DOCTOR OF SCIENCES

The Complex System of International Relations

VINA CERVANTES, Viviana Marcela

Award date:
2020

Awarding institution:
University of Namur

UNIVERSITÉ DE NAMUR

FACULTÉ DES SCIENCES

DÉPARTEMENT DE MATHÉMATIQUE

The Complex system of International Relations

A thesis submitted by
Viviana Marcela Viña Cervantes
in fulfillment of the requirements for the degree of Doctor in Sciences

Composition of the Jury

Renaud LAMBIOTTE (Supervisor)
Michele COSCIA (Supervisor)
Timoteo CARLETTI
Jean-Charles DELVENNE
Joseph WINKIN (President)

December 2020

Cover design : ©Presses universitaires de Namur

Rempart de la Vierge, 13
B-5000 Namur (Belgique)

Printed in Belgium.

ISBN : 978-2-39029-012-4
Registration of copyright : D/2018/1881/18

Université de Namur
Faculté des Sciences
rue de Bruxelles, 61, B-5000 Namur (Belgique)

Abstract

The study of international relations is becoming more and more relevant in society today. Effective communication between countries is essential for making beneficial relationships and ensuring a safer world as a result.
For different reasons but all framed in the study of the complexity of international relations, we have created a multilayer network where the nodes represent the countries and the links are the different types of relations that the countries can establish among themselves.
The main goal of this work is not the study and understanding of historical, economic or political reasons why the relations between the countries are in a certain way but the relations themselves and use them as an input for predictive models.

In the development of this thesis, we have addressed 2 fundamental problems regarding relations between countries. In both cases, we have made use of the tools that the study of complex networks offer us, in order to study, in each case, a specific problem or situation.

In the first case, we consider the global trade system as a dynamic ecosystem in which exporters struggle for resources in the same than an organism in its attempt to monopolize a given subset of resources in an ecosystem, in this case: the markets in which they export.
In this work, we adopt a multilayer network approach to describe this struggle. We connect two countries with a directed link if the source country's appearance in a market correlates with the target country's disappearing. A market is defined as a country- product combination in a given decade. Each market is a layer in the network. We show that, by analyzing the countries' network roles in each layer, we are able to classify them as out-competing, transitioning or displaced. This classification is a meaningful one: when testing the future export patterns of these countries, we show that out-competing countries have distinctly stronger growth rates than the other two classes.

In the second case, we have created a model for forecasting global conflicts. One of the oldest and most important forecasting tasks focuses on international conflicts. *Can we predict when and where a new war will start? Can we estimate how big the conflict will be?* These questions are particularly relevant today, since developments in network analysis allow us to use a com-

plex perspective of the international relations system. In this case, we have not focused on estimating the size or global impact of conflicts, but we are more interested in finding out whether international conflicts follow social balance. In order to do that, we have modeled international relations as a multilayer network in which each layer is a different type of relationship that two countries can establish with each other: trade, migration, sanctions, etc. Conflicts (wars) is one of these layers and our prediction goal, making this a multi-layer link prediction problem. We have also considered the sign of relationships, in this way in this multilayer system, some networks represent positive relationships while others are negative.
In addition to demonstrating that conflicts between countries follow the laws of social balance, our model is able to predict with high precision not only how conflictive a year will be but also the countries that will be involved.

Thèse de doctorat en Sciences Mathématiques (Ph.D. thesis in Mathematics)
Date: 18/12/2020
Département de Mathématique
Promoteurs (Advisors): Renaud LAMBIOTTE, Michele COSCIA

Acknowledgments

First of all, I would like to express my sincere gratitude to my advisors Dr. Renaud Lambiotte and Dr. Michele Coscia. Since the beginning of my doctoral studies, they have always allowed and helped me to develop my own ideas and to become an independent researcher. I am very grateful for having the opportunity to work and learn from them, it has truly been a privilege.

I would also like to thank the rest of the jury members, Dr. Joseph Winkin, Dr. Jean-Charles Delvenne, and Dr. Timoteo Carletti, for their detailed evaluation of my work, their suggestions, and appropriate questions. I want to highlight in a very special way the support and valuable friendship of Teo during all these years. To my colleagues with whom I had the opportunity to collaborate, discuss, or just scribble on a blackboard, thank you very much.

To my boyfriend, Bastiaan, thank you for helping me to make so many dreams come true, including this one. Thank you for being the shoulder that I can always lean on. I was so lucky to have found you on the train that day.
To my children, Thomas and William, I promise you guys that all the time together that we have sacrificed, all those plans canceled because I had to work on this project will be worth it. I am so grateful for having you in my life, you have made everything easier, you are my biggest motivation.

A mis padres y hermanos, gracias por su apoyo desde la distancia. Gracias por hacerme sentir cerca a ustedes, especialmente en los últimos meses y ayudarme a lidiar con ese guayabo que no se va. Sobretodo muchas gracias por demostrarme siempre lo orgullosos que están de mí.

To my colleague, and close friend Carlos Quintero, thank you for all your help and collaboration in all the technical aspects of my thesis. Thank you, Charlie, because I had always counted on you at any time, I really appreciate it.

I want to thank my best friends Ini and Maria for their invaluable friendship, their unconditional support, and for always celebrating even the smallest of my achievements.

Finally, to all the people who have become important to me since I came to live in this country, thank you very much for being part of this.

"Chaos was the law of nature; Order was the dream of man."

Henry Adams

Contents

1 Introduction **1**
- 1.1 Why study International relations? . . . 1
- 1.2 Networks and International relations . . . 2
- 1.3 Economic networks and Ecology . . . 4
- 1.4 Networks and International Conflicts (Wars) . . . 4
- 1.5 Contributions . . . 5
- 1.6 Structure of the Thesis . . . 6
- 1.7 Publications . . . 7
- 1.8 International Conferences . . . 7
- 1.9 Schools and visits . . . 8

2 Theoretical Background **9**
- 2.1 Fundamentals of Network Theory . . . 9
  - 2.1.1 Networks . . . 9
  - 2.1.2 Weighted networks . . . 10
- 2.2 Degree distribution . . . 11
  - 2.2.1 Degree . . . 11
  - 2.2.2 Degree distribution . . . 11
- 2.3 Adjacency matrices . . . 13
- 2.4 Clustering coefficient . . . 15
- 2.5 Node Roles . . . 18
- 2.6 Node similarity . . . 20
- 2.7 Link Prediction . . . 23
- 2.8 Signed networks . . . 25
- 2.9 Multilayer networks . . . 27
  - 2.9.1 Incidence matrices for Multilayer networks . . . 29

2.10 Dynamic Graphs 30
2.10.1 Representations 31

3 Introduction to Machine Learning 33
3.1 Types of Learning 34
3.2 Supervised Learning algorithms 36
3.3 Unsupervised Learning Algorithms 40
3.3.1 Clustering Algorithms 41
3.4 Precision and Recall 43
3.5 Confusion Matrix 45
3.6 Capacity, Overfitting and Underfitting 46
3.7 Feature Importance 50
3.7.1 Feature Importance from Decision Trees 50

4 The Struggle for Existence in the World Market Ecosystem 53
4.1 Methods 56
4.1.1 Data & Cleaning 56
4.1.2 Inferring Competition Relationships 58
4.1.3 Detecting Roles 61
4.2 Results 65
4.2.1 Competition Network Statistical Analysis 65
4.2.2 Role Clusters 68
4.2.3 Prediction 70
4.2.4 Validation 77
4.3 Discussion 81

5 Forecasting International Conflicts via Machine Learning and Multilayer Social Balance 83
5.1 Model 84
5.2 Results 87
5.2.1 Social Balance 87
5.2.2 Forecasting Conflicts 90
5.3 Discussion 97
5.4 Materials and Methods 99
5.4.1 Data 99
5.4.2 Preprocessing 101
5.5 Additional Results 103
5.5.1 Predictions vs Conflicts 103
5.5.2 Conflicts vs Visa requirement 106
5.5.3 Conflicts vs Sanctions 106

6 Discussion **109**

Bibliography **114**

Chapter 1

Introduction

1.1 Why study International relations?

The entire world can be considered as constructed by one large network in which the nodes are the countries and the edges can represent any kind of relationship that the countries can establish between them. But, the first question that one can ask is: *Why is it important to study International relations?* (Petermann 2006). In a world where terrorist attacks come without warning and where poverty kills thousands of people each day only because of international system's mistakes, the study of the international relations is not just important but essential study field.
We are all part of international relations because of our identities, religion, cultural backgrounds and the places where we live (Sharp 2018). Modern international relations give us deep cultural understanding that is a foundation for interaction with cultures with different values and beliefs.

The study of international relations is becoming more and more relevant in society today. Effective communication between countries is essential for making beneficial relationships and ensuring a safer world as a result. In short, international relations are all about power and weakness, war and peace, conflicts and cooperation.
It is important to make clear that the main goal of this work is not the study and understanding of historical, economic or political reasons why the relations between the countries are in a certain way but the relations themselves and use them as an input for predictive models.

1.2 Networks and International relations

International relations research has regarded networks as a particular mode of organization, distinguished from markets or state hierarchies. In contrast, network analysis permits the investigation and measurement of network structures like emergent properties of persistent patterns of relations among agents that can define, enable, and constrain those agents. Network analysis offers a set of theories, typically drawn from contexts outside international relations, that relate structures to outcomes.

The value of network analysis in international relations has been demonstrated in the precise description of international networks, investigation of network effects on key international outcomes, testing of existing network theory in the context of international relations, and development of new sources of data.

Network analysis complements existing structural approaches to international relations that focus on actor attributes and static equilibria. Instead, it emphasizes how material and social relationships create structures among actors through dynamic processes. It also provides methods for measuring these structures, allows for the operationalization of processes such as socialization and diffusion, and opens new avenues for reconsidering core concepts in international relations.

In recent years, there has been more and more interest in studying economy-related questions by means of network science (Barabási et al. 2016, Borgatti et al. 2009, Emmert-Streib et al. 2018, Jackson 2010, Schweitzer et al. 2009). A reason for this interest builds on the realization that the behaviour of the economy cannot be investigated by individually studying the constituting components of it but only by considering the interplay between all relevant parts. This is in strong contrast to the standard economic theory (Allen & Babus n.d., Kirman 1997, Meng et al. 2017, Nagurney & Siokos 1997).

The discipline of international relations has not, until recently, treated networks as structures that can constrain and enable individual agents and influence international outcomes (Keohane & Nye 1974, 1977). Networks are significant actors in international politics and represent a specific mode of international interaction and governance.

From a practical point of view, the digitalization of our society, e.g., of the stock market or the availability of business records, enabled the empirical estimation and creation of different types of economic networks. This is similar to other fields like biology, chemistry or sociology (Bonchev 1991, Dehmer et al. 2011, Emmert-Streib & Dehmer 2011, Freeman 2004, Newman 2018*a*). That means, without the need of making theoretical assumptions about the structure of economic networks, data can be used for their construction. This is in contrast to, e.g., simulation-based approaches allowing to generate network topologies with certain characteristics, e.g., scale-free or small-world networks (Barabási & Albert 1999*a*, Watts & Strogatz 1998*b*, Dehmer & Emmert-Streib 2009).

Network analysis concerns relationships defined by links among nodes (or agents). Nodes can be individuals or corporate actors, such as organizations and states. Network analysis addresses the associations among nodes rather than the attributes of particular nodes. It is grounded in three principles: nodes and their behaviours are mutually dependent, not autonomous; ties between nodes can be channels for transmission of both material (for example, weapons, money, or disease) and non-material products (for example, information, beliefs, and norms); and persistent patterns of association among nodes create structures that can define, enable, or restrict the behaviour of nodes (Wasserman et al. 1994). In network analysis, networks are defined as any set or sets of ties between any set or sets of nodes; unlike the study of network forms of organization, no assumptions are made about the homogeneity or other characteristics of the nodes or ties. Consequently, network analysis can be used to analyse any kind of ties, including market and hierarchical relations. Beyond these basic principles, network analysis enables calculation and mapping of structural properties of nodes, groups, or the entire network; predictions regarding the creation, growth, and dissolution of these networks; and investigation of the effects of networks on actors' behaviour.

Network analysis has a long history in the behavioural sciences, although its incorporation into international relations has been slow and uneven. Moreover, studies that have used the tools of network analysis have not always produced significant insights; a few have emphasized method over substance, and most have mapped but not explained some aspect of international politics, assuming rather than demonstrating the causal mechanisms through which networks actually constrain and enable their members.

The value of network analysis has already been demonstrated in more precise description of international networks, investigation of network effects on key international outcomes, tests of existing network theory in the context of international relations, and development of new sources of data.

1.3 Economic networks and Ecology

According to Robert May and collaborators in (May et al. 2008), there is common ground in analyzing financial systems and ecosystems, especially in the need to identify conditions that dispose of a system to be knocked from seeming stability into another, less happy state.
But, to what extent can the study of ecosystems inform the design of financial networks in, for instance, their robustness against perturbation? Ecosystems are robust by their continued existence. They have survived eons of change — continental drift, climate fluctuations, movement and evolution of constituent species — and show some remarkable constancies in a structure that have persisted for hundreds of millions of years: witness, for example, the constancy in predator-prey ratios in different situations (Bambach et al. 2002).

Identifying structural attributes shared by these diverse systems that have survived rare systemic events, or have indeed been shaped by them, could provide clues about which characteristics of complex systems correlate with a high degree of robustness.

Following this argument, in Chapter 4 we propose an analogy between predators competing for prey in an ecosystem and exporting countries competing for market dominance.

1.4 Networks and International Conflicts (Wars)

The enormous costs of war make it imperative to understand the conditions under which wars can be prevented. Forecasting large geopolitical changes is a key task to improve the resilience of human society. One of the oldest and most important forecast tasks (not without criticism (Cederman Weidmann 2017)) focuses on international conflicts (Richardson 1960, Holsti Holsti 1991, Pettersson Wallensteen 2015).

Between 1823 and 2003, 40% of wars with more than 1,000 casualties involved more than two countries, and many of the most destructive (e.g., the World Wars, Korean War, Vietnam, First and Second Congo Wars, etc.) involved multilateral conflicts. Most importantly, this is a network problem (Jackson & Nei 2015).

In Chapter 5, we frame conflict forecasting as a link prediction problem (Liben-Nowell Kleinberg 2007, Lü Zhou 2011). Specifically, we model international relations as a multilayer network (Kivelä et al. 2014b, Berlingerio et al. 2013).

1.5 Contributions

The theoretical background and connecting thread of the works presented in this thesis is the study and analysis of the effects on the structure and dynamics of networks after the appearance and disappearance of links. We have considered both cases, positive and negative links. Depending on the symbol (positive or negative) of these links, we have brought their appearance or disappearance to a political-economic context.

In the elaboration of this thesis, we have addressed two relevant problems, both of them framed in the study of international relations. In both cases, we consider the world as a large network composed of countries and different types of relationships between them.

The first of these is the creation of a model that allows us to predict which countries will lead the export market in the coming years in terms of how much they are capable of displacing (disappearance of trade links) others from the international market. The second is the creation of a model to predict conflicts (appearance of negative links) between countries using a multilayer network system and machine learning.

Below you will find a description of how this thesis is organized.

1.6 Structure of the Thesis

Chapters 2 and 3:

Theoretical Background and **Introduction to Machine Learning**.
In the first two chapters, we provide an introduction of the most relevant concepts that we used in the development of this Thesis. In Chapter 2 we provide a brief description of the basic concepts of the Complex Systems Theory, then we introduce the Multilayer systems and some of their properties. Chapter 3 is a short introduction to Machine Learning.

Chapter 4:

The Struggle for Existence in the World Market Ecosystem.
In this chapter we propose a model based on an analogy between the competition between predators in a ecosystem and the competition between exporter countries for leading a market.
In the global trade market, there is constant competition between exporting countries. We consider exporter countries like predators with the pair importer-product as preys. We are particularly interested in cases when this competition affects dramatically the export of one of the countries.
We focus on longitudinal, multiplex data on commercial relations, to test the presence of predator countries in the course of time. We consider their relation to the complexity of products and national economies, predict which countries are leading the global trade network and which countries are at risk in different dimensions.

Chapter 5:

Forecasting International Conflicts via Machine Learning and Multilayer Social Balance.
In this chapter we propose a model using machine learning and social balance theory in order to predict wars between countries and to proof that that international conflicts follow social balance. In this chapter, by conflicts, we mean specifically wars.

The motivation of this work is that understanding the determinants of war can lead us to create a more resilient global society. Here, we model international relationships as a multilayer signed network: each type of relationship between countries – trade, conflicts, sanctions – is a layer, and layers can be

of two types: positive or negative. We then frame conflict prediction as a multilayer signed link prediction problem: given the status of the network at time t, will two countries connect in the conflict layer at time t + 1? We propose to extend social balance theory to the multilayer signed network scenario and we use it as the key feature in a machine learning framework.
We find that multilayer social balance increases the accuracy of the machine learning task, and that the average time a false positive takes to turn into a true positive is 1.45 years.

1.7 Publications

Journal publications

- V. Vina-Cervantes, M. Coscia, and R. Lambiotte. The struggle for existence in the world market ecosystem. PloS one, 13(10):e0203915, 2018.

In redaction

- V. Vina-Cervantes, M. Coscia, V. Salnikov and R. Lambiotte. Forecasting of international conflicts via Machine Learning and multilayer social balance.
- V.Vina-Cervantes, M. Coscia, T. Carletti, F. Gargiulo, R. Lambiotte. Fingerprints in the world-trade market.

1.8 International Conferences

- Conference on Complex Systems 2016, Amsterdam, 19-22 Sept 2016
- Complenet 2017, Dubrovnik, 21-24 March 2017
- Crossroads in Complex systems, 5-8 June 2017
- 7th International Conference on Complex Networks and their applications, 11-13 Dec 2018
- Latin American Conference 2.0 on Complex Networks, 5-9 Aug 2019

1.9 Schools and visits

- CISM / CECI 2015/2016, 21-22 Oct 2015
- Interdisciplinary training school for PhD students 2016: L'alimentation en question, 17-19 May 2016
- 2nd Unamur UCL Winter school on Networks in Economics and Finance, 12-14 Dec 2016
- Mediterranean school on Complex Networks, 3-8 Sept 2017
- Research internship University of Vienna with Dr. Stephan Turner, 12-24 April 2016
- Research internship at Harvard University with Dr. Michele Coscia, 28 Jun-7 Jul 2017

Chapter 2

Theoretical Background

2.1 Fundamentals of Network Theory

2.1.1 Networks

In the most basic sense, a *Network* is any collection of objects in which some pairs of these objects are connected by links. This definition is very flexible: depending on the setting, many different forms of relationships or connections can be used to define links (Easley & Kleinberg 2010). Because of this flexibility, it is easy to find networks in many domains.

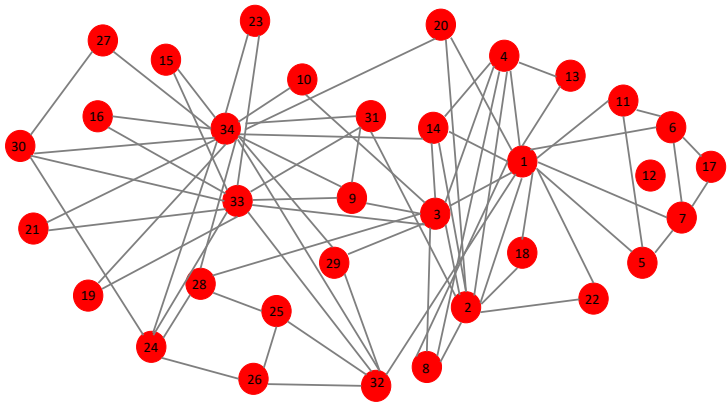
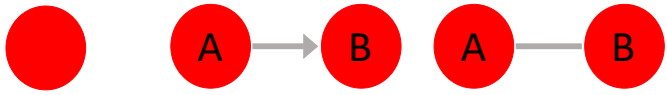


Figure 2.1 – The social network of friendships within a 34-person karate club (Zachary 1977)

As a good example of what a network looks like, Fig 2.1 depicts the social network among 34 people in a university karate club studied by the anthropologist Wayne Zachary in the 1970s. The people are represented by small circles, with lines joining the pairs of people who are friends outside the context of the club.

The interacting components of a network, are usually depicted as a graph of vertices (*Nodes*) connected by *edges*.



(a) (b) (c)

Figure 2.2 – (a) A node. (b) A directed edge, there is only a path from **A**, the origin, to **B**, the destination. (c) An undirected edge, the path between **A** and **B** is bidirectional, meaning that origin and destination are not fixed.

The use of a graph to represent Networks, is a very practical and easy way to represent any kind of interactions between any kind of elements and has applications in many disciplines including statistical physics, particle physics, computer science, electrical engineering, biology, economics, finance, operations research, climatology, ecology and sociology. In our particular case, we are going to use networks in order to represent and study International relations. In this case, the nodes will represent countries and the edges any kind of relations between countries.

2.1.2 Weighted networks

Since all the connections in a network are not necessarily equally important or equally strong, there are networks with and extra parameter that indicates the weight of each connection. These networks are known as *weighted networks*.

Weighted networks are often used to model real objects and processes, these can be also directed. For example, the network in Fig 2.3 can be considered as a map, where the nodes are cities and the edges are roads. The weight of each edge is the distance between two cities.

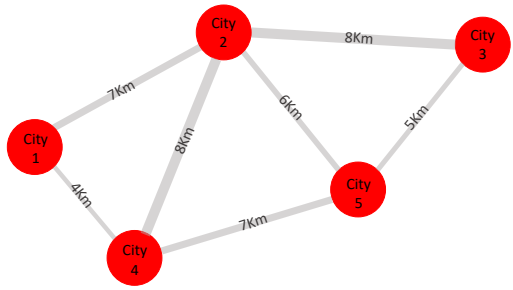


Figure 2.3 – Simple weighted network in which the nodes represent cities and the edges the different distances between each other

2.2 Degree distribution

2.2.1 Degree

The number of connections of each node is known as the *degree*. If a network is directed, meaning that edges point in one direction from one node to another node, then nodes have two different degrees, the *In-degree*, which is the number of incoming edges, and the *Out-degree*, which is the number of outgoing edges.

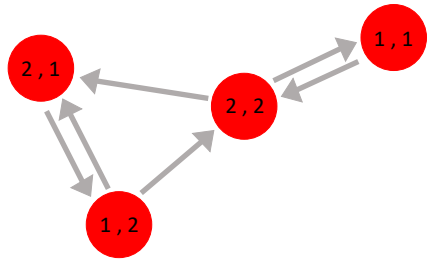


Figure 2.4 – A directed network with links labeled (indegree, outdegree)

2.2.2 Degree distribution

The degree of a node only provides information about that specific node. The average degree of a network gives information about the whole structure. By counting how many nodes have each degree, we form the *Degree distribution* $P(k)$.

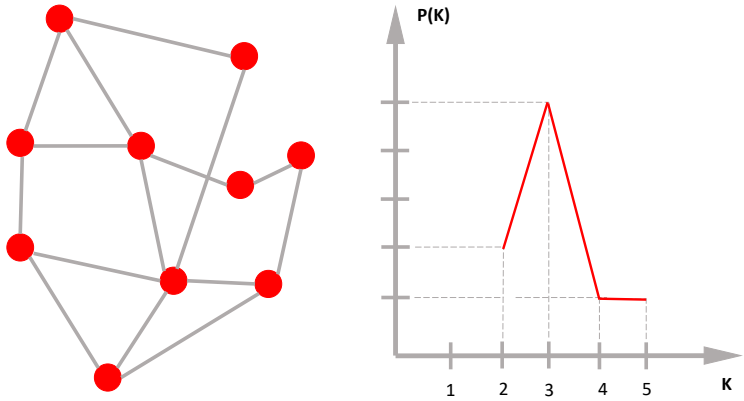


Figure 2.5 – Degree distribution plot (right) of the network on the left.

The degree distribution of a network is then defined to be the fraction of nodes in the network with degree k, which also means, the probability for a node to have a degree equal to k. Thus, if there are n nodes in total in a network and n_k of them have degree k, we have $P(k) = n_k/n$. This clearly captures only a small amount of information about a network, but that information still gives important clues into the structure of a network. For example, in the simplest types of networks, one would find that most nodes in the network had similar degrees (see the first pair of plots, below). However, real-world networks usually have very different degree distributions. In a real-world network, most nodes have a relatively small degree, but a few nodes will have a very large degree, being connected to many other nodes. These large-degree nodes are often referred to as hubs, in analogy to the transportation networks such as one connecting airports, where some very large hub airports have connections to many others.

Some networks, notably the Internet, the world wide web, and some social networks are found to have degree distributions that approximately follow a power law: $P(k) \sim k^{(-\gamma)}$ where γ is a constant. Such networks are called scale-free networks and have attracted particular attention for their structural and dynamical properties, (Barabási & Albert 1999*b*, Albert & Barabási 2000, Dorogovtsev et al. 2001).

- CCDF

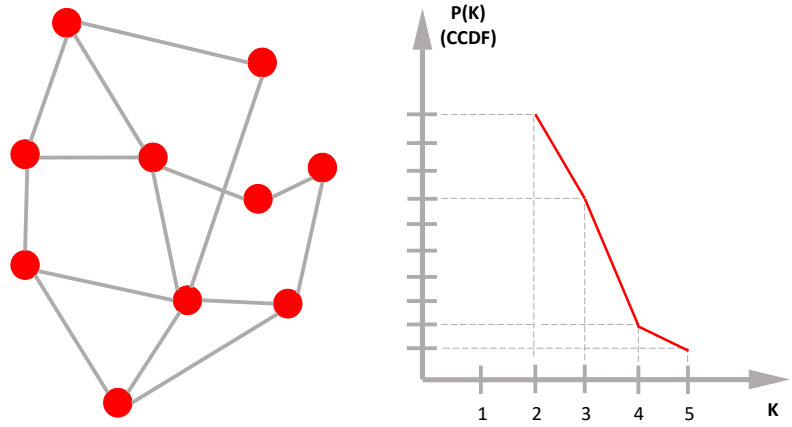


Figure 2.6 – CCDF plot (right) of the network on the left. Note that this is the same network of Fig 2.5

We mentioned a few lines above that the degree distribution was the probability for a node to have a degree equal to k. This concept is not often used in the research but the *Cumulative Degree Distribution CDF*, and even more, the *Complementary Cumulative Degree Distribution CCDF* (Figure 2.6).

The CCDF describes the probability that a random node n will have a degree equal to or higher than k.

2.3 Adjacency matrices

An *Adjacency Matrix* is a square matrix used to represent a network. In an adjacency matrix, a grid is set up that lists all the nodes on both the X-axis (horizontal) and the Y-axis (vertical). Then, values are filled in to the matrix to indicate if there is or is not an edge between every pair of nodes. Typically, 0 indicates no edge and 1 indicates an edge.

Notice a couple of things about the matrix in Fig 2.7. First, the diagonal is all zeroes because there are no edges between a node and itself in our example. Some networks do allow for self-loops. For example, in an email network, if a person emails himself, there could be a link from one node to itself, and thus there would be a 1 on the diagonal. Second, the matrix is symmetric. The numbers in the first row are the same as the numbers in the first column. The

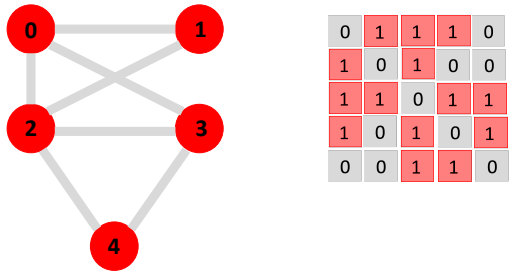


Figure 2.7 – Graph and matrixial representation of a undirected network. The adjacency matrix is a (0,1)-matrix with zeros on its diagonal. The adjacency matrix is symmetric.

numbers in the second row are the same as the numbers in the second column. This is because the graph is *undirected*.

If we have a directed network, the matrix will not necessarily be symmetric. For example, consider the small network in Fig 2.8. In this case, we have some edges without a reciprocal one.

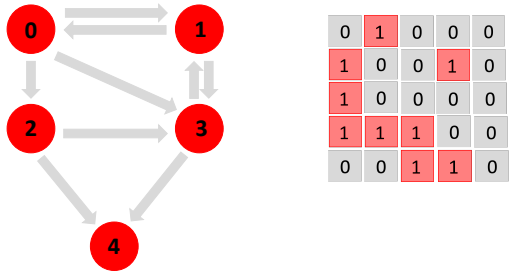


Figure 2.8 – Graph and matricial representation of a directed network. The adjacency matrix is a (0,1)-matrix with zeros on its diagonal. The adjacency matrix isn't symmetric.

In the adjacency matrix, the rows and columns are each defined to be the vertices in the graph. According to that, the adjacency matrix is defined as

$$A_{ij} =$$

$$\begin{cases} 1 & \text{if there exists an edge connecting the nodes } u_i \text{ and } u_j, \\ 0 & \text{Otherwise} \end{cases}$$

2.4 Clustering coefficient

In graph theory, the *Clustering coefficient* is a measure of the degree to which the nodes in a graph tend to cluster together. Evidence suggests that in most real-world networks, and in particular social networks, nodes tend to create tightly knit groups characterised by a relatively high density of ties; this likelihood tends to be greater than the average probability of a tie randomly established between two nodes (Paul & Leinhardt 1971, Watts & Strogatz 1998*a*).

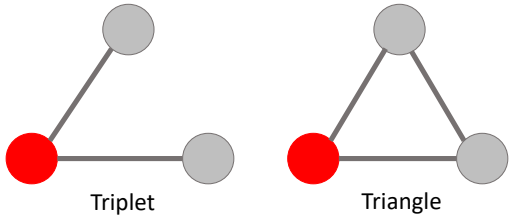


Figure 2.9 – Graphical representation of a Triplet (left) and a Triangle (right)

This measure can be global and local. The global measure gives an overall indication of the clustering in the network, whereas the local indicates the embeddedness of the single nodes.

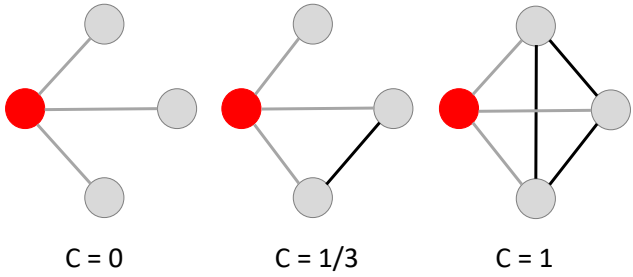


Figure 2.10 – Clustering coefficient.

The global clustering coefficient is based on triplets of nodes. A triplet is three nodes that are connected by either two (open triplet) or three (closed triplet) undirected ties. A triangle graph therefore includes three closed triplets,

one centered on each of the nodes, this means the three triplets in a triangle come from overlapping selections of nodes.

The global clustering coefficient is the number of closed triplets (or 3 x triangles) over the total number of triplets (both open and closed) (Luce & Perry 1949). This measure gives an indication of the clustering in the whole network (global), and can be applied to both undirected and directed networks.
The global clustering coefficient is defined as: $CC = 3\#Triangles/\#Triplets$. You may be wondering why do we multiply the number of triangles by three? The reason is that in a triangle there are, obviously, three nodes, so we have to count the triangles from the three different perspectives.

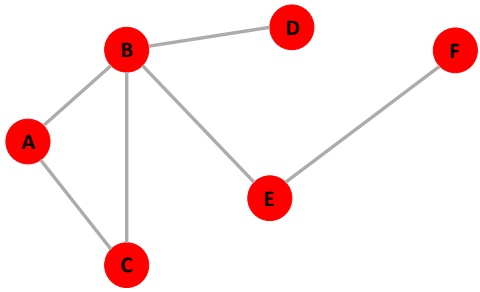


Figure 2.11 – For this network the binary global clustering coefficient would be 3 over 9, or 0.33. The closed triplets are $B \rightarrow A \leftarrow C$; $A \rightarrow B \leftarrow C$; $A \rightarrow C \leftarrow B$; whereas the open triplets are $A \rightarrow B \leftarrow D$; $A \rightarrow B \leftarrow E$; $C \rightarrow B \leftarrow D$; $C \rightarrow B \leftarrow E$; $D \rightarrow B \leftarrow E$; $B \rightarrow E \leftarrow F$.

- Network average clustering coefficient

On the other hand, the *local clustering coefficient* is a measure of the clustering coefficient of a node given by following formula:

$$C_i = \frac{2L_i}{k_i(k_i - 1)}$$

where k_i is the degree of node i and L_i is the number of edges between the k_i neighbors of node i.

The overall level of clustering in a network is measured by Watts and Strogatz (Watts & Strogatz 1998*a*) as the average of the clustering coefficients of all the vertices n :

$$\bar{C} = \frac{1}{n}\sum_{i=1}^{n} C_i$$

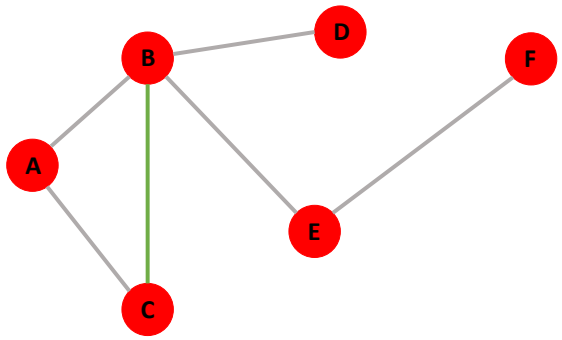


Figure 2.12 – Green Line is the edge of neighbours for Node A. Note that is the same network of Fig 2.11

Consider for example the graph presented in Fig 2.12. Let's calculate clustering coefficient for each node,

- **For node A:** $k_A = 2, L_A = 1, C_A = \frac{2*1}{2*(2-1)} = \frac{2}{2} \Rightarrow C_A = 1$
- **For node B:** $k_B = 4, L_B = 1, C_B = \frac{2*1}{4*(4-1)} = \frac{1}{6} \Rightarrow C_B = 0.17$
- **For node C:** $k_C = 2, L_C = 1, C_C = \frac{2*1}{2*(2-1)} = \frac{2}{2} \Rightarrow C_C = 1$
- **For node D:** $k_D = 1, L_D = 0 \Rightarrow C_D = 0$
- **For node E:** $k_E = 2, L_E = 0 \Rightarrow C_E = 0$
- **For node F:** $k_F = 1, L_F = 0 \Rightarrow C_F = 0$

For average clustering coefficient

$$\bar{C} = \frac{1}{n}\sum_{i=1}^{n} C_i \Rightarrow \frac{1}{6}(1+0.17+1+0+0+0) \Rightarrow 0.36$$

As we can see, the values of the global clustering coefficient (0.33) and the average clustering coefficient (0.36) are, by definition, different.

2.5 Node Roles

Role is a concept that is used to describe the behaviour of a node in relationship to its neighbours and to the network at large. A *node role* is a subjective characterization of the part it plays in a network structure. Knowing the role of a node is important for many link meaning applications. For example, in Web search, nodes that are deemed to be authorities on a given topic are often found to be most relevant to the user's queries. There are a number of metrics that can be used to assign roles to individual nodes in a network, including degree, that can be used to assess a node's popularity, closeness, and betweenness, that can be used to assess its centrality (Scripps et al. 2007).

There are different roles that we can find in a network, and the names for the roles vary according to the study's context. In social network analysis, for example, people are characterized according to their network positions. In this context we can find *Connectors*, which are the nodes with many direct connections, *Isolates*, nodes with no or just a single connection, *Influencers* which are closely connected to the remaining network and *Brokers* which tie together otherwise disconnected nodes.

If we want to divide the nodes according the connectivity between them, we have *Broadcasters*, which convey information from the group to the outside, *Connectors*, that connect nodes that are in two different groups, *Coordinators* that connects nodes within the same group, and *Gatekeepers* that buffer between their group and other groups.

In Chapter 4 we use a different methodology proposed by Cooper and Barahona in (Cooper & Barahona 2010). They introduce an alternative measure for the grouping of nodes in directed networks. Given that the defining characteristic of directed graphs is the implicit existence of flows, they propose to group nodes according to their role in the network, defined in terms of the overall pattern of incoming and outgoing flows. Essentially, the profile of paths for each node is a vector that is computed from the powers of the adjacency matrix weighted with a scale parameter to yield a similarity matrix, defined by the distances between such node vectors. This matrix is then clustered to find groupings of nodes with similar profiles of reachability flows at all lengths.
In this analysis, nodes are grouped according to a quantitative measure that reflects the mixture of 'hub' vs. 'authority' characteristics of each node with respect to all paths in the graph. This definition is inspired by a vast array of literature from the social sciences, dealing with structural and regular equiva-

lence (De Nooy et al. 2018, Borgatti & Everett 1993, Freeman n.d., Reichardt & White 2007), and from computer science, where alternative algorith- mic measures of similarity have been considered (Ninove et al. 2007, Jeh & Widom 2002, Kleinberg 1999*b*, Leicht et al. 2006).
For instance, in a food-web, two predators are not likely to be linked directly although both perform the same function and would be canonically grouped within the same trophic level. Hence, role similarity can uncover a coarse-grained functional representation for networks where the dominating feature is the transfer of an underlying quantity (e.g., information, energy, matter, etc). This role-based representation is relevant in fields such as ecology, economics, social sciences and cellular metabolism, where it can aid in the assignment of a putative function to uncharacterised nodes and in establishing functional relations between seemingly distant network elements, that's the reason why we consider this approach the appropriate one for the dynamics that we consider in the problem addressed in Chapter 4.

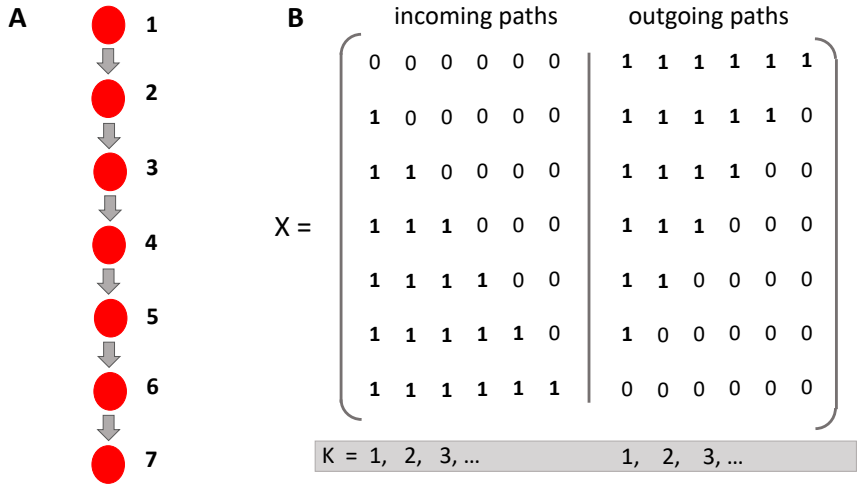


Figure 2.13 – Role clustering for the directed path graph in (**A**). The construction of the flow matrix **X** (shown in (**B**) for $\beta = 1$)

The measure is defined by Cooper and Barahona as follows. Consider a directed graph with N nodes and adjacency matrix A, which is in general asymmetric. The number of outgoing paths of length k for node i is given by the i-th coordinate of the vector $[A^k\mathbf{1}]$, where 1 is the $N1$ vector of ones. Similarly, the number of incoming paths of length k for node i is: $[A^{T^k}\mathbf{1}]_i$. Note that the case $k = 1$ corresponds to the out-degree and in-degree which, from this

perspective, represent the number of paths of length one originating or terminating at the node.

They now construct a matrix that compiles the incoming and outgoing paths of all lengths up to k_{max} by appending the column vectors indexed by path length and scaled by the factors β^k:

$$X = \begin{bmatrix} \mathbf{x}_1 \\ . \\ . \\ . \\ \mathbf{x}_N \end{bmatrix} \equiv [\underbrace{\ldots (\beta A^T)^k \mathbf{1} \ldots}_{k_{max}} | \underbrace{\ldots (\beta A)^k \mathbf{1} \ldots}_{k_{max}}],$$

where $\beta = \alpha/\lambda_1$, with λ_1 the largest eigenvalue of the adjacency matrix and

$0 \leq \alpha \leq 1$. The parameter α is a scale factor that allows us to tune the weight of the local environment (short paths) relative to the global network structure (long paths).

The presence of the factors βk ensures the convergence of the sequence of the columns due to the asymptotic limit $\lim_{k \to +\infty} \frac{\|A^{k+1}\|}{\|A^k\|} \to \lambda_1$. (Leicht et al. 2006)

Each row vector of X contains the flow profile of a node in terms of the scaled number of incoming and outgoing paths of all lengths starting and ending at that node (see Fig 2.13). Their criterion to group nodes together is that they have similar flow profiles. This can be quantified via a distance between the vectors $\mathbf{x}_i$.

2.6 Node similarity

Similarity in network analysis occurs when two nodes (or other more elaborate structures) fall in the same equivalence class.

There are three fundamental approaches to constructing measures of network similarity: structural equivalence, automorphic equivalence, and regular equivalence (Newman 2018*b*). There is a hierarchy of the three equivalence concepts: any set of structural equivalences are also automorphic and regular equivalences. Any set of automorphic equivalences are also regular equivalences. Not all regular equivalences are necessarily automorphic or structural;

and not all automorphic equivalences are necessarily structural.(Hanneman & Riddle 2005)

- **Structural equivalence** Two vertices of a network are structurally equivalent if they share many of the same neighbors.

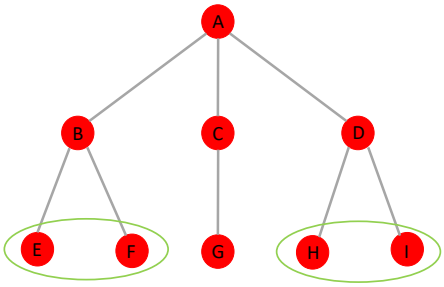


Figure 2.14 – Structural equivalence. There is no actor who has exactly the same set of ties as actor A, so actor A is in a class by itself. The same is true for actors B, C, D and G.

Each of these nodes in Fig 2.14 has a unique set of edges to other nodes. E and F, however, fall in the same structural equivalence class. Each has only one edge; and that tie is to B. Since E and F have exactly the same pattern of edges with all the vertices, they are structurally equivalent. The same is true in the case of H and I.

- **Automorphic equivalence** Formally "Two vertices are automorphically equivalent if all the vertices can be relabeled to form an isomorphic graph with the labels of u and v interchanged. Two automorphically equivalent vertices share exactly the same label-independent properties" (Hanneman & Riddle 2005).

  In Fig 2.15 actor A is the central headquarter, actors B, C, and D are managers. Actors E, F and H, I are workers at smaller stores; G is the lone worker at another store. There are five automorphic equivalence classes: $\{A\}$, $\{B,D\}$, $\{C\}$, $\{E,F,H,I\}$, and $\{G\}$. Note that the less strict definition of "equivalence" has reduced the number of classes.
  More intuitively, actors are automorphically equivalent if we can permute the graph in such a way that exchanging the two actors has no effect on the distances among all actors in the graph.

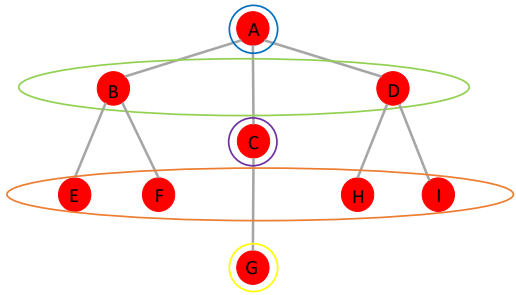


Figure 2.15 – Automorphic equivalence. Organizational structure of a company.

- **Regular equivalence** "Two actors are regularly equivalent if they are equally related to equivalent others" (Hanneman & Riddle 2005). In other words, regularly equivalent vertices are vertices that, while they do not necessarily share neighbours, have neighbours who are themselves similar.

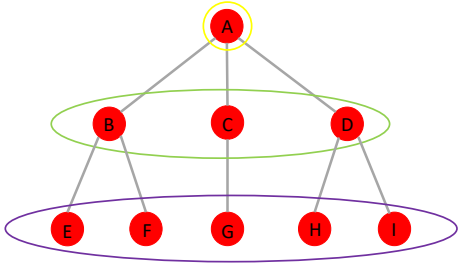


Figure 2.16 – Regular equivalence. There are three regular equivalence classes. The first is actor *A*; the second is composed of the three actors *B*, *C*, and *D*; the third consists of the remaining five actors *E*, *F*, *G*, *H*, and *I*.

Two mothers, for example, are equivalent, because each has a similar pattern of connections with a husband, children, etc. The two mothers do not have ties to the same husband or the same children, so they are not structurally equivalent. Because different mothers may have different numbers of husbands and children, they will not be automorphically equivalent. But they are similar because they have the same relationships with some member or members of another set of actors (who are themselves regarded as equivalent because of the similarity of their ties to a member of the set "mother").

2.7 Link Prediction

Given a snapshot of a network, can we infer which new interactions among its members are likely to occur in the near future?

Among the most fundamental network analysis tools, are those designed for *link prediction*. Intuitively, such tools analyse the topology of a given network in order to predict the connections that are most likely to form in the future.
These tools can also be used to analyse the observed network topology to identify connections that

are hidden from the observer, either due to data scarcity, or due to the deliberate concealment of information.

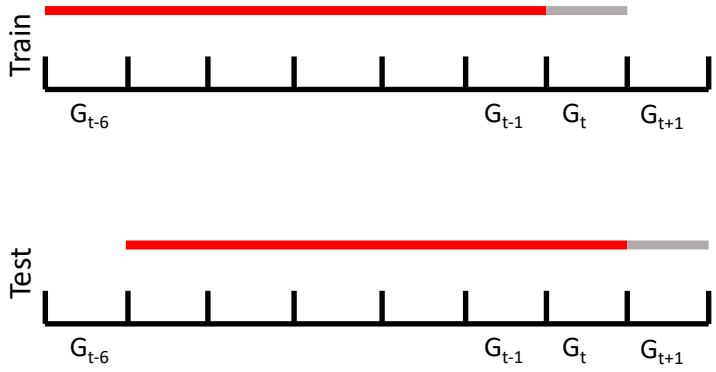


Figure 2.17 – Link prediction and evaluation. The red color indicates snapshots of the network from which link prediction features are calculated (feature network)The gray color indicates snapshots of the network from which the link prediction instances are labeled (label network). We can observe all links at or before time t, and we aim to predict future links that will occur at time $t+1$.

According to (Yang et al. 2015), link prediction entails all the complexities of evaluating ordinary binary classification for imbalanced class distributions, but it also includes several new parameters and intricacies that make it fundamentally different. Real-world networks are often very large and sparse, involving many millions or billions of nodes and roughly the same number of edges.
Due to the resulting computational burden, test set sampling is common in link prediction evaluation (Liben-Nowell & Kleinberg 2007, Al Hasan et al.

2006, Murata & Moriyasu 2007). Such sampling, when not properly reflective of the original distribution, can greatly increase the likelihood of biased evaluations that do not meaningfully indicate the true performance of link predictors. The selected evaluation metric can have a tremendous bearing on the apparent quality and ranking of predictors even with proper testing distributions (Raeder et al. 2010). The directionality of links also introduces issues that do not exist in typical classification tasks. Finally, for tasks involving network evolution, such as predicting the appearance of links in the future, the classification process involves temporal aspects. Training and testing set constructs must appropriately address these nuances.

There are a variety of methods for predicting connections, such as , *Graph Distance*, *Common Neighbours*, *Preferential Attachment* and so on. Here we are not going to explain these methods, instead, we will focus in the methods used in this thesis, which are related to supervised machine learning, this important concept and the methods will be explained in detail in the next chapter.

To better describe the intricacies of evaluation in the link prediction problem, the first step is to depict the framework for evaluation (O'Madadhain et al. 2005, O'Madadhain et al. 2005, Lichtenwalter et al. 2010) (Fig.2.17). Computations occur within network snapshots based on particular segments of data. Comparisons among predictors require that evaluation encompasses precisely the same set of instances whether the predictor is unsupervised or supervised (We explain these two concepts in Chapter 3). We construct four network snapshots:

- Training features: data from some period in the past, G_{t-x} up to G_{t-1}, from which we derive feature vectors for training data.
- Training labels: data from G_t, the last training-observable period, from which we derive class labels, whether the link forms or not, for the training feature vectors.
- Testing features: data from some period in the past up to G_t, from which we derive feature vectors for testing data. Sometimes it may be ideal to maintain the window size that we use for the training feature vector, so we commence the snapshot at G_{t-x+1}. In other cases, we might want to be sure not to ignore effects of previously existing links, so we commence the snapshot at G_{t-x}.
- Testing labels: data from G_{t+1}, from which we derive class labels for

the testing feature vector. This data is strictly excluded from inclusion in any training data.

A classifier is constructed from the training data and evaluated on the testing data. There are always strictly divided training and testing sets, because G_{t+1} is never observable in training.

In chapters 3 and 5 we are going to explain more in detail the methodology used in this work for link prediction.

2.8 Signed networks

As we mentioned a few lines above, the edges in a graph may be directed or weighted; either the nodes or edges may have types, categories, or labels of some kind; nodes may have positions in space; edges may have lengths or capacities, and so forth. According to that it is not difficult to intuit that the edges can be either positive or negative depending the kind of relationship that they are representing (Newman 2018*b*, Harary 1953, Wasserman & Faust 1994).

The most common example is a social network that represents patterns of both amity and enmity among a group of individuals: positive edges represent friendship, negative ones animosity.

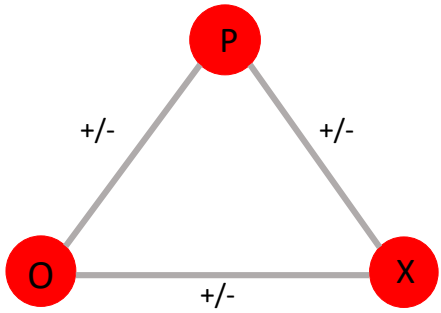


Figure 2.18 – Signed Networks

Studies of signed networks go back at least to the classic work of Harary in the 1950s, who argued, largely informal rather than empirical grounds, that certain patterns of signs should be more common than others—the enemy of my enemy should be my friend, for example (Harary 1953).

Networks that display such regularities are said to be structurally balanced, or just balanced for short. A natural question to ask is whether real signed networks are in fact balanced. Despite a considerable amount of research on this issue, however, the jury is still out. Some researchers have claimed that real networks are balanced, at least partially, while others have claimed that they are not. Social balance theory—in its strong form (Cartwright & Harary 1956) —claims that positive triads are *balanced* whereas negative triads are *unbalanced*.

P – O	P – X	O – X	Relation characteristic
+	+	+	**Balanced**
+	+	-	Imbalanced
+	-	+	Imbalanced
+	-	-	**Balanced**
-	+	+	Imbalanced
-	+	-	**Balanced**
-	-	+	**Balanced**
-	-	-	Imbalanced

Table 21 – Possible combination of relation sign in a triad and the relation's characteristic.

The balanced and unbalanced can be understood easily using again the analogy of friends and enemies (Szell et al. 2010).

In the relation of three people or triad Fig 2.18, balance state occurs when all sign multiplication of its sentiment relation charges positive. In this way, balance state will occur when there are sentiment relations with signs all positive $(+ \times + \times + = +)$, or two negatives and one positive $(- \times - \times + = +)$. This model is popular as *pox* model where *p* is focal individual, *o* is object, issue or a person, and *x* is object or other individual. Sentiment relation *p* and *x* is determined by an attitude of *p* and *x* toward *o*. If the multiplication of signs of these relations is positive, then the balance state is achieved (Heider, 1946). There are 8 possible configurations of the current sentiment relations described in the Table 21. In the table we can see that there are 4 patterns of balance state and 4 others imbalance one in the relation of three agents (triad).

2.9 Multilayer networks

Since, as we have discussed, there different kinds of edges, with different meanings, weights and directions and they can be either positive of negative, to represent systems consisting of networks with multiple types of links, or other similar features, we consider structures that have layers in addition to nodes and links. Networks in real life integrate all these kind of edges, for example, if we consider ourselves as elements of a big network including all the different relationship that we can establish with other people, we will have friendship edges, family edges, colleagues edges and so on. So we can conclude that our relationships shape a multilayer network Fig 2.19.

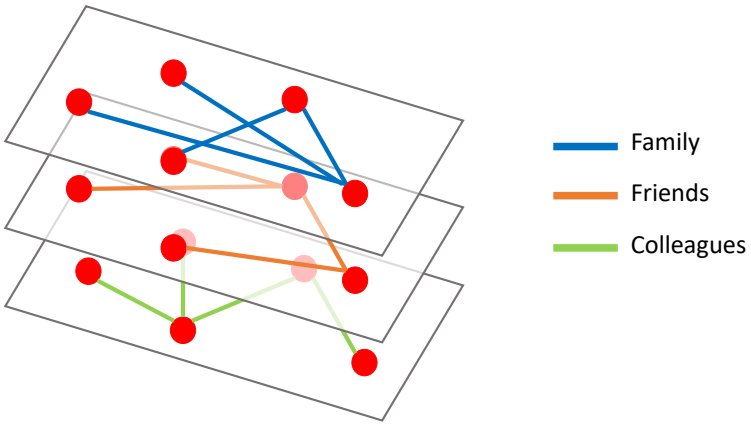


Figure 2.19 – Multilayer social network

In its more general form, in a multilayer framework a node u in layer α can be connected to any node v in any layer β. Layers will represent aspects or features that characterize the nodes or the links that belong to that layer. (Aleta & Moreno 2019)

Multilayer networks encode two major classes of systems (Bianconi 2015): multiplex networks and network of networks. Multiplex networks are networks where the same set of nodes is represented in every layer, although the interaction between nodes might be different in each one. As an example, two nodes might be connected in one layer an might not in other. This is the case of online social systems, where a given user might have a Twitter account (layer 1) and a Facebook profile (layer 2). The set of followers/friends does not in

general coincide for both layers, thus leading to two different intra-layer adjacency matrices. On the contrary, a network of networks is instead formed by networks that are interlaced to each other but formed by different types of nodes. Although these two are the most common kinds of multilayer networks, Kivela et al. (Kivelä et al. 2014*a*) present several other types and denominations of multilayer networks. From now on we will always refer to Multiplex networks.

We provide now a precise definition of a multilayer network based on our above description Fig 2.19. A multilayer network has a set of nodes V just like a normal network (i.e., a graph). In addition, we need to have layers. However, because we want to be able to include multiple aspects in a multilayer network, we cannot restrict ourselves to having a single set of layers. For example, in a network in which the first aspect is interaction type and the second one is time, we need one set of layers for interaction types and a second set of layers for time. A multilayer network can have any number d of aspects, and we need to define a sequence $\mathbf{L} = L_a{}_{a=1}^{d}$ of sets of elementary layers such that there is one set of elementary layers L_a for each aspect a.
Using the sequence of sets of elementary layers, we can construct a set of layers in a multilayer network by assembling a set of all of the combinations of elementary layers using a Cartesian product $L_1 \times \cdots \times L_d$. We want to allow nodes to be absent in some of the layers. That is, for each choice of a node and layer, we need to indicate whether the node is present in that layer. To do so, we first construct a set $V \times L_1 \times \cdots \times L_d$ of all of these combinations and then define a subset $V_M \subseteq V \times L_1 \times \cdots \times L_d$ that contains only the node-layer combinations in which a node is present in the corresponding layer. We will often use the term nodelayer tuple (or simply node-layer) to indicate a node that exists on a specific layer. Thus, the node-layer $(u, \alpha_1, \cdots, \alpha_d)$ represents node u on layer $(\alpha_1, \cdots, \alpha_d)$.

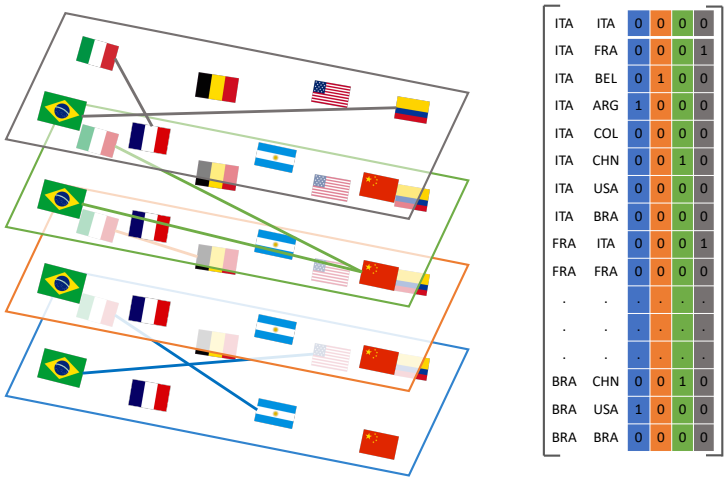
In a multilayer network, we also need to define connections between pairs of node-layer tuples. As with monoplex networks, we will use the term adjacency to describe a direct connection via an edge between a pair of node-layers and the term incidence to describe the connection between a node-layer and an edge. Two edges that are incident to the same node-layer are also "incident" to each other. We want to allow all of the possible types of edges that can occur between any pair of node-layers — including ones in which a node is adjacent to a copy of itself in some other layer as well as ones in which a node is adjacent to some other node from another layer. In normal networks (i.e., graphs), the adjacencies are defined by an edge set $E \subseteq V \times V$, in which the first el-

ement in each edge is the starting node and the second element is the ending node. In multilayer networks, we also need to specify the starting and ending layers for each edge. We thus define an edge set E_M of a multilayer network as a set of pairs of possible combinations of nodes and elementary layers. That is, $E_M \subseteq V_M \times V_M$.
Using the components that we set up above, we define a *multilayer network* as a quadruplet $M = (V_M, E_M, V, \mathbf{L})$. If the number of aspects is zero (i.e., if $d = 0$), then the multilayer network M reduces to a monoplex (i.e., single-layer) network. In that case, $V_M = V$, so the set V_M becomes redundant. (By convention, the product term in the set $V \times L_1 \times \cdots \times L_d$ does not exist if $d = 0$.) and convenient to use different semantics for edges that cross layers than for edges.

According to (Buldyrev et al. 2010), when a system of two interdependent networks is considered, where nodes in one network have a bidirectional coupling with nodes in the other, the percolation properties are significantly affected (*Percolation describes a phase transition process of network failure, whose critical point distinguishes the network from connected to disconnected. Percolation theory makes use of statistical physics principles and graph theory to analyse such change in the structure of a complex network.*). Due to coupling, not only the transition threshold is increased but also the order of the transition changes. The presence of interdependency between nodes in different networks, such that if one of the nodes is inactive the other can't function as well, leads to catastrophic effects when some nodes are removed from the system. A cascade of events is then ignited leading to an abrupt decomposition of the mutually connected giant component. These interesting results raise new questions in the field of complex networks. As a natural follow up, it is interesting to understand what happens when different types of networks are coupled, with special focus on real networks. The effect of different types of inter-networks connections is also relevant.

2.9.1 Incidence matrices for Multilayer networks

For single networks, we introduced the concept of adjacency matrix (Fig 2.7), which it is a square matrix (number of nodes $\times$ number of nodes) and in which we have non-zero elements if there is a connection between the corresponding nodes. In an adjacency matrix all the edges are considered to indicate the same the type of relationship between the nodes. Since in a multilayer network the edges indicate different relationships depending the layer that they belong to, the way to have the information of all the layers in the same two-dimensional matrix is through what is known as the *Incidence matrix*.



ITA	ITA	0	0	0	0
ITA	FRA	0	0	0	1
ITA	BEL	0	1	0	0
ITA	ARG	1	0	0	0
ITA	COL	0	0	0	0
ITA	CHN	0	0	1	0
ITA	USA	0	0	0	0
ITA	BRA	0	0	0	0
FRA	ITA	0	0	0	1
FRA	FRA	0	0	0	0
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
BRA	CHN	0	0	1	0
BRA	USA	1	0	0	0
BRA	BRA	0	0	0	0

Figure 2.20 – Graph and matricial representation of a multilayer network in which the nodes represent countries and each layer (each column) a different relation between them.

The incidence matrix of an undirected graph is a $n \times m$ matrix B, where n and m are the numbers of vertices and edges respectively, such that $B_{i,j} = 1$ if the vertex v_i and edge e_j are incident and 0 otherwise.

In our case, in Chapter 5, the nodes of our multilayer network will represent countries and each layer represent a different kind of relation that the countries can establish between them. In the incidence matrix is, each row correspond to a pair of countries and each column give information about if there is or not a link between the pair of countries (See Fig 2.20).

2.10 Dynamic Graphs

Real networks are, in principle, dynamic, that's mean that they vary over time. Of course the number of nodes can also change but let's focus in the temporal change of the connections between the nodes. Each link carries information on when it is active, along with other possible characteristics such as a weight. Time-varying networks are of particular relevance to spreading processes, like the spread of information and disease, since each link is a contact opportunity and the time ordering of contacts is included.
Examples of time-varying networks include communication networks, where each link is relatively short or instantaneous, such as phone calls or e-mails.

(Karsai et al. 2014), (Eckmann et al. 2004) Information spreads over both networks, and some computer viruses spread over the second. Networks of physical proximity, encoding who encounters whom and when, can be represented as time-varying networks.(Eagle & Pentland 2006) Some diseases, such as airborne pathogens, spread through physical proximity. Real-world data on time resolved physical proximity networks has been used to improve epidemic modelling.(Stehlé et al. 2011) Neural networks and brain networks can be represented as time-varying networks since the activation of neurons are time-correlated.(Holme & Saramäki 2012)
Time-varying networks are characterized by intermittent activation at the scale of individual links like the three graph of the same network presented in Fig 2.21. This is in contrast to various models of network evolution, which may include an overall time dependence at the scale of the network as a whole.

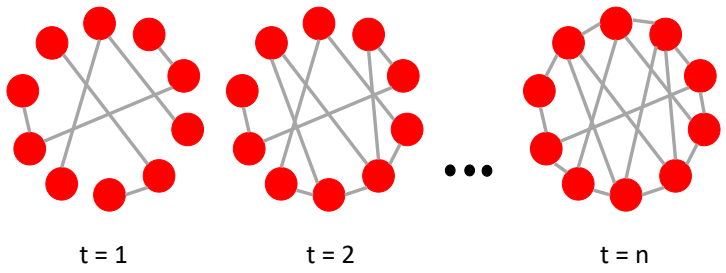


Figure 2.21 – Example of a dynamic (temporal) network. Each graph represent the network in a different time

2.10.1 Representations

There are three common representations for time-varying network data:

- *Contact sequences*, if the duration of interactions are negligible, the network can be represented as a set C of contacts (i, j, t) where i and j are the nodes and t the time of the interaction. Alternatively, it can be represented as an edge list E where each edge e is a pair of nodes and has a set of active times $T_e = \{t_1, \cdots t_n\}$.

- *Interval graphs*, if the duration of interactions are non-negligible, T_e becomes a set of intervals over which the edge e is active.

$$T_e = \{(t_1, t_1^{'}), \cdots (t_n, t_n^{'})\}$$

- *Snapshots* – time-varying networks can also be represented as a series of static networks, one for each time step.

Chapter 3

Introduction to Machine Learning

Machine learning can be broadly defined as computational methods using experience to improve performance or to make accurate predictions.
Here, *experience* refers to the past information available to the learner, which typically takes the form of electronic data collected and made available for analysis. This data could be in the form of digitized human-labelled training sets, or other types of information obtained via interaction with the environment. In all cases, its quality and size are crucial to the success of the predictions made by the learner (Mohri et al. 2018).

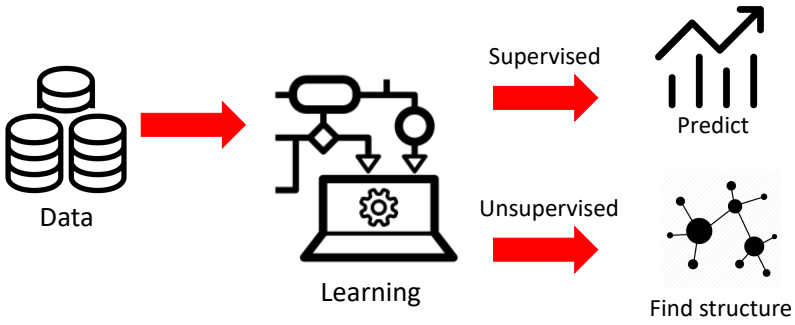


Figure 3.1 – Generic operation of Machine Learning algorithms

In 1959, Arthur Samuel, a pioneer in the field of machine learning defined it as the *field of study that gives computers the ability to learn without being explicitly programmed* (Samuel 1959).
A more formal definition was given by Tom Mitchell (Nigam et al. 1998) as a computer program is said to learn from experience (E) with respect to some task (T) and some performance measure (P), if its performance on T, as measured by P, improves with experience E then the program is called a machine learning program.

3.1 Types of Learning

There are two major settings in which we wish to learn a function, Fig 3.1. In one, called *supervised learning*, we know (sometimes only approximately) the values of the input (x) and output (y) variables connected by a mapping function (f) for the m samples in the training set, Ξ. We assume that if we can find a hypothesis, h, that closely agrees with f for the members of Ξ, then this hypothesis will be a good guess for f—especially if Ξ is large.(Nilsson 1996)

As the name implies, this type of machine learning requires a certain level of supervision of machine learning models, or bots. This is done by teaching the model, or loading it with knowledge, so that it can *learn* the desirable behaviours it should perform. *Supervised learning* problems can be further grouped into regression (the prediction of trends in labelled data to determine future outcomes) and classification (the organisation of labelled data) problems. Both rely on knowing what kind of data they are dealing with in a controlled environment, with regular testing guiding the learning process.

The generation process is based on classifying the elements of the training set and comparing the result with the label associated with each element. This process is carried out iteratively to adjust the predictive model. Fig 3.2 is a sketch of the supervised learning process.
In this work we will focus in two of the best-known supervised learning algorithms, *Decision Trees* and *Random Forest* (See Section 3.2).

In the other setting, termed *unsupervised learning*, we simply have a training set of vectors without function values for them. The problem in this case, typically, is to partition the training set into subsets, $\Xi_1, \cdots, \Xi_R$, in some appropriate way. (We can still regard the problem as one of learning a function;

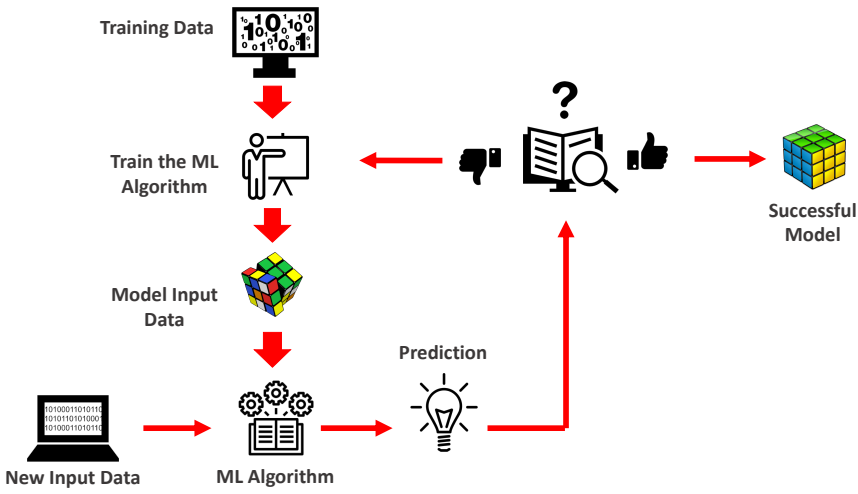


Figure 3.2 – General outline of supervised learning processes

the value of the function is the name of the subset to which an input vector belongs.) Unsupervised learning methods have application in taxonomic problems in which it is desired to invent ways to classify data into meaningful categories.

Unlike supervised learning, this type of learning is different in that it isn't observed by anyone and happens on its own terms. We let the model, discover information that may not be visible to the human eye or brain. Unsupervised learning relies on machine learning algorithms to make conclusions and predictions on sets of unlabelled data. As such, its biggest strength lies in the ability to find patterns and groupings from unknown types of data.

Unsupervised learning is the training of machine using information that is neither classified nor labelled and allowing the algorithm to act on that information without guidance. Here the task of machine is to group unsorted information according to similarities, patterns and differences without any prior training of data.
Unlike supervised learning, no teacher is provided that means no training will be given to the machine. Therefore machine is restricted to find the hidden structure in unlabelled data by itself.

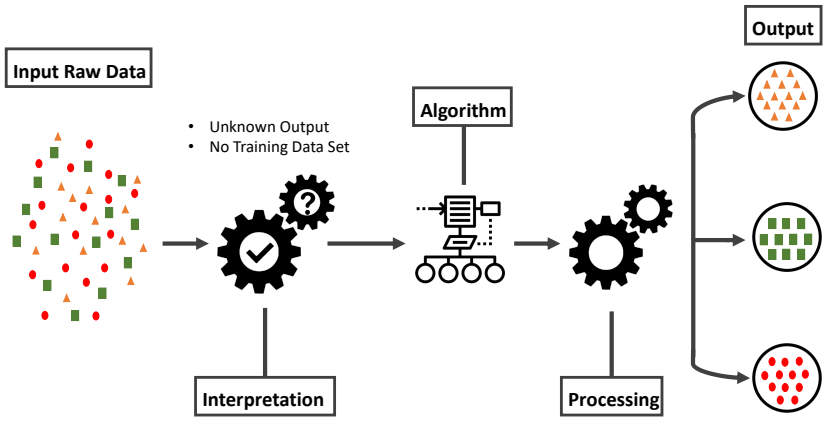


Figure 3.3 – General outline of Unsupervised learning processes

Consider for example the Fig 3.3, in this case the machine has no idea about the features of triangles, circles and squares, so we can't categorize it in triangles, circles or squares. *Clustering* is an important concept when it comes to this kind of problems. (We explain this concept in detail in the section 3.3.1) It mainly deals with finding a structure or pattern in a collection of uncategorized data. Clustering algorithms will process your data and find natural clusters(groups) if they exist in the data. You can also modify how many clusters your algorithms should identify. It allows you to adjust the granularity of these groups.

3.2 Supervised Learning algorithms

As we mentioned before, in this work we will focus in two of the best-known supervised learning algorithms, *Decision Trees* and *Random Forest*.

- ***Decision Tree***. A decision tree (generally defined) is a tree whose internal nodes are tests (on input patterns) and whose leaf nodes are categories (of patterns). We show an example in Fig. 3.4. A decision tree assigns a class number (or output) to an input pattern by filtering the pattern down through the tests in the tree. Each test has mutually exclusive and exhaustive outcomes. For example, test $T2$ in the tree of Fig. 3.4 has three outcomes; the left-most one assigns the input pattern to *class*

3, the middle one sends the input pattern down to test $T4$, and the rightmost one assigns the pattern to *class 1*. We follow the usual convention of depicting the leaf nodes by the class number (Quinlan 1986, 1993). Note that in discussing decision trees we are not limited to implementing Boolean functions—they are useful for general, categorically valued functions.

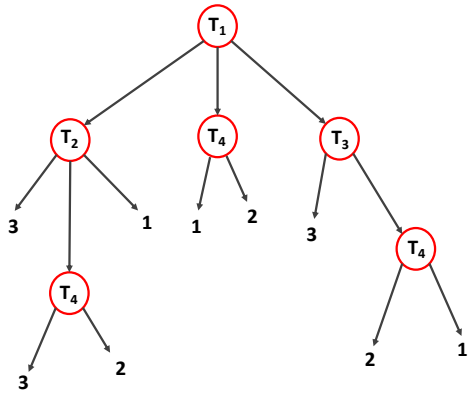


Figure 3.4 – A Decision Tree

There are several dimensions along which decision trees might differ:

- The tests might be multivariate (testing on several features of the input at once) or univariate (testing on only one of the features).
- The tests might have two outcomes or more than two. (If all of the tests have two outcomes, we have a binary decision tree.)
- The features or attributes might be categorical or numeric. (Binary-valued ones can be regarded as either.)
- We might have two classes or more than two. If we have two classes and binary inputs, the tree implements a Boolean function, and is called a Boolean decision tree.

It is straightforward to represent the function implemented by a univariate Boolean decision tree in Disjunctive Normal Form DNF.
A logical formula is considered to be in DNF if it is a disjunction of one or more conjunctions of one or more literals (Davey & Priestley 1990). A DNF formula is in full disjunctive normal form if each of its variables

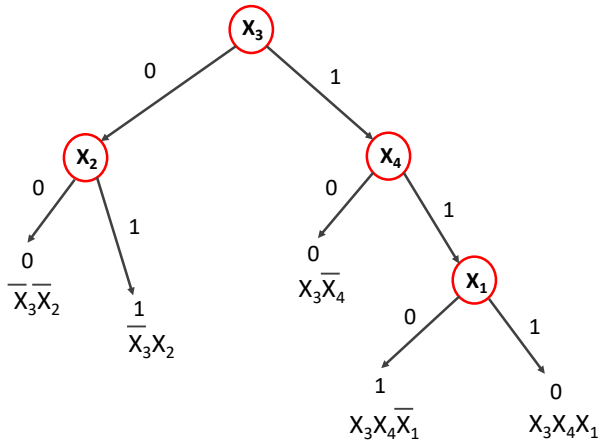


Figure 3.5 – A Decision Tree Implementing a DNF Function

appears exactly once in every conjunction.

The basic elements in such formula are called *terms*. A term is simply a conjunctions of literals. That is,

$$T = \ell_1 \wedge \ell_2 \wedge \cdots \wedge \ell_t$$

where all ℓ_i's are literals and $\wedge$ denotes the boolean AND operation. A term is *satisfied* (that is, gets the value TRUE) if and only if each of the literals in the term gets the value TRUE. Therefore, each term is just a function from $\{0,1\}^n$ to $\{0,1\}$. We denote by TERMS_n the class of all possible terms over the literals $x_1,\ldots,x_n,\bar{x}_1,\ldots,\bar{x}_n$. When n is clear from the context we may use TERMS for short. Also, for simplicity of notations we omit the $\wedge$ symbols from the term. That is, we use $\bar{x}_3 x_2$ to denote $\bar{x}_3 \wedge x_2$ (left branch of tree in Figure 3.5).

A DNF formula is a disjunction of terms. That is,

$$F = T_1 \vee T_2 \vee \cdots \vee T_k$$

where each T_i is a term and $\vee$ denotes the boolean OR operation. The formula is *satisfied* (that is, gets the value TRUE) if and only if at least one of its terms is satisfied. Again, each such formula defines a function from $\{0,1\}^n$ to $\{0,1\}$.

For example, for the decision tree in Figure 3.5 we get the following 2-term DNF (corresponding to the 2 nodes in the tree which are labelled by 1):

$$\bar{x}_3 x_2 \vee x_3 x_4 \bar{x}_1$$

- ***Random Forest***. As the name suggests, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction (see Fig. 3.6).

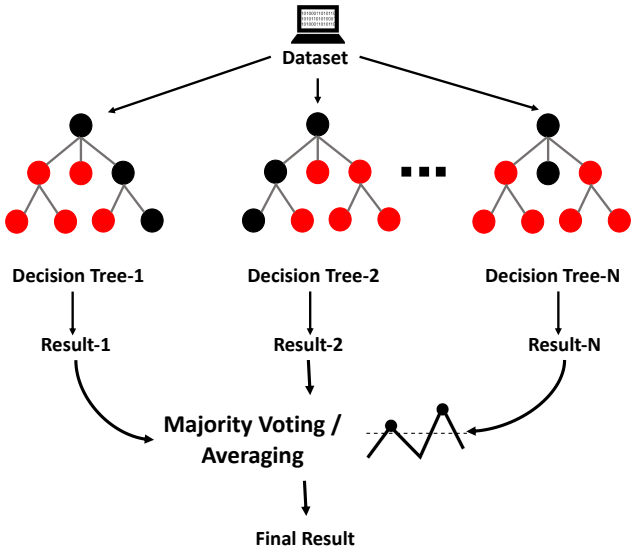


Figure 3.6 – Random Forest

The fundamental concept behind random forest is a simple but powerful one — the wisdom of crowds. In data science speak, the reason that the random forest model works so well is: A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models.

The training algorithm for random forests applies the general technique of bootstrap aggregating, or bagging, to tree learners.
Bootstrap aggregating, also called ***bagging*** (from bootstrap aggregating), is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used

in statistical classification and regression. It also reduces variance and helps to avoid overfitting (See Section 3.6). Although it is usually applied to decision tree methods, it can be used with any type of method. Bagging is a special case of the model averaging approach (GRU n.d.).

Given a training set $X = x_1, \ldots, x_n$ with responses $Y = y_1, \ldots, y_n$, bagging repeatedly (B times) selects a random sample with replacement of the training set and fits trees to these samples:
For $b = 1, \ldots, B$:

1. Sample, with replacement, n training examples from X, Y; call these X_b, Y_b.
2. Train a classification or regression tree f_b on X_b, Y_b.

After training, predictions for unseen samples x' can be made by averaging the predictions from all the individual regression trees on x':

$$\hat{f} = \frac{1}{B} \sum_{b=1}^{B} f_b(x')$$

or by taking the majority vote in the case of classification trees.
The above procedure describes the original bagging algorithm for trees. Random forests differ in only one way from this general scheme: they use a modified tree learning algorithm that selects, at each candidate split in the learning process, a random subset of the features. This process is sometimes called "feature bagging". The reason for doing this is the correlation of the trees in an ordinary bootstrap sample: if one or a few features are very strong predictors for the response variable (target output), these features will be selected in many of the B trees, causing them to become correlated. An analysis of how bagging and random subspace projection contribute to accuracy gains under different conditions is given by Ho in (Ho 2002).

3.3 Unsupervised Learning Algorithms

Unsupervised learning problems can be further grouped into clustering and association problems.

- ***Clustering***: is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some

sense) to each other than to those in other groups (clusters). In this work we will focus in this kind of algorithms.

- ***Association***: is a method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness

3.3.1 Clustering Algorithms

Up to know, we have explored supervised Machine Learning algorithms and techniques to develop models where the data had labels previously known. In other words, our data had some target variables with specific values that we used to train our models (Cacciari et al. 2008).
However, when dealing with real-world problems, most of the time, data will not come with predefined labels, so we will want to develop machine learning models that can classify correctly this data, by finding by themselves some commonality in the features, that will be used to predict the classes on new data.

In basic terms, the objective of clustering is to find different groups within the elements in the data. To do so, clustering algorithms find the structure in the data so that elements of the same cluster (or group) are more similar to each other than to those from different clusters.

These unsupervised learning algorithms have an incredible wide range of applications and are quite useful to solve real world problems such as anomaly detection, recommending systems, documents grouping, or finding customers with common interests based on their purchases. In this work we will focus in two very similar clustering algorithms: K-means and Spherical K-means.

3.3.1.1 Clustering Algortithms

- ***K-Means Clustering***: K-means (MacQueen et al. 1967) is one of the simplest unsupervised learning algorithms to solve the clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as

much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early groupage is done. At this point we need to re-calculate k new centroids as barycenters of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move anymore.
Finally, this algorithm aims at minimizing an objective function, in this case a squared error function. The objective function

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \| x_i^{(j)} - C_j \|^2$$

where $\| x_i^{(j)} - C_j \|^2$ is a normally the Euclidean distance between a data point $x_i^{(j)}$ and the cluster centre C_j, is an indicator of the distance of the n data points from their respective cluster centres.

- ***Spherical K-means Clustering***: For high-dimensional data such as text documents (represented as TF-IDF vectors, which is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus (Rajaraman & Ullman 2011)) and market baskets, *Cosine Similarity* has been shown to be a superior measure to Euclidean distance (Strehl et al. 2000). The implication is that the direction of a document vector is more important than the magnitude.
The spherical k-means algorithm aims to maximize the average cosine similarity objective.

  In Chapter 4 , we use Cosine similarity (or cosine distance) to clusters country with similar performance in the Trade Market. Let's remember it's definition.

  The ***Cosine similarity*** is a measure of similarity that can be used to compare documents or, say, give a ranking of documents with respect to a given vector of query words. Let **A** and **B** (vectors in Fig 3.7) be two vectors for comparison. Using the cosine measure as a similarity function, we have

$$sim(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{\| \mathbf{A} \| \| \mathbf{B} \|}$$

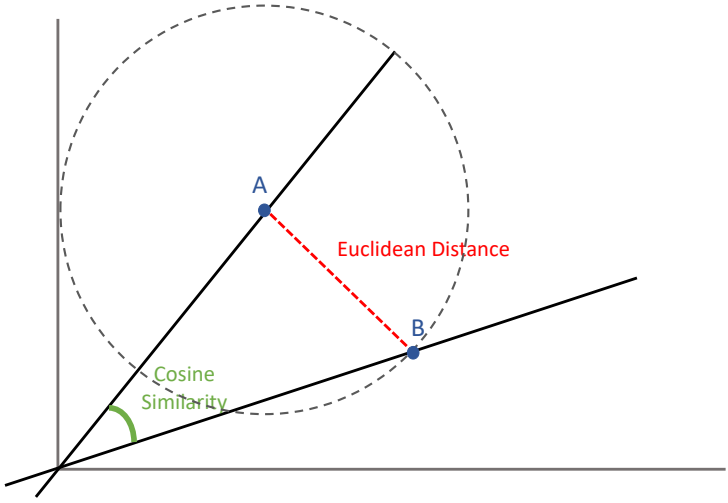


Figure 3.7 – Euclidean Distance and Cosine Similarity between vectors **A** and **B**. Euclidean Distance measures how far apart two points in a n-dimensional space are, i.e. it measures the length of a straight line from point A to point B. Cosine Similarity measures their similarity in orientation, i.e. the angle between two vectors **A** and **B** with vertex at zero

where $\| \mathbf{A} \|$ is the Euclidean norm of vector $\mathbf{A} = (a_1, a_2, \ldots a_p)$, defined as $\sqrt{x_1^2 + x_2^2 + \ldots x_p^2}$. Conceptually it is the length of the vector. Similarly, $\| \mathbf{B} \|$ is the Euclidean norm of vector **B**. The measure computes the cosine of the angle between vectors **A** and **B**. The cosine of 0° is 1, and it is less than 1 for any angle in the interval $(0, \pi]$ radians. It is thus a judgment of orientation and not magnitude: two vectors with the same orientation have a cosine similarity of 1, two vectors oriented at 90° relative to each other have a similarity of 0, and two vectors diametrically opposed have a similarity of -1, independent of their magnitude. The cosine similarity is particularly used in positive space, where the outcome is neatly bounded in $[0, 1]$.

3.4 Precision and Recall

Machine learning models ought to be able to give accurate predictions in order to create real value for a given organization.
While training a model is a key step, how the model generalizes on unseen data is an equally important aspect that should be considered in every machine learning pipeline. We need to know whether it actually works and, consequently, if we can trust its predictions. Could the model be merely memorizing

the data it is fed with, and therefore unable to make good predictions on future samples, or samples that it hasn't seen before? Model evaluation aims to estimate the generalization accuracy of a model on future (unseen/out-of-sample) data.
After the creation and training and testing of the model we have the prediction – the results of the model – and reality – the test set. We want to know how much these two sets overlap. There are four possible cases:

- A ***True Positive*** is an outcome where the model correctly predicts the positive class.
- Similarly, a ***True Negative*** is an outcome where the model correctly predicts the negative class.
- A ***False Positive*** is an outcome where the model incorrectly predicts the positive class.
- And a ***False Negative*** is an outcome where the model incorrectly predicts the negative class.

In the evaluation of a model the concepts of Precision and recall are extremely important. While precision refers to the percentage of the results which are relevant, recall refers to the percentage of total relevant results correctly classified by the algorithm (See Figure 3.8).

Recall or Sensitivity (as it is called in Psychology) is the proportion of *Real Positive* cases that are correctly *Predicted Positive*. It tends not to be very highly valued in Information Retrieval (on the assumptions that there are many relevant documents, that it doesn't really matter which subset we find, that we can't know anything about the relevance of documents that aren't returned) (Powers 2011).
However, Recall has been shown to have a major weight in predicting success in several context, for example, in a Medical context Recall is primary (Fraser & Marcu 2007).

In some contexts it is referred to as True Positive Rate *(TPR)*. Recall is defined, with its various common appellations, by the equation:

$$Recall = TPR = \frac{TP}{TP + FN}$$

Conversely, ***Precision*** or Confidence (as it is called in Data Mining) denotes the proportion of *Predicted Positive* cases that are correctly *Real Positives*. This is what Machine Learning, Data Mining and Information Retrieval

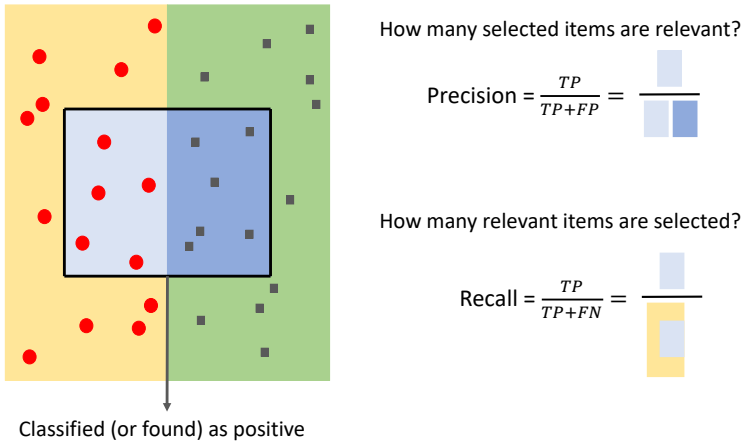


Figure 3.8 – In machine learning evaluation, ***Precision*** (also called positive predictive value) is the fraction of relevant instances among the retrieved instances, while ***Recall*** (also known as sensitivity) is the fraction of the total amount of relevant instances that were actually retrieved. Both precision and recall are therefore based on an understanding and measure of relevance.

focus on. It can however analogously be called True Positive Accuracy *(TPA)*, being a measure of accuracy of Predicted Positives in contrast with the rate of discovery of Real Positives *(TPR)*. Precision is defined as:

$$Precision = TPA = \frac{TP}{TP+FP}$$

These two concepts (Recall and Precision) play a relevant role in the evaluation of the predictive model created in Chapter 5 in which we have created a model using Machine Learning in order to predict International conflicts.

3.5 Confusion Matrix

The Confusion Matrix is a metric used to quantify the performance of a machine learning classifier. It is used when there are two or more classes as the output of the classifier.
Confusion matrices are used to visualize important predictive analytics like recall, specificity, accuracy, and precision. These are useful because they give direct comparisons of values like True Positives, False Positives, True Negatives and False Negatives. In contrast, other machine learning classification metrics like “Accuracy” give less useful information, as Accuracy is simply

the difference between correct predictions divided by the total number of predictions.

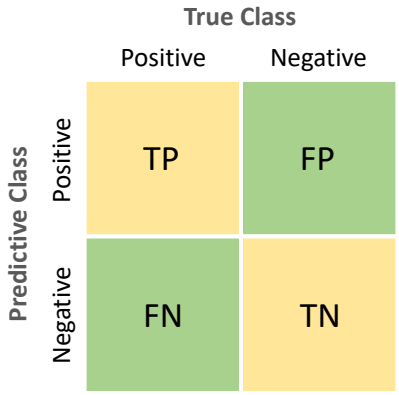


Figure 3.9 – Confusion matrix. From the top-left corner, clockwise: (**TP**) True Positives, (**FP**) False Positives, (**TN**) True Negatives, (**FN**) False Negatives.

The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix. It shows the ways in which your classification model is confused when it makes predictions. It gives us insight not only into the errors being made by a classifier but more importantly the types of errors that are being made.

3.6 Capacity, Overfitting and Underfitting

The central challenge in machine learning is that we must perform well on new, previously unseen inputs—not just those on which our model was trained. The ability to perform well on previously unobserved inputs is called *generalization*. Typically, when training a machine learning model, we have access to a training set, we can compute some error measure on the training set called the *training error*, and we reduce this training error. So far, what we have described is simply an optimization problem. What separates machine learning from optimization is that we want the *generalization error*, also called the test error, to be low as well.
The generalization error is defined as the expected value of the error on a new input. Here the expectation is taken across different possible inputs, drawn from the distribution of inputs we expect the system to encounter in practice. We typically estimate the generalization error of a machine learning model by measuring its performance on a test set of examples that were collected sepa-

rately from the training set.

When we use a machine learning algorithm, we do not fix the parameters ahead of time, then sample both datasets. We sample the training set, then use it to choose the parameters to reduce training set error, then sample the test set. Under this process, the expected test error is greater than or equal to the expected value of training error.

Suppose that we have a training set consisting of a set of points $x_1, \ldots, x_n$ and real values y_i associated with each point x_i. We assume that there is a function with noise $y = f(x) + \varepsilon$, where the noise, ε , has zero mean and variance σ^2.

We want to find a function $\hat{f}(x;D)$, that approximates the true function $f(x)$ as well as possible, by means of some learning algorithm based on a training dataset (sample) $D = \{(x_1, y_1) \ldots, (x_n, y_n)\}$. We make "as well as possible" precise by measuring the mean squared error between y and $\hat{f}(x;D)$: we want $(y - \hat{f}(x;D))^2$ to be minimal, both for $x_1, \ldots, x_n$ and for points outside of our sample. Of course, we cannot hope to do so perfectly, since the y_i contain noise ε; this means we must be prepared to accept an irreducible error in any function we come up with.

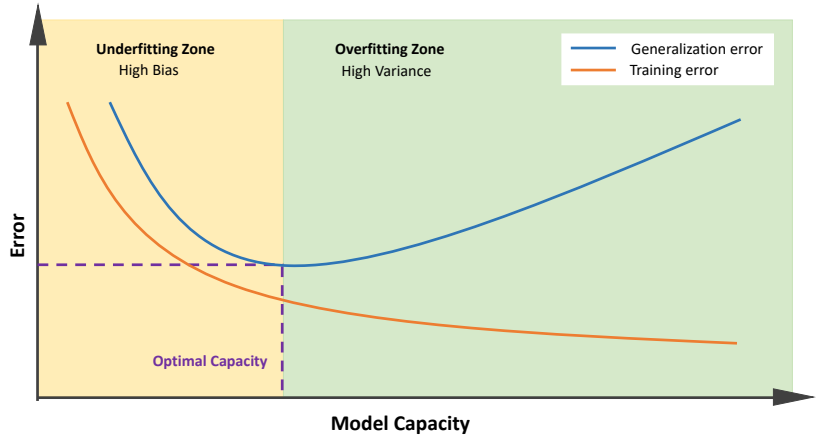


Figure 3.10 – Machine learning model capacity and error on a bias-variance spectrum. Bias refers to the error that is introduced by constructing an excessively simple model with poor prediction accuracy. Variance, on the other hand, refers to the error incurred by building an excessively complex model, attuned to noise in the training data.

Finding an $\hat{f}$ that generalizes to points outside of the training set can be done with any of the countless algorithms used for supervised learning. It turns out that whichever function $\hat{f}$ we select, we can decompose its expected error on an unseen sample x as follows:

$$\mathrm{E}_D\left[\left(y-\hat{f}(x;D)\right)^2\right] = \left(\mathrm{Bias}_D\left[\hat{f}(x;D)\right]\right)^2 + \mathrm{Variance}_D\left[\hat{f}(x;D)\right] + \sigma^2$$

where Bias is the error due to the expected value from the estimated model being different from the expected value of the true model.

$$\mathrm{Bias}_D\left[\hat{f}(x;D)\right] = \mathrm{E}_D\left[\hat{f}(x;D)\right] - f(x)$$

And Variance is the error due to the model being sensitive to small fluctuations in the data set.

$$\mathrm{Variance}_D\left[\hat{f}(x;D)\right] = \mathrm{E}_D[\left(\mathrm{E}_D[\hat{f}(x;D)] - \hat{f}(x;D)\right)^2].$$

The factors determining how well a machine learning algorithm will perform are its ability to:

- Make the training error small (low Bias).
- Make the gap between training and test error small (low Variance).

These two factors correspond to the two central challenges in machine learning: ***underfitting*** and ***overfitting***. ***Underfitting*** occurs when the model is not able to obtain a sufficiently low error value on the training set, it means a high Bias. ***Overfitting*** occurs when the gap between the training error and test error is too large even is the training error is low. In other words, overfitting occurs when the bias is low, but Variance is high.

In Figure 3.10 we plot the training error vs. model capacity to show the bias-variance spectrum. Two distinct zones exist, namely, overfitting and underfitting zones. During model training, there may exist a hinge point that represents the achievable optimal capacity of the model. Prior to this point, the model has high bias and is underfitted. Subsequent to the hinge point, if trained, the model will eventually become overfitted and have high variance.

We can control whether a model is more likely to overfit or underfit by altering its ***capacity***. Informally, a model's capacity is its ability to fit a wide variety of functions. Models with low capacity may struggle to fit the training set. Models with high capacity can overfit by memorizing properties of the training set that do not serve them well on the test set.

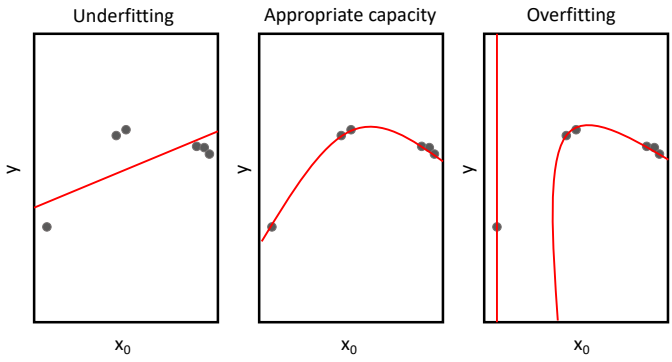


Figure 3.11 – We fit three models to this example training set. The training data was generated synthetically, by randomly sampling x values and choosing y deterministically by evaluating a quadratic function. *(Left)* A linear function fit to the data suffers from underfitting—it cannot capture the curvature that is present in the data. *(Center)* A quadratic function fit to the data generalizes well to unseen points. It does not suffer from a significant amount of overfitting or underfitting. *(Right)* A polynomial of degree 9 fit to the data suffers from overfitting.

One way to control the capacity of a learning algorithm is by choosing its *hypothesis space*, the set of functions that the learning algorithm is allowed to select as being the solution. For example, the linear regression algorithm has the set of all linear functions of its input as its hypothesis space (See Figure 3.11). We can generalize linear regression to include polynomials, rather than just linear functions, in its hypothesis space. Doing so increases the model's capacity as long as we don't cross the border between a good capacity model and overfitted model.

Our modern ideas about improving the generalization of machine learning models invoke a principle of parsimony that is now most widely known as ***Occam's razor*** (c. 1287-1347). This principle states that among competing hypotheses that explain known observations equally well, one should choose the "simplest" one. This principle is implicitly used also for dealing with noise, in order to avoid overfitting a noisy training set by rule truncation or by pruning of decision trees. This idea was formalized and made more precise in the 20th century by the founders of statistical learning theory (Vapnik & Chervonenkis 2015, 1982, Blumer et al. 1989, Cortes & Vapnik 1995). In Figure 3.11 for example, both plots in the center and right fit the data, following Occam's razor principle, we must choose the model of the center, which is the simplest

one that fit the data.

3.7 Feature Importance

Feature importance refers to a class of techniques for assigning scores to input features to a predictive model that indicates the relative importance of each feature when making a prediction. It can be calculated for problems that involve predicting a numerical value, called regression, and those problems that involve predicting a class label, called classification.
Feature importance scores are useful and can be used in a range of situations in a predictive modelling problem, such as:

- ***Better understanding the data***: The relative scores can highlight which features may be most relevant to the target, and the converse, which features are the least relevant.
- ***Better understanding a model***: Inspecting the importance score provides insight into a specific model and which features are the most important and least important to the model when making a prediction.
- ***Improving predictive model***: This can be achieved by using the importance scores to select those features to delete (lowest scores) or those features to keep (highest scores).

There are many ways to calculate feature importance scores and many models that can be used for this purpose. The three main types of more advanced feature importance are: feature importance from model coefficients, from decision trees and from permutation testing. In this work we will focus in Feature importance from decision trees.

3.7.1 Feature Importance from Decision Trees

Decision trees recursively split features with regard to their target variable's purity. The algorithm is designed to find the optimal point of the most predictive feature in order to split 1 dataset into 2. These 2 new dataset's target variable will be more pure than the original dataset's.
"Pure" is the key word here, however. What does that word mean, exactly? In a general sense "purity" can be thought of as how homogenized a group is. But homogeneity can mean different things depending on which mathematical backbone your decision tree runs on. The 2 most popular backbones for decision tree's decisions are Gini Impurity and Information Entropy.

Gini Impurity is a measurement of the likelihood of an incorrect classification of a new instance of a random variable, if that new instance were randomly classified according to the distribution of class labels from the data set. Gini impurity is lower bounded by 0, with 0 occurring if the data set contains only one class.
The Gini Impurity of a given node m with $k \in K$ being the classes is defined as follows:

$$GI_m = 1 - \sum_{k \in K} (p_k(m))^2 \tag{3.7.1}$$

where $p_k(m)$ is the class frequency of class k in node m.

On the other hand, ***Entropy*** is more computationally heavy due to the log in the equation. Like Gini, The basic idea is to gauge the disorder of a grouping by the target variable. Instead of utilizing simple probabilities, this method takes the log base2 of the probabilities (you can use any log base, however, as long as you're consistent). The entropy equation uses logarithms because of many advantageous properties.

$$Entropy_m = -\sum_{k \in K} p_k(m) \log_2 p_k(m) \tag{3.7.2}$$

In this work, we have used Gini impurity as a criterion for selecting split points.
Before to define the Feature importance, let's introduce this formula for the Impurity reduction (or Gini Gain) $Gain_m$ for a given node m, based in Gini Impurity (GI):

$$Gain_m = GI_m - (weight_l \cdot GI_l - weight_r \cdot GI_r) \tag{3.7.3}$$

For this node you can compute the Gini Impurity GI_m (see equation 3.7.1). Now this node has two children: a left node l and a right node r. For these the Gini Impurities are GI_l and GI_r respectively. The weights are defined as the share of the parents examples in a child node (e.g. $weight_l = N_l/N_m$ where N is the number of examples in a node or leaf).
If you substract the Gini impurities of the child nodes from the Gini Impurity of node m, then you know by how much the split performed by this node has decreased the overall Gini Impurity.

Now, to derive the total impurity reduction of a given feature f in tree t you need to sum across all nodes $m \in M_f^t$ which perform a split on that feature t and divide it by the total impurity reduction number of all nodes of that tree:

$$Importance_f^{(t)} = \frac{\sum_{m \in M_f^t} Gain_m}{\sum_f \sum_{m \in M_f^t} Gain_m} \tag{3.7.4}$$

(Note that due to this normalization step your feature importance's sum up to 1).
Eventually, the total importance of a feature f is calculated across all trees t in your random forest with a total number of trees T:

$$Importance_f = \frac{1}{T} \sum_{t=1}^{T} Importance_f^{(t)} \tag{3.7.5}$$

Chapter 4

The Struggle for Existence in the World Market Ecosystem

Global trade can be considered as a complex system, whose sophisticated behavior emerges from its many interacting parts – countries exporting products in different importing markets. This systemic view has been adopted in the past and it proved to be an effective one. Diversity and product relatedness in the export basket of countries and regions has been used as proxy of their economic solidity (Hidalgo et al. 2007, Neffke et al. 2011, Saviotti & Frenken 2008, Ženka et al. 2014). Different economic complexity indexes have proven to be incredibly successful in predicting future economic growth, better than traditional indicators such as years of schooling or the quality of public institutions (e.g. in terms of resistance to corruption) (Hausmann et al. 2014, Tacchella et al. 2012, Cristelli et al. 2013, Poncet & de Waldemar 2013). The complexity approach illustrated how knowledge flows across neighbouring countries (Bahar et al. 2014, Martin 2015), and how these dynamics allow us to predict structural change (Bustos et al. 2012, Klimek et al. 2012), suggesting new avenues for development (Stein et al. 2014).

Here, we enrich the literature on complexity and economic development by further investigating its relationship with ecology. Traditionally, export patterns are considered as static and only locally related to the other countries in the world. We draw relations among countries by inferring potential competition among them across time. We see a pair of importer-product, for instance the car market in the US, as an evolving trade "niche", with exporters appearing and disappearing like fit and unfit organisms in an ecosystem. In

our analysis, the fitness of an economy in a niche correlates with its ability to displace (out-compete) unfit economies. If this happens consistently in many other car importing countries, then the fit economy should be able to grow its car exporting business in the future.

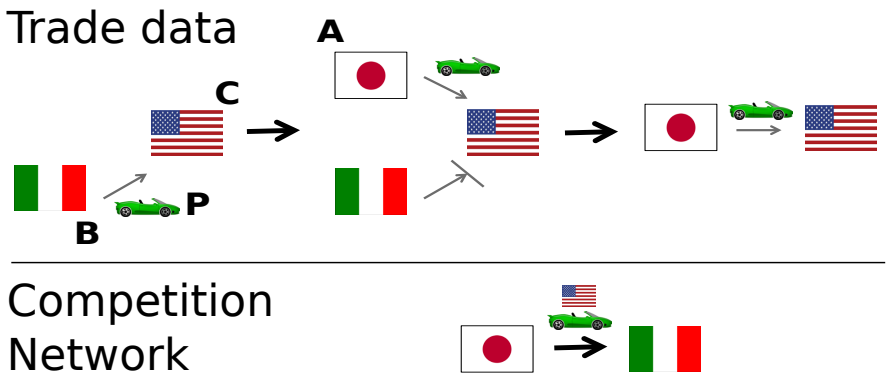


Figure 4.1 – An example of a displacement relation. From left to right we observe a pattern in the yearly car import data from the US. In the first year, only Italy is present. In the second year, Japan appears in the market. In the third year, Italy disappears from the market. This pattern from the trade data is represented in the competition network as a directed edge from the displacer (Japan) to the displaced (Italy). The edge is labelled with its layer: the car market in the United States.

We test this theory by creating a competition network, connecting country a to country b if a's appearance in a market preceded b's disappearance, as illustrated in Fig 4.1. Since we have different products and different years in which these relationships can be established, we use a multilayer network model (Berlingerio et al. 2013, Kivelä et al. 2014*b*). Our competition network is a peculiar structure, because traditionally networks are used to express positive relations, while in our case the relation is negative (competition). Negative relationships are relatively less explored than positive ones, and previous works showed they obey to different dynamics (Leskovec et al. 2010*b*, Szell et al. 2010), whether they are studied using social balance or status theory frameworks (Fişek et al. 1991, Willer 1999). For instance, negative edges are much less prone to generate triangles and high clustering (Easley & Kleinberg 2010, Davis 1967). They also allow for the emergence of more complex network motifs (Bachi et al. 2012).

World's markets are highly dynamic, with exporters frequently appearing and disappearing in a niche. This means that basic statistical properties of our competition network are not enough to unveil the potential displacement patterns. In the competition network there is a very strong correlation be-

tween out- and in-degree, which record the number of displacements a country caused and to which it was subject, respectively. As a consequence, it is not possible to detect if a country tends to out-compete more than it is out-competed. Moreover, in the competition network we observe a number of unexpected properties, such as reciprocity – countries repeatedly displacing each other – and triangles – cycles of countries where the displacers are displaced by their displaced's displaced. To tackle these issues, we need to employ non-local analysis techniques, and take into account indirect patterns in the directed graph. This is an approach frequently used in network science, from ranking a node's structural importance (Kleinberg 1999*a*), to the measurement of node similarity (Blondel et al. 2004).

In this work, we choose to borrow the tools of a third non-local node-centric network analysis: role detection. In the role detection literature, different connectivity patterns are used to classify nodes in particular network roles (Henderson et al. 2012). One specific and very popular case is the one of community detection, which aims at finding densely connected modules (Coscia et al. 2011). Node roles have been used to describe a wide range of phenomena, from metabolic networks (Guimera & Amaral 2005) to the connectivity in the brain (Meunier et al. 2010, Sporns & Betzel 2016). Specifically, we borrow the approach described in (Cooper & Barahona 2010). In this method, we compute a feature vector for each node describing the size of the out/in neighbourhoods at a given network distance. Through this vector, we redefine a fit economy from "able to out-compete many countries" to "able to out-compete many countries who are able to out-compete many countries" – up to six degrees of separation. We perform the same operation for in-degree roles (an unfit exporter is an exporter who is "displaced by countries who are prone to be displaced themselves").

We define three roles for exporters: "Out-competing" countries are countries which consistently score high in out-degree roles and low in in-degree roles; "Displaced" countries are countries scoring the opposite (low in out-degree roles and high in in-degree roles); and "Transitioning" countries, whose scores in both roles are comparable.

This classification is a meaningful one. We test it by predicting the future export patterns of countries. *Countries classified as out-competing in a particular product in a particular decade show significant export growth patterns in that product in the following decade.* This means, for instance, that if

Japan is classified as out-competing in worldwide car exports in the 1960-1970 decade, then its car exports are going to grow significantly in the 1970-1980 decade. This result is consistent across decades – with the exception of the last decade for the lack of a long enough time span to test the data – and across different product types – with the exception of the ones dominated by profitable natural resources such as crude oil.

Even if we are not observing direct competition relationships, due to the correlative nature of our edge creation process, the resulting roles are informative of future patterns in global trade. Our method can be used to detect emerging countries in the global market for a particular product.

The common point of all the works developed during these years and presented in this thesis, is the study of the appearance and disappearance of links in the networks and their effects on both their dynamics and their reorganization. In this particular chapter, we have the disappearance of positive (commercial) links between countries that are replaced by other ones, and based on this, we study the effects of these disappearance and appearance of links through a competition network.

The aim of this section is to describe the process starting from raw trade data to the creation and analysis of multilayer competition network. We start by describing the data sources and the cleaning phase. We then provide an informal example, before detailing out the full procedure.

4.1 Methods

4.1.1 Data & Cleaning

The data contains the entire set of worldwide trade relationships from 1962 until 2013. The data has been collected by the UN Comtrade organization (`https://comtrade.un.org/`), and cleaned by CEPII (Mayer et al. 2008). A product is defined as a 4 digit SITC category. A product can be, for instance, poultry meat for eating (code 0123), or ferro-manganese (code 6714).

UN Comtrade gathers data about all sovereign countries and territories in the world. Many of these sovereign entities are very small and cause wide fluctuations in the observations. For this reason, we focus only on larger and more stable countries. We drop countries with less than 300k inhabitants and/or with a total GDP lower than 300 million US dollars. Given our large time span, we

also have data about countries who do not exist anymore (for instance, Yugoslavia). We drop the observations involving them too.

Even if the data is gathered at a 4-digit level of detail, we find that this is too granular for our analytic aims. We exploit the fact that SITC is a hierarchical classification: all products whose code starts with the same digit are related to each other. Thus we aggregate the trade data at the 1-digit level, summing up the trade flows of all products classified under the first digit.
The meaning of each digit are: SITC 1: *Beverages and tobacco*, SITC 2: *Crude materials except fuels*, SITC 3: *Mineral fuels, lubricants and related materials*, SITC 4: *Animal and vegetable oils, fat and waxes*, SITC 5: *Chemical and related products*, SITC 6: *Manufactured goods clasified chiefly by material*, SITC 7: *Machinery and transport equipment* and SITC 8: *Miscellaneous manufactured articles*.

Finally, we represent the data as a four dimensional tensor $\mathbb{T}_{p,i,e,y}$. The dimensions of the tensor are: product (p), importer (i), exporter (e) and year (y). Basically, $\mathbb{T}_{p,i,e,y}$ can be seen as a set of matrices $T^{p,i}_{e,y}$, one for each pair of product p and importer i. The matrix contains, for each exporter e a timeline vector recording, for each year y, the amount of trade in p flowing from e to i. So, each $T^{p,i}_{e,y}$ is a $e \times y$ matrix.

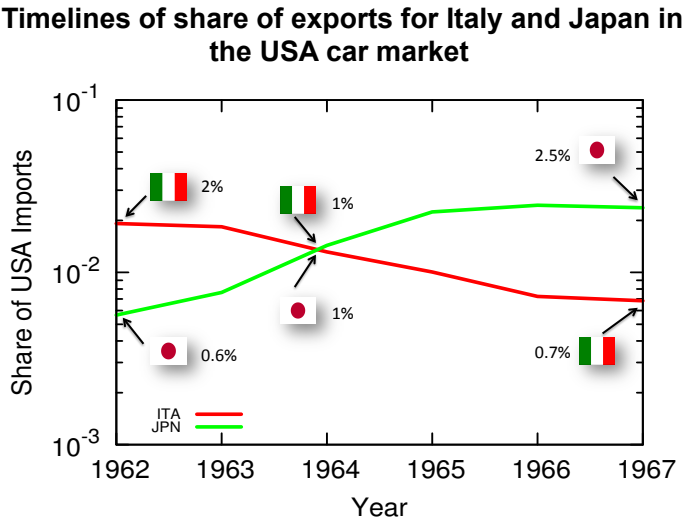


Figure 4.2 – Timeline of Japan's and Italy's exports in the US car market in the 60s.

4.1.2 Inferring Competition Relationships

To better understand how the procedure works, let us start with an example detailing how a single edge in our multilayer directed network is established. We consider the car market in the United States. We focus on the export patterns of two countries in a potential relationship of competition: Japan and Italy. Fig 4.2 depicts the share of US car market of Japan and Italy, from 1962 to 1967.

Each step corresponds to a parameter in our methodology, which is reported between parenthesis, and which we formally define in the rest of this section:

1. Detect whether there is an anti-correlation between the export patterns of the two countries (δ);
2. Detect whether one of the two countries appeared from the market, while the other disappeared (κ);
3. Detect whether the disappearing country did not reappear in the market immediately after the event (λ).

In the first step we calculate the correlation coefficients of Japan's and Italy's export timelines. The two timelines have a 1.67 correlation distance. If we assume that this is higher than our δ parameter, we can say that there is a potential competition edge between Japan and Italy.

In the second step, we check if either country appeared in the market while the other disappeared. This is regulated by the parameter κ, which defines the relative market share below which an exporter is considered to have "disappeared". We do not set $\kappa = 0$, because a complete disappearance is a rare event. If we set $\kappa = 1\%$, we can say that Italy disappeared while Japan appeared, suggesting that the competition edge runs from Japan to Italy.

Finally, we check if Italy was absent from the US car market for at least λ years. Assuming $\lambda = 2$, also this final test is positive. We then draw a directed edge in our competition network from Japan to Italy.
We now describe more formally each step in the following subsections. We remind that the same operation is performed for each product 1-digit class and for each decade separately (1960-70, 1970-80, 1980-90, 1990-2000).

Step #1: Detecting the Potential Edges

To detect the candidate relationships (i.e. the edges), we slice $\mathbb{T}_{p,i,e,y}$ such that we consider each pair of product and importer country independently from all other pairs, i.e. we analyze one $T^{p,i}_{e,y}$ matrix at a time. We column normalize each $T^{p,i}_{e,y}$, such that each entry will report the share e exported of p to i in y. We then calculate the row-wise correlation distance between each pair of exporting countries:

$$d_{e_1,e_2} = 1 - corr(\overrightarrow{e_1}, \overrightarrow{e_2}),$$

where e_1 and e_2 are two exporting countries, $\overrightarrow{e_1}$ and $\overrightarrow{e_2}$ are the vectors of $T^{p,i}_{e,y}$ corresponding to them, and *corr* is a function calculating the Pearson correlation of two vectors. d_{e_1,e_2} establishes the distance in the trends of e_1 and e_2, regardless of their relative volume. Remember that here we are interested in linking countries that are *dissimilar* to each other, so we perform an operation that is opposite to what is usually done in network science: two countries with very different market shares are not connected with an edge if their trends are similar.

d_{e_1,e_2} takes values between 0 ($\overrightarrow{e_1}$ and $\overrightarrow{e_2}$ are perfectly correlated) and 2 ($\overrightarrow{e_1}$ and $\overrightarrow{e_2}$ are perfectly anti-correlated). d_{e_1,e_2} equals to 1 for linearly uncorrelated vectors. The δ threshold establishes the value below which we discard the potential edge. Given the value domain of d_{e_1,e_2}, δ must be higher than 1 (otherwise we would consider positively correlated vectors).

Step #2: Detecting the Potential Edge Direction

To establish if the anti-correlation of exports can lead to a potential competition edge – and its direction – we have several requirements to satisfy:

1. i must have not stopped importing p;
2. Either e_1 or e_2 has to have ceased to export p to i – this is the potential displaced exporter;
3. Whenever e_1 ceased to export p to i, e_2 still has to be exporting the product, and vice versa – this is the potential out-competitor exporter;
4. The potential displaced exporter must have been exporting p previously.

To satisfy requirements #1, #2 and #3, we use our second threshold, κ, which represents the minimum export share to be considered still exporting p to i. If an exporter e has less than κ market share of p in i, then e in this context is considered to have ceased exporting. Being κ a relative threshold, we can make sure that the size of the importing market is not affecting our definition of relationship, which would make too easy to have competition relationships in small countries and small products.

Requirement #1 is now satisfied automatically: it is impossible to have a share of export larger than κ if the denominator is 0 (i.e. i did not import p), because the fraction would be undefined.

Each candidate edge is a quadruple (p, i, e_1, e_2). For each p and i, we binarize $\overrightarrow{e_1}$ and $\overrightarrow{e_2}$ as follows:

$$e_y^* = \begin{cases} 1 & \text{if } e_y > \kappa \\ 0 & \text{otherwise.} \end{cases}$$

where e_y is $\overrightarrow{e}$'s value at time y. Then we calculate $\overrightarrow{e_1}^* \oplus \overrightarrow{e_2}^*$, which is the XOR product of the two vectors: the result is `true` for a year y if in y e_1 exported more than κ share of p to i and e_2 did not, and vice versa. This satisfies requirements #2 and #3.

We satisfy requirement #4 by removing the first streak of `true` values in $\overrightarrow{e_1}^* \oplus \overrightarrow{e_2}^*$. The first streak of `true` values represents a period in which either e_1 or e_2 did not start exporting p to i yet. Thus, we cannot talk about either of them being displaced, because they did not have a chance to interact with each other yet.

We can now easily detect the edge direction. The country which disappeared from the importing market – say e_2 – is the displaced one and it is thus on the receiving end of the edge, which originates from the other country – in our case e_1.

Step #3: Establishing the Edge

Before adding the edge to the multilayer competition network we have to ensure that the displaced exporter has actually been displaced. We test this by checking if the cessation of its exports has been longer than a certain number

of years.

We satisfy this requirement by using our third parameter, λ, which represents the minimum number of years needed to declare a potential displaced exporter out of the market. This means that the displaced country has to cease exporting at least κ share of p to i for λ consecutive years, while its out-competitor consistently stays above the κ threshold in the same period. This means that we have to find at least λ consecutive `true` values in $\overrightarrow{e_1}^* \oplus \overrightarrow{e_2}^*$.

The result of these three steps is another tensor, $\mathbb{D}_{p,d,e_1,e_2}$. $\mathbb{D}_{p,d,e_1,e_2}$ is a directed multilayer network, where each layer represents a pair of product p and decade d. For simplicity, $\mathbb{D}$ is collapsed over the importer dimension i using a logical OR operator. In other words, each layer contains an directed graph connecting two countries ($e_1 \rightarrow e_2$) if their trends in exporting p during d satisfy all posited requirements for at least one importer i:

- $\overrightarrow{e_1}$ and $\overrightarrow{e_2}$ are strongly anti-correlated (correlation distance $> \delta$);
- $\overrightarrow{e_2}$ contains at least λ consecutive values $< \kappa$ not at its beginning;
- The corresponding $\overrightarrow{e_1}$ values are $\geq \kappa$.

We then say that e_1 is an out-competitor of e_2 in product p. The edges are weighted according to in how many importers i this competition relationship has been established.

4.1.3 Detecting Roles

We now turn to the detection of node roles in the multilayer competition network. We follow closely the methodology delineated in (Cooper & Barahona 2010). In that paper, Cooper and Barahona propose to group nodes according to their role in the network, defined in terms of the overall pattern of incoming and outgoing flows. According to this, we expect to find three categories of countries: out-performing, displaced and transitioning. The roles emerge by looking at the path profile of each node. A path profile is a vector computed from the powers of the adjacency matrix weighted with a scale parameter. Then, we define path profile templates and we cluster nodes according to the similarity their path profiles have when compared to the templates.

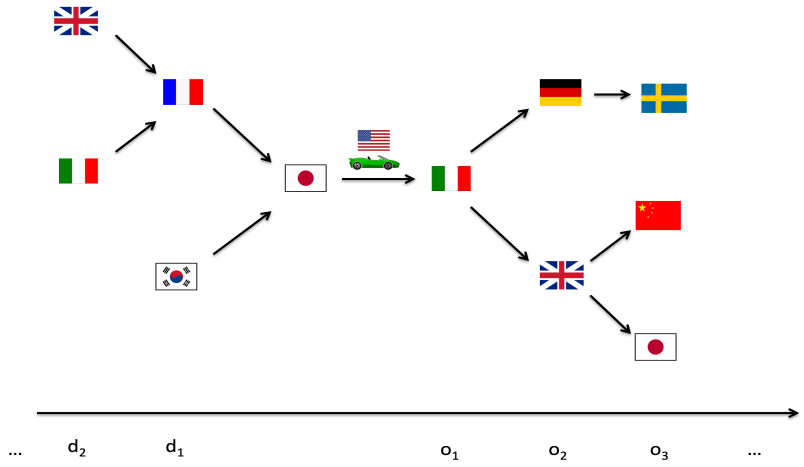


Figure 4.3 – Sequence of displacement events for Japan, located at the degree of separation 0 (the one between d1 (displacement at 1 degree of separation) and o1 (out-performing at 1 degree of separation), represented with arrows in the direction of displacement. We consider up to 6 degrees of separation in each direction.

Consider a directed network with N nodes and an asymmetric adjacency matrix M. Consider its $[M^k\mathbf{1}]$ vector, where $\mathbf{1}$ is the $N \times 1$ vector of ones. The i-th entry of this vector is the number of displacement events happening in all chains of length k originating from node i. For $k = 1$, $[M^k\mathbf{1}]$ is equivalent to the out-degree vector of M. In the same way, the number of displacement events happening in all chains of length k ending in node i is $[M'^k\mathbf{1}]_i$, where M' is the transpose of M. For $k = 1$, this is equivalent to the in-degree vector of M.

We construct a matrix that compiles the incoming and outgoing paths of all lengths up to k_{max} by appending the column vectors indexed by path length and scaled by the factors βk:

$$X = \begin{bmatrix} \mathbf{x}_1 \\ . \\ . \\ . \\ \mathbf{x}_N \end{bmatrix} \equiv [\underbrace{\ldots(\beta M')^k\mathbf{1}\ldots}_{k_{max}} \mid \underbrace{\ldots(\beta M)^k\mathbf{1}\ldots}_{k_{max}}],$$

where $\beta = \alpha/\lambda_1$, with λ_1 being the largest eigenvalue of the adjacency matrix and $\alpha > 0$. α governs how much weight we put on local or global flow structure. Setting $\alpha \sim 0$ means that in- and out-degrees dominate over the other values when calculating roles. Given the issues caused by using in- and out-degree that we will describe in the next section, we aim at doing the exact opposite, and thus we set $\alpha = 1$. Note that one could set an $\alpha > 1$, however that would mean that the farther relationships (mediated by more than one edge) have more weight than the more proximate ones, which we believe not to be reasonable. We also consider up to 6 degrees of separation in each direction, i.e. $k_{max} = 6$.

By following this methodology, each row vector of X contains the flow profile of a node in terms of the scaled number of displacement paths of all lengths starting and ending at that node. Following (Cooper & Barahona 2010), we group nodes if they have similar flow profiles. Nodes in the same cluster have similar flow profiles, thus they play a similar role in terms of the flow in the directed graph. To detect such nodes, we calculate the distance of each country from a synthetic template of a perfect out-competing, transitioning, and displaced exporter. We assign the country to the closest template according to the cosine distance described in Fig 3.7. The objective is to minimize the average cosine distance between a country and its template.

To create our templates we need to ensure that each element in each row vector in X takes value between 0 and 1:

$$X^* = \frac{X - min(X)}{max(X) - min(X)}.$$

Note that this operation is done row by row, i.e. $min(X)$ and $max(X)$ are calculated only considering the values of each row separately. In this way, each country is a vector of values between 0 and 1 included. If we would took the global $min(X)$ and $max(X)$, only one country could span the full domain value, narrowing down the values of all other countries, and thus making the result dominated by outliers. As a result of this operation, a hypothetical country i could be described by the following vector:

$$X_i^* = [\underbrace{0.97, 0.94, 1, 0.99, 0.92, 0.77}_{d}, \underbrace{0.42, 0.11, 0.14, 0.13, 0, 0.09}_{o}].$$

Here, the first k_{max} values are the displaced (d) role scores, while the latter k_{max} values are the out-competing (o) role scores. As a convention, we always list first the d scores in decreasing order and then the o scores in increasing order, so that the two middle values of the vector are always d_1 and o_1, i.e. the normalized in-degree and the out-degree. Generally speaking, the d_n entry in the ith row of matrix X^* is the (normalized) number of paths of length n ending at node i. A high score in displaced roles means that the country tends to be displaced by countries that are displaced themselves. The opposite is true for the out-competing role scores. Since we know that all scores must take value between 0 and 1, creating a cluster template is now trivial:

$$\mathscr{O} = [\underbrace{0,0,0,0,0,0}_{d},\underbrace{1,1,1,1,1,1}_{o}]$$

is an hypothetically perfect out-competing exporter, with zero in-degree and maximum out-degree. With the same logic, we can define the perfect displaced country $\mathscr{D}$, and the middle point, the transitioning country $\mathscr{T}$:

$$\mathscr{D} = [\underbrace{1,1,1,1,1,1}_{d},\underbrace{0,0,0,0,0,0}_{o}],$$

$$\mathscr{T} = [\underbrace{0.5,0.5,0.5,0.5,0.5,0.5}_{d},\underbrace{0.5,0.5,0.5,0.5,0.5,0.5}_{o}].$$

For each country, we calculate the cosine distance from these hypothetical perfect scenarios. We chose the cosine distance, because the intensity of the vector is not important: what matters is its direction. We assign the country to the closest template, i.e. the one scoring the lowest cosine distance among the three. The average leftover cosine distance (we also refer to this like 'Energy') is a measure of how good the clustering was, i.e. how similar each country is to its assigned template.

If an exporter has a high values for the out-degree roles and low ones for in-degree roles, then it is assigned to the "Out-competing" cluster. Vice versa, low values for the out-degree roles and high ones for in-degree roles will place the country in the "Displaced" cluster. In all other cases, when the out- and in-degree roles have comparable values, the exporter is classified as "Transitioning".

4.2 Results

4.2.1 Competition Network Statistical Analysis

The fundamental assumption of this paper is that the competition network that we build using the methodology discussed in the previous section contains information that will allow us to predict an exporter's future performance in the global market. If a country can out-compete many other countries in a product, then it is expected to export more of that product. The first question one might ask is: why do we need to calculate node roles? The number of times an exporter out-competes its rivals is simply its out-degree. Could this simpler statistical property inform us about export dynamics?
There are two reasons why this is not the case. The first reason is that out- and in-degree in the competition network are highly correlated. The second reason is that the competition network's structure is more complex than one would assume.

Fig 4.4 shows the out- and in-degree correlation. On the left we show the out-degree distribution per country, and in the middle the in-degree distribution. We can see that both distributions are very similar. In fact, the top and the bottom countries in these distributions are almost the same, sometimes in a slightly different order. On the right, we show the correlation directly. It is not possible, from this picture, to characterize any country as predominantly out-competing its rivals, because the same country will have an almost equal amount of cases in which it is displaced. So, In the case that we want to classify the countries according to their in and out-degrees, we would find that some countries are part of both classes, which makes it impossible to classify them in terms of how many times they have displaced or have been displaced. It is for this reason that in our study we have clustered the countries using the technique proposed by Cooper and Barahona in (Cooper & Barahona 2010) .

Regarding the second reason, we observe a number of topological properties that we would not expect to find in a competition network. The first one is reciprocity. When country a displaces country b in a niche, we would expect it to do so because fitter for that particular market. Yet, we observe a large number of reciprocal edges. This means that, after some time, country b reappears in the niche and displaces country a. Across our problem space (for all decades, products and parameter choice) the median reciprocity was 11.38%.

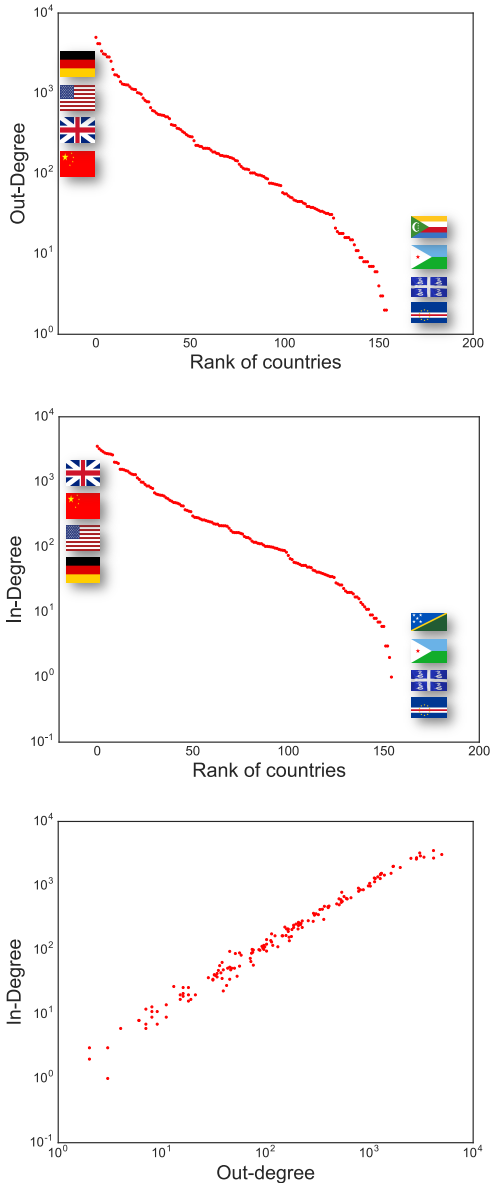


Figure 4.4 – Out- and in-degree distributions. (Top) Countries are sorted and ranked in the x-axis according to their out-degree, i.e. the number of times they out-compete another country – reported in the y-axis. (Middle) The same plot, replacing the out-degree measure with the in-degree one, i.e. the number of times the country was displaced from a niche. (Bottom) The relationship between out-degree (x-axis) and in-degree (y-axis). Each observation is a country.

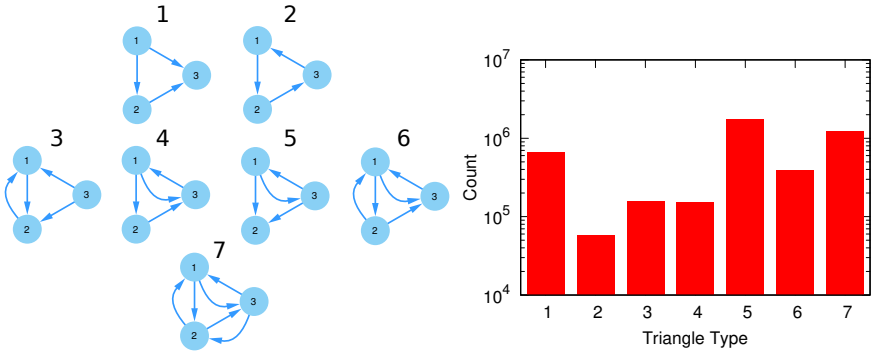


Figure 4.5 – (Left) All possible triangles in a directed graph. (Right) Frequency of different types of directed triangles in the multilayer network.

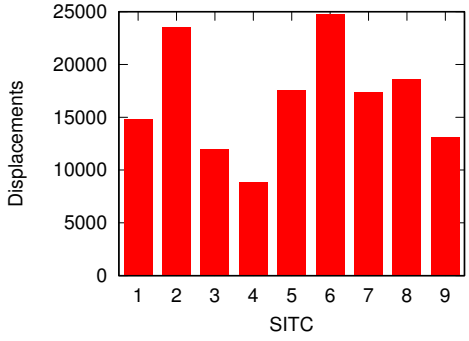


Figure 4.6 – Distribution of number of displacements per one digit SITC product.

The second surprising topological feature is the presence of a high number of triangles. Triangles are surprising because we would not expect a displaced country to displace a displacer. Yet, this happens frequently. Fig 4.5 shows on the left the seven possible types of triangles that can appear in a directed network. On the right, it depicts the counts of each type of triangle in $\sim$100 randomly chosen networks across all decades, products and parameter choices. Triangle types 5 and 7 are the most common, 7 being the case in which all three exporters are displacer of each other.

Fig 4.6 shows the distribution of number of displacements per one digit SITC product. We can see that there are products that are more dynamic than others.

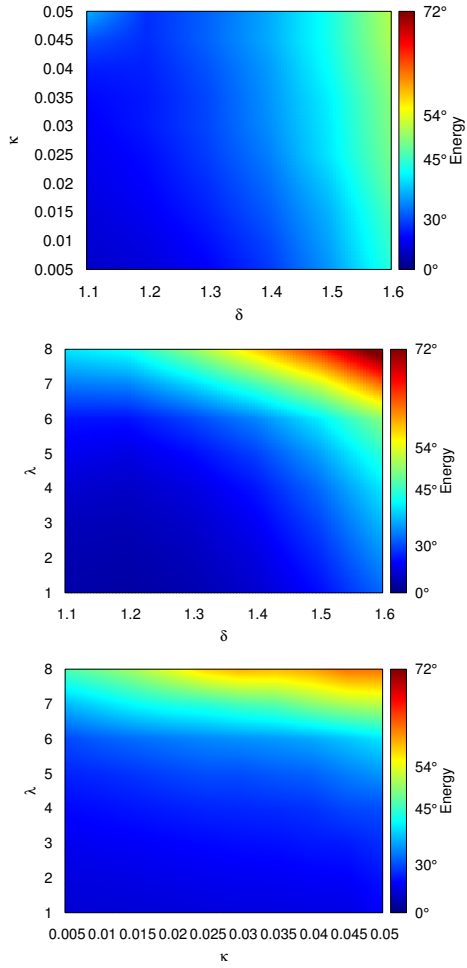


Figure 4.7 – The average cosine distance landscape of the parameter space (projected over the omitted dimension). From top to bottom: δ-κ (λ omitted); δ-λ (κ omitted); κ-λ (δ omitted).

4.2.2 Role Clusters

Before performing the clustering and the prediction task, we need to determine the optimal parameter choice, and evaluate the robustness of our results to this choice. Many topological properties of the multilayer networks are dependent on our choice of parameters. We investigate the direct effect on clustering quality of the three parameters δ, κ, λ. For each combination of parameter we calculate the average cosine distance between a country and the cluster template to which it is the most similar.

Since we have three parameters, the space of this search is three dimensional. To explore it, we project it into three two dimensional slices. We fix two parameters and then we calculate the average cosine distance (Energy) across the omitted dimensions. Fig 4.7 reports the result.

From the figure, we can see that the most important parameter that creates a rugged landscape is λ – the length of a displaced exporter disappearance necessary to determine whether it is really out of the market. This is intuitive: since we are considering a decade-long period, if we require long disappearances (e.g. 8 years) the interval in which the displacement could happen becomes very narrow (e.g. only the first two years of the period). As a consequence, there are going to be very few edges in our competition network, and displacements happening after $(10 - \lambda)$ years from the beginning of the decade are going to be ignored.

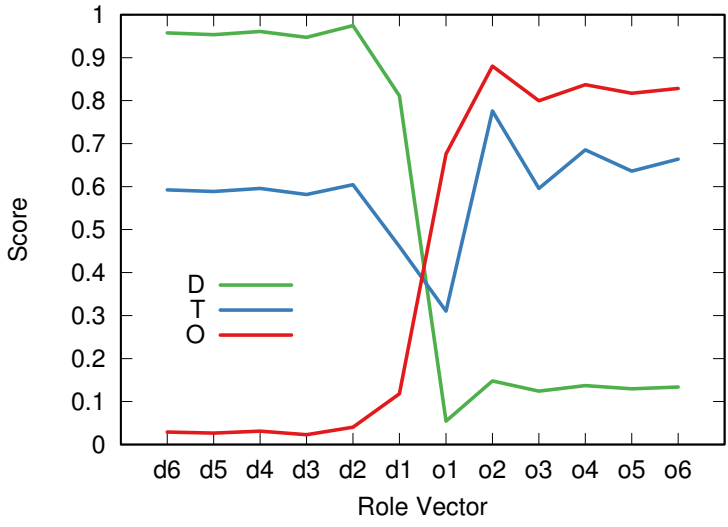


Figure 4.8 – The average role-feature scores per cluster in our example. O = Out-competing, D = Displaced, T = Transitioning.

On the other hand, the δ-κ space is very smooth, showing that results are going to be consistent no matter the level of correlation distance we require (δ) or the disappearance threshold (κ). Between the two, δ seems to be more important (there is a weak left to right gradient). Again, this is unsurprising for the same reason as before: the higher the δ the more demanding we are in

our edge creation process. For $\delta > 1.5$ we start having degenerate networks which are sparser and sparser, and where triangles are impossible.

Once we fix δ, κ, λ such as to minimize the clustering average cosine distance, we obtain our final clusters, dividing countries in out-competing, transitioning and displaced for each decade and product category. As discussed in the methods section, we have three templates and countries are matched to the template most similar to them. Here, we visualize one instance of such clustering. We average the role scores for all countries in each cluster. Fig 4.8 depicts the result.

From the figure, we can see that the clustering procedure is able to capture the essence of the network roles. Countries in the out-competing cluster have small displaced role scores on average, and high out-competing scores. The converse is true for countries in the displaced cluster. As for the transitioning countries, they tend to have high scores in both role classes. The only exception is their low score in the first displaced role. This means that transitioning countries tend to have low in-degree, although that in-degree is generated from countries with a very high in-degree – otherwise also the other displaced scores would be low.

4.2.3 Prediction

Once we fix δ, κ, λ such that we obtain the lowest residual average cosine distance, we can perform a simple predictive task. We calculate the clusters using exclusively data from a given decade, say 1971 to 1980. Then, we look at the exports of each country in that product in the next decade – from 1981 to 1990. We calculate the slope of the decade trend, normalized with the maximum export value of the top exporter in that product in that period. In this way, we have for each country its competition network cluster for a decade and its corresponding export growth in the following decade. We then calculate the mean export growth rate for each country cluster. We also calculate the standard error of the mean. This is an out-of-sample prediction, since there is no information that is used both for calculating the clusters and the growth rate: the sets of years considered are disjoint.

We perform this operation for all decades for all product classes. Table 41 reports the results – in the previous section we provided the legend for each product code. For each decade and product class (first two columns) we test if

Decade	SITC	Out-competing	Transitioning	Displaced	R^2
1960-1970	1	0.048***	0.005**	0.004**	0.151
1960-1970	2	0.039***	0.005**	0.004***	0.169
1960-1970	3	0.043***	0.005*	0.004*	0.167
1960-1970	4	0.043***	0.001*	0.003	0.129
1960-1970	5	0.144***	0.017***	0.006***	0.266
1960-1970	6	0.074***	-0.001*	0.006*	0.298
1960-1970	7	0.094***	0.007***	0.002***	0.381
1960-1970	8	0.068***	0.012**	0.008**	0.186
1970-1980	1	0.025***	0.005*	0.001**	0.165
1970-1980	2	0.024***	0.008**	0.001***	0.310
1970-1980	3	-0.019***	-0.000	-0.000	0.068
1970-1980	4	0.027***	-0.000*	0.001*	0.170
1970-1980	5	0.131***	0.002***	0.007***	0.347
1970-1980	6	0.052***	0.004***	0.003***	0.325
1970-1980	7	0.063***	0.007***	0.002***	0.401
1970-1980	8	0.107***	0.012***	0.004***	0.515
1980-1990	1	0.021***	0.002***	0.002***	0.260
1980-1990	2	0.009***	0.002*	0.001**	0.163
1980-1990	3	0.008***	-0.000	0.002	0.075
1980-1990	4	0.046***	n/a	0.002***	0.325
1980-1990	5	0.076***	0.014***	0.005***	0.305
1980-1990	6	0.027***	0.011	0.002**	0.301
1980-1990	7	0.061***	0.004***	0.002***	0.464
1980-1990	8	0.029***	0.002**	0.002**	0.148
1990-2000	1	0.042***	0.008*	0.005*	0.161
1990-2000	2	0.024***	0.004***	0.001***	0.279
1990-2000	3	0.026***	0.005*	0.003*	0.180
1990-2000	4	0.055***	0.002***	0.002***	0.183
1990-2000	5	0.103***	0.003***	0.005***	0.330
1990-2000	6	0.077***	0.007***	0.005***	0.282
1990-2000	7	0.055***	0.013*	0.001**	0.227
1990-2000	8	0.024***	0.003	0.001*	0.141
2000-2010	1	0.026***	0.006	0.005	0.063
2000-2010	2	-0.013***	-0.000*	-0.000*	0.270
2000-2010	3	0.003	-0.000	-0.002	0.005
2000-2010	4	-0.027***	0.001	0.000	0.050
2000-2010	5	0.013	0.004	-0.001	0.025
2000-2010	6	-0.015*	-0.002	-0.003	0.017
2000-2010	7	0.014	0.003	0.000	0.018
2000-2010	8	0.015**	0.001	0.001	0.037

Table 41 – The mean export growths per country. For each decade and product class (first two columns) we test if the corresponding clusters have an export value growth in the following decade in the same product significantly higher than zero. From left to right the means of: out-competing, transitioning, and displaced clusters. Last column is the R^2 of a regression using the clusters as fixed effects. (*** 3σ, ** 2.5σ, * 2σ)

Decade	SITC	Hi-Slope	Med-Slope	Lo-Slope	R^2
1960-1970	1	0.037***	0.003***	0.001***	0.136
1960-1970	2	0.027***	0.003***	0.001***	0.125
1960-1970	3	0.037***	0.002***	0.001***	0.175
1960-1970	4	0.041***	0.001***	0.001***	0.142
1960-1970	5	0.080***	0.002***	0.000***	0.151
1960-1970	6	0.054***	0.002***	0.000***	0.248
1960-1970	7	0.032***	0.000***	0.000***	0.117
1960-1970	8	0.054***	0.002***	0.000***	0.220
1970-1980	1	0.018***	0.001***	0.000***	0.119
1970-1980	2	0.018***	0.002***	0.000***	0.219
1970-1980	3	-0.012**	0.000*	0.000*	0.043
1970-1980	4	0.018***	0.000***	0.000***	0.109
1970-1980	5	0.060***	0.003***	0.000***	0.163
1970-1980	6	0.037***	0.001***	0.000***	0.250
1970-1980	7	0.022***	0.000***	0.000***	0.157
1970-1980	8	0.049***	0.003***	0.000***	0.220
1980-1990	1	0.014***	0.001***	0.001***	0.203
1980-1990	2	0.007***	0.001***	0.000***	0.150
1980-1990	3	0.005***	0.000**	0.005	0.055
1980-1990	4	0.022***	0.001***	0.004**	0.096
1980-1990	5	0.043***	0.001***	0.000***	0.217
1980-1990	6	0.018***	0.001***	0.001***	0.226
1980-1990	7	0.027***	0.000***	0.000***	0.219
1980-1990	8	0.014***	0.001**	0.000***	0.083
1990-2000	1	0.035***	0.003***	0.002***	0.174
1990-2000	2	0.018***	0.002***	0.001***	0.176
1990-2000	3	0.026***	0.000***	0.003***	0.218
1990-2000	4	0.031***	0.001***	0.001***	0.093
1990-2000	5	0.074***	0.002***	0.000***	0.247
1990-2000	6	0.048***	0.002***	0.001***	0.188
1990-2000	7	0.031***	0.000***	0.000***	0.120
1990-2000	8	0.012***	0.000**	0.001**	0.060
2000-2010	1	0.022***	0.001***	0.004**	0.072
2000-2010	2	-0.010***	-0.001***	0.000***	0.171
2000-2010	3	-0.000	0.001	-0.001	0.000
2000-2010	4	-0.014*	0.000	-0.000	0.021
2000-2010	5	0.003	0.000	0.000	0.002
2000-2010	6	-0.012*	-0.002	-0.000	0.021
2000-2010	7	0.007	0.000	-0.000	0.009
2000-2010	8	0.009**	0.000	0.000	0.027

Table 42 – The mean export growths per country. For each decade and product class (first two columns) we test if the corresponding clusters have an export value growth in the following decade in the same product significantly higher than zero. Here, the clusters are based on the slope in the previous decade. From left to right the means of: highest, medium and lowest slopes clusters. Last column is the R^2 of a regression using the clusters as fixed effects. (*** 3σ, ** 2.5σ, * 2σ)

Decade	Out-competing	Transitioning	Displaced	R^2
1960-1970	0.079***	0.014***	0.007***	0.269
1970-1980	0.045***	0.001**	0.004*	0.187
1980-1990	0.030***	0.000*	0.001*	0.300
1990-2000	0.058***	0.028*	0.006***	0.230
2000-2010	0.010	0.006	-0.001	0.018

Table 43 – The mean export growths per country, aggregated to the total export of the country. The coefficients can be interpreted as discussed in the caption of Table 41.

the corresponding clusters have an export value growth in the following decade in the same product significantly higher than zero. From left to right the means of: out-competing, transitioning, and displaced clusters. Last column is the R^2 of a regression using the clusters as fixed effects. (*** 3σ, ** 2.5σ, * 2σ, we have assumed that the error is normally distributed).
Let us consider decade 1960-1970 in product 4 (fourth row). The row tells us that the countries in the out-competing cluster grew on average 4.3% per year and the ones in the transitioning cluster by .1% per year. Since the displaced cluster's growth average was less than two standard errors from zero, we cannot be sure that their observed growth rate is significantly different from zero. The transitioning cluster was at least 2 standard errors away from zero (i.e. there is a 1 in 22 chance that the result could be observed if the null hypothesis is true); while the out-competing estimate is more than 3 standard errors from zero (1 in 370 chance of observing such result from the null hypothesis).

Almost all cases considered show that countries in the out-competing clusters performed well, given that their average slope is significantly higher than zero (which would imply no growth). Both the displaced and the transitioning countries have a slope significantly lower than the countries in the out-competing cluster. In many cases they still experienced export growth, but that export growth was significantly lower than the one experienced by the out-competing countries.

There are two main deviations from this rule. The first involves product 3, which shows negative coefficients and/or lower R^2. This is unsurprising, given that SITC category 3 is dominated by the product with the highest trade traffic: crude oil. Since its dynamics are more related to geological discoveries than to the ability of countries to compete, it is expected to show counter-intuitive patterns. The second exception is for all estimates using the 2000-2010 data for calculating the clusters. Also in this case this failure can be attributed to

external causes. The trade data we have runs only until 2013. 2011-2013 is too short of a period to detect reliable trends, thus the test data is not good enough to evaluate our clustering.

One could ask if we couldn't have reached the same results simply doing an autocorrelation of the growth of the exports. To answer this question, we have made the same exercise (Table 42) with the difference that instead of using the competition network to cluster the countries as out-competing, transitioning and displaced, we have clustered them into three clusters based on their commercial growth in the previous decade as high, medium and low slope. Considering that R^2 is a statistical measure of how well the regression predictions approximate the real data points and comparing the corresponding values in both tables, we see that Table 41 shows higher values of R^2 in almost all decades and product categories (the only exception corresponds to the SITC 3 which, as we mentioned few lines above, is dominated by crude oil). According to this, the clustering using the competition networks predicts better the commercial performance of the countries in the following decade.

The method works at different levels of data granularity. To test this, we repeat the full analysis, collapsing the one-digit product categories to a single product, which stands for the entire export basket of a country. Once we perform the analysis, we still find that the out-competing countries grow their exports significantly more than the transitioning and displaced countries. Table 43 reports the coefficients, per decade. For instance, in the 1970-1980 decade, the countries which were classified as out-competing in 1960-1970 grew on average almost 8%. The transitioning countries grew 1.4%, while the displaced countries grew only 0.7%. Just as in the previous case, we fail to predict the last decade for lack of long enough data.

We pick some interesting cases to represent graphically: our best, most average and worst prediction among the ones reported in Table 41. Fig 4.9 depicts the slope distribution in each cluster as box plots.

In our best case, the out-competing cluster was able to correctly capture all the eleven fastest growing countries in the manufacturing sector in the 80s. The twelfth country, Thailand, had less than a third the average export growth rate in the sector ($\sim 4.92\%$) than the average of the top countries.

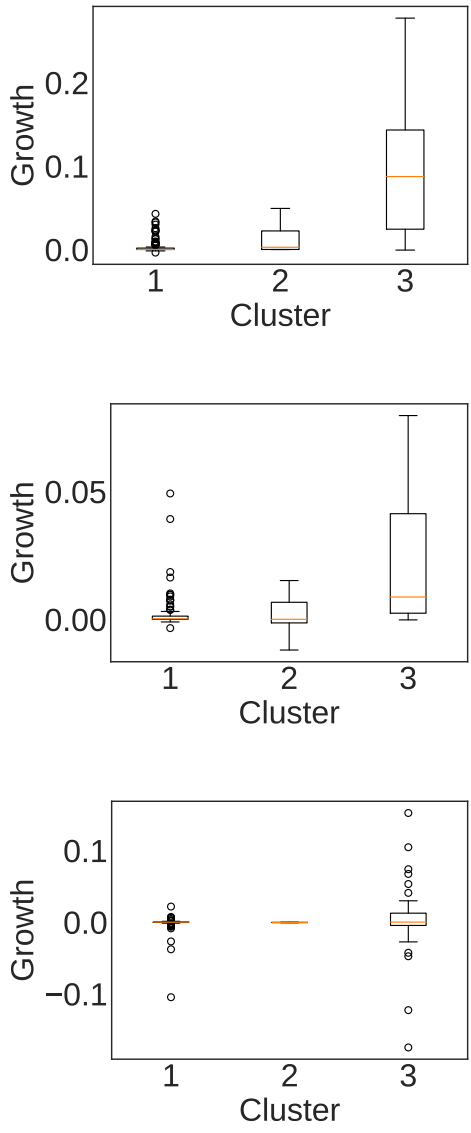


Figure 4.9 – The distribution of growth rates for countries classified in the different clusters. Box plots report medians (horizontal lines at the median of each box), whiskers (vertical lines extending to the most extreme, non-outlier data points), caps (horizontal lines at the ends of the whiskers) and fliers (points representing data that extend beyond the whiskers). From top to bottom: 1981-90 growth in SITC product 8 (Miscellaneous manufacturing); 1991-2000 growth in SITC product 1 (Beverages and tobacco); and 2011-13 growth in SITC product 3 (Mineral fuels, lubricants and related materials)

To give a better sense of this data we focus on one case from this example. Product 8 includes all manufacturing sectors, except machines (which is product 7) or manufactory chiefly focused on a single material (product 6). This category includes many products with very related machine-intensive production process, for instance a variety of garments. One of the rising economies in this sector in the 80s was China. China grew across the board in this sector, and displaced many countries in many markets. For instance, in 1986, China provided only .64% of watches imported in the United States, while France provided 1.1% (`http://atlas.media.mit.edu/en/visualize/tree_map/sitc/import/usa/show/8851/1986/`). By the end of the decade, in 1990 China rose almost tenfold in the market to provide 5.4% of US imported watches, while France halved to .58% (`http://atlas.media.mit.edu/en/visualize/tree_map/sitc/import/usa/show/8851/1990/`).

For the average case, we focus on the nine fastest growing countries, of which the out-performing cluster captured seven. The out-performing cluster captured all four countries that had an average yearly growth rate higher than 5%. Finally, the last plot shows a case in which the clustering did not manage to make sense of the export patterns. This is due to the fact that every country is an outlier in this product category, due to the importance of oil. The discovery of a large reservoir or the drying up of another one is unpredictable using the past trade patterns, and so we expect our methodology to fail in this case.

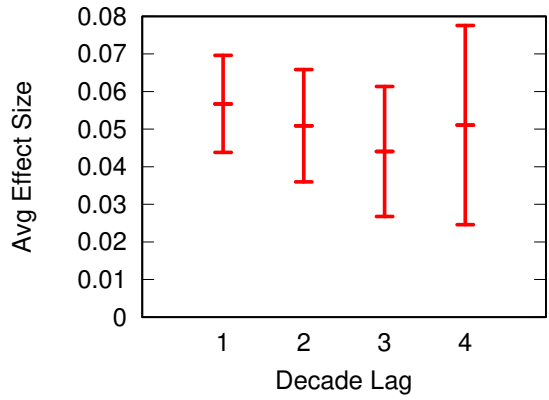


Figure 4.10 – The aggregate coefficient values across all product across all decades (y axis) for increasing decade lag (x axis). A decade lag equalling one means that we predict the decade after the data used to calculate the cluster (i.e. the main result of the paper). A decade lag equalling two means we predict two decades after the cluster data: if we had cluster data from 1960 we predict 1980 growth; if we had cluster data from 1980, we predict 2000 growth.

One could argue that we are capturing a random fluctuation in world trade trends. A displacement event might be a fluke of a country entering into a market niche and then exiting after some time. If this objection would be true, we should expect to observe reversion to the mean. In other words, if we use 1960 clusters to predict 1970 trade shares, then 1980 trade shares are expected to shrink by the same amount they grew in 1970. This is not the case.

Fig 4.10 shows the aggregate coefficient values across all products across all decades for increasing decade lag. In the figure, we exclude product 3 and clusters from 2000, for the reason explained above. The figure shows average and standard error of the regression coefficients, per decade lag. For instance, the first distribution (marked 1) is the average and standard error of the "Out-competing" column of Table 41. The second distribution reports the same for the regression coefficients predicting growth rates two decades away: for instance, we calculate the clusters using the 1960-1970 data and we predict the growth rate in the 1980-1990 period, i.e. two decades away. We see no sign of mean reversion. In fact, clusters from 1960 still predict – on average – a significant increase in market share in 2000, four decades later. The standard error range increase, as expected: the further away the prediction, the more uncertainty there is.

4.2.4 Validation

Here we validate the role detection methodology against a series of possible objections. The first issue we address is the arbitrariness of the role detection parameters.

In the paper we delineate a procedure to choose the δ, κ, and λ parameters. The role detection strategy introduces other parameters that influence the result, such as k_{max} and α. However, we do not provide an equivalent procedure to choose them. The reason to fix $k_{max} = 6$ and $alpha = 1$ comes from their meaning. k_{max} should be set equal to the network's diameter, because paths longer than the diameter do not provide any additional topological information. On the other hand, $\alpha = 1$ is the most reasonable choice because it gives each role an equal weight: choosing a different weight for different role would require a reason which we cannot provide.

What is the impact of these choices on the quality of our prediction? We pick product 1 in the 1960 decade to perform such exploration. Fig 4.11 shows

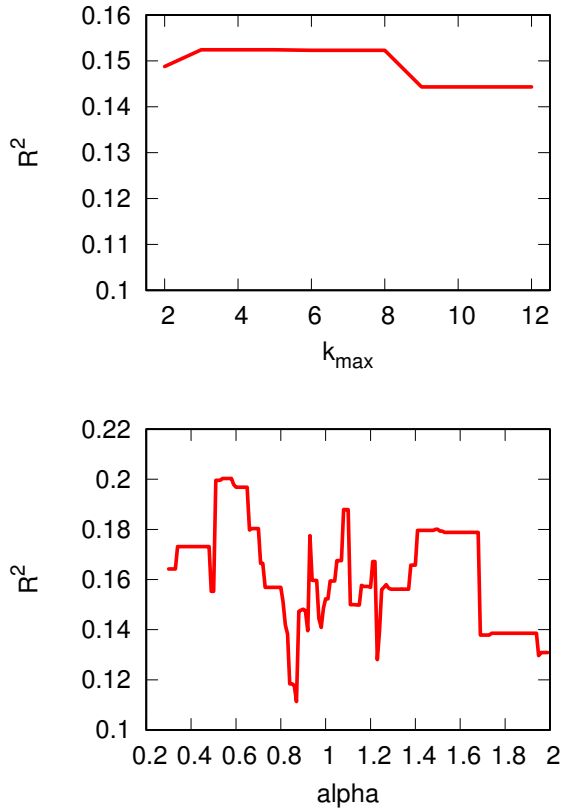


Figure 4.11 – Robustness tests for k_{max} (Top) and α (Bottom). For different choices of these parameters, we report the effect in the R^2 of the prediction. We focus on product 1 in the 1960-1970 decade.

their effect on the R^2 of our prediction. Note that, since this test involves directly our predictive task, it cannot be used to find the optimal parameter choices, because that would imply overfitting. If our best prediction comes with, say, $k_{max} = 4$ we cannot set k_{max} to that value, because there would be no way to know this before running the test.

Fig 4.11 (left) shows that k_{max} has a minimal impact on the prediction quality. Any value between 3 and 8 is acceptable. Performance deteriorates for high values, as more and more noisy information from long paths is included, while it also deteriorates for small values, when not enough information from the network is included.
Fig 4.11 (right) shows that the impact of α is more difficult to interpret. As a result, there is no specific guidance whether to choose $\alpha < 1$ or $\alpha > 1$.

We now move to addressing the issue that our methodology is a correlative analysis. Correlations arise randomly even for null phenomena, provided there are enough of them. If we generate hundreds of random countries with random export patterns, some of them will have anti-correlations strong enough to clear our δ threshold.

To address this concern we pick 100 random triplets of exporter-importer-product. For each exporter we generate an expected export value using a zero-inflated Poisson negative binomial model – meaning that the export value is directly proportional to the total amount it exported of that product, and inversely proportional to the importer-exporter geographical distance, controlling for the fact that trade data is sparse and with a heavy tail distribution, as suggested in (Burger et al. 2009). Then we apply our methodology to detect displacements. The expectation is that if our methodology is capturing some real phenomenon, then it should detect more displacements from the observed data than from the random data. This expectation is confirmed, since on average we observe two times more displacements than random expectation.

Still, this means that we expect half of inferred displacements to be noise. This is related to our second validation analysis. Noise connections link countries at random. In such networks, there are no non-local phenomena. Our role detection strategy operates under the assumption that the competition network is non-random, and that the kth role score is meaningful. If a random network with the same in- and out-degree distribution – but without any non-local phenomena – would return comparable kth role scores, then it means that the competition network could be dominated by the noisy connections.

To address this issue we generated 80 random networks which preserve the exact in- and out-degree distributions. Each random network is generated by picking pairs of edges at random and changing their endpoints, following (Hanhijärvi et al. 2009). We perform our analysis and we obtain the out-competing, transitioning and displaced clusters for our shuffled networks. We then calculate the adjusted mutual information between the shuffled network clusters and the observed ones. The average adjusted mutual information we obtained is equal to $.1 \pm .02$ (the theoretical maximum for identical clusters is 1, and 0 means completely independent clusters). We consider this as an argument supporting our clustering, given that shuffled networks with no non-local interactions return clusters which are not related with the ones we observe.

Moreover, the clusters obtained from the shuffled networks do not divide countries well when it comes to their export growth. We replicate the result for the 1960-1970 decade in product 1 (first row of Table 41). The clusters from the shuffled network returned very similar growth rates with each other, and significantly different from the non-shuffled network ones: 1.47% (shuffled) vs 4.8% (observed) for out-competing, 1.45% (shuffled) vs 0.5% (observed) for transitioning, and 1.24% (shuffled) vs 0.4% (observed) for displaced. The shuffled network preserved the in- and out-degrees but disrupted non-local dynamics, and this analysis proved that this disruption significantly affects the ability of sorting through the countries.

A third robustness check involves our clustering procedure. Since we compare the exporter role vectors to templates, our clustering is supervised, i.e. we impose what the clusters should look like. On the one hand, this enhances the interpretability of the extracted clusters, on the other hand it might introduce biases. We test for possible biases by designing an unsupervised version of the clustering.

In this version, we still fix the number of desired clusters to three (out-competing, transitioning, displaced), but we do not provide templates. Rather, we run a kMeans algorithm on the role matrix. We then correlate the results of the supervised and unsupervised clustering. We perform this test on a subset of our parameter space. We obtain a correlation of $.932 \pm .032$. Since we obtain a very high correlation, we conclude that using a supervised strategy did not introduce significant bias: the extracted clusters are virtually indistinguishable from the ones extracted with an unsupervised technique.

Finally, we test whether the role detection and the clustering procedure are necessary at all. When motivating the method we use, we showed that the outdegree and the indegree are highly correlated, thus they cannot be used for prediction. However, one could use their difference for making the prediction. In Fig 4.12 we show the predictive power such operation has. We predict the growth in export with the logarithm of the outdegree/indegree ratio. In all cases but two, such test returns worse results than the role detection method – shown in the figure below the identity line.

Moreover, by clustering the role scores we are compressing their information: we go from a vector of 12 numbers to a single variable that can have only

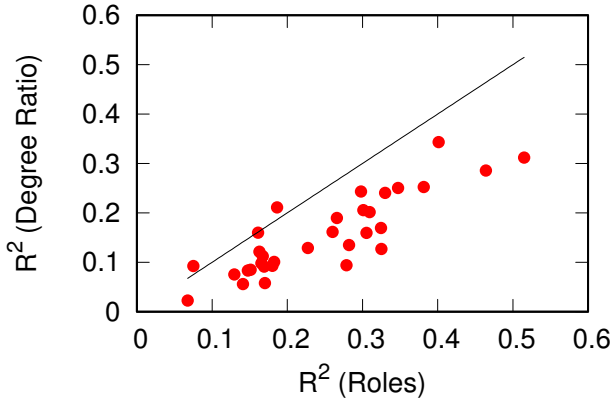


Figure 4.12 – The relationship between the R^2 export growth prediction using the role clusters (x-axis) and using the logarithm outdegree/indegree ratio (y-axis). Each observation is a decade-product combination: the x-axis values are the R^2 values reported in Table 41, excluding product 3 and the 2000-2010 decade. The black line is the identity line: observations below the line are the ones for which the role clusters performed better than the log degree ratio.

three values (out-competing, transitioning, displaced). We do so because we believe that the role vectors might have fluctuations that might introduce noise, and that noise will cancel out if we cluster the vectors. To verify if this is the case, we test the same linear regression model we used in the previous section, using the 12 role scores instead of the cluster labels. Every single model has lower R^2 than the corresponding model using the cluster labels (average $-.073 \pm .044$). We can conclude that the clusters are indeed improving the quality of the prediction.

4.3 Discussion

In this work, we adopted an ecosystem approach to the analysis of the global trade patterns. We see exporters as organisms competing for resources in different market niches. A market niche is a country importing a product. The assumption is that exporters want to out-compete other exporters, attempting to occupy the entire market niche. The appearance of a new exporter in a niche can be followed by the disappearance of another country. This is what we call a displacement event. We create a formal definition of displacements and we systematically collect all of them along a period spanning fifty years. A displacement event can be represented as a directed edge going from the

out-competing exporter to the displaced one. We call the collection of all displacements a "competition network", which is a weighted directed multilayer network, where each layer is a product class.

While the in- and out-degree of a node in a competition network have an intuitive interpretation – being the number of displacements experienced and caused by an exporter, respectively –, we show that in practice these measures cannot be used for classifying countries. The reason is their very high correlation. To fix this issue, we calculate network roles based on in- and out-degree flows. By clustering nodes according to their role score, we are able to classify them in three categories: out-competing, transitioning, and displaced. We show that these classes can be used to predict the future performance of an exporter in a particular market, in term of growth of total export value.

Apart from its useful application in predicting the economic performance of countries in the following years, we consider that our model, based on the creation of a competition network, could be used as a ranking method in any type of competition. For example, it could be used as an alternative to the Elo rating system which is a method for calculating the relative skill levels of players in zero-sum games such as chess (Hvattum & Arntzen 2010).

Our methodology has several issues. First, it does not consider actual displacements: the edge creation process is correlative by design, so we are not really capturing if the appearance of a new exporter really caused the disappearance of another. Second, it cannot be applied to all product classes: our predictions fail when considering natural resources composing the vast majority of some countries' exports, such as crude oil. Finally, we have not built a formal theory of why the competition network roles are predictive: we do not control for confounding factors that might drive both growth in exports and the position of a country in the network.

Chapter 5

Forecasting International Conflicts via Machine Learning and Multilayer Social Balance

Forecasting large geopolitical changes is a key task to improve the resilience of human society. One of the oldest and most important forecast tasks (not without criticism (Cederman & Weidmann 2017)) focuses on international conflicts (Richardson 1960, Holsti & Holsti 1991, Pettersson & Wallensteen 2015): *Can we predict when and where a new war will start? Can we estimate how large the conflict will be?* These questions are particularly relevant today, as developments in network analysis allows us to use a complex system perspective of international relations (Clauset 2019, Pomeroy et al. 2018, Risi et al. 2019). This perspective can integrate traditional geopolitical studies (Goldstone et al. 2010, Brandt et al. 2011, Mach et al. 2019) and it is the perspective we adopt in this work. Here we are interested in narrowing down to the forecasting of the outbreak of a conflict, without focusing on estimating its size or global impact. We frame conflict forecasting as a link prediction problem (Liben-Nowell & Kleinberg 2007, Lü & Zhou 2011). Specifically, we model international relations as a multilayer network (Kivelä et al. 2014*b*, Berlingerio et al. 2013). The nodes of the network are countries. Each layer is a different type of relation that two countries can establish between each other: trade, migration, sanctions, and so on. Conflicts (wars) are one of these layers and our prediction target, making this a multilayer link prediction problem (Rossetti et al. 2011, De Bacco et al. 2017, Hristova et al. 2016).

Differently from the classical multilayer link prediction, in this chapter we also consider the sign of a relationship. Some layers in the network represent positive relationships (trade, flight passenger commutes) while others are negative (sanctions, visa requirements). Thus, we use elements from social balance theory (Heider 2013, Antal et al. 2005) in signed networks to augment our link prediction strategy. Social balance theory has been applied to simple signed networks (Leskovec et al. 2010*b*,*a*, Bachi et al. 2012, Antal et al. 2006, Moore 1978), but not to multilayer ones as we do here. Signed multilayer networks have been analysed in the past (Szell et al. 2010), but without focusing on the link prediction problem.
The application of (signed) network analysis to international relations is not new (Easley & Kleinberg 2010, Vina-Cervantes et al. 2018) although it rarely focused on the prediction of outbreaks of new conflicts.

To sum up, we build a machine learning framework to solve the signed multilayer link prediction problem and we apply it to the prediction of new conflict links in a network of inter-country global relations. We do so by building a multilayer social balance measure and use it as an input feature for the decision tree learning algorithm (Safavian & Landgrebe 1991).
We find that international conflicts follow the social balance theory, as our feature positively contributes to the prediction task – specifically boosting the recall performance. This means that we are able to predict that, in the following year, two countries will be in conflict due to their relations with other common neighbours in the other layers of the network in the current year. Moreover, if we look at false positives, we can estimate that, on average, in 55% of cases a false positive will turn into a true positive in ~ 2.72 years. This means that in 55% of the cases in which we predict a conflict between two countries at year y and this conflict does not happen (false positive), the conflict happens (false positive becomes true positive) on average 2.72 years after the predicted year, that is at year $y+2.72$.

5.1 Model

We model international relations as an evolving multilayer network. For the rest of the chapter, we use the following convention. The multilayer network is $G = \{G^1, G^2, ..., G^t\}$, with G^t being the network at time t. When not relevant, we omit the temporal superscript t. Each snapshot of the network is $G = (V, E, L)$ with V being the set of nodes (V_l specifies the nodes with at least one connection in layer l), E the set of edges (E_l specifies the connections in

layer l), and L the set of layers. $c_1, c_2 \in V$ are nodes (countries) in the network, $l_1, l_2 \in L$ are layers, and N_c^l is the set of neighbours of node c in layer l.

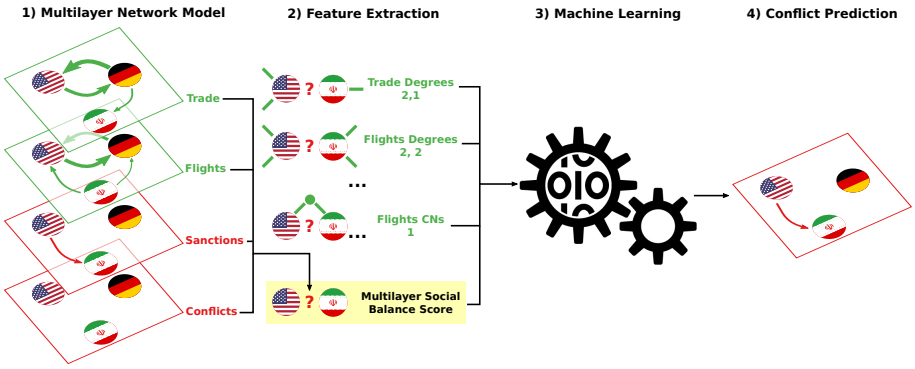


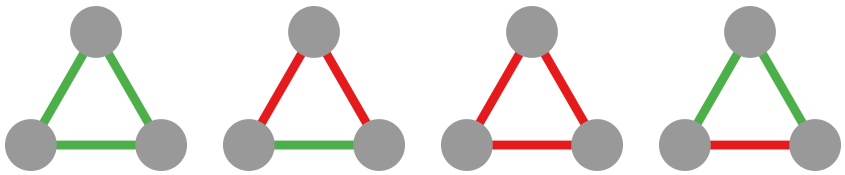
Figure 5.1 – Our workflow. From left to right: multilayer signed network model, feature extraction (highlighting the main contribution of this work: the multilayer social balance score), machine learning, output prediction.

The nodes of the network are individual countries. Each layer contains a different type of international relationship two countries can establish between each other. Some of these relations are positive (trade, migration, air passenger flows), some are negative (visa requirements, sanctions, conflicts) and one of them is neutral (Geographical border). Our main question is:

Given the status of the network at time t, would countries c_1 and c_2 connect on the conflict layer at time $t+1$?

We answer the question by building a machine learning framework. Figure 5.1 shows the workflow. First, we build the multilayer network, collecting data from public datasets – described in the Material and Methods section. We have datasets spanning different time periods: here we focus on the 1962-2012 period, as it is the one for which we have the largest overlap in our data sources. We drop data from years outside this period. Then, we extract a series of features from the multilayer network. We feed a decision tree with these features, which provides a binary label for each pair of countries c_1 and c_2. The label equals to 1 if the algorithm predicts an edge in the conflict layer between the countries, 0 otherwise.

The majority of features are simple features such as the degree and the number of common neighbors of c_1 and c_2 in each layer -described in Chapter 2. However, we focus in particular on a non-trivial feature we extract from the combination of the entire multilayer structure around c_1 and c_2: their multilayer social balance.



(a) PPP (b) PNN (c) NNN (d) PPN

Figure 5.2 – Signed triangles, with positive (green) and negative (red) relationships. Balanced triangles: (a) and (b) – odd number of positive relationships. (c) and (d) are classically considered unbalanced, although, under certain circumstances, (c) can be considered a balanced or neutral configuration.

Classical social balance theory studies triangles in signed networks. Figure 5.2 shows four possible triangles in an undirected signed network. Some of these triangles are "balanced" meaning that the relationships are in equilibrium. In real world signed systems, we expect to find an overexpression of balanced triangles and an under expression of unbalanced ones. This expectation is usually met and thus can be used for link prediction: in a signed network, missing links are more likely to carry the balanced sign.

However, in a multilayer network, depending on which layers we focus, we can build a number of different triangles, some of which can be balanced and some of which that are not. Figure 5.3 shows an example.

We solve the problem by considering all possible pairs of layers l_1 and l_2 and counting the number of potential triangles that a missing edge between c_1 and c_2 could close. If l_1 and l_2 carry the same sign (both positive or both negative) we expect to find a positive edge between c_1 and c_2, otherwise a negative one. Formally:

$$MLBal_{c_1,c_2} = \sum_{l_1,l_2 \in L} \sum_{z \in (N_{c_1}^{l_1} \cap N_{c_2}^{l_2}) \cup (N_{c_1}^{l_2} \cap N_{c_2}^{l_1})} l_1 l_2,$$

where the product of two layers ($l_1 l_2$) is the arithmetic product of their signs: 1 if they have the same sign, -1 if they are of unlike sign.

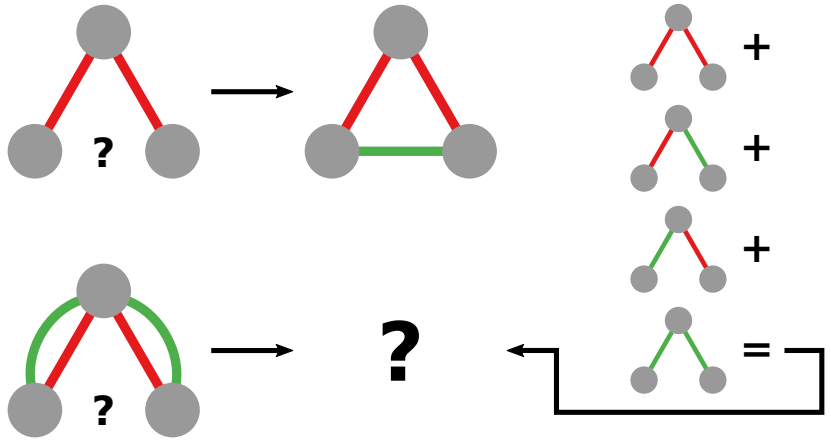


Figure 5.3 – The multilayer social balance problem. In a monolayer signed network, the sign of the next edge is simply the one balancing the triangle. In a multilayer signed network, we need to look at all possible triangle combinations across layers.

5.2 Results

5.2.1 Social Balance

Our multilayer signed network has eight layers. Four of them are positive: alliances, trade, migration, and flight traffic. Three are negative: visa requirements, sanctions, and conflicts. One is neutral: whether the countries share a border or not. See Materials & Methods section for information about data sources and cleaning. Figure 5.4 shows the number of nodes and edges for the allies, trade, sanctions, and conflicts networks – for which we have yearly snapshots. This figure simply shows a count of nodes per year regardless of whether they are new nodes or old nodes re-entering the network. Table 51 shows the same counts for the flights, visas and migration layers – for which we have a single or few snapshots, and also for the geographical border, which is normally the same over the years.

Note from Figure 5.4 that the conflict layer is extremely sparse, with $|E| \sim |V|$ or even $|E| < |V|$ for many years, meaning an average degree of two or lower. The allies layer is even sparser, and shows interesting peaks corresponding to the first (1991) and second (2003) Gulf wars, and the Balkan war (1998), which were NATO-wide conflicts. This shows how NATO was quiet on the frontal country-level action during the actual Cold War, and started being openly active only after the fall of the Soviet Union.

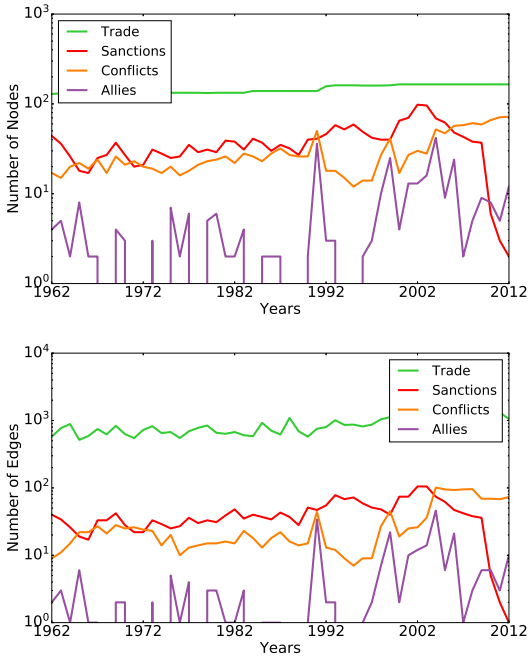


Figure 5.4 – Number of nodes and edges per year for the trade, allies, sanctions, and conflicts layers.

Figure 5.5 shows the complementary of the cumulative distribution of the average degree per time step in the network (that is: total number of edges across all ts, divided by the number of ts). The figure shows a degree of heterogeneity in network density – with conflict and allies being the sparser layers and visa being the one with most edges.

Layer	# Nodes	# Edges
Flights	214	1,565
Visas	245	30,849
Boder	152	568
Migration 1990	173	1,566
Migration 2000	173	1,171
Migration 2010	171	1,281

Table 51 – Number of nodes and edges per year for the Flights, Visas, Border and Migration networks.

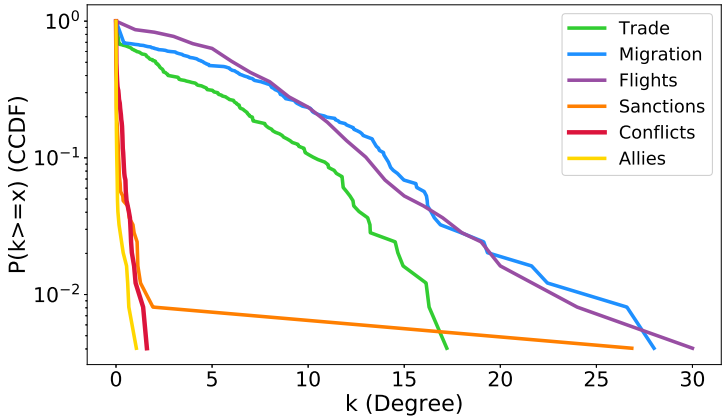


Figure 5.5 – Complementary cumulative distribution (CCDF) of the average yearly degree across all layers of the network for the period from 1962 to 2012. CCDF is the probability (y-axis) that a node has an average degree equal to, or higher than, a certain value (x-axis).

Since we are planning to apply social balance theory to predict conflicts, we should check whether social balance theory applies in the multilayer network of international relations. We thus count the number of triangles of each type we find in the network across all snapshots – note that we ignore the border sharing layer as it is not signed. There are 38.2M balanced triangles in the multilayer network (PPP + PNN) against 147.5M unbalanced triangles (PPN + NNN). We need to compare the counts with the *expected* number of triangles of each type, given the number of nodes and edges in each layer.

The expected number of triangles with two edges in layer l_1 and one edge in layer l_2 is given by:

$$T_{l_1,l_2} = \binom{|V_{l_1,l_2}|}{3} p^2_{l_1,V_{l_1,l_2}} p_{l_2,V_{l_1,l_2}},$$

where: $V_{l_1,l_2} = V_{l_1} \cap V_{l_2}$ is the set of nodes that l_1 and l_2 have in common (there can be no multilayer triangle if the three nodes involved are not present in both layers), and $p_{l_1,V_{l_1,l_2}}$ is the probability that two nodes in V_{l_1,l_2} connect in layer l_1 (similarly for $p_{l_2,V_{l_1,l_2}}$). Note that it is possible that $l_1 = l_2$ and, in that case, the formula reduces to the expected number of triangles in a single layer network: $\binom{|V|}{3} p^3$.

We find that the multilayer network of international relations follows social balance. The ratio between observed versus expected balanced triangles is 1.52, while the same ratio for unbalanced triangles is only 0.31: there is one and a half balanced triangles per expected one, but only one unbalanced triangle per three expected.

Note that, among the balanced types, PPP is vastly more overexpressed (2.68) than PNN (1.49), meaning that the number of observed PPP triangles is 2.86 times higher than what we would expect if the triangles would close with a randomly selected sign, versus 1.49 for PNN. This means that, in our particular network, in which we have not only more positive layers but also two (Conflicts and Sanctions) of the three negative layers are very sparse in comparison with the rest, social balance is much more suited for the prediction of positive, rather than negative, links. Since we are interested in predicting a negative link type (conflicts), we expect social balance to underperform.

5.2.2 Forecasting Conflicts

For the period between 1962 and 2012 we have around 1500 conflicts between countries. For Trade, for example, around 45000 links and for migration, around 73000 links. So, as we can see, conflicts data is very sparse. For this reason, for the machine learning purpose, we have decided to split our data chronologically and not using k-fold validation.
K-fold validation involves partitioning a sample of data into complementary subsets, performing the analysis on one subset, and validating the analysis on the other subset and in most methods multiple rounds of k-fold validation are performed using different partitions, and the validation results are combined (e.g. averaged) over the rounds to give an estimate of the model's predictive performance.

If we choose to go for k-fold validation the risk that we take is randomly select training datasets that include a big portion of the conflict links, so our model will overestimate the frequency of conflicts and therefore we will have a large number of false positive. Basically the decision of splitting the data chronologically was made to respect and deal with the sparseness of the conflicts.

The decision tree we use to forecast conflicts was trained using the multilayer network for the period from 1962 to 2005 to have the other 7 years (2005 to 2012) to test it. This achieves a precision of 0.62 and a recall of 0.86 (how these values are calculated is explained in Section 3.4).

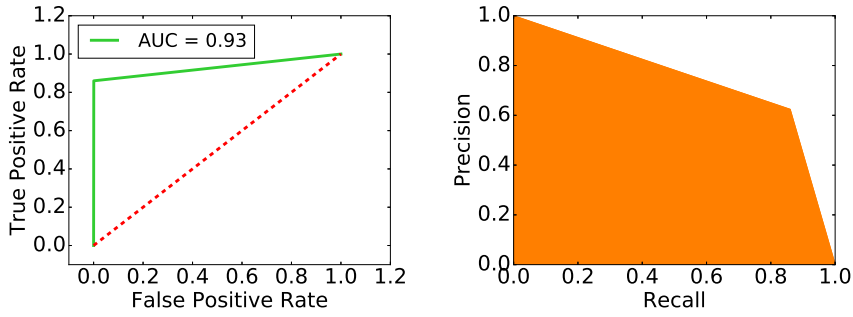


Figure 5.6 – ROC (a) and Precision-Recall (b) curves for Decision Tree predictive model.

Figure 5.6(a) is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve and AUC represents degree or measure of separability. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting correctly the different classes. This curve plots two parameters: True Positive Rate (TPR) which is a synonym for Recall and False Positive Rate (FPR) which is the ratio between the number of negative events wrongly categorized as positive (false positives) and the total number of actual negative events (regardless of classification). Figure 5.6(b) shows the trade off between precision and recall for different threshold. A high area under the curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate. High scores for both show that the classifier is returning accurate results (high precision), as well as returning a majority of all positive results (high recall).

To predict a conflict edge at time $t+1$, we drop the feature recording whether the two countries were in conflict at time t, due to very strong autocorrelations: multiyear conflicts should not be used to predict themselves. We also drop the edge presence in the allied layer, as it's trivial that allies will declare wars to aggressors of their allies. We have also created a "short" version of the data, where conflicts are compressed for their duration. In other words, for a conflict spanning 1992-1996, in the short table we only have a conflict edge at $t = 1992$, rather than as in the long table at $t = 1992, 1993, 1994, 1995, 1996$. This seems to hurt recall, while allowing us to be more precise (precision 0.88, recall 0.38). This is unsurprising, as we already noted that the conflict layer is extremely sparse, and it is doubly so in the "short" version of the data.

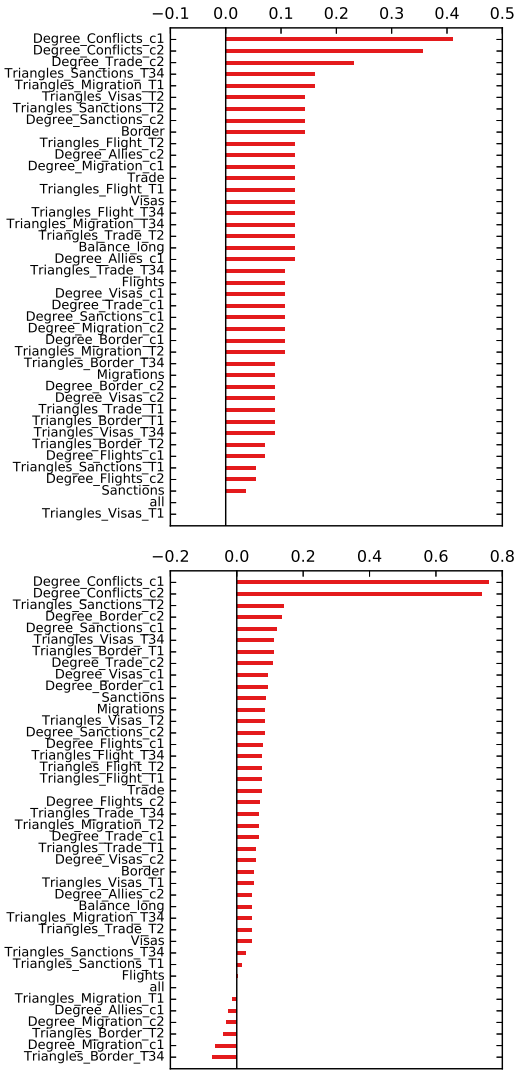


Figure 5.7 – Recall (top) and precision (bottom) scores for Decision Tree predictive model. Label legend: Degree_*l*_c*x* = degree of country x in layer *l*; Triangles_*l*_T*t* = number of triangles of type *t* between the two countries in layer *l* (see Material and Methods for a depiction of the different triangle feature types); Balance long = our multilayer balance score; all = combination of all features (with contribution zero by default); everything else = presence of an edge in layer *l*.

Figure 5.7 shows the feature importance to the precision and recall evaluations for the short version of the data. ***Feature importance*** refers to techniques that assign a score to input features based on how useful they are at predicting a target variable. This concept is explained in details in section 3.7.

In Figure 5.7 we can see that our balance score is one of the features that play an important role in the prediction of conflicts. However, in this case in which we consider all the attributes (including degrees, triangles and so on) in our model, balance score is far from being one of the most important features.

In Figure 5.7 we can see that the two most important features are the Conflict degrees (since our conflict data is indirect, the suffixes $c1$ and $c2$ have the same meaning: Countries involves in conflicts), and for Recall also the Trade degree for the importer country $c2$. According to the figure, Conflict degrees add $0.7-0.8$ to precision and around 0.4 to recall and Trade degree add around 0.25 to recall. Testing our model using the same training set, we have found that the precision and recall have values of 1, which means that when we consider all the attributes, we are overfitting our model. So we have decided to run our model discarding the features that are the best candidates as being the cause of overfitting.
Once you decide to drop a class of features, you don't get to pick and choose which ones you drop inside that class, you have to drop all of them. So, in our case, we have dropped the Degree class.
After discard the degree values in our model and we have found that the Balance score is the second most important feature for Precision and the fourth for Recall (see Fig 5.8).
These results demonstrate that not only International conflicts follow social balance but that Balance score is a powerful feature for conflicts prediction.

We perform a number of robustness checks.

- As is implied by the names "Tree" and "Forest," a Random Forest is essentially a collection of Decision Trees (In Section 3.2 we have explained in detail these to types of supervised learning).
  A decision tree is built on an entire dataset, using all the features/variables of interest, whereas a random forest randomly selects observations/rows and specific features/variables to build multiple decision trees from and then averages the results. After a large number of trees are built using this method, each tree "votes" or chooses the class, and the class receiving the most votes by a simple majority is the "winner" or predicted class. This is the reason why in most of the cases Random Forest is more accurate. So, one could ask why aren't we using random forest directly instead of decision tree. In this case in which the main goal is to proof the importance of Balance score in the prediction of conflicts, and in general, in all the cases in which explainability between variable

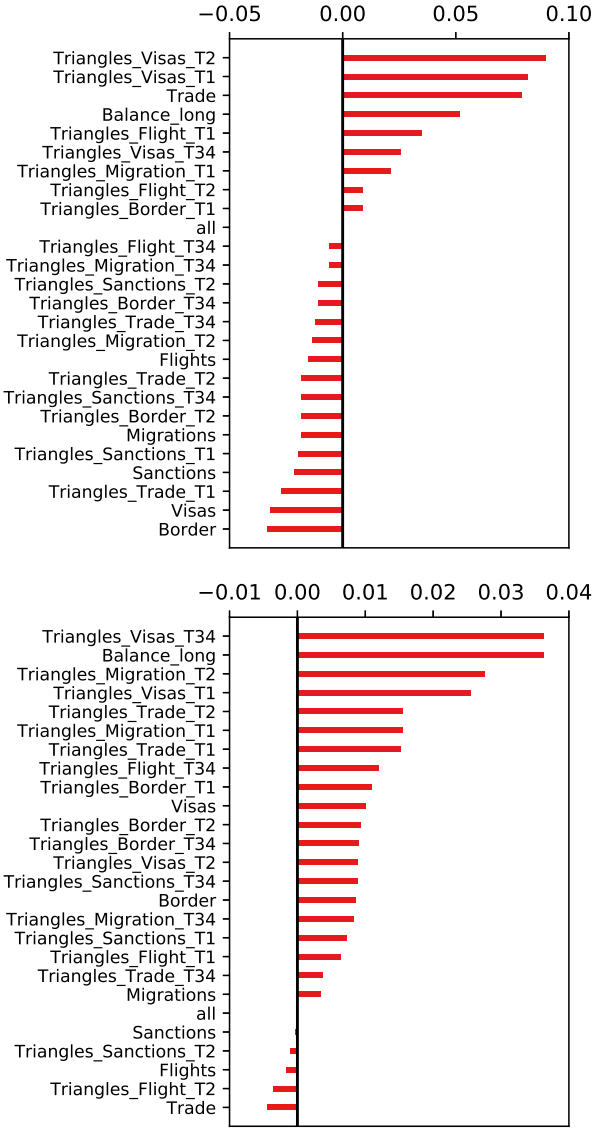


Figure 5.8 – Recall (top) and precision (bottom) scores for Decision Tree predictive model. In this case we have discarded the Degree Class. Label legend: Degree_*l*_c*x* = degree of country x in layer *l*; Triangles_*l*_T*t* = number of triangles of type *t* between the two countries in layer *l* (see Material and Methods for a depiction of the different triangle feature types); Balance long = our multilayer balance score; all = combination of all features (with contribution zero by default); everything else = presence of an edge in layer *l*.

is prioritised over accuracy, the results are easier to analyse when using decision tree as machine learning (Liaw et al. 2002).

Since random forest chooses features randomly during the training process, it does not depend highly on any specific set of features.

This randomized feature selection makes random forest much more accurate than a decision tree, so it is suitable for situations when we have a large dataset, and interpretability is not a major concern, but as we mentioned before, the main goal of this work isn't the accuracy but to show how international conflicts follow social balance and how our Balance score is a key feature in the prediction of conflicts.

Nonetheless, we have tested a Random Forest classifier, Figure 5.9 shows the feature contributions.
Since we have rather strong correlations between variables and randomforests are not 'overfitting' the differences, it can be the case, that the performance is the same. Moreover as a random forest has multiple trees, some of them can even ignore the feature of interest and take a correlated one. This would explain the fact that most of the features have the same Importance score.
When a dataset has two (or more) correlated features, from the point of view of the model, any of these correlated features can be used as the predictor, with no concrete preference of one over the others. But once one of them is used, the importance of others is significantly reduced since effectively the impurity they can remove is already removed by the first feature. As a consequence, they will have a lower reported importance (this would explain the difference in the importance for the same feature using the different classifiers, see Figure 5.8 and Figure 5.9). When interpreting the data, it can lead to the incorrect conclusion showing equal or similar importance to correlated features (like we see in Figure 5.9) and/or showing that one of the variables is a strong predictor while the others in the same group are unimportant, while actually they are very close in terms of their relationship with the response variable.

On the other hand, random forest is a more cautious predictor, minimizing false positives and even more false negatives (precision 0.82, recall 0.86), and the contribution of multilayer social balance holds.

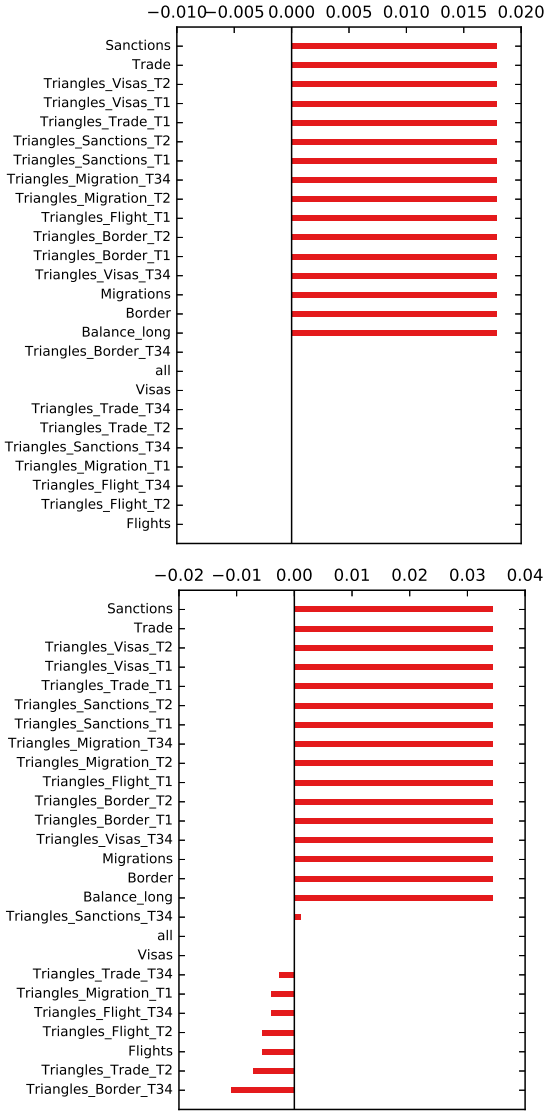


Figure 5.9 – Recall (top) and precision (bottom) scores for Random Forest predictive model., Label legend: Degree_*l*_c*x* = degree of country x in layer *l*; Triangles_*l*_T*t* = number of triangles of type *t* between the two countries in layer *l* (see Material and Methods for a depiction of the different triangle feature types); Balance long = our multilayer balance score; all = combination of all features (with contribution zero by default); everything else = presence of an edge in layer *l*.

- We relax the prediction task, by not limiting our prediction to the $t+1$ snapshot. In practice, every time we predict a conflict, we look arbitrarily far in the future for the beginning of the conflict. We find that, if a false positive will turn into a true positive – which happens in 55% of cases –, it does so on average in 2.72 years. Figure 5.10 shows a timeline of such predictions: the first dot is the year of our prediction and the second dot is the year of the actual conflict.

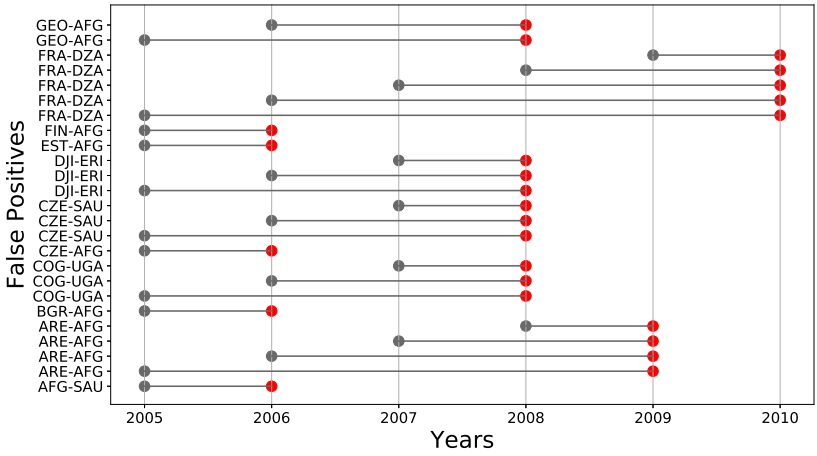


Figure 5.10 – Our model has 179 false positive. This mean that our model has predicted 179 times a conflict that didn't happen. In this Figure we present only 25 of the 99 (55%) cases in which a false positive became true positive after an average delay of 2.72 years. Grey dots represent the year in which the conflict was predicted to happen and the red dots the year in which the conflict happened.

5.3 Discussion

In this chapter we aim to forecast the outbreak of international conflicts. We do so assuming that international relations follow social balance theory: countries will tend to form coherent groups of allies. We extend social balance theory to the multilayer signed network model, creating a multilayer balance measure. When we use multilayer social balance as a feature of a machine learning algorithm, we see that it positively contributes to the prediction of links in the conflict layer. This is robust to the choice of machine learning algorithm, to different views of the data, and we also show that false positives tend to transform in true positives given a lag of less than three years on average.

Our model has a number of limitations. First, we do not really use any theory of international relationship in our work. Even if forecasting and prediction are related to the same concept, that is future oriented, there is a fine line that differentiates them. Prediction is concerned with estimating the outcomes for unseen data. For this purpose, you fit a model to a training data set, which results in an estimator $f(x)$ that can make predictions for new samples x. Forecasting is a sub-discipline of prediction in which we are making predictions about the future, on the basis of time-series data. According to this, although the choice of using social balance has some theoretical grounding, we eschew from any causal implication: that is, we are not predicting conflicts, only forecasting them.
Second, the data collection is limited: countries can establish more relationships between them than the ones we see here (co-signing international treaties, having diplomatic missions, etc). As a result, our social balance measure is necessarily incomplete. Third, given the structure of the multilayer network, social balance is much more suitable for the prediction of positive, rather than negative, links. (See section 5.2.1) Since we are predicting a negative relationship – conflicts – we are not using the best feature possible, although we did not know this beforehand.

Notwithstanding these limitations, even it wasn't the main goal of this work, we have achieved good precision and recall values. This is even more impressive given the extreme sparsity of international conflict data. There was already some prior work showing that social balance holds at the international level (Antal et al. 2006, Moore 1978), but this is the first proof that it does so even considering multiple different international relationships, not just alliances and military pacts.

We can extend the present work in multiple ways. First we can address some of the limitations. Specifically, we could include more possible relationships between countries: economic ties, business travel, diplomatic relationships, hacking attacks, trade tariffs, and so on. Second, we could use the multilayer social balance theory to try and predict positive, rather than negative, outcomes: a task for which the measure might be more suited. Finally, we could apply multilayer social balance theory also to micro systems as well, rather than looking at the level of international relationships between state actors.

5.4 Materials and Methods

5.4.1 Data

Our multilayer signed network includes eight layers, trade, migration, flights, border, visas, sanctions, conflicts and allies. Since our main goal is to show that international conflicts follow social balance, we had to assign signs (positive or negative) to each of the layers in order to create different kind of triangles between them. According to this we assume trade, migration, flights and allies as positive relations. We consider that if two countries exchange products, passengers in flights for business, tourism or any other purpose, if inhabitants of one country migrate to the other, and specially, if two countries are allies in a war, it is a sign of good relation between them, or at least it is an indication that the countries don't have any kind of restriction between them. The border layer remained as neutral because we cannot say whether sharing a border is a positive or a negative relationship. Sanctions, Visas and Conflicts are the negative layers of our system.
Here we report our data sources for each of them. Note that some layers are directed and some are undirected. For consistency, all directed layers are rendered undirected by making the edges symmetric.

5.4.1.1 Positive Layers

Trade. The trade data comes originally from the CEPII dataset (Gaulier & Zignago 2010), further cleaned and publicly available from `http://atlas.cid.harvard.edu/engage#data-download`). The original dataset splits yearly country-country trade flows into different product (SITC). Links are weighted accordingly to the trade intensity value in millions of dollars. The data spans from 1962 to 2013. For this data we had to aggregate all the categories of products because we were only interested in the total amount of the trades between the each pair of countries per year. Since this layer is dense, we perform further data cleaning (via backboning) as we detail below.

Migration. For the migration data we rely on official statistics from the UN about the stock of migrants originating from one country and living in the other (DESA 2013). We have four snapshots: 1990, 2000, 2010 and 2013. We associate each year with the closest year in the data. For instance, the migration layer at time $t = 1988$ uses data from the 1990 snapshot. We find this appropriate given the strong autocorrelation in the data. As Figure 5.11(left) shows, the correlation of the logarithm of stocks between snapshots is always higher than 0.96, even at 23 years of distance.

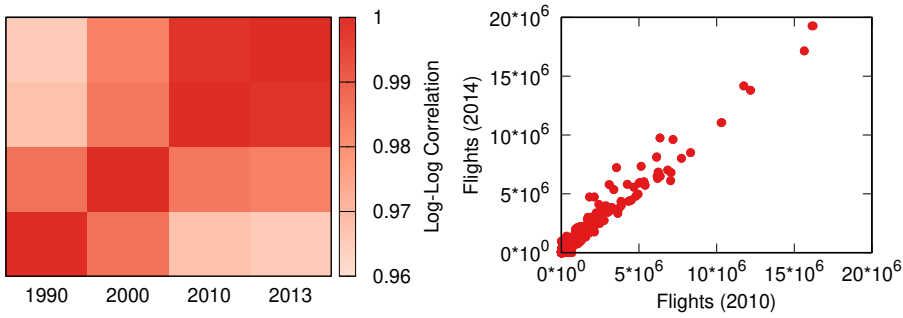


Figure 5.11 – The strong autocorrelations in our data. (Left) The migration layer across years. (Right) The flight layer: 2010 versus 2014 installed seating capacity (one observation per country pair).

Flights. To create the flight layer we obtained data from OAG `https://www.oag.com/`. OAG collects data of all flights between all airport and provided us with the total installed seating capacity for planes flying from country c_1 to country c_2. We only have information for the years 2010 and 2014. Just like in the migration layer, we interpolate the data, on the basis of its strong autocorrelation (Figure 5.11(right)). We also performed network backboning on this layer, as we detail below.

Allies. The allies information comes from the same source as the conflict layer. For each conflict, the dataset includes the full coalition on both sides. If two countries are on the same side, we establish an link in the allies layer. We discuss the data source more in depth below.

5.4.1.2 Negative Layers

Visas. We collected information about the current visa regime between pairs of countries from Wikipedia (`https://en.wikipedia.org/wiki/Category:Visa_requirements_by_nationality`). The source reports different types of visa regimes ("visa required", "Electronic Entry Visa", "Visa on arrival", etc). Since this is a negative network, we establish a link between countries if the country requires a visa or outright refuses to admit citizens of the other country. This is a necessarily static dataset that only includes information about the current status, which is a large limitation of our study. However, it is the only way we can include such layer, as historic data about visa regimes is notoriously sparse and noisy (De Haas & Czaika 2013).

Sanctions. We use the Threat and Imposition of Economic Sanctions (TIES) dataset, version 4 (Morgan et al. 2014). For each sanction we have the beginning and end date (if any) and the countries involved. The data spans

from 1945 to 2005. We have an edge in the layer at time t if, in year t, the two countries had a sanction relationship (either a new one or the continuation of an old one).

Conflicts. The conflict data comes from the Uppsala Conflict Data Program (Gleditsch et al. 2002, Pettersson & Eck 2018) (`https://ucdp.uu.se/`). The data includes both country-level wars as well as civil wars and internal threats. We drop intra-country conflicts from the data, focusing exclusively on global level events – when there was at least a state-level actor on both sides of a war. The data spans from 1939 to 2016. The data includes information about the coalitions on each side. We create a biclique connecting all countries in a coalition with all countries in the opposite coalition – while we create coalition cliques per conflict in the allied layer.

5.4.1.3 Neutral Layer

Geographical distance. Conflicts are more likely to happen between neighbouring countries for a number of logistic reasons. Therefore, we augment our machine learning features by also considering information on whether the two countries share a border. This is a layer in our network, although it does not carry a sign, as we cannot say whether sharing a border is a positive or a negative relationship. The data comes from the CEPII dataset (Mayer & Zignago 2011).

5.4.2 Preprocessing

5.4.2.1 Data Cleaning

As mentioned in the main text, for all layers with yearly snapshots, we only consider the 1962-2012 interval.

The trade, migration, and flight networks are weighted and noisy: the likelihood of a weak and insignificant relationship between any two countries is high. We thus need to estimate statistical significance and drop the edges which do not represent robust signals between countries. We do so via network backboning, specifically using the noise corrected backboning technique (Coscia & Neffke 2017). The backboning algorithm generates an expectation of the edge weight given the propensities of nodes c_1 and c_2 of connecting. It then allows to filter edges with a parameter δ which represent the minimum t-score level we allow in the network. All edges with a significance lower than δ are removed. We fix the following δ values per layer: flights 100, migration 2.95, trade 6.16.

We chose the noise corrected backboning because it is the backboning technique which maintains the largest number of nodes in the network while at the same time reducing the most the number of spurious edges. In the flight layer, the backboned version has 214 nodes (against 216 original nodes in the non-backboned version), and 1,574 edges (against 4,391). Figure 5.12 shows the yearly evolution of the number of nodes and edges before and after backboning for the trade and migration layers.

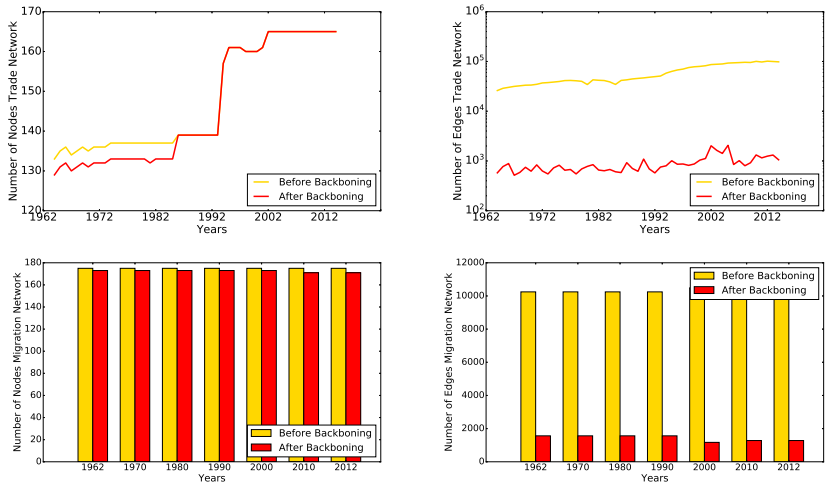


Figure 5.12 – Number of nodes and edges before and after the backboning process for the trade and migration layers.

5.4.2.2 Feature Extraction

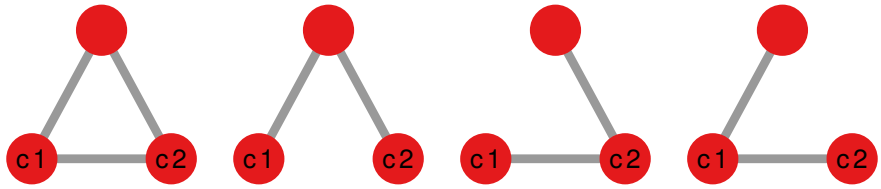
There are three classes of features we feed to the machine learning classifier. Each layer provides one or more features in all three classes, unless otherwise specified. The fourth feature class is a multilayer feature – the multilayer social balance – which we already described.

Feature Class #1. The first feature class we extract is the presence of an edge in a layer. Thus this feature uses the correlation (or the anti-correlation) of a given layer with the conflict network. For instance, if there was a trade link between c_1 and c_2 at time t, we will have a 1 in this feature when trying to predict a conflict at time $t+1$. We omit this feature for the conflict layer, as ongoing conflicts spanning multiple years would make this feature overwhelm everything else – and it is not interesting to predict that an ongoing conflict will

continue. We also omit this feature in the allies layer, as it is equally obvious that two allies will engage in the conflict together.

Feature Class #2. The second feature class we extract is the degree of the nodes in a layer. This generates two features: the degree of c_1 and the degree of c_2. So, if c_1 had seven trade links at time t, the feature equals to 7 when predicting conflicts at time $t+1$.

Feature Class #3. The final feature class at the layer level is the number of triangles in which c_1 and c_2 are involved, There are four types of possible triangles, generating three features – as two triangle types are equivalent. Figure 5.13 shows the possible type of triangles:



(a) T1 (b) T2 (c) T3 (d) T4

Figure 5.13 – The four types of triangles we count to generate the "Triangle_x" features.

- Triangle T1 is a closed triangle: the number of common neighbors between c_1 and c_2 if they are connected.
- Triangle T2 is a triangle open on the c_1-c_2 side: the number of common neighbors between c_1 and c_2 if they are not connected.
- Triangles T3 and T4 are triangles open on a different side: the combined degree of c_1 and c_2 if they are connected, minus their common neighbors.

5.5 Additional Results

5.5.1 Predictions vs Conflicts

Our model was able to assess the number of conflicts expected to happen in the following year. See Figure 5.14 for details. We can see that our model using both Decision tree and Random Forest as predictive method is, in general, timid. This means that, the predictions are in the vast majority of the years under the right number of conflicts.
As we can see, when it comes to predicting the number of conflicts that will

happen in a year, or in other words, how bad a year is going to be based on the number of conflicts that will happen, Decision tree is a better predictor. In the figure we can see how in certain years the number of predictions is almost equal to the real number of conflicts. On the other hand, random forest in its predictions remains always below the real number of conflicts.

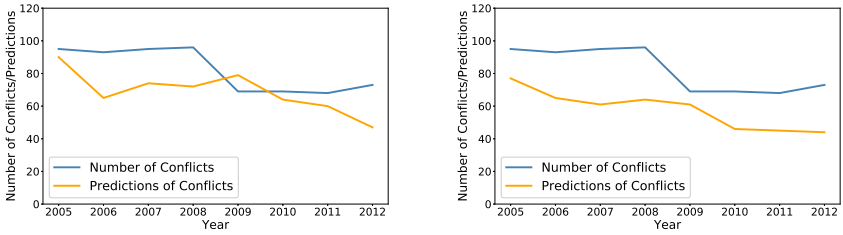


Figure 5.14 – Number of conflicts vs the number of conflicts predicted per year for Decision Tree (Left) and Random Forest (Right) predictive model.

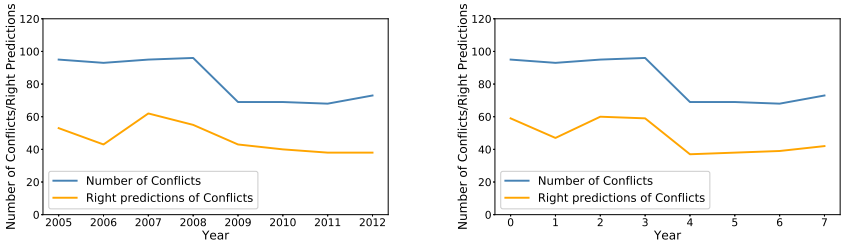


Figure 5.15 – Number of conflicts vs right number of conflicts predicted (True Positive) per year for Decision Tree (Left) and Random Forest (Right) predictive model.

Figure 5.15 shows the evolution of our forecasting accuracy per year. We can see that in both cases (Decision Tree on the left and Random Forest on the right), although the right predictions curve follows more or less the conflicts curve, the predictive models are timid, in both cases the prediction curve is below the conflicts curve during all the time frame.

The percentage of false positives for the Decision Tree and Random Forest are 32% and 17% respectively. On the other hand, the percentage of false positives that become true positive after a delay of 2.72 and 2.46 are 55% and 89% for Decision Tree and Random Forest respectively.

Figure 5.16 shows the percentage of right predictions per year for both De-

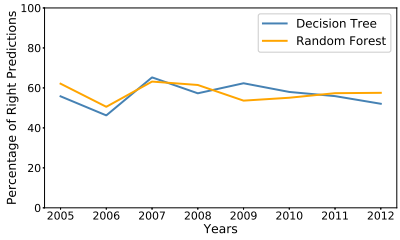


Figure 5.16 – Percentage of right predictions (True Positives) per year for Decision Tree and Random Forest predictive models

cision tree and Random Forest predictive method. As we can see, both of them are very similar and go from $45\% - 65\%$ approximately.

On the other hand, since conflict data is very scarce, only 1,564 conflicts in 50 years, our model is very well trained to predict when a conflict will NOT happen. This is demonstrated by the high percentages of true negatives obtained for both predictive models, 99.93% and 99.98% for Decision tree and Random Forest respectively. This means that both models have extremely high accuracies.

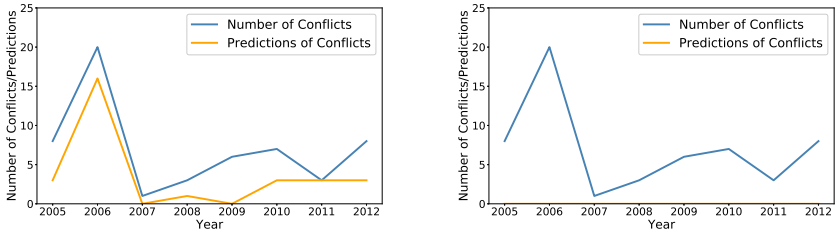


Figure 5.17 – Number of conflicts vs the number of conflicts predicted per year for Decision Tree (Left) and Random Forest (Right) predictive model for the "short" version of the data.

Figure 5.17 shows the number of conflicts per year vs the number of conflicts predicted for the "short" version of the data. Remember that in the short version of the data we consider the initial year of each conflict. As we can see, even if it is a very sparse data (only 56 conflicts in the period of 7 years), Decision Tree predictive model has a remarkable accuracy. On the other hand Random Forest didn't predict any conflict for this version of the data.

5.5.2 Conflicts vs Visa requirement

The visa layer is particularly problematic, given that it shows a static snapshot of the current status, which was different in the past, however, from 1962 to 2012 there were 1564 conflicts and 1273 were between countries with a visa requirement link, it means that 81.4% of the conflicts (wars) during this period were between countries that in certain way didn't have the best relation. Figure 5.18 shows this effect. Virtually (96.8%) all conflicts in the 2001-2005 period were between countries with a visa requirement link. This is expected as those snapshots are the closest temporally to our visa data. However, only $\sim 50\%$ of conflict in 1996 were between countries with a visa link. We can expect that those countries had a visa requirement at that time, which has been relaxed by now.

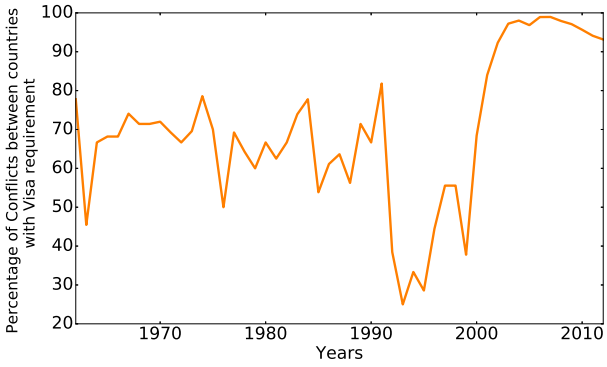


Figure 5.18 – Percentage of Conflicts between countries that have a Visa requirement.

Interestingly enough, this does not seem to hurt our precision. In fact, if anything, precision goes down as we move closer to the present, as Figure 5.16 shows. This is probably due to many confounders factors – as mentioned before, the number of actual conflicts makes the prediction harder – but it shows that the static nature of the visa layer is not necessarily a significant problem.

5.5.3 Conflicts vs Sanctions

In (McCormack & Pascoe 2017) the authors demonstrate that through their use as tools of military containment, sanctions play an unappreciated role in international politics. They also show that sanctions can be used to smooth shifts in relative power that would otherwise lead to preventive war.

The question of how states might use sanctions in order to display their resolve over a disputed issue in international politics has been central to the sanctions literature. Because sanctions are costly, a theory explaining their imposition as an alternative to war might look to their signalling properties – because of their costliness, sanctions or the threat of sanctions provide a way for states to demonstrate their resolve, either through sinking costs or tying hands (Banks et al. 1991, Fearon 1997). If sanctions serve a signalling purpose, the sanctioner might use the costliness of sanctions to demonstrate its own resolve on the issue at hand. Economic sanctions may include various forms of trade barriers, tariffs, and restrictions on financial transactions (Haidar 2017).

Conflicts	**Sanctions**
Afghanistan	United States
Syria	South Africa
Iraq	Germany
Saudi Arabia	Egypt
United States	Russia
Viet Nam	Viet Nam
United Kingdom	China
Somalia	Italy
North Korea	Israel
Israel	Romania
Portugal	Japan
Palestine	India
Australia	Brazil
Uganda	Chile
France	Argentina

Table 52 – Top 15 of countries more involved in Conflicts (left) and Sanctions (right)

In our dataset, from 1564 conflicts (remember that in this work when we mention conflicts we refer specifically to wars) that happened in the period from 1962 to 2012 only 4 of them were preceded or followed by a sanction. Table 52 shows the Top 15 of the countries involved the most in Conflicts (left) and Sanctions (right). Conflicts network isn't a directed network, so we just counted the number of conflicts per country and make a single table. By the other hand, Sanctions network is directed, so what we have done is to split it between sanction-sending and sanction-receiving countries and make the Top countries list considering the both roles. We found that only 3 countries are in both lists, United States, Viet Nam and Israel.

Although the efficacy of sanctions is debatable and sanctions can have unintended consequences (Lee 2018), according to our results, they are an effective measure so that countries that have some type of problem between them, solve them in a civilized way and avoid reaching conflicts in which weapons are involved.

Table 52 shows us that there are different kind of countries in both lists, countries with different economic, historical, social and geographical conditions. Our results show that the countries that receive or impose sanctions tend not to engage, in general, in wars and that sanctions can be used to offset shifts in military power that would otherwise cause commitment problem-driven wars. This argument indicates that sanctions can be used as a tool for peace (McCormack & Pascoe 2017).

Chapter 6

Discussion

In this work we have made an analysis, using the tools that the study of networks provide us, of different aspects regarding relations between countries. We have considered the world as a great multilayer network in which the nodes represent the countries and the links have different meanings depending on the problem addressed. The versatility of networks allowed us to consider links from commercial relationships to competencies in an ecosystem. In the two works presented in this thesis, the analysis of multilayer systems has been fundamental, since this has allowed us to analyze the competitions for exports between countries considering all the different categories of products and in another work to create a conflicts forecasting model that consider as input some of the relationships that countries establish between them.

In the first work we have adopted an ecosystem approach to the analysis of the global trade patterns. We have represented exporters as organisms competing for resources in different market niches. A market niche was defined as country importing a product. The fundamental assumption was that exporters want to out-compete other exporters, attempting to occupy the entire market niche. The appearance of a new exporter in a niche can be followed by the disappearance of another country. This is what we have called a displacement event. We have created a formal definition of displacements and we have systematically collected all of them along a period spanning fifty years. A displacement event was represented as a directed edge going from the out-competing exporter to the displaced one. We have called the collection of all displacements a *"competition network"*, which is a weighted directed multilayer network, where each layer is a product class.

While the in- and out-degree of a node (country) in a competition network have an intuitive interpretation – being the number of displacements experienced and caused by an exporter, respectively –, we have showed that in practice these measures cannot be used for classifying countries. The reason is their very high correlation. To fix this issue, we have calculated network roles based on in- and out-degree flows. By clustering nodes according to their role score, we were able to classify them in three categories: out- competing, transitioning, and displaced. We showed that these classes can be used to predict the future performance of an exporter in a particular market, in terms of the growth of total export value.
This classification is a meaningful one: when testing the future export patterns of these countries, we show that out-competing countries have distinctly stronger growth rates than the other two classes.

This methodology has some issues and could lead to future developments. First, it does not consider actual displacements: the edge creation process is correlative by design, so we are not really capturing if the appearance of a new exporter really caused the disappearance of another. We have not built a formal theory of why the competition network roles are predictive: we do not control for confounding factors that might drive both growth in exports and the position of a country in the network.

In the second work presented in this thesis, we have used a multilayer signed network model to forecast the outbreak of international conflicts. We did so assuming that international relations follow social balance theory: countries will tend to form coherent groups of allies. We have extended social balance theory creating a multilayer balance measure. When we use multilayer social balance as a feature of a machine learning algorithm, we see that it positively contributes to the prediction of links in the conflict layer.
Apart from that, our machine learning model can forecast with high precision how conflictive the next years will be, and not only that but also the countries that will be involved. We also showed that false positives tend to transform into true positives given a lag of less than three years on average.

However, this work has also some limitations. First, we do not really use any theory of international relationship in our work. We are actually not predicting conflicts, only forecasting them. Second, the data collection is limited. As a result, our social balance measure is necessarily incomplete. Third, given

the structure of the multilayer network, social balance is much more suitable for the prediction of positive, rather than negative, links and since we are predicting a negative relationship – conflicts – we are not using the best feature possible, although we did not know this beforehand.
Notwithstanding these limitations, we achieve good precision and recall values. This is even more impressive given the extreme sparsity of international conflict data.

In the development of this thesis we have made use of the tools offered by the study of networks to scrutinize relations between countries from a different perspective. Our goal hasn't been to understand the historical, political or economic reasons why the relations between countries are in certain way, but to use them as input for predictive models. We consider the countries as elements of a big multilayer complex network in which the links between them can represent any kind of relations that the countries can establish between them. This perspective is easy to implement and it has been proved, with the accuracy of the predictive models presented in this thesis, to be an effective one.
We open with the aforementioned works the possibility of further exploring international relations from this perspective.

Future Work

Fingerprints in the world-trade market

As we have mentioned before, the common thread of the works presented in this thesis is the study and analysis of the effects of the appearance and disappearance of both positive and negative links on the dynamics and organization of networks. Continuing in this same line of thought, we have begun the study of the traces, whether positive or negative, of certain economic and historical events that in one way or another have impacted actual society. We wanted to include different types of events, both positive, such as economic agreements and negative, such as wars. We present these preliminary results as a preview of what could be a very interesting future work in the same line of research that we have handled throughout these years in the preparation of this thesis.

So far the events that we have studied are: two of the most important economical agreements in the last 20 years, the creation of the Euro as unique currency for a very select group of countries in Europe, the signing of the eco-

nomical agreement between countries in South America, MERCOSUR and the Gulf War and the war between Iran and Iraq.

The fundamental questions that we try to answer with this work are: *Do political tensions have economic consequences? Do economic agreements have a positive impact on trade as expected?*

According to Franko, "*the theoretical underpinning of free trade is the theory of comparative advantage, which states that countries should trade those goods that they can most efficiently produce to maximize global output*." (Franko 2018).

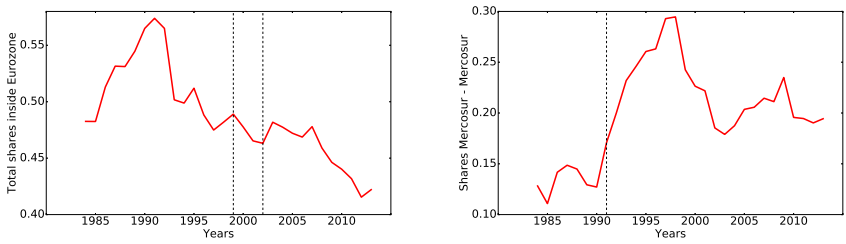


Figure 6.1 – Total shares inside of Eurozone (left) and Mercosur (right).

Figure 6.1 shows the evolution in time of the total export market shares (including all the different categories of products, See section 4.1.1) within the corresponding countries belonging to each agreement, before, during and after the two economical events that we are considering, the creation of the Euro and MERCOSUR. In this case, the export market shares are calculated by dividing the exports between the countries within the agreement by the total exports of the countries within the agreement to the rest of the world (including the countries within the agreement), and it is expressed as percentage.

Let's start from the fact that, although in both cases the main goal has been the reinforcement of the commercial relations between the involved countries, these two events have different economic purposes and were created under different conditions and within different historical, social, economic and cultural contexts. Therefore their economic effects are not expected to be the same.
However, considering that these events occurred with the purpose of increasing trade within their corresponding involved countries, we can observe in Figure 6.1 that this expected behaviour is not obvious and that if we want to delve into the effects of these events and discover if the economies involved have been really benefited independently, a graph of the total export shares within the zones is not a enough source of information. Trade system is complex, and

must be treated and studied as such (Bhattacharya et al. 2008).

The other two events that we have considered are two of the most important wars of the end of the last century, the Gulf War and the war between Iran and Iraq. In the context of the study of networks, a war between countries is represented with the appearance of new negative links between the nodes (countries) participating in that war. The effects of war are widely spread and can be long term or short term (Machel et al. 1996). Effects of war include mass destruction of cities and have long lasting effects on a country's economy. (Olmsted 2015)

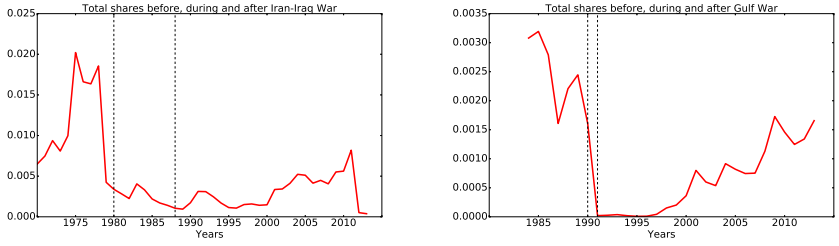


Figure 6.2 – Total shares Before, During and After Iran-Iraq (left) and Gulf (right) Wars

Once more, we have to start by clarifying that the reasons that led to these wars are totally different. Despite this, the economic and political effects on the countries that participate in wars have common denominators and for this reason it is natural to expect a more or less similar effect in both cases (Olmsted 2015).
Figure 6.2 shows us how the negative effects of wars on trade among participating countries can occur abruptly as in the case of the Gulf War or in a more slightly progressive way and can even not cancel completely the trade relations between the countries involved as in the case of Iran-Iraq war. Once again, the complexity of the relationships between countries makes it impossible to measure the consequences of the wars in the network considering only the total changes in exports between the participating countries.

The fundamental objective of this work raises the study and analysis of the effects of the appearance of new positive and negative links in a network brought to an economic and political context (Schweitzer et al. 2009). Since nodes (countries in the case) and links are parts of a chaotic complex system, the appearance of these new links in the network affects to each of the elements in a different way, for this reason, if we want to delve into the details of the effects on the structure and dynamics of temporary networks after the

appearance of new links, the study of the network as a whole is insufficient. The differences in the effects lie in the structure itself of the network (clustering coefficient, degree distribution, etc.) and the initial role of each node, for example, nodes with a greater number of connections will be affected differently from nodes with few connections (Maoz 2010). To evaluate the effects of these appearances it is important and necessary to make use of the tools provided by the study of temporal networks to analyze structural changes and reorganizations of previously existing connections (Lambiotte & Masuda 2016, Johansson 1995).

Bibliography

Al Hasan, M., Chaoji, V., Salem, S. & Zaki, M. (2006), Link prediction using supervised learning, *in* 'SDM06: workshop on link analysis, counter-terrorism and security'.

Albert, R. & Barabási, A.-L. (2000), 'Topology of evolving networks: Local events and universality', *Phys. Rev. E* **85**(0), 5234–5237.

Aleta, A. & Moreno, Y. (2019), 'Multilayer networks in a nutshell', *Annual Review of Condensed Matter Physics* **10**(1), 45–62.

Allen, F. & Babus, A. (n.d.), Networks in finance (august 2008), Technical report, Wharton Financial Institutions Center Working Paper.

Antal, T., Krapivsky, P. L. & Redner, S. (2005), 'Dynamics of social balance on networks', *Physical Review E* **72**(3), 036121.

Antal, T., Krapivsky, P. L. & Redner, S. (2006), 'Social balance on networks: The dynamics of friendship and enmity', *Physica D: Nonlinear Phenomena* **224**(1-2), 130–136.

Bachi, G., Coscia, M., Monreale, A. & Giannotti, F. (2012), Classifying trust/distrust relationships in online social networks, *in* 'Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)', IEEE, pp. 552–557.

Bahar, D., Hausmann, R. & Hidalgo, C. A. (2014), 'Neighbors and the evolution of the comparative advantage of nations: Evidence of international knowledge diffusion?', *Journal of International Economics* **92**(1), 111–123.

Bambach, R. K., Knoll, A. H. & Sepkoski, J. J. (2002), 'Anatomical and ecological constraints on phanerozoic animal diversity in the marine realm', *Proceedings of the National Academy of Sciences* **99**(10), 6854–6859.

Banks, J. S. et al. (1991), *Signaling games in political science*, Vol. 46, Psychology Press.

Barabási, A.-L. & Albert, R. (1999*a*), 'Emergence of scaling in random networks', *science* **286**(5439), 509–512.

Barabási, A.-L. & Albert, R. (1999*b*), 'Emergence of scaling in random networks', *Science* **286**(5439), 509–512.

Barabási, A.-L. et al. (2016), *Network science*, Cambridge university press.

Berlingerio, M., Coscia, M., Giannotti, F., Monreale, A. & Pedreschi, D. (2013), 'Multidimensional networks: foundations of structural analysis', *World Wide Web* **16**(5-6), 567–593.

Bhattacharya, K., Mukherjee, G., Saramäki, J., Kaski, K. & Manna, S. S. (2008), 'The international trade network: weighted network analysis and modelling', *Journal of Statistical Mechanics: Theory and Experiment* **2008**(02), P02002.

Bianconi, G. (2015), 'Interdisciplinary and physics challenges of network theory', *EPL (Europhysics Letters)* **111**(5), 56001.

Blondel, V. D., Gajardo, A., Heymans, M., Senellart, P. & Van Dooren, P. (2004), 'A measure of similarity between graph vertices: Applications to synonym extraction and web searching', *SIAM review* **46**(4), 647–666.

Blumer, A., Ehrenfeucht, A., Haussler, D. & Warmuth, M. K. (1989), 'Learnability and the vapnik-chervonenkis dimension', *Journal of the ACM (JACM)* **36**(4), 929–965.

Bonchev, D. (1991), *Chemical graph theory: introduction and fundamentals*, Vol. 1, CRC Press.

Borgatti, S. P. & Everett, M. G. (1993), 'Two algorithms for computing regular equivalence', *Social networks* **15**(4), 361–376.

Borgatti, S. P., Mehra, A., Brass, D. J. & Labianca, G. (2009), 'Network analysis in the social sciences', *science* **323**(5916), 892–895.

Brandt, P. T., Freeman, J. R. & Schrodt, P. A. (2011), 'Real time, time series forecasting of inter-and intra-state political conflict', *Conflict Management and Peace Science* **28**(1), 41–64.

Buldyrev, S. V., Parshani, R., Paul, G., Stanley, H. E. & Havlin, S. (2010), 'Catastrophic cascade of failures in interdependent networks', *Nature* **464**(7291), 1025.

Burger, M., Van Oort, F. & Linders, G.-J. (2009), 'On the specification of the gravity model of trade: zeros, excess zeros and zero-inflated estimation', *Spatial Economic Analysis* **4**(2), 167–190.

Bustos, S., Gomez, C., Hausmann, R. & Hidalgo, C. A. (2012), 'The dynamics of nestedness predicts the evolution of industrial ecosystems', *PloS one* **7**(11), e49393.

Cacciari, M., Salam, G. P. & Soyez, G. (2008), 'The anti-kt jet clustering algorithm', *Journal of High Energy Physics* **2008**(04), 063.

Cartwright, D. & Harary, F. (1956), 'Structural balance: a generalization of heider's theory.', *Psychological review* **63**(5), 277.

Cederman, L.-E. & Weidmann, N. B. (2017), 'Predicting armed conflict: Time to adjust our expectations?', *Science* **355**(6324), 474–476.

Clauset, A. (2019), 'On the frequency and severity of interstate wars', *arXiv preprint arXiv:1901.05086* .

Cooper, K. & Barahona, M. (2010), 'Role-based similarity in directed networks', *arXiv preprint arXiv:1012.2726* .

Cortes, C. & Vapnik, V. (1995), 'Machine learning', *Support vector networks* **20**, 273–297.

Coscia, M., Giannotti, F. & Pedreschi, D. (2011), 'A classification for community discovery methods in complex networks', *Statistical Analysis and Data Mining* **4**(5), 512–546.

Coscia, M. & Neffke, F. M. (2017), Network backboning with noisy data, *in* '2017 IEEE 33rd International Conference on Data Engineering (ICDE)', IEEE, pp. 425–436.

Cristelli, M., Gabrielli, A., Tacchella, A., Caldarelli, G. & Pietronero, L. (2013), 'Measuring the intangibles: A metrics for the economic complexity of countries and products', *PloS one* **8**(8), e70726.

Davey, B. & Priestley, H. (1990), 'Introduction to lattices and order cambridge univ', *Press, Cambridge* .

Davis, J. A. (1967), 'Clustering and structural balance in graphs', *Human relations* **20**(2), 181–187.

De Bacco, C., Power, E. A., Larremore, D. B. & Moore, C. (2017), 'Community detection, link prediction, and layer interdependence in multilayer networks', *Physical Review E* **95**(4), 042317.

De Haas, H. & Czaika, M. (2013), 'Measuring migration policies: Some conceptual and methodological reflections', *Journal of Migration and Law* **6**(1), 47–65.

De Nooy, W., Mrvar, A. & Batagelj, V. (2018), *Exploratory social network analysis with Pajek: Revised and expanded edition for updated software*, Vol. 46, Cambridge University Press.

Dehmer, M. & Emmert-Streib, F. (2009), *Analysis of complex networks: from biology to linguistics*, John Wiley & Sons.

Dehmer, M., Emmert-Streib, F., Graber, A. & Salvador, A. (2011), *Applied Statistics for network biology*, Wiley Online Library.

DESA, U. (2013), 'Trends in international migrant stock: Migrants by destination and origin', *United Nations database, POP/DB/MIG/Stock/Rev* **2013**.

Dorogovtsev, S. N., Mendes, J. F. F. & Samukhin, A. N. (2001), 'Size-dependent degree distribution of a scale-free growing network', *Phys. Rev. E* **63**(4), 062101.

Eagle, N. & Pentland, A. S. (2006), 'Reality mining: sensing complex social systems', *Personal and ubiquitous computing* **10**(4), 255–268.

Easley, D. & Kleinberg, J. (2010), *Networks, crowds, and markets: Reasoning about a highly connected world*, Cambridge University Press.

Eckmann, J.-P., Moses, E. & Sergi, D. (2004), 'Entropy of dialogues creates coherent structures in e-mail traffic', *Proceedings of the National Academy of Sciences* **101**(40), 14333–14337.

Emmert-Streib, F. & Dehmer, M. (2011), 'Networks for systems biology: conceptual connection of data and function', *IET systems biology* **5**(3), 185–207.

Emmert-Streib, F., Tripathi, S., Yli-Harja, O. & Dehmer, M. (2018), 'Understanding the world economy in terms of networks: A survey of data-based network science approaches on economic networks', *Frontiers in Applied Mathematics and Statistics* **4**, 37.

Fearon, J. D. (1997), 'Signaling foreign policy interests: Tying hands versus sinking costs', *Journal of Conflict Resolution* **41**(1), 68–90.

Fişek, M. H., Berger, J. & Norman, R. Z. (1991), 'Participation in heterogeneous and homogeneous groups: A theoretical integration', *American Journal of Sociology* **97**(1), 114–142.

Franko, P. (2018), *The puzzle of Latin American economic development*, Rowman & Littlefield.

Fraser, A. & Marcu, D. (2007), 'Measuring word alignment quality for statistical machine translation', *Computational Linguistics* **33**(3), 293–303.

Freeman, L. (2004), 'The development of social network analysis', *A Study in the Sociology of Science* **1**, 687.

Freeman, L. C. (n.d.), '1978-1979."', *Centrality in Social Networks." Social Networks1* pp. 215–39.

Gaulier, G. & Zignago, S. (2010), 'Baci: international trade database at the product-level (the 1994-2007 version)', *CEPII Working Paper 2010-23.* .

Gleditsch, N. P., Wallensteen, P., Eriksson, M., Sollenberg, M. & Strand, H. (2002), 'Armed conflict 1946-2001: A new dataset', *Journal of peace research* **39**(5), 615–637.

Goldstone, J. A., Bates, R. H., Epstein, D. L., Gurr, T. R., Lustik, M. B., Marshall, M. G., Ulfelder, J. & Woodward, M. (2010), 'A global model for forecasting political instability', *American Journal of Political Science* **54**(1), 190–208.

GRU, R. L. (n.d.), 'Bootstrap aggregating'.

Guimera, R. & Amaral, L. A. N. (2005), 'Functional cartography of complex metabolic networks', *Nature* **433**(7028), 895–900.

Haidar, J. I. (2017), 'Sanctions and export deflection: evidence from iran', *Economic Policy* **32**(90), 319–355.

Hanhijärvi, S., Garriga, G. C. & Puolamäki, K. (2009), Randomization techniques for graphs, *in* 'Proceedings of the 2009 SIAM International Conference on Data Mining', SIAM, pp. 780–791.

Hanneman, R. & Riddle, M. (2005), *Introduction to Social Network Methods*, University of California, Riverside.

Harary, F. (1953), 'On the notion of balance of a signed graph', *The Michigan Mathematical Journal* **2**(2), 143–146.

Hausmann, R., Hidalgo, C. A., Bustos, S., Coscia, M., Simoes, A. & Yildirim, M. A. (2014), *The atlas of economic complexity: Mapping paths to prosperity*, Mit Press.

Heider, F. (2013), *The psychology of interpersonal relations*, Psychology Press.

Henderson, K., Gallagher, B., Eliassi-Rad, T., Tong, H., Basu, S., Akoglu, L., Koutra, D., Faloutsos, C. & Li, L. (2012), Rolx: structural role extraction & mining in large graphs, *in* 'Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining', ACM, pp. 1231–1239.

Hidalgo, C. A., Klinger, B., Barabási, A.-L. & Hausmann, R. (2007), 'The product space conditions the development of nations', *Science* **317**(5837), 482–487.

Ho, T. K. (2002), 'A data complexity analysis of comparative advantages of decision forest constructors', *Pattern Analysis & Applications* **5**(2), 102–112.

Holme, P. & Saramäki, J. (2012), 'Temporal networks', *Physics reports* **519**(3), 97–125.

Holsti, K. J. & Holsti, K. J. (1991), *Peace and war: Armed conflicts and international order, 1648-1989*, Vol. 14, Cambridge University Press.

Hristova, D., Noulas, A., Brown, C., Musolesi, M. & Mascolo, C. (2016), 'A multilayer approach to multiplexity and link prediction in online geo-social networks', *EPJ Data Science* **5**(1), 24.

Hvattum, L. M. & Arntzen, H. (2010), 'Using elo ratings for match result prediction in association football', *International Journal of forecasting* **26**(3), 460–470.

Jackson, M. O. (2010), *Social and economic networks*, Princeton university press.

Jackson, M. O. & Nei, S. (2015), 'Networks of military alliances, wars, and international trade', *Proceedings of the National Academy of Sciences* **112**(50), 15277–15284.

Jeh, G. & Widom, J. (2002), Simrank: a measure of structural-context similarity, *in* 'Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining', pp. 538–543.

Johansson, B. (1995), The dynamics of economic networks, *in* 'Networks in Action', Springer, pp. 287–308.

Karsai, M., Perra, N. & Vespignani, A. (2014), 'Time varying networks and the weakness of strong ties', *Scientific Reports* .

Keohane, R. O. & Nye, J. S. (1974), 'Transgovernmental relations and international organizations', *World politics* **27**(1), 39–62.

Keohane, R. O. & Nye, J. S. (1977), *Power and interdependence*, Boston.

Kirman, A. (1997), 'The economy as an evolving network', *Journal of evolutionary economics* **7**(4), 339–353.

Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y. & Porter, M. A. (2014*a*), 'Multilayer networks', *Journal of Complex Networks* **2**(3), 203–271.

Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y. & Porter, M. A. (2014*b*), 'Multilayer networks', *Journal of complex networks* **2**(3), 203–271.

Kleinberg, J. M. (1999*a*), 'Authoritative sources in a hyperlinked environment', *Journal of the ACM (JACM)* **46**(5), 604–632.

Kleinberg, J. M. (1999*b*), 'Hubs, authorities, and communities', *ACM computing surveys (CSUR)* **31**(4es), 5–es.

Klimek, P., Hausmann, R. & Thurner, S. (2012), 'Empirical confirmation of creative destruction from world trade data', *PloS one* **7**(6), e38924.

Lambiotte, R. & Masuda, N. (2016), *A guide to temporal networks*, Vol. 4, World Scientific.

Lee, Y. S. (2018), 'International isolation and regional inequality: Evidence from sanctions on north korea', *Journal of Urban Economics* **103**, 34–51.

Leicht, E., Holme, P. & Newman, M. (2006), 'Modularity-maximizing network communities via mathematical programming', *Phys. Rev. E* **73**, 026120.

Leskovec, J., Huttenlocher, D. & Kleinberg, J. (2010*a*), Predicting positive and negative links in online social networks, *in* 'Proceedings of the 19th international conference on World wide web', ACM, pp. 641–650.

Leskovec, J., Huttenlocher, D. & Kleinberg, J. (2010*b*), Signed networks in social media, *in* 'Proceedings of the SIGCHI conference on human factors in computing systems', ACM, pp. 1361–1370.

Liaw, A., Wiener, M. et al. (2002), 'Classification and regression by randomforest', *R news* **2**(3), 18–22.

Liben-Nowell, D. & Kleinberg, J. (2007), 'The link-prediction problem for social networks', *Journal of the American society for information science and technology* **58**(7), 1019–1031.

Lichtenwalter, R. N., Lussier, J. T. & Chawla, N. V. (2010), New perspectives and methods in link prediction, *in* 'Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining', pp. 243–252.

Lü, L. & Zhou, T. (2011), 'Link prediction in complex networks: A survey', *Physica A: statistical mechanics and its applications* **390**(6), 1150–1170.

Luce, R. & Perry, A. (1949), 'A method of matrix analysis of group structure', *PSYCHOMETRIKA* .

Mach, K. J., Kraan, C. M., Adger, W. N., Buhaug, H., Burke, M., Fearon, J. D., Field, C. B., Hendrix, C. S., Maystadt, J.-F., O'Loughlin, J., Roessler, P., Scheffran, J., Schultz, K. A. & von Uexkull, N. (2019), 'Climate as a risk factor for armed conflict', *Nature* .

Machel, G. et al. (1996), *Impact of armed conflict on children*, UN.

MacQueen, J. et al. (1967), Some methods for classification and analysis of multivariate observations, *in* 'Proceedings of the fifth Berkeley symposium on mathematical statistics and probability', Vol. 1, Oakland, CA, USA, pp. 281–297.

Maoz, Z. (2010), *Networks of nations: The evolution, structure, and impact of international networks, 1816–2001*, Vol. 32, Cambridge University Press.

Martin, R. (2015), 'Rebalancing the spatial economy: the challenge for regional theory', *Territory, Politics, Governance* **3**(3), 235–272.

May, R. M., Levin, S. A. & Sugihara, G. (2008), 'Ecology for bankers', *Nature* **451**(7181), 893–894.

Mayer, T., Paillacar, R. & Zignago, S. (2008), 'Tradeprod. the cepii trade, production and bilateral protection database: Explanatory notes', *CEPII working paper* .

Mayer, T. & Zignago, S. (2011), 'Notes on cepii's distances measures: The geodist database', *CEPII Working Paper No. 2011-25* .

McCormack, D. & Pascoe, H. (2017), 'Sanctions and preventive war', *Journal of Conflict Resolution* **61**(8), 1711–1739.

Meng, H., Xu, H.-C., Zhou, W.-X. & Sornette, D. (2017), 'Symmetric thermal optimal path and time-dependent lead-lag relationship: novel statistical tests and application to uk and us real-estate and monetary policies', *Quantitative Finance* **17**(6), 959–977.

Meunier, D., Lambiotte, R. & Bullmore, E. T. (2010), 'Modular and hierarchically modular organization of brain networks', *Frontiers in neuroscience* **4**, 200.

Mohri, M., Rostamizadeh, A. & Talwalkar, A. (2018), *Foundations of machine learning*, MIT press.

Moore, M. (1978), 'An international application of heider's balance theory', *European Journal of Social Psychology* **8**(3), 401–405.

Morgan, T. C., Bapat, N. & Kobayashi, Y. (2014), 'Threat and imposition of economic sanctions 1945–2005: Updating the ties dataset', *Conflict Management and Peace Science* **31**(5), 541–558.

Murata, T. & Moriyasu, S. (2007), Link prediction of social networks based on weighted proximity measures, *in* 'IEEE/WIC/ACM International Conference on Web Intelligence (WI'07)', IEEE, pp. 85–88.

Nagurney, A. & Siokos, S. (1997), Variational inequalities, *in* 'Financial Networks', Springer, pp. 49–73.

Neffke, F., Henning, M. & Boschma, R. (2011), 'How do regions diversify over time? industry relatedness and the development of new growth paths in regions', *Economic Geography* **87**(3), 237–265.

Newman, M. (2018*a*), *Networks*, Oxford university press.

Newman, M. E. J. (2018*b*), *Networks*, Oxford University Press.

Nigam, K., McCallum, A., Thrun, S., Mitchell, T. et al. (1998), 'Learning to classify text from labeled and unlabeled documents', *AAAI/IAAI* **792**(6).

Nilsson, N. J. (1996), 'Introduction to machine learning: An early draft of a proposed textbook'.

Ninove, L. et al. (2007), Graph similarity algorithms, *in* 'Seminar Presented at Department of Mathematical Engineering, University of Catholique de Louvain'.

Olmsted, P. J. C. (2015), Globalization denied: Gender and poverty in iraq and palestine, *in* 'Wages of Empire', Routledge, pp. 186–197.

O'Madadhain, J., Hutchins, J. & Smyth, P. (2005), 'Prediction and ranking algorithms for event-based network data', *ACM SIGKDD explorations newsletter* **7**(2), 23–30.

O'Madadhain, J., Smyth, P. & Adamic, L. (2005), Learning predictive models for link formation, *in* 'International sunbelt social network conference'.

Paul, H. & Leinhardt, S. (1971), 'Transitivity in structural models of small groups', *Comparative Group Studies* **2**(2), 107–124.

Petermann, J.-H. (2006), *Scientific Approaches to the Study of International Relations.*, Grin.

Pettersson, T. & Eck, K. (2018), 'Organized violence, 1989–2017', *Journal of Peace Research* p. 0022343318784101.

Pettersson, T. & Wallensteen, P. (2015), 'Armed conflicts, 1946–2014', *Journal of peace research* **52**(4), 536–550.

Pomeroy, C., Dasandi, N. & Mikhaylov, S. J. (2018), 'Multiplex communities and the emergence of international conflict', *arXiv preprint arXiv:1806.00615* .

Poncet, S. & de Waldemar, F. S. (2013), 'Export upgrading and growth: the prerequisite of domestic embeddedness', *World Development* **51**(2), 104–118.

Powers, D. M. (2011), 'Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation'.

Quinlan, J. R. (1986), 'Induction of decision trees', *Machine learning* **1**(1), 81–106.

Quinlan, J. R. (1993), *Programs for Machine Learning C4. 5*.

Raeder, T., Hoens, T. R. & Chawla, N. V. (2010), Consequences of variability in classifier performance estimates, *in* '2010 IEEE International Conference on Data Mining', IEEE, pp. 421–430.

Rajaraman, A. & Ullman, J. D. (2011), *Mining of massive datasets*, Cambridge University Press.

Reichardt, J. & White, D. R. (2007), 'Role models for complex networks', *The European Physical Journal B* **60**(2), 217–224.

Richardson, L. F. (1960), *Statistics of deadly quarrels*, Boxwood Press, Pittsburgh.

Risi, J., Sharma, A., Shah, R., Connelly, M. & Watts, D. J. (2019), 'Predicting history', *Nature Human Behaviour* .

Rossetti, G., Berlingerio, M. & Giannotti, F. (2011), Scalable link prediction on multidimensional networks, *in* 'Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on', IEEE, pp. 979–986.

Safavian, S. R. & Landgrebe, D. (1991), 'A survey of decision tree classifier methodology', *IEEE transactions on systems, man, and cybernetics* **21**(3), 660–674.

Samuel, A. L. (1959), 'Some studies in machine learning using the game of checkers', *IBM Journal of research and development* **3**(3), 210–229.

Saviotti, P. P. & Frenken, K. (2008), 'Export variety and the economic performance of countries', *Journal of Evolutionary Economics* **18**(2), 201–218.

Schweitzer, F., Fagiolo, G., Sornette, D., Vega-Redondo, F. & White, D. R. (2009), 'Economic networks: What do we know and what do we need to know?', *Advances in Complex Systems* **12**(04n05), 407–422.

Scripps, J., Tan, P.-N. & Esfahanian, A.-H. (2007), Node roles and community structure in networks, *in* 'Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis', pp. 26–35.

Sharp, P. (2018), *Introducing International Relations.*, Routledge.

Sporns, O. & Betzel, R. F. (2016), 'Modular brain networks', *Annual review of psychology* **67**, 613–640.

Stehlé, J., Voirin, N., Barrat, A., Cattuto, C., Colizza, V., Isella, L., Régis, C., Pinton, J.-F., Khanafer, N., Van den Broeck, W. et al. (2011), 'Simulation of an seir infectious disease model on the dynamic contact network of conference attendees', *BMC medicine* **9**(1), 87.

Stein, E., Crespi, G. et al. (2014), *Rethinking Productive Development: Sound Policies and Institutions for Economic Transformation*, Springer.

Strehl, A., Ghosh, J. & Mooney, R. (2000), Impact of similarity measures on web-page clustering, *in* 'Workshop on artificial intelligence for web search (AAAI 2000)', Vol. 58, p. 64.

Szell, M., Lambiotte, R. & Thurner, S. (2010), 'Multirelational organization of large-scale social networks in an online world', *Proceedings of the National Academy of Sciences* **107**(31), 13636–13641.

Tacchella, A., Cristelli, M., Caldarelli, G., Gabrielli, A. & Pietronero, L. (2012), 'A new metrics for countries' fitness and products' complexity', *Scientific reports* **2**(2), 723.

Vapnik, V. N. & Chervonenkis, A. Y. (1982), 'Necessary and sufficient conditions for the uniform convergence of means to their expectations', *Theory of Probability & Its Applications* **26**(3), 532–553.

Vapnik, V. N. & Chervonenkis, A. Y. (2015), On the uniform convergence of relative frequencies of events to their probabilities, *in* 'Measures of complexity', Springer, pp. 11–30.

Vina-Cervantes, V., Coscia, M. & Lambiotte, R. (2018), 'The struggle for existence in the world market ecosystem', *PloS one* **13**(10), e0203915.

Wasserman, S. & Faust, K. (1994), *Social Network Analysis*, Cambridge University Press.

Wasserman, S., Faust, K. et al. (1994), *Social network analysis: Methods and applications*, Vol. 8, Cambridge university press.

Watts, D. J. & Strogatz, S. H. (1998*a*), 'Collective dynamics of 'small-world' networks', *Nature* .

Watts, D. J. & Strogatz, S. H. (1998*b*), 'Collective dynamics of 'small-world'networks', *nature* **393**(6684), 440.

Willer, D. (1999), *Network exchange theory*, Greenwood Publishing Group.

Yang, Y., Lichtenwalter, R. N. & Chawla, N. V. (2015), 'Evaluating link prediction methods', *Knowledge and Information Systems* **45**(3), 751–782.

Zachary, W. W. (1977), 'An information flow model for conflict and fission in small groups', *Journal of anthropological research* **33**(4), 452–473.

Ženka, J., Novotnỳ, J. & Csank, P. (2014), 'Regional competitiveness in central european countries: in search of a useful conceptual framework', *European Planning Studies* **22**(1), 164–183.