular automata model: Sciddica s 3–hex. **Natural Hazards and Earth System Sciences**, Copernicus GmbH, v. 3, n. 6, p. 545–559, 2003. Citation on page 91.

D'AMBROSIO, D.; SPATARO, W.; IOVINE, G. Parallel genetic algorithms for optimising cellular automata models of natural complex phenomena: an application to debris flows. **Computers & Geosciences**, Elsevier, v. 32, n. 7, p. 861–875, 2006. Citation on page 64.

DARABOS, C.; GIACOBINI, M.; TOMASSINI, M. Performance and robustness of cellular automata computation on irregular networks. **Advances in Complex Systems**, World Scientific, v. 10, n. supp01, p. 85–110, 2007. Citations on pages 79 and 236.

DARABOS, C.; MOORE, J. H. Critical properties of cellular automata with evolving network topologies. In: IEEE. **Evolutionary Computation (CEC), 2015 IEEE Congress on**. [S.l.], 2015. p. 2105–2112. Citation on page 79.

DARABOS, C.; TOMASSINI, M.; CUNTO, F. D.; PROVERO, P.; MOORE, J. H.; GIACOBINI, M. Toward robust network based complex systems: from evolutionary cellular automata to biological models. **Intelligenza Artificiale**, IOS Press, v. 5, n. 1, 2011. Citations on pages 79 and 236.

DAS, R.; MITCHELL, M.; CRUTCHFIELD, J. P. A genetic algorithm discovers particle-based computation in cellular automata. In: **Parallel problem solving from nature—PPSN III**. [S.l.]: Springer, 1994. p. 344–353. Citation on page 79.

DAVATZIKOS, C.; TAO, X.; SHEN, D. Hierarchical active shape models, using the wavelet transform. **IEEE transactions on medical imaging**, IEEE, v. 22, n. 3, p. 414–423, 2003. Citation on page 183.

De LIGNE, L.; ULZURRUN, G. V. D.; BAETENS, J. M.; BULCKE, J. V.; ACKER, J. V.; DE BAETS, B. Analysis of spatio-temporal fungal growth dynamics under different environmental conditions. **IMA Fungus**. In press, 2019. Citations on pages 176 and 180.

DEMONGEOT, J.; GOLÈS, E.; TCHUENTE, M. **Dynamical systems and cellular automata**. [S.l.]: Academic Pr, 1985. Citation on page 69.

DOROGOVTSSEV, S. N.; MENDES, J. F. Evolution of networks. **Advances in physics**, Taylor & Francis, v. 51, n. 4, p. 1079–1187, 2002. Citations on pages 36, 52, 153, 154, and 162.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern classification**. [S.l.]: John Wiley & Sons, 2012. Citations on pages 58, 59, and 61.

DUGAS, A. F.; JALALPOUR, M.; GEL, Y.; LEVIN, S.; TORCASO, F.; IGUSA, T.; ROTHMAN, R. E. Influenza forecasting with Google flu trends. **PloS One**, Public Library of Science, v. 8, n. 2, p. e56176, 2013. Citations on pages 204 and 205.

EBENMAN, B.; JONSSON, T. Using community viability analysis to identify fragile systems and keystone species. **Trends in Ecology & Evolution**, Elsevier, v. 20, n. 10, p. 568–575, 2005. Citation on page 129.

ECDC. **European Centre for Disease Prevention and Control. Implementation of the Council Recommendation on seasonal influenza vaccination (2009/1019/EU)**. 2009. [Online; accessed 23-September-2017]. Available: <https://ecdc.europa.eu/sites/portal/files/media/en/publications/Publications/Implementation-seasonal-influenza-vaccination-Council-Recommendation-Jan-2014.pdf>. Citations on pages 203 and 218.

ECDC Summary. **Summarising the 2014-2015 influenza season in Europe**. 2015. [Online; accessed 14-March-2018]. Available: <https://ecdc.europa.eu/en/news-events/summarising-2014-2015-influenza-season-europe>. Citation on page 204.

ECKMANN, J.-P.; RUELLE, D. Ergodic theory of chaos and strange attractors. In: **The Theory of Chaotic Attractors**. [S.l.]: Springer, 1985. p. 273–312. Citation on page 74.

EPPSTEIN, D. Growth and decay in life-like cellular automata. In: **Game of Life Cellular Automata**. [S.l.]: Springer, 2010. p. 71–97. Citation on page 70.

ERDÖS, P.; RÉNYI, A. On the evolution of random graphs. **Publ. Math. Inst. Hungar. Acad. Sci.**, Citeseer, v. 5, p. 17–61, 1960. Citations on pages 36, 44, and 47.

ESTRADA, E. **The structure of complex networks: theory and applications**. [S.l.]: Oxford University Press, 2012. Citation on page 43.

FAN, Y.; SHEN, D.; GUR, R. C.; GUR, R. E.; DAVATZIKOS, C. Compare: classification of morphological patterns using adaptive regional elements. **IEEE transactions on medical imaging**, IEEE, v. 26, n. 1, p. 93–105, 2007. Citation on page 183.

FILHO, A. H.; MACHICAO, J.; BRUNO, O. M. A hierarchical model of metabolic machinery based on the kcore decomposition of plant metabolic networks. **PloS one**, Public Library of Science, v. 13, n. 5, p. e0195843, 2018. Citation on page 167.

FLORINDO, J. B.; LANDINI, G.; FILHO, H. A.; BRUNO, O. M. Analysis of stomata distribution patterns for quantification of the foliar plasticity of *Tradescantia zebrina*. In: IOP PUBLISHING. **Journal of Physics: Conference Series**. [S.l.], 2015. v. 633, n. 1, p. 012113. Citation on page 172.

FORTUNATO, S. Community detection in graphs. **Physics Reports**, Elsevier, v. 486, n. 3, p. 75–174, 2010. Citation on page 36.

GARAS, A.; ARGYRAKIS, P.; ROZENBLAT, C.; TOMASSINI, M.; HAVLIN, S. Worldwide spreading of economic crisis. **New journal of Physics**, IOP Publishing, v. 12, n. 11, p. 113043, 2010. Citation on page 205.

GARDNER, M. Mathematical games: The fantastic combinations of John Conway's new solitaire game "life". **Scientific American**, vol. 223, n. 4, p. 120–123, 1970. Citations on pages 64 and 92.

GASTNER, M. T.; NEWMAN, M. E. The spatial structure of networks. **The European Physical Journal B-Condensed Matter and Complex Systems**, Springer, v. 49, n. 2, p. 247–252, 2006. Citation on page 44.

GENEROUS, N.; FAIRCHILD, G.; DESHPANDE, A.; VALLE, S. Y. D.; PRIEDHORSKY, R. Global disease monitoring and forecasting with Wikipedia. **PLoS Computational Biology**, Public Library of Science, v. 10, n. 11, p. e1003892, 2014. Citation on page 204.

GERIG, G.; STYNER, M.; SHENTON, M. E.; LIEBERMAN, J. A. Shape versus size: Improved understanding of the morphology of brain structures. In: SPRINGER. **International Conference on Medical Image Computing and Computer-Assisted Intervention**. [S.l.], 2001. p. 24–32. Citation on page 183.

GFT. **Google Flu Trends**. 2017. [Online; accessed 27-July-2017]. Available: <<https://www.google.org/flutrends/about/>>. Citation on page 204.

GILLESPIE, D. T. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. **Journal of computational physics**, Elsevier, v. 22, n. 4, p. 403–434, 1976. Citation on page 133.

GINSBERG, J.; MOHEBBI, M. H.; PATEL, R. S.; BRAMMER, L.; SMOLINSKI, M. S.; BRILLIANT, L. Detecting influenza epidemics using search engine query data. **Nature**, Nature Publishing Group, v. 457, n. 7232, p. 1012–1014, 2009. Citation on page 204.

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. Citation on page 169.

GITTINS, R. **Canonical analysis: a review with applications in ecology**. [S.l.]: Springer Science & Business Media, 2012. Citation on page 158.

GIUȘCĂ, B. **Creative Commons The problem of the Seven Bridges of Königsberg**. 2005. This file is licensed under the Creative Commons Attribution-Share Alike 3.0 Unported license. Available: <https://commons.wikimedia.org/wiki/File:Konigsberg_bridges.png>. Citation on page 44.

GOG, A.; CHIRA, C. Dynamics of networks evolved for cellular automata computation. **Hybrid Artificial Intelligent Systems**, Springer, p. 359–368, 2012. Citation on page 80.

GONCALVES, W. N.; BACKES, A. R.; MARTINEZ, A. S.; BRUNO, O. M. Texture descriptor based on partially self-avoiding deterministic walker on networks. **Expert Systems with Applications**, Elsevier, v. 39, n. 15, p. 11818–11829, 2012. Citations on pages 86 and 151.

GONCALVES, W. N.; BRUNO, O. M. Dynamic texture analysis and segmentation using deterministic partially self-avoiding walks. **Expert Systems with Applications**, Elsevier, v. 40, n. 11, p. 4283–4300, 2013. Citations on pages 58 and 86.

GONÇALVES, W. N.; MARTINEZ, A. S.; BRUNO, O. M. Complex network classification using partially self-avoiding deterministic walks. **Chaos: An Interdisciplinary Journal of Nonlinear Science**, AIP Publishing, v. 22, n. 3, p. 033139, 2012. Citations on pages 58, 86, and 151.

GREENBERG, J. M.; HASTINGS, S. Spatial patterns for discrete models of diffusion in excitable media. **SIAM Journal on Applied Mathematics**, SIAM, v. 34, n. 3, p. 515–523, 1978. Citation on page 83.

GRENFELL, B. Chance and chaos in measles dynamics. **Journal of the Royal Statistical Society. Series B (Methodological)**, JSTOR, v. 54, p. 383–398, 1992. Citation on page 208.

GRENFELL, B. T.; BJØRNSTAD, O. N.; FINKENSTÄDT, B. F. Dynamics of measles epidemics: scaling noise, determinism, and predictability with the TSIR model. **Ecological Monographs**, Wiley Online Library, v. 72, n. 2, p. 185–202, 2002. Citation on page 208.

HAMMING, R. W. Error detecting and error correcting codes. **Bell System technical journal**, Wiley Online Library, v. 29, n. 2, p. 147–160, 1950. Citation on page 71.

HAYKIN, S. **Neural Networks and Learning Machines, 3a edição**. [S.l.]: Prentice Hall., 2008. Citation on page 60.

HEIMANN, T.; MEINZER, H.-P. Statistical shape models for 3d medical image segmentation: a review. **Medical image analysis**, Elsevier, v. 13, n. 4, p. 543–563, 2009. Citation on page 183.

HELD, L.; PAUL, M. Modeling seasonality in space-time infectious disease surveillance data. **Biometrical Journal**, Wiley Online Library, v. 54, n. 6, p. 824–843, 2012. Citation on page 211.

HEMMING, J.; RATH, T. Pa—precision agriculture: Computer-vision-based weed identification under field conditions using controlled lighting. **Journal of agricultural engineering research**, Elsevier, v. 78, n. 3, p. 233–243, 2001. Citation on page 183.

HETHCOTE, H. W. The mathematics of infectious diseases. **SIAM Review**, SIAM, v. 42, n. 4, p. 599–653, 2000. Citation on page 208.

HEYDEN, J. v. d. **Creative Commons Map of Belgium, its districts and major cities**. 2006. This file is licensed under the Creative Commons Attribution-Share Alike 3.0 Unported license, 2.5 Generic, 2.0 Generic and 1.0 Generic license. Available: <<https://commons.wikimedia.org/wiki/File:Belgium.geohive.png>>. Citation on page 220.

HEYMANS, M.; SINGH, A. K. Deriving phylogenetic trees from the similarity analysis of metabolic pathways. **Bioinformatics**, Oxford University Press, v. 19, n. suppl_1, p. i138–i146, 2003. Citation on page 167.

HURST, H. E. {Long-term storage capacity of reservoirs}. **Trans. Amer. Soc. Civil Eng.**, v. 116, p. 770–808, 1951. Citation on page 73.

ILACHINSKI, A. **Cellular automata: a discrete universe**. [S.l.]: World Scientific Publishing Company, 2001. Citations on pages 64 and 71.

Influenza - Sciensano. **Flu surveillance system in Belgium**. 2017. [Online; accessed on 30-September-2017]. Available: <<https://epidemio.wiv-isp.be/ID/diseases/Pages/Influenza.aspx>>. Citations on pages 203, 204, and 206.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. **ACM computing surveys (CSUR)**, Acm, v. 31, n. 3, p. 264–323, 1999. Citation on page 61.

JEONG, H.; TOMBOR, B.; ALBERT, R.; OLTVAI, Z. N.; BARABÁSI, A.-L. The large-scale organization of metabolic networks. **Nature**, Nature Publishing Group, v. 407, n. 6804, p. 651–654, 2000. Citations on pages 36, 57, 152, 153, 166, 167, and 168.

JOHNSTON, N. The b36/s125 2x2 life-like cellular automaton. In: **Game of Life Cellular Automata**. [S.l.]: Springer, 2010. p. 99–114. Citation on page 70.

KALE, M.; BUTAR, F. B. Fractal analysis of time series and distribution properties of hurst exponent. **J. Math. Sci. Math. Ed**, v. 5, p. 8–19, 2011. Citation on page 73.

KANE, M. J.; PRICE, N.; SCOTCH, M.; RABINOWITZ, P. Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks. **BMC Bioinformatics**, BioMed Central, v. 15, n. 1, p. 276, 2014. Citations on pages 204 and 211.

KANEHISA, M.; GOTO, S.; KAWASHIMA, S.; OKUNO, Y.; HATTORI, M. The kegg resource for deciphering the genome. **Nucleic acids research**, Oxford University Press, v. 32, n. suppl_1, p. D277–D280, 2004. Citation on page 166.

KAUFFMAN, S. A. Metabolic stability and epigenesis in randomly constructed genetic nets. **Journal of theoretical biology**, Elsevier, v. 22, n. 3, p. 437–467, 1969. Citation on page 83.

KEERTHI, S.; SHEVADE, S.; BHATTACHARYYA, C.; MURTHY, K. Improvements to platt's smo algorithm for svm classifier design. **Neural Computation**, v. 13, n. 3, p. 637–649, 2001. Citation on page 193.

KEPHART, J. O.; WHITE, S. R.; CHESS, D. M. Computers and epidemiology. **IEEE Spectrum**, IEEE, v. 30, n. 5, p. 20–26, 1993. Citation on page 205.

KERMACK, W. O.; MCKENDRICK, A. G. A contribution to the mathematical theory of epidemics. **Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences**, The Royal Society, v. 115, n. 772, p. 700–721, 1927. Citations on pages 205, 207, and 208.

KERR, B.; RILEY, M. A.; FELDMAN, M. W.; BOHANNAN, B. J. Local dispersal promotes biodiversity in a real-life game of rock–paper–scissors. **Nature**, Nature Publishing Group, v. 418, n. 6894, p. 171, 2002. Citation on page 129.

KIRKUP, B. C.; RILEY, M. A. Antibiotic-mediated antagonism leads to a bacterial game of rock–paper–scissors in vivo. **Nature**, Nature Publishing Group, v. 428, n. 6981, p. 412, 2004. Citation on page 129.

KOSCHÜTZKI, D.; LEHMANN, K. A.; PEETERS, L.; RICHTER, S.; TENFELDE-PODEHL, D.; ZLOTOWSKI, O. Centrality indices. In: **Network analysis**. [S.l.]: Springer, 2005. p. 16–61. Citation on page 55.

KROGAN, N. J.; CAGNEY, G.; YU, H.; ZHONG, G.; GUO, X.; IGNATCHENKO, A.; LI, J.; PU, S.; DATTA, N.; TIKUISIS, A. P. *et al.* Global landscape of protein complexes in the yeast *saccharomyces cerevisiae*. **Nature**, Nature Publishing Group, v. 440, n. 7084, p. 637, 2006. Citation on page 36.

KRUJA, E.; MARKS, J.; BLAIR, A.; WATERS, R. A short note on the history of graph drawing. In: SPRINGER. **International Symposium on Graph Drawing**. [S.l.], 2001. p. 272–286. Citation on page 43.

LACASA, L.; LUQUE, B.; BALLESTEROS, F.; LUQUE, J.; NUNO, J. C. From time series to complex networks: The visibility graph. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 105, n. 13, p. 4972–4975, 2008. Citation on page 57.

LAIRD, R. A.; SCHAMP, B. S. Competitive intransitivity promotes species coexistence. **The American Naturalist**, The University of Chicago Press, v. 168, n. 2, p. 182–193, 2006. Citation on page 129.

_____. Competitive intransitivity, population interaction structure, and strategy coexistence. **Journal of theoretical biology**, Elsevier, v. 365, p. 149–158, 2015. Citations on pages 129, 130, 135, and 136.

LANGTON, C. G. Computation at the edge of chaos: phase transitions and emergent computation. **Physica D: Nonlinear Phenomena**, Elsevier, v. 42, n. 1-3, p. 12–37, 1990. Citation on page 71.

LAVINE, J. S.; KING, A. A.; BJØRNSTAD, O. N. Natural immune boosting in pertussis dynamics and the potential for long-term vaccine failure. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 108, n. 17, p. 7259–7264, 2011. Citation on page 208.

LEMPEL, A.; ZIV, J. On the complexity of finite sequences. **IEEE Trans. Inf. Theor.**, v. 22, n. 1, p. 75–81, 1976. Citation on page 73.

LESKOVEC, J.; KREVL, A. **SNAP Datasets: Stanford Large Network Dataset Collection**. 2014. Available: <<http://snap.stanford.edu/data>>. Accessed: 08/08/2018. Citations on pages 152, 154, and 169.

LI, W.; PACKARD, N. The structure of the elementary cellular automata rule space. **Complex systems**, Citeseer, v. 4, n. 3, p. 281–297, 1990. Citation on page [68](#).

LIMA, G. F.; MARTINEZ, A. S.; KINOUCI, O. Deterministic walks in random media. **Physical Review Letters**, APS, v. 87, n. 1, p. 010603, 2001. Citation on page [85](#).

MACÊDO, H. B.; OLIVEIRA, G.; RIBEIRO, C. H. Dynamic behaviour of chaotic cellular automata—a comparative entropy analysis of regular lattices and small-world structures. In: **IEEE. Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on**. [S.l.], 2013. p. 1566–1571. Citation on page [80](#).

MACHICAO, J. **Padrões e pseudo-aleatoriedade usando sistemas complexos**. Phd Thesis (PhD Thesis) — University of São Paulo, 2017. Citation on page [116](#).

MACHICAO, J.; MARCO, A. G.; BRUNO, O. M. Chaotic encryption method based on life-like cellular automata. **Expert Systems with Applications**, Elsevier, v. 39, n. 16, p. 12626–12635, 2012. Citation on page [70](#).

MACQUEEN, J. *et al.* Some methods for classification and analysis of multivariate observations. In: OAKLAND, CA, USA. **Proceedings of the fifth Berkeley symposium on mathematical statistics and probability**. [S.l.], 1967. v. 1, n. 14, p. 281–297. Citation on page [61](#).

MAK, T. K.; MANGTANI, P.; LEESE, J.; WATSON, J. M.; PFEIFER, D. Influenza vaccination in pregnancy: current evidence and selected national policies. **The Lancet Infectious Diseases**, Elsevier, v. 8, n. 1, p. 44–52, 2008. Citation on page [203](#).

MALLET, D. G.; PILLIS, L. G. D. A cellular automata model of tumor–immune system interactions. **Journal of theoretical biology**, Elsevier, v. 239, n. 3, p. 334–350, 2006. Citation on page [37](#).

MARGOLUS, N. Physics-like models of computation. **Physica D: Nonlinear Phenomena**, North-Holland, v. 10, n. 1-2, p. 81–95, 1984. Citation on page [64](#).

MARON, M. E. Automatic indexing: an experimental inquiry. **Journal of the ACM (JACM)**, ACM, v. 8, n. 3, p. 404–417, 1961. Citation on page [193](#).

MARR, C.; HÜTT, M.-T. Topology regulates pattern formation capacity of binary cellular automata on graphs. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 354, p. 641–662, 2005. Citations on pages [80](#), [81](#), [92](#), [95](#), and [236](#).

_____. Outer-totalistic cellular automata on graphs. **Physics Letters A**, Elsevier, v. 373, n. 5, p. 546–549, 2009. Citations on pages [37](#), [81](#), [92](#), [95](#), and [236](#).

_____. Cellular automata on graphs: topological properties of ER graphs evolved towards low-entropy dynamics. **Entropy**, Molecular Diversity Preservation International, v. 14, n. 6, p. 993–1010, 2012. Citations on pages [37](#), [72](#), [73](#), [82](#), and [236](#).

MARR, C.; MÜLLER-LINOW, M.; HÜTT, M.-T. Regularizing capacity of metabolic networks. **Physical Review E**, APS, v. 75, n. 4, p. 041917, 2007. Citations on pages [81](#) and [236](#).

MARTINEZ, G. J.; ADAMATZKY, A.; ALONSO-SANZ, R. Designing complex dynamics in cellular automata with memory. **International Journal of Bifurcation and Chaos**, World Scientific, v. 23, n. 10, p. 1330035, 2013. Citation on page [68](#).

MCAULEY, J. J.; LESKOVEC, J. Learning to discover social circles in ego networks. In: **NIPS**. [S.l.: s.n.], 2012. v. 2012, p. 548–56. Citation on page [169](#).

MCIVER, D. J.; BROWNSTEIN, J. S. Wikipedia usage estimates prevalence of influenza-like illness in the united states in near real-time. **PLoS Computational Biology**, Public Library of Science, v. 10, n. 4, p. e1003581, 2014. Citation on page [204](#).

MENDES, R. S.; EVANGELISTA, L. R.; THOMAZ, S. M.; AGOSTINHO, A. A.; GOMES, L. C. A unified index to measure ecological diversity and species rarity. **Ecography**, Wiley Online Library, v. 31, n. 4, p. 450–456, 2008. Citation on page [129](#).

MICHALSKI, R. S.; CARBONELL, J. G.; MITCHELL, T. M. **Machine learning: An artificial intelligence approach**. [S.l.]: Springer Science & Business Media, 2013. Citation on page [59](#).

MILGRAM, S. The small world problem. **Psychology today**, New York, v. 2, n. 1, p. 60–67, 1967. Citations on pages [44](#) and [49](#).

MIRANDA, G. H.; MACHICAO, J.; BRUNO, O. M. An optimized shape descriptor based on structural properties of networks. **Digital Signal Processing**, Elsevier, v. 82, p. 216–229, 2018. Citations on pages [176](#), [183](#), [186](#), [188](#), [189](#), [191](#), [193](#), [194](#), [195](#), [197](#), [198](#), [199](#), [200](#), and [201](#).

MIRANDA, G. H. B.; BOSSUYT, N.; BAETENS, J. M.; BRUNO, O. M.; DE BAETS, B. Real-time prediction of influenza outbreaks in belgium. **Epidemics**, *accepted for publication*, Elsevier, 2019. Citations on pages [206](#), [207](#), [208](#), [209](#), [210](#), [211](#), [212](#), [214](#), [216](#), [217](#), and [219](#).

MIRANDA, G. H. B.; MACHICAO, J.; BAETENS, J. M.; BRUNO, O. M.; DE BAETS, B. A family of network automata based on neighborhood density. In: **24th International Workshop on Cellular Automata and Discrete Complex Systems (AUTOMATA 2018)**. [S.l.: s.n.], 2018. p. 1566–1571. Citation on page [115](#).

MIRANDA, G. H. B.; MACHICAO, J.; BRUNO, O. M. Exploring spatio-temporal dynamics of cellular automata for pattern recognition in networks. **Scientific reports**, Nature Publishing Group, v. 6, p. 37329, 2016. Citations on pages [58](#), [113](#), [116](#), [119](#), [120](#), [121](#), [151](#), [152](#), [158](#), [159](#), [160](#), [161](#), [162](#), [163](#), [164](#), [165](#), [166](#), [168](#), [169](#), [170](#), [171](#), [174](#), and [175](#).

_____. Network analysis using spatio-temporal patterns. **Journal of Physics: Conference Series**, IOP Publishing, v. 738, p. 012011, 2016. Citations on pages [92](#), [93](#), [94](#), [95](#), [96](#), and [151](#).

MITCHELL, M. **An introduction to genetic algorithms**. [S.l.]: MIT press, 1998. Citation on page [239](#).

_____. Complex systems: Network thinking. **Artificial Intelligence**, Elsevier, v. 170, n. 18, p. 1194–1212, 2006. Citation on page [35](#).

_____. **Complexity: A guided tour**. [S.l.]: Oxford University Press, 2009. Citations on pages [35](#), [36](#), and [72](#).

MOKHTARIAN, F.; BOBER, M. **Curvature scale space representation: theory, applications, and MPEG-7 standardization**. [S.l.]: Springer Science & Business Media, 2013. Citation on page [183](#).

MOORE, E. F. Machine models of self-reproduction. In: AMERICAN MATHEMATICAL SOCIETY NEW YORK. **Proceedings of symposia in applied mathematics**. [S.l.], 1962. v. 14, n. 1962, p. 17–33. Citation on page [63](#).

MURRAY, J. D. **Mathematical biology. II Spatial models and biomedical applications {Interdisciplinary Applied Mathematics V. 18}**. [S.l.]: Springer-Verlag New York Incorporated New York, 2001. Citation on page [218](#).

MURRAY, W. H. The application of epidemiology to computer viruses. **Computers & Security**, Elsevier Advanced Technology, v. 7, n. 2, p. 139–145, 1988. Citation on page [205](#).

MYHILL, J. The converse of moore's garden-of-eden theorem. **Proceedings of the american mathematical society**, v. 14, n. 4, p. 685–686, 1963. Citation on page [63](#).

NEKOVEE, M.; MORENO, Y.; BIANCONI, G.; MARSILI, M. Theory of rumour spreading in complex social networks. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 374, n. 1, p. 457–470, 2007. Citation on page [169](#).

NELDER, J. A.; MEAD, R. A simplex method for function minimization. **The Computer Journal**, Oxford University Press, v. 7, n. 4, p. 308–313, 1965. Citations on pages [209](#) and [213](#).

NETO, J. C.; MEYER, G. E.; JONES, D. D.; SAMAL, A. K. Plant species identification using elliptic fourier leaf shape analysis. **Computers and electronics in agriculture**, Elsevier, v. 50, n. 2, p. 121–134, 2006. Citation on page [183](#).

NEUMANN, J. V.; BURKS, A. W. *et al.* Theory of self-reproducing automata. **IEEE Transactions on Neural Networks**, v. 5, n. 1, p. 3–14, 1966. Citation on page [63](#).

NEWMAN, M. E. The structure and function of complex networks. **SIAM review**, SIAM, v. 45, n. 2, p. 167–256, 2003. Citations on pages [36](#), [48](#), and [53](#).

_____. A measure of betweenness centrality based on random walks. **Social networks**, Elsevier, v. 27, n. 1, p. 39–54, 2005. Citation on page [56](#).

NEWMANN, M.; BARABÁSI, A.-L.; WATTS, D. The structure and dynamics of networks. **Princeton Studies in Complexity**, 2006. Citations on pages [43](#) and [130](#).

NIAZI, M.; HUSSAIN, A. Agent-based tools for modeling and simulation of self-organization in peer-to-peer, ad hoc, and other complex networks. **IEEE Communications Magazine**, IEEE, v. 47, n. 3, 2009. Citation on page [44](#).

NOKLEBY, H.; NICOLL, A. Risk groups and other target groups-preliminary ECDC guidance for developing influenza vaccination recommendations for the season 2010-11. **Eurosurveillance**, v. 15, n. 12, p. 19525, 2010. Citation on page [203](#).

NSOESIE, E.; MARARTHE, M.; BROWNSTEIN, J. Forecasting peaks of seasonal influenza epidemics. **PLoS Currents**, Public Library of Science, v. 5, 2013. Citations on pages [204](#) and [205](#).

NSOESIE, E. O.; BROWNSTEIN, J. S.; RAMAKRISHNAN, N.; MARATHE, M. V. A systematic review of studies on forecasting the dynamics of influenza outbreaks. **Influenza and Other Respiratory Viruses**, Wiley Online Library, v. 8, n. 3, p. 309–316, 2014. Citations on pages [203](#) and [204](#).

OKABE, A.; BOOTS, B.; SUGIHARA, K.; CHIU, S. N. **Spatial tessellations: concepts and applications of Voronoi diagrams**. [S.l.]: John Wiley & Sons, 2009. Citations on pages [77](#) and [91](#).

OSOWSKI, S. *et al.* Fourier and wavelet descriptors for shape recognition using neural networks—a comparative study. **Pattern Recognition**, Elsevier, v. 35, n. 9, p. 1949–1957, 2002. Citations on pages 183 and 201.

OTT, E. **Chaos in dynamical systems**. [S.l.]: Cambridge university press, 2002. Citation on page 74.

OVERBEEK, R.; AL. *et.* WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. **Nucleic Acids Res.**, v. 28, n. 1, p. 123–125, 2000. Citation on page 167.

PASTOR-SATORRAS, R.; VESPIGNANI, A. Epidemic spreading in scale-free networks. **Physical review letters**, APS, v. 86, n. 14, p. 3200, 2001. Citation on page 205.

_____. Epidemics and immunization in scale-free networks. **Handbook of graphs and networks: from the genome to the internet**, Wiley Online Library, p. 111–130, 2005. Citations on pages 36 and 205.

PAUL, M. J.; DREDZE, M.; BRONIATOWSKI, D. Twitter improves influenza forecasting. **PLoS Currents**, Public Library of Science, v. 6, 2014. Citation on page 204.

PICKOVER, C. A. **The math book: from Pythagoras to the 57th dimension, 250 milestones in the history of mathematics**. [S.l.]: Sterling Publishing Company, Inc., 2009. Citation on page 64.

PIN, S. P.; GOLMARD, J. L.; COTTO, E.; ROTHAN-TONDEUR, M.; CHAMI, K.; PIETTE, F. Excess winter mortality in france: influence of temperature, influenza like illness, and residential care status. **Journal of the American Medical Directors Association**, Elsevier, v. 13, n. 3, p. 309–e1, 2012. Citations on pages 203 and 204.

PLATT, J. Fast training of support vector machines using sequential minimal optimization. In: SCHOELKOPF, B.; BURGESS, C.; SMOLA, A. (Ed.). **Advances in Kernel Methods - Support Vector Learning**. MIT Press, 1998. Available: <<http://research.microsoft.com/~jplatt/smo.html>>. Citation on page 193.

PLOTZE, R. d. O.; FALVO, M.; PÁDUA, J. G.; BERNACCI, L. C.; VIEIRA, M. L. C.; OLIVEIRA, G. C. X.; BRUNO, O. M. Leaf shape analysis using the multiscale minkowski fractal dimension, a new morphometric method: a study with passiflora (passifloraceae). **Canadian Journal of Botany**, NRC Research Press, v. 83, n. 3, p. 287–301, 2005. Citations on pages 183 and 201.

QUINLAN, J. R. Simplifying decision trees. **International journal of man-machine studies**, Elsevier, v. 27, n. 3, p. 221–234, 1987. Citation on page 60.

RANGAYYAN, R. M.; MUDIGONDA, N. R.; DESAUTELS, J. L. Boundary modelling and shape analysis methods for classification of mammographic masses. **Medical and Biological Engineering and Computing**, Springer, v. 38, n. 5, p. 487–496, 2000. Citation on page 183.

REICHENBACH, T.; MOBILIA, M.; FREY, E. Mobility promotes and jeopardizes biodiversity in rock–paper–scissors games. **Nature**, Nature Publishing Group, v. 448, n. 7157, p. 1046, 2007. Citations on pages 129, 130, 131, 132, 133, and 138.

_____. Self-organization of mobile populations in cyclic competition. **Journal of Theoretical Biology**, Elsevier, v. 254, n. 2, p. 368–383, 2008. Citations on pages 130 and 132.

REICHERT, T. A.; SIMONSEN, L.; SHARMA, A.; PARDO, S. A.; FEDSON, D. S.; MILLER, M. A. Influenza and the winter increase in mortality in the United States, 1959–1999. **American Journal of Epidemiology**, Oxford University Press, v. 160, n. 5, p. 492–502, 2004. Citations on pages 203 and 204.

RODRIGUEZ-BLANCO, T.; VILA-CORCOLES, A.; DIEGO, C. de; OCHOA-GONDAR, O.; VALDIVIESO, E.; BOBE, F.; MORRO, A.; HERÁNDEZ, N.; MARTIN, A.; CALAMOTE, F. *et al.* Relationship between annual influenza vaccination and winter mortality in diabetic people over 65 years. **Human Vaccines & Immunotherapeutics**, Taylor & Francis, v. 8, n. 3, p. 363–370, 2012. Citation on page 203.

RULQUIN, C.; ARENZON, J. J. Globally synchronized oscillations in complex cyclic games. **Physical Review E**, APS, v. 89, n. 3, p. 032133, 2014. Citation on page 129.

RYAN, J.; ZOELLNER, Y.; GRADL, B.; PALACHE, B.; MEDEMA, J. Establishing the health and economic impact of influenza vaccination within the European Union 25 countries. **Vaccine**, Elsevier, v. 24, n. 47, p. 6812–6822, 2006. Citation on page 203.

SÁ JUNIOR, J. J. d. M.; BACKES, A. R. Shape classification using line segment statistics. **Information Sciences**, Elsevier, v. 305, p. 349–356, 2015. Citation on page 202.

SCABINI, L. F.; FISTAROL, D. O.; CANTERO, S. V.; GONÇALVES, W. N.; MACHADO, B. B.; JR, J. F. R. Angular descriptors of complex networks: A novel approach for boundary shape analysis. **Expert Systems with Applications**, Elsevier, v. 89, p. 362–373, 2017. Citation on page 183.

SCHIFF, J. L. **Cellular automata: a discrete view of the world**. [S.l.]: John Wiley & Sons, 2011. Citation on page 64.

SCHRANDT, R. G.; ULAM, S. M. **RECURSIVELY DEFINED GEOMETRICAL OBJECTS AND PATTERNS OF GROWTH**. [S.l.], 1967. Citation on page 63.

SCHREIBER, S. J.; KILLINGBACK, T. P. Spatial heterogeneity promotes coexistence of rock–paper–scissors metacommunities. **Theoretical population biology**, Elsevier, v. 86, p. 1–11, 2013. Citation on page 136.

SCHÜTT, M.; CLAUSSEN, J. C. Stabilization of biodiversity in the coevolutionary rock-paper-scissors game on complex networks. **arXiv preprint arXiv:1003.2922**, 2010. Citations on pages 129 and 130.

Sciensano. **Scientific Institute of Public Health**. 2017. [Online; accessed 17-October-2017]. Available: <<https://www.sciensano.be/en>>. Citations on pages 206 and 218.

SEBASTIAN, T. B.; KLEIN, P. N.; KIMIA, B. B. Recognition of shapes by editing their shock graphs. **'IEEE' Transactions on pattern analysis and machine intelligence**, IEEE, v. 26, n. 5, p. 550–571, 2004. Citation on page 192.

SELLE, D.; PREIM, B.; SCHENK, A.; PEITGEN, H.-O. Analysis of vasculature for liver surgical planning. **IEEE transactions on medical imaging**, IEEE, v. 21, n. 11, p. 1344–1357, 2002. Citation on page 183.

Serviço Nacional de Saúde. **Saúde 24**. 2017. [Online; accessed 17-October-2017]. Available: <<http://www.saude24.pt/>>. Citation on page 205.

SHALIZI, C. R. Methods and techniques of complex systems science: An overview. In: **Complex systems science in biomedicine**. [S.l.]: Springer, 2006. p. 33–114. Citations on pages [37](#) and [65](#).

SHAMAN, J.; KARSPECK, A. Forecasting seasonal outbreaks of influenza. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 109, n. 50, p. 20425–20430, 2012. Citations on pages [204](#) and [205](#).

SHANNON, C. A mathematical theory of communication. **Bell System Technical Journal, The**, v. 27, n. 3, p. 379–423, 1948. ISSN 0005-8580. Citation on page [72](#).

SHARVIT, D.; CHAN, J.; TEK, H.; KIMIA, B. B. Symmetry-based indexing of image databases. In: IEEE. **Content-Based Access of Image and Video Libraries, 1998. Proceedings. IEEE Workshop on**. [S.l.], 1998. p. 56–62. Citation on page [192](#).

SILVA, T. C.; ZHAO, L. Network-based high level data classification. **IEEE Transactions on Neural Networks and Learning Systems**, IEEE, v. 23, n. 6, p. 954–970, 2012. Citation on page [58](#).

_____. High-level pattern-based classification via tourist walks in networks. **Information Sciences**, Elsevier, v. 294, p. 109–126, 2015. Citation on page [84](#).

_____. **Machine learning in complex networks**. [S.l.]: Springer, 2016. Citation on page [58](#).

SINERVO, B.; HEULIN, B.; SURGET-GROBA, Y.; CLOBERT, J.; MILES, D. B.; CORL, A.; CHAINE, A.; DAVIS, A. Models of density-dependent genic selection and a new rock-paper-scissors social system. **The American Naturalist**, The University of Chicago Press, v. 170, n. 5, p. 663–680, 2007. Citation on page [129](#).

SMITH, D. M.; ONNELA, J.-P.; LEE, C. F.; FRICKER, M. D.; JOHNSON, N. F. Network automata: Coupling structure and function in dynamic networks. **Advances in Complex Systems**, World Scientific, v. 14, n. 03, p. 317–339, 2011. Citation on page [80](#).

SOARES, J. V.; LEANDRO, J. J.; CESAR, R. M.; JELINEK, H. F.; CREE, M. J. Retinal vessel segmentation using the 2-d gabor wavelet and supervised classification. **IEEE Transactions on medical Imaging**, IEEE, v. 25, n. 9, p. 1214–1222, 2006. Citation on page [183](#).

SOEBIYANTO, R. P.; ADIMI, F.; KIANG, R. K. Modeling and predicting seasonal influenza transmission in warm regions using climatological parameters. **PloS One**, Public Library of Science, v. 5, n. 3, p. e9450, 2010. Citations on pages [204](#) and [211](#).

STANLEY, H. E.; BULDYREV, S. V. Statistical physics: The salesman and the tourist. **Nature**, Nature Publishing Group, v. 413, n. 6854, p. 373–374, 2001. Citation on page [85](#).

STROGATZ, S. H. Exploring complex networks. **Nature**, Nature Publishing Group, v. 410, n. 6825, p. 268–276, 2001. Citation on page [49](#).

SUI, Y.; SHAO, F.; WANG, C.; SUN, R.; JI, J. Complex network modeling of spectral remotely sensed imagery: A case study of massive green algae blooms detection based on modis data. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 464, p. 138–148, 2016. Citation on page [183](#).

SZABÓ, G.; SZOLNOKI, A.; IZSÁK, R. Rock-scissors-paper game on regular small-world networks. **Journal of physics A: Mathematical and General**, IOP Publishing, v. 37, n. 7, p. 2599, 2004. Citations on pages [130](#), [136](#), and [137](#).

SZOLNOKI, A.; SZABÓ, G. Phase transitions for rock-scissors-paper game on different networks. **Physical Review E**, APS, v. 70, n. 3, p. 037102, 2004. Citations on pages [130](#) and [136](#).

TOFFOLI, T. Cellular automata as an alternative to (rather than an approximation of) differential equations in modeling physics. **Physica D: Nonlinear Phenomena**, Elsevier, v. 10, n. 1-2, p. 117–127, 1984. Citation on page [64](#).

TOMASSINI, M. Generalized automata networks. In: **Cellular Automata**. [S.l.]: Springer, 2006. p. 14–28. Citations on pages [37](#) and [79](#).

TOMASSINI, M.; GIACOBINI, M.; DARABOS, C. Evolution and dynamics of small-world cellular automata. **Complex Systems**, [Champaign, IL, USA: Complex Systems Publications, Inc., c1987-, v. 15, n. 4, p. 261–284, 2005. Citations on pages [37](#), [79](#), and [236](#).

TORRES, R. d. S.; FALCAO, A. X.; COSTA, L. d. F. A graph-based approach for multiscale shape analysis. **Pattern Recognition**, Elsevier, v. 37, n. 6, p. 1163–1174, 2004. Citations on pages [183](#) and [201](#).

TSAI, A.; YEZZI, A.; WELLS, W.; TEMPANY, C.; TUCKER, D.; FAN, A.; GRIMSON, W. E.; WILLSKY, A. A shape-based approach to the segmentation of medical imagery using level sets. **IEEE transactions on medical imaging**, IEEE, v. 22, n. 2, p. 137–154, 2003. Citation on page [183](#).

TVERSKY, A. Intransitivity of preferences. **Psychological review**, American Psychological Association, v. 76, n. 1, p. 31, 1969. Citation on page [129](#).

UHART, M.; BRICOUT, H.; CLAY, E.; LARGERON, N. Public health and economic impact of seasonal influenza vaccination with quadrivalent influenza vaccines compared to trivalent influenza vaccines in europe. **Human Vaccines & Immunotherapeutics**, Taylor & Francis, v. 12, n. 9, p. 2259–2268, 2016. Citation on page [203](#).

ULZURRUN, G. V.-D. de; BAETENS, J.; BULCKE, J. Van den; LOPEZ-MOLINA, C.; WINDT, I. D.; DE BAETS, B. Automated image-based analysis of spatio-temporal fungal dynamics. **Fungal Genetics and Biology**, Elsevier, v. 84, p. 12–25, 2015. Citations on pages [37](#), [154](#), [175](#), and [176](#).

VAPNIK, V. **Statistical learning theory**. 1998. [S.l.]: Wiley, New York, 1998. Citation on page [60](#).

VEGA, T.; LOZANO, J. E.; MEERHOFF, T.; SNACKEN, R.; MOTT, J.; LEJARAZU, R. Ortiz de; NUNES, B. Influenza surveillance in europe: establishing epidemic thresholds by the moving epidemic method. **Influenza and other respiratory viruses**, Wiley Online Library, v. 7, n. 4, p. 546–558, 2013. Citation on page [207](#).

VEGA, T.; LOZANO, J. E.; MEERHOFF, T.; SNACKEN, R.; BEAUTÉ, J.; JORGENSEN, P.; LEJARAZU, R. Ortiz de; DOMEGAN, L.; MOSSONG, J.; NIELSEN, J. *et al.* Influenza surveillance in europe: comparing intensity levels calculated using the moving epidemic method. **Influenza and other respiratory viruses**, Wiley Online Library, v. 9, n. 5, p. 234–246, 2015. Citation on page [207](#).

VIBOUD, C.; BOËLLE, P.-Y.; CARRAT, F.; VALLERON, A.-J.; FLAHAULT, A. Prediction of the spread of influenza epidemics by the method of analogues. **American Journal of Epidemiology**, Oxford University Press, v. 158, n. 10, p. 996–1006, 2003. Citation on page [211](#).

VICHNIAC, G. Y. Simulating physics with cellular automata. **Physica D: Nonlinear Phenomena**, Elsevier, v. 10, n. 1-2, p. 96–116, 1984. Citation on page 64.

WATTS, D. J. **Small worlds: the dynamics of networks between order and randomness**. [S.l.]: Princeton university press, 1999. Citations on pages 37, 78, 92, 136, and 236.

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 36, 44, 49, 58, and 190.

WAXMAN, B. M. Routing of multipoint connections. **IEEE journal on selected areas in communications**, IEEE, v. 6, n. 9, p. 1617–1622, 1988. Citation on page 52.

WENG, L.; MENCZER, F.; AHN, Y.-Y. Virality prediction and community structure in social networks. **Scientific reports**, Nature Publishing Group, v. 3, p. 2522, 2013. Citation on page 205.

WHITE, S. H.; REY, A. M. del; SÁNCHEZ, G. R. Modeling epidemics using cellular automata. **Applied Mathematics and Computation**, Elsevier, v. 186, n. 1, p. 193–202, 2007. Citation on page 37.

WHO. **Influenza (Seasonal)**. 2016. [Online; accessed on 19-July-2017]. Available: <<http://www.who.int/mediacentre/factsheets/fs211/en/>>. Citation on page 203.

WHO Review. **WHO. Review of the 2014–2015 influenza season in the northern hemisphere**. 2015. [Online; accessed 14-March-2018]. Available: <<http://www.who.int/wer/2015/wer9023.pdf?ua=1>>. Citation on page 204.

WILLIAMS, R. J.; BERLOW, E. L.; DUNNE, J. A.; BARABÁSI, A.-L.; MARTINEZ, N. D. Two degrees of separation in complex food webs. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 99, n. 20, p. 12913–12916, 2002. Citation on page 36.

WOLFRAM, S. Statistical mechanics of cellular automata. **Reviews of modern physics**, APS, v. 55, n. 3, p. 601, 1983. Citation on page 66.

_____. Cellular automata as models of complexity. **Nature**, v. 311, n. 5985, p. 419–424, 1984. Citations on pages 37 and 64.

_____. Universality and complexity in cellular automata. **Physica D: Nonlinear Phenomena**, Elsevier, v. 10, n. 1, p. 1–35, 1984. Citations on pages 66, 68, 71, and 94.

_____. Complex systems theory. In: **Emerging Syntheses in Science: Proceedings of the Founding Workshops of the Santa Fe Institute** (Addison-Wesley, 1988). [S.l.: s.n.], 1985. Citations on pages 63 and 64.

_____. **A new kind of science**. [S.l.]: Wolfram media Champaign, 2002. Citation on page 64.

WON, M.; MARQUES-PITA, M.; LOURO, C.; GONÇALVES-SÁ, J. Early and real-time detection of seasonal influenza onset. **PLoS Computational Biology**, Public Library of Science, v. 13, n. 2, p. e1005330, 2017. Citation on page 205.

WU, A.-C.; XU, X.-J.; WANG, Y.-H. Excitable greenberg-hastings cellular automaton model on scale-free networks. **Physical Review E**, APS, v. 75, n. 3, p. 032901, 2007. Citation on page 83.

WU, Z.; LU, X.; DENG, Y. Image edge detection based on local dimension: A complex networks approach. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 440, p. 9–18, 2015. Citation on page [183](#).

WUENSCHÉ, A.; LESSER, M. **Global Dynamics Of Cellular Automata: An Atlas Of Basin Of Attraction Fields Of One-dimensional Cellular Automata**. [S.l.]: Andrew Wuensche, 1992. Citation on page [68](#).

YING, C.-y.; HUA, D.-y.; WANG, L.-y. Phase transitions for a rock–scissors–paper model with long-range-directed interactions. **Journal of Physics A: Mathematical and Theoretical**, IOP Publishing, v. 40, n. 17, p. 4477, 2007. Citation on page [136](#).

ZHANG, D.; LU, G. *et al.* A comparative study on shape retrieval using fourier descriptors with different shape signatures. In: **Proc. of international conference on intelligent multimedia and distance education (ICIMADE01)**. [S.l.: s.n.], 2001. p. 1–9. Citation on page [183](#).

ZHANG, G.-Y.; CHEN, Y.; QI, W.-K.; QING, S.-M. Four-state rock-paper-scissors games in constrained newman-watts networks. **Physical Review E**, APS, v. 79, n. 6, p. 062901, 2009. Citations on pages [130](#) and [136](#).

ZHANG, J.; SMALL, M. Complex network from pseudoperiodic time series: Topology versus dynamics. **Physical review letters**, APS, v. 96, n. 23, p. 238701, 2006. Citation on page [57](#).

ZHENJIANG, M. Zernike moment-based image shape analysis and its application. **Pattern Recognition Letters**, Elsevier, v. 21, n. 2, p. 169–177, 2000. Citation on page [201](#).

ZHOU, H.; LIPOWSKY, R. Dynamic pattern evolution on scale-free networks. **Proceedings of the National Academy of Sciences of the United States of America**, National Acad Sciences, v. 102, n. 29, p. 10052–10057, 2005. Citation on page [84](#).

BIRTH-SURVIVAL NETWORK AUTOMATA: PARTIALLY-ORDERED SETS & MEASURES

Figure 104 – Average Lyapunov exponent over 50 simulations of the BSNA on a small-world network. The rules are presented as a Hasse diagram, using the Inclusion relationship (5.3.1). Parameters $N = 400$, $\langle k \rangle = 8$ and $R = 2$.

Figure 105 – Average Lyapunov exponent over 50 simulations of the BSNA on a scale-free network. The rules are presented as a Hasse diagram, using the Inclusion relationship (5.3.1). Parameters $N = 400$, $\langle k \rangle = 8$ and $R = 2$.

Figure 106 – Average Lyapunov exponent over 50 simulations of the BSNA on a geographical network. The rules are presented as a Hasse diagram, using the Inclusion relationship (5.3.1). Parameters $N = 400$, $\langle k \rangle = 8$ and $R = 2$.

Figure 107 – Average Lempel-Ziv complexity over 50 simulations of the BSNA on a small-world network. The rules are presented as a Hasse diagram, using the Inclusion relationship (5.3.1). Parameters $N = 400$, $\langle k \rangle = 8$ and $R = 2$.

Figure 108 – Average Lempel-Ziv complexity over 50 simulations of the BSNA on a scale-free network. The rules are presented as a Hasse diagram, using the Inclusion relationship (5.3.1). Parameters $N = 400$, $\langle k \rangle = 8$ and $R = 2$.

Figure 109 – Average Lempel-Ziv complexity over 50 simulations of the BSNA on a geographical network. The rules are presented as a Hasse diagram, using the Inclusion relationship (5.3.1). Parameters $N = 400$, $\langle k \rangle = 8$ and $R = 2$.

Figure 110 – Average word length over 50 simulations of the BSNA on a small-world network. The rules are presented as a Hasse diagram, using the Inclusion relationship (5.3.1). Parameters $N = 400$, $\langle k \rangle = 8$ and $R = 2$.

Figure 111 – Average word length over 50 simulations of the BSNA on a scale-free network. The rules are presented as a Hasse diagram, using the Inclusion relationship (5.3.1). Parameters $N = 400$, $\langle k \rangle = 8$ and $R = 2$.

Figure 112 – Average word length over 50 simulations of the BSNA on a geographical network. The rules are presented as a Hasse diagram, using the Inclusion relationship (5.3.1). Parameters $N = 400$, $\langle k \rangle = 8$ and $R = 2$.

Inclusion relationship (5.3.1). Parameters $N = 400$, $\langle k \rangle = 8$ and $R = 3$.

Figure 115 – Average Lyapunov exponent over 50 simulations of the BSNA on a geographical network. The rules are presented as a Hasse diagram, using the Inclusion relationship (5.3.1). Parameters $N = 400$, $\langle k \rangle = 8$ and $R = 3$.

Figure 116 – Average Lempel-Ziv complexity over 50 simulations of the BSNA on a regular network. The rules are presented as a Hasse diagram, using the Inclusion relationship (5.3.1). Parameters $N = 400$, $\langle k \rangle = 8$ and $R = 3$.

Figure 117 – Average Lempel-Ziv complexity over 50 simulations of the BSNA on a random network. The rules are presented as a Hasse diagram, using the Inclusion relationship (5.3.1). Parameters $N = 400$, $\langle k \rangle = 8$ and $R = 3$.

Figure 118 – Average Lempel-Ziv complexity over 50 simulations of the BSNA on a small-world network. The rules are presented as a Hasse diagram, using the Inclusion relationship (5.3.1). Parameters $N = 400$, $\langle k \rangle = 8$ and $R = 3$.

Figure 119 – Average Lempel-Ziv complexity over 50 simulations of the BSNA on a scale-free network. The rules are presented as a Hasse diagram, using the Inclusion relationship (5.3.1). Parameters $N = 400$, $\langle k \rangle = 8$ and $R = 3$.

Figure 120 – Average Lempel-Ziv complexity over 50 simulations of the BSNA on a geographical network. The rules are presented as a Hasse diagram, using the Inclusion relationship (5.3.1). Parameters $N = 400$, $\langle k \rangle = 8$ and $R = 3$.

Figure 121 – Average word length over 50 simulations of the BSNA on a regular network. The rules are presented as a Hasse diagram, using the Inclusion relationship (5.3.1). Parameters $N = 400$, $\langle k \rangle = 8$ and $R = 3$.

Figure 123 – Average word length over 50 simulations of the BSNA on a small-world network. The rules are presented as a Hasse diagram, using the Inclusion relationship (5.3.1). Parameters $N = 400$, $\langle k \rangle = 8$ and $R = 3$.

Figure 124 – Average word length over 50 simulations of the BSNA on a scale-free network. The rules are presented as a Hasse diagram, using the Inclusion relationship (5.3.1). Parameters $N = 400$, $\langle k \rangle = 8$ and $R = 3$.

Figure 125 – Average word length over 50 simulations of the BSNA on a geographical network. The rules are presented as a Hasse diagram, using the Inclusion relationship (5.3.1). Parameters $N = 400$, $\langle k \rangle = 8$ and $R = 3$.

RPS DYNAMICS ON NETWORKS

This appendix complements the results discussed in Chapter 7. We present the corresponding plots of the measures extracted from the evolution of the rock-paper-scissors on the different network topologies analyzed in this thesis. The referred measures are (i) the probability of extinction (section B.1) and (ii) the average time to the first extinction (section B.2). We present all the combinations of network model (random, small-world, geographic) for different network sizes ($N = 400, 900, 1600, 10000$) and different average degree ($\langle k \rangle = 4, 6, 8, 10$).

B.1 Probability of Extinction

Figure 126 – Probability of extinction as a function of mobility for random networks. Each plot corresponds to a different network size: (a) $N = 400$, (b) $N = 900$, (c) $N = 1600$ and (d) $N = 10000$. Within each plot, each curve corresponds to different average degrees, $\langle k \rangle$.

Source: Elaborated by the author.

Figure 127 – Probability of extinction as a function of mobility for geographic networks. Each plot corresponds to a different network size: (a) $N = 400$, (b) $N = 900$, (c) $N = 1600$ and (d) $N = 10000$. Within each plot, each curve corresponds to different average degrees, $\langle k \rangle$.

Source: Elaborated by the author.

Figure 128 – Probability of extinction as a function of mobility for small-world networks. Each plot corresponds to a different network size: (a) $N = 400$, (b) $N = 900$, (c) $N = 1600$ and (d) $N = 10000$. Within each plot, each curve corresponds to different average degrees, $\langle k \rangle$.

Source: Elaborated by the author.

B.2 Time to extinction

Figure 129 – Average time to the first extinction as a function of mobility for random networks. Each plot corresponds to a different network size: (a) $N = 400$, (b) $N = 900$, (c) $N = 1600$ and (d) $N = 10000$. Within each plot, each curve corresponds to different average degrees, $\langle k \rangle$.

Source: Elaborated by the author.

Figure 130 – Average time to the first extinction as a function of mobility for geographical networks. Each plot corresponds to a different network size: (a) $N = 400$, (b) $N = 900$, (c) $N = 1600$ and (d) $N = 10000$. Within each plot, each curve corresponds to different average degrees, $\langle k \rangle$.

Source: Elaborated by the author.

Figure 131 – Average time to the first extinction as a function of mobility for small-world networks. Each plot corresponds to a different network size: (a) $N = 400$, (b) $N = 900$, (c) $N = 1600$ and (d) $N = 10000$. Within each plot, each curve corresponds to different average degrees, $\langle k \rangle$.

Source: Elaborated by the author.

ILI PREDICTION PLOTS FOR THE BELGIAN EPIDEMIC SEASONS

C.1 Prediction of ILI numbers using the global approach

This appendix presents the results concerning the predicted numbers of ILI cases for the following epidemic seasons: 2011–2012, 2012–2013, 2013–2014 and 2015–2016. The figures presented here complements the results of Chapter 10 for the global analysis, on which the data of all the Belgian arrondissements were considered. Similarly to the plots presented in that chapter, Figs 132, 133, 134 and 135 show the predicted numbers of ILI cases for season 2010–2011 for different starting weeks of the simulated period, as well as the normalized ILI in the population. We can observe, for instance, that predictions for season 2013–2014 (Fig. 134) were the ones with the best agreement between the normalized and the predicted ILI numbers. Season 2012–2013 (Fig. 133) presents a steep increase until the ILI peak and the predictions before the peak were no good independently of the start of the simulated period.

Figure 132 – Simulated (gray) and normalized (black) numbers of ILI cases in Belgium for season 2011–2012. The blue dot corresponds to the epidemic onset and the red dot to the current week (curr. week) up to which the normalized ILI data were used for model calibration. Simulated period starting (a) the week before the onset (b) 2 weeks before the onset and (c) 3 weeks before the onset.

Source: Elaborated by the author.

Figure 133 – Simulated (gray) and normalized (black) numbers of ILI cases in Belgium for season 2012–2013. The blue dot corresponds to the epidemic onset and the red dot to the current week (curr. week) up to which the normalized ILI data were used for model calibration. Simulated period starting (a) the week before the onset (b) 2 weeks before the onset and (c) 3 weeks before the onset.

Source: Elaborated by the author.

Figure 134 – Simulated (gray) and normalized (black) numbers of ILI cases in Belgium for season 2013–2014. The blue dot corresponds to the epidemic onset and the red dot to the current week (curr. week) up to which the normalized ILI data were used for model calibration. Simulated period starting (a) the week before the onset (b) 2 weeks before the onset and (c) 3 weeks before the onset.

Source: Elaborated by the author.

Figure 135 – Simulated (gray) and normalized (black) numbers of ILI cases in Belgium for season 2015–2016. The blue dot corresponds to the epidemic onset and the red dot to the current week (curr. week) up to which the normalized ILI data were used for model calibration. Simulated period starting (a) the week before the onset (b) 2 weeks before the onset and (c) 3 weeks before the onset.

Source: Elaborated by the author.

C.2 Prediction of ILI numbers using the local approach

In this section we present the analysis performed for the Brussels district for two other districts: Antwerpen and Gent. Similarly to the plots presented in section 10.6.2.1, we can observe the predictions obtained using the confidence region approach (Figs. 136, 137, 138 and 139), and, the network of commuting data together with the confidence region (Figs. 140, 141, 142 and 143).

Figure 136 – Simulated (gray) and normalized (black) numbers of ILI cases in Antwerpen district for season 2010–2011 up to 2012–2013 obtained through of the confidence region approach. The blue dot corresponds to the epidemic onset and the red dot to the current week (curr. week) up to which the normalized ILI data were used for model calibration. The simulated period starts one week before the onset.

Source: Elaborated by the author.

Figure 137 – Simulated (gray) and normalized (black) numbers of ILI cases in Antwerpen district for season 2013–2014 up to 2015–2016 obtained through of the confidence region approach. The blue dot corresponds to the epidemic onset and the red dot to the current week (curr. week) up to which the normalized ILI data were used for model calibration. The simulated period starts one week before the onset.

Source: Elaborated by the author.

Figure 138 – Simulated (gray) and normalized (black) numbers of ILI cases in Gent district for season 2010–2011 up to 2012–2013 obtained through of the confidence region approach. The blue dot corresponds to the epidemic onset and the red dot to the current week (curr. week) up to which the normalized ILI data were used for model calibration. The simulated period starts one week before the onset.

Source: Elaborated by the author.

Figure 139 – Simulated (gray) and normalized (black) numbers of ILI cases in Gent district for season 2013–2014 up to 2015–2016 obtained through of the confidence region approach. The blue dot corresponds to the epidemic onset and the red dot to the current week (curr. week) up to which the normalized ILI data were used for model calibration. The simulated period starts one week before the onset.

Source: Elaborated by the author.

Figure 140 – Simulated (gray) and normalized (black) numbers of ILI cases in Antwerpen district for season 2010–2011 up to 2012–2013 obtained through of the confidence region approach and the commuting data. The blue dot corresponds to the epidemic onset and the red dot to the current week (curr. week) up to which the normalized ILI data were used for model calibration. The simulated period starts one week before the onset.

Source: Elaborated by the author.

Figure 141 – Simulated (gray) and normalized (black) numbers of ILI cases in Antwerpen district for season 2013–2014 up to 2015–2016 obtained through of the confidence region approach and the commuting data. The blue dot corresponds to the epidemic onset and the red dot to the current week (curr. week) up to which the normalized ILI data were used for model calibration. The simulated period starts one week before the onset.

Source: Elaborated by the author.

Figure 142 – Simulated (gray) and normalized (black) numbers of ILI cases in Gent district for season 2010–2011 up to 2012–2013 obtained through of the confidence region approach and the commuting data. The blue dot corresponds to the epidemic onset and the red dot to the current week (curr. week) up to which the normalized ILI data were used for model calibration. The simulated period starts one week before the onset.

Source: Elaborated by the author.

Figure 143 – Simulated (gray) and normalized (black) numbers of ILI cases in Gent district for season 2013–2014 up to 2015–2016 obtained through of the confidence region approach and the commuting data. The blue dot corresponds to the epidemic onset and the red dot to the current week (curr. week) up to which the normalized ILI data were used for model calibration. The simulated period starts one week before the onset.

Source: Elaborated by the author.