Lecture Notes in Economics and Mathematical Systems 683

Pasquale Commendatore
Mariano Matilla-García
Luis M. Varela
Jose S. Cánovas *Editors*

Complex Networks and Dynamics

Social and Economic Interactions

COST
EUROPEAN COOPERATION IN SCIENCE AND TECHNOLOGY

Springer

Lecture Notes in Economics and Mathematical Systems 683

More information about this series at http://www.springer.com/series/300

About COST

COST (European Cooperation in Science and Technology) is a pan-European intergovernmental framework. Its mission is to enable breakthrough scientific and technological developments leading to new concepts and products and thereby contribute to strengthening Europe's research and innovation capacities.

It allows researchers, engineers and scholars to jointly develop their own ideas and take new initiatives across all fields of science and technology, while promoting multi- and interdisciplinary approaches. COST aims at fostering a better integration of less research-intensive countries to the knowledge hubs of the European Research Area. The COST Association, an international not-for-profit association under Belgian Law, integrates all management, governing and administrative functions necessary for the operation of the framework. The COST Association has currently 36 member countries (www.cost.eu).

EUROPEAN COOPERATION
IN SCIENCE AND TECHNOLOGY

COST is supported by the EU Framework Programme Horizon 2020

Acknowledgement

This book is based upon work from COST Action ISCH IS1104 "The EU in the new complex geography of economic systems: models, tools and policy evaluation", supported by COST (European Cooperation in Science and Technology).

Pasquale Commendatore • Mariano Matilla-García •
Luis M. Varela • Jose S. Cánovas
Editors

Complex Networks and Dynamics

Social and Economic Interactions

Springer

Editors
Pasquale Commendatore
Department of Law
University of Naples "Federico II"
Naples, Italy

Mariano Matilla-García
Faculty of Economics
UNED
Madrid, Spain

Luis M. Varela
Faculty of Physics
University of Santiago de Compostela
Santiago de Compostela, Spain

Jose S. Cánovas
Department of Applied Mathematics
Technical University of Cartagena
Cartagena, Spain

ISSN 0075-8442 ISSN 2196-9957 (electronic)
Lecture Notes in Economics and Mathematical Systems
ISBN 978-3-319-40801-9 ISBN 978-3-319-40803-3 (eBook)
DOI 10.1007/978-3-319-40803-3

Library of Congress Control Number: 2016949373

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

Contents

Part III Dynamical Systems

List of Contributors

Roberto Basile Department of Economics, Second University of Naples, Caserta, Campania, Italy

Spiros Bougheas School of Economics, University of Nottingham, Nottingham, UK

Jose S. Cánovas Departamento de Matemática Aplicada y Estadística, Universidad Politécnica de Cartagena, Cartagena, Spain

Pasquale Commendatore Department of Law, University of Naples 'Federico II', Naples, Italy

Anna Maria D'Arcangelis Department of Economics and Business, University of Tuscia, Viterbo, Italy

Luca De Benedictis Department of Law and Economics, University of Macerata, Macerata, Italy

Giorgio Fagiolo Istituto di Economia, Scuola Superiore Sant'Anna, Pisa, Italy

Giovanny Guerrero Universidad Central del Ecuador, Ciudadela Universitaria, Quito, Ecuador

María Muñoz Guillermo Departamento de Matemática Aplicada y Estadística, Universidad Politécnica de Cartagena, Cartagena, Spain

Alan Kirman GREQAM, EHESS, Aix-Marseille University, Marseille, France

Ingrid Kubin Department of Economics, Institute for International Economics and Development, WU Vienna University of Economics and Business, Vienna, Austria

José Antonio Langa Departamento de Ecuaciones Diferenciales y Análisis Numérico, Seville University, Sevilla, Spain

Dunia López-Pintado CORE, Université catholique de Louvain, Louvain-la-Neuve, Belgium

Department of Economics, Universidad Pablo de Olavide, Sevilla, Spain

Mariano Matilla-García Departamento de Economía Aplicada Cuantitativa, Universidad Nacional de Educación a Distancia (UNED), Madrid, Spain

Ugo Merlone Department of Psychology, Center for Cognitive Science, University of Turin, Torino, Italy

Jesús Mur Department of Economic Analysis, Faculty of Economics, Campus Rio Ebro University of Zaragoza, Zaragoza, Spain

Tönu Puu CERUM, Umeå University, Umeå, Sweden

Davide Radi School of Economics and Management, LIUC - Università Cattaneo, Castellanza, Lombardy, Italy

Antonio Rodriguez-Moral Universidad Nacional de Educación a Distancia (UNED), Madrid, Spain

Angelo Romano Department of Psychology, University of Turin, Torino, Italy

Giulia Rotundo Department of Methods and Models for Economics, Territory and Finance, La Sapienza University of Rome, Rome, Italy

Ian Stewart Mathematics Institute, University of Warwick, Coventry, UK

Antonio Suárez Departamento de Ecuaciones Diferenciales y Análisis Numérico, Seville University, Sevilla, Spain

Luis M. Varela Grupo de Nanomateriais, Fotónica e Materia Branda, Departamento de Física da Materia Condensada, Universidade de Santiago de Compostela, Santiago de Compostela, Spain

Marc Vorsatz Departamento de Análisis Económico II, Universidad Nacional de Educación a Distancia (UNED), Madrid, Spain

Fundación de Estudios de Economía Aplicada (FEDEA), Madrid, Spain

Introduction

Abstract This collected volume represents an output of a training school held in Madrid on 12–14 February 2014, under the coverage of the ISCH COST Action IS1104 project on "The EU in the new complex geography of economic systems: models, tools and policy evaluation". It focuses on complex networks and on dynamic methods applied to the study of economic and social interactions and offers an ample view of current research on complex network analysis. Three different and interdisciplinary, but complementary, aspects of networks are put together in a single piece, namely, (1) theoretical features of complex networking, (2) applied network analysis to social and economic issues and (3) dynamical aspects of systems and networks.

In 12–14 February 2014, under the coverage of the ISCH COST Action IS1104 project on 'The EU in the new complex geography of economic systems: models, tools and policy evaluation'—a research group financed by the European Commission (http://www.gecomplexity-cost.eu/home/)—a very active training school in Madrid at the Faculty of Economics of UNED (National Distance Education University) (http://www.dmae.upct.es/~training/index.html) took place.

This training school was focused on complex networks theory and applications to economics and to economic dynamics generated by differential and difference equations. The courses have been delivered by experts in these fields, some of them within the Action as well as very well-recognized specialists outside the Action.

This collected volume represents an output of that meeting. It is focused on complex networks and dynamic methods applied to the study of economic and social interactions. Some of the excellent papers discussed in Madrid's training school form the main structure of the book. However, in order to add further value to the volume, we invited other leading experts to contribute.

One of the most interesting aspects of this volume is that it offers an ample view of current research on complex network analysis. Three different and interdisciplinary, but complementary, aspects of networks are put together in a single piece, namely, (1) theoretical features of complex networking, (2) applied network analysis to social and economic issues and (3) dynamical aspects of systems and networks. These topological objects have become a major tool for the description of agent-based systems, and indeed they can be considered to be the skeleton of every complex system. The structure and the dynamics of these systems can be regarded as emergent properties of the set of connections (the links) between their constituent parts (the nodes). Hence, almost every discipline dealing with complex systems—particularly economics, but also sociology, biology and even physics—has included complex network analysis in its toolbox, opening new trends in these disciplines.

More specifically, this formalism is essential for every contemporary social science. Social forces, whether manifested through neighbourhoods, schools, communities or firms, impact much of what constitutes modern economic life. A non-trivial part of these interactions takes place within a social network. In a social network, agents have their own reference group that may influence their behaviour. In turn, the agents' attributes and their behaviour affect the formation and the structure of the social network.

Moreover, network structures involving economic actors have often a territorial dimension and emerge across space at different aggregation levels: at the macro-level, countries represent the nodes and their interactions—in the form of commodity and financial trade or migration flows—constitute their links. At the meso-level, economic entities in between the micro- and the macro-layers such as industries, local institutions and markets are taken as the nodes and their interactions—input-output interlinkages, partnerships and decisional processes, local exchanges of information and of goods and factors—are the links. Finally at the micro-level, single economic agents such as firms, households or even individuals are the nodes and their interactions or relationships—family and friendship ties, goods and knowledge exchanges and so on—are their links

As it has been commented, scholarly interest in social interactions has rapidly expanded in many areas of economics and has led to numerous methodological and empirical advances. For econometricians and those scholars centred in measurement, a primary challenge is the complete integration of the measurement of social interactions with the joint processes of group formation and subsequent behaviours within groups. Much of the methodological progress focused on the identification of social interaction, and, in turn, such identification relies on features of different types of data.

Accordingly, phenomena as diffusion processes, contagion and consensus have attracted the interest not only of economists but also of sociologists and social psychologists. Salient studied examples include friendship networks among adolescents, coauthorship networks among scientists and trade networks between countries but also diffusion of innovations in markets or of any kind of information (markets, rumours, etc.) As stated above, these phenomena take place in an agent-based system, so complex network analysis is the basic mathematical tool for their description. Hence, several efforts have been tributed to diffusion phenomena in complex networks that combine the statistical mechanics of complex networks themselves with the theory of dynamical systems, which describe the usually nonlinear dynamical phenomena.

Dynamical systems have a long history, but its development in the last decades has been exponential. In continuous time, via differential and partial differential equations, and discrete time, via difference equations, the applications of dynamical systems to physics, biology, economy and other branches of science have increased as well. In most of the papers focused in applications of dynamical systems, there is a dichotomy between order and complexity. On one hand, it is expected that models behave in a predictive way, but on the other hand, it is well known that this order may not appear for all the models, which often generates trajectories which are called chaotic or complex. Although this complexity has been understood for several cases, for instance, one-dimensional difference equations, we must remark that it is a branch of mathematics that presents a huge number of interesting open problems.

A Brief Summary of the Book

The book is divided into three parts. Part I gives a general introduction to complex networks looking at the main tools of analysis both from the theoretical and empirical perspectives. In chapter 'Complex Network Analysis and Nonlinear Dynamics', **Luis M. Varela and Giulia Rotundo** review complex network and nonlinear dynamical models and methods relevant for socioeconomic issues and pertinent to the theme of new economic geography starting from their foundations. Applications of the formalism to economics, finance, epidemic spreading of innovations and regional trade and developments are briefly reviewed, as well as to other relevant socioeconomic issues. In chapter 'An Overview of Diffusion in Complex Networks', **Dunia López-Pintado** assesses a series of theoretical contributions on diffusion in random networks, starting with reference epidemiological models such as the susceptible-infected-susceptible model, which describes the spread of an infectious disease in a population. The interaction structure is considered as a heterogeneous sampling process characterized by the degree distribution, and the author provides a detailed characterization of the diffusion threshold, above which the spreading of the infectious agent to a significant fraction of the population and its persistence is secured. Finally, a general diffusion model is introduced to describe the diffusion of a new product, idea, behaviour, etc. In chapter 'Opinion Dynamics on Networks', **Ugo Merlone, Davide Radi and Angelo Romano** survey some of the most recent contributions on opinion dynamics, illustrate Galam's model of rumour diffusion and extend it to consider more general networks. These networks are used to describe more complicated social spaces where agents can interact and exchange opinions about the rumour at stake even if seated at different tables of the social space. In chapter 'Econometric Aspects of Social Networks', **Mariano Matilla-García and Jesus Mur** aim to show why and how is it possible to use econometric analysis tools in the empirical study of social networks. Current technical limitations and open challenges are considered. The authors draw also connections between the recent econometric literature (identification and estimation) on networks and the literature on spatial econometrics. Finally, in chapter 'An Overview of the Measurement of Segregation: Classical Approaches and Social Network Analysis', **Antonio Rodriguez-Moral and Marc Vorsatz** present a comprehensive overview of the literature on the measurement of segregation. Particularly they review those contributions focusing on two specific dimensions of segregation, the evenness and exposure dimensions—

two of the five dimensions of segregation in the multidimensional framework defined by Massey and Denton (The dimensions of residential segregation. Soc Forces 67(2):281–315, 1988). Some of the most relevant segregation measures, developed within the classical statistical approach and within the framework of social network analysis, are also discussed.

Part II deals with applications of complex network analysis to theoretical and empirical issues in economics. In chapter ‘An Investigation of Interregional Trade Network Structures’, **Roberto Basile, Pasquale Commendatore, Luca De Benedictis and Ingrid Kubin** investigate the interregional trade network structures linking the European regions at the NUTS-2 level. Using social network analysis, the authors analyse the regional data and identify the topological properties of the European Regional Network deriving its main local and global centrality measures. Focusing on the local structures of the network, the authors apply the triadic census approach detecting the frequency at which specific patterns of trade emerge. The unit of analysis is a triad, i.e. a group of three regions and their possible trade relationships. The empirical analysis is complemented by a three-region new economic geography model. This model tries to clarify, from the theoretical point of view, the relationship between trade costs, market competition and the patterns of trade in a three-region context identified by the triadic census analysis. In chapter ‘The Empirics of Macroeconomic Networks: A Critical Review’, **Giorgio Fagiolo** deals with the recent empirical literature applying complex network analysis to the analysis of macroeconomic systems. The author focuses on three macroeconomic networks describing interactions between world countries: international trade, finance and migration or mobility. He examines the empirical evidence concerning the topological properties of such networks. Moreover, he considers the implications of such properties on the dynamics of countries performance and shock diffusion via a comparative analysis of the different properties of the three types of network under study. In chapter ‘Bank Insolvencies, Priority Claims and Systemic Risk’, **Spiros Bougheas and Alan Kirman** review the interdisciplinary literature—spurred by the 2008 global financial crisis—interested in the design of bankruptcy resolution procedure and priority rules for banks. The authors assess various arguments put forward by both economists and financial law scholars focusing on the allocation of priority rights among the bank creditors. It is shown that, within a deeply connected interbank network, this choice has substantial consequence on the total loss of the economy due to existence of systemic risk inherent in deeply connected interbank system. Finally, in chapter ‘Complex Networks in Finance’, **Anna Maria D’Arcangelis and Giulia Rotundo** survey the literature that applies the complex network approach to the analysis of financial data dealing with correlation matrix and systemic risk. These authors also explore the complex network structure of integrated ownership and control of corporations and interbank networks. Finally, they contribute to the literature that applies complex network analysis in the area of investments and managed portfolios, providing new results concerning the ownership distribution of stockholdings through mutual funds across European countries.

Part III, which concludes the book, is dedicated to the theory of dynamical systems and networks dynamics and their applications to economic issues. Chapter 'A Formal Setting for Network Dynamics' by **Ian Stewart** is an introduction to coupled cell networks, a formal setting in which to analyse general features of dynamical systems that are coupled together in a network. Such networks are common in many areas of application. The nodes ('cells') of the network represent system variables, and directed edges ('arrows') represent how variables influence each other. Cells and arrows are assigned types, which determine the form of admissible differential equations—i.e. those compatible with the network structure. In chapter 'Dynamics on Large Sets and Its Applications to Oligopoly Dynamics' **Jose S. Cánovas and Maria Muñoz Guillermo** consider discrete dynamical systems with a high-dimensional phase space. The authors explain some techniques which allow them to apply dynamical systems analysis to oligopoly models. Chapter 'Attracting Complex Networks' by **Giovanny Guerrero, Jose A. Langa and Antonio Suarez** focuses on the concept of global attractor, which is crucial when the dynamics on complex networks is considered. In the last decades, there has been an intensive research in the geometrical characterization of global attractors. However, there still exists a weak connection between the asymptotic dynamics of a complex network and the structure of associated global attractors. Finally, in chapter 'Good Old Economic Geography', **Tönu Puu** discusses classical economic modelling in continuous two-dimensional geographical space. This author examines some ingenious models due to Harold Hotelling and Martin Beckmann concerning population growth and migration and spatial market equilibrium, respectively. He also introduces some models of business cycles and discusses the issue of the shape of market areas, focusing on transversality and stability of structures

Acknowledgements First of all we would like to acknowledge the help of the scientific and local organization committees and of the speakers and students participating in the training school on complex networks and dynamics held in Madrid on 12–14 February 2014 at UNED and promoted by the COST Action IS1104 on 'The EU in the new complex geography of economic systems: models, tools and policy evaluation'. Especially, the venue provided by UNED has greatly contributed to the creation of an atmosphere of deep and fruitful discussion.

Our special thanks go to the external anonymous referees for their careful work which helped greatly to improve the quality of the chapters.

We are very grateful to the editors of the series 'Lecture Notes in Economics and Mathematical Systems' and especially to Walter Trockel and Herbert Dawid. Finally, we would like to give our appreciation to Katharina Wetzel-Vandai, Ruth Milewski and Sivachandran Ravanan for their patience and constructive presence during the editorial process.

Part I
Complex Networks

Complex Network Analysis and Nonlinear Dynamics

Luis M. Varela and Giulia Rotundo

Abstract This chapter aims at reviewing complex network and nonlinear dynamical models and methods that were either developed for or applied to socioeconomic issues, and pertinent to the theme of New Economic Geography. After an introduction to the foundations of the field of complex networks, the present summary introduces some applications of complex networks to economics, finance, epidemic spreading of innovations, and regional trade and developments. The chapter also reviews results involving applications of complex networks to other relevant socioeconomic issues.

Keywords Complex networks • Computer simulations • Market and financial models • Regional trade and development • Social networks • Statistical mechanics

1 Introduction

Complex networks have been proved during the latest decades as an essential formalism in order to describe the many situations in which agent-based models are required (Albert and Barabási 2002; Newman 2003; Pastor-Satorras et al. 2003; Boccaletti et al. 2006; Newman et al. 2006). Starting with the papers of Watts and Strogatz (1998) and of Barabási and Albert (1999) in the late nineties, this formalism has been accepted in the basic tool set of a growing number of disciplines like physics, biology, computer science, sociology, epidemiology, and economics among

L.M. Varela (✉)
Grupo de Nanomateriais, Fotónica e Materia Branda, Departamento de Física da Materia Condensada, Universidade de Santiago de Compostela, Campus Vida s/n. E-15782, Santiago de Compostela, Spain
e-mail: luismiguel.varela@usc.es

G. Rotundo
Department of Methods and Models for Economics, Territory and Finance, La Sapienza University of Rome, via del Castro Laurenziano 9, I-00161 Rome, Italy
e-mail: giulia.rotundo@gmail.com

P. Commendatore et al. (eds.), *Complex Networks and Dynamics*, Lecture Notes in Economics and Mathematical Systems 683,
DOI 10.1007/978-3-319-40803-3_1

others. This is so, because the skeleton of any complex system can be described by a network arising from the interactions of the many parts of which it is made up of. Hence, network theory is now considered to be one of the pillars in the theory of complex systems.

Complex system theory investigates how relationships between parts give rise to the collective (emergent) behaviors of a system and how the system interacts and forms relationships with its environment. The definition of complex systems is not unanimous, but one of the most popular visions states that a complex system is one made up of many parts with nonlinear interaction between them. These systems are open, nested systems, with memory and feedback loops, which may show emergent patterns and are structurally founded in dynamic networks of multiplicity. This definition emphasizes the importance of both the network of the system constituents and the nonlinear coupling between them. Therefore, one should expect that the relevant dynamics of complex systems is inherently nonlinear, and so that a combined study of complex networks and dynamic systems is a need for the development of the field of complex systems. Undoubtedly, structure decisively conditions the behavior of a dynamic system, and hence the interactions among the massive amount of their individual parts must be collectively considered, which is the main role of the network analysis. Indeed, the relationship between the dynamics and the topology of complex networks is a nowadays a central issue of network theory and, therefore, in its applications (see, e.g. Wang 2002 and Boccaletti et al. 2006 and references therein). Networks of coupled dynamical systems have been extensively considered in nonlinear dynamics since they can exhibit non trivial emergent phenomena, such as autowaves, Turing patterns, spiral waves, and spatio-temporal chaos, with applications to a plethora of large-scale real systems (Chua 1998). Specifically, synchronization phenomena have been extensively considered, as well as the evolution of the networks themselves (rewiring, growth, and so on) (Boccaletti et al. 2006).

In the present chapter, we briefly review the fundamental structural and dynamic properties of complex networks, and we consider some applications of this formalism to the understanding of nonlinear phenomena, including epidemiological models, synchronization in networks and financial contagions, games on networks, and clustering of network nodes and geographical information. Moreover, we present some novel results for epidemiological models of flu and sexually transmitted diseases (STDs), and we review the mean-field theory of contagion in networks.

The chapter is divided in four sections. The first one is this introduction, the second contains a review of the theory of complex networks; in Sect. 3 we present some applications of this theory to the analysis of nonlinear phenomena, including some novel results for flu and STDs and, finally, Sect. 4 summarizes the main conclusions of the chapter.

2 Theory of Complex Networks

In this section we briefly review some of the main results in the theory of complex networks developed during the last two decades. The theory of complex networks has evolved from the branch of mathematics known as graph theory, within which these objects were traditionally considered. Probably the best known of all of them—besides the regular graph or lattice—was Erdös and Renyi's random graph (Erdös and Rényi 1959).

Probably the most significant result in the contemporary theory of networks was the introduction of small-world networks by Watts and Strogatz (1998) (WS), trying to describe the transition from a regular lattice to a random graph. Beyond its mathematical interest, this new type of network successfully gives account of the "six degrees of separation" concept, uncovered by the social psychologist Milgram in the late 1960s (Milgram 1967). This is a well-known emergent property of many social networks in which the average path length (i.e. the average number of steps one has to take across the network for connecting two given individuals) is notably smaller than that in a regular lattice. Purely random networks also exhibit the small-world property, but, as we shall see, they give rise to far sparsely connected networks (low clusterization) to accurately describe social networks.

The second major advance in the theory of complex networks took place the year after Watts and Strogatz's paper in Nature, when Barabási and Albert (1999) introduced scale-free networks, which mimic many situations where the networks show vertex connectivities that follow a scale-free power-law distribution. They were able to prove that this emergent property followed from the continuous expansion of the network due to the addition of new vertices, and the preferential attachment of the new nodes to those sites that are already well connected. Of course, this new type of network showed the small-world property, but, in contrast with WS networks, they were able to describe situations where a high degree of heterogeneity between the agents' connectivity exists. This new type of network completed the network ecosystem, and the theory was then in place to describe any possible connection pattern in agent-based systems, which opened the door to the explosive expansion of the theory of complex networks themselves and their applications. In what follows we summarize some of the most important results reported since then for the structure and dynamics of these fascinating objects.

A network (graph) is formally described as a pair (V, E), where V is a set of nodes (vertices), and E is a set of links (edges) defined by two nodes that represent the source and the end of the link Suppress. The assignment of labels to the elements of the network is called graph labelling, and the special case of coloring is among the most important labelling schemes. A graph is said to be k-colorable if it can be assigned a k-colouring, i.e. a labelling of the graph's vertices with colors such that no two vertices sharing an edge have the same color. The smallest number of colors needed to color a graph G is called its chromatic number, and is often denoted as $\chi(G)$. Associated to this concept is that of chromatic polynomial, which counts the number of ways a graph can be coloured using no more than a given number of colors.

Networks can be classified according to several criteria:

1. According to the directionality of the links, they can be either directed or undirected graphs.
2. Networks can be classified as weighted or unweighted depending on whether different weights are associated to their edges.
3. Sparse and fully connected networks differ in the fraction of interconnected nodes.
4. Depending on their time evolution, networks can be classified into static and evolving.
5. Depending on their node degree probability distribution, they can be classified as: regular (deterministic), exponentially distributed Watts-Strogatz networks, scale-free Albert-Barabási network, fully random graphs, general uncorrelated networks and others.

Another important type of networks are bipartite networks, formed by two disjoint sets of nodes U and V such that every node in U is connected to a node in V. The adjacency matrix of a bipartite network is of the form:

$$A_{ij} = \begin{pmatrix} 0 & B \\ B & 0 \end{pmatrix} \tag{1}$$

Some important properties of these networks are:

(a) A graph is bipartite if and only if it does not contain an odd cycle.
(b) A graph is bipartite if and only if its chromatic number is less than or equal to 2.
(c) The spectrum of a graph is symmetric if and only if it's a bipartite graph.

The most frequently employed method for the representation of the vertex connectivity in networks with N nodes is the *adjacency matrix* $A = (a_{ij}) \in \mathbb{R}^{N \times N}$, whose rows and columns are the nodes of the network, and whose term $a_{ij} > 0$ corresponds to the weight of the link from node i to node j, which in an unweighted network is always set to $a_{ij} = 1$ for the non null elements (if the network is directional, other criteria have been introduced). In this case, and when the network is undirected -the links connecting two edges acts bidirectionally- the matrix A is symmetric $a_{ji} = a_{ji}$ for all nodes i, j. The absence of links is given by zero elements $a_{ij} = 0$. This matrix contains much useful information about the network including the so called spectrum of the graph, given by the set of eigenvalues of A_{ij}. On the other hand, a path from node i_{h_1} to node i_{h_k} in the network is a sequence of unique nodes $i_{h_1}, i_{h_2}, \cdots, i_{h_k}$ such that $a_{i_{h_j}, i_{h_j}+1} \neq 0$, and can be detected through the power of the adjacency matrix A^{h_k}. These paths are relevant for studying diffusion processes as well as the relevance of nodes.

This matrix allows the calculation of others related to it such as the degree matrix,

$$D_{i,j} := \begin{cases} \deg(v_i) & \text{if} \quad i = j \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

or the Laplacian matrix, $L_{ij} = D_{ij} - A_{ij}$ or

$$L_{i,j} := \begin{cases} \deg(v_i) & \text{if } i = j \\ -1 & \text{if } i \neq j \text{ and } v_i \text{ is adjacent to } v_j \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Spectral graph theory is the study of the properties of a graph by means of the characteristic polynomial, eigenvalues, and eigenvectors of its adjacency matrix or Laplacian matrix. Some interesting results are: (1) zero is always an eigenvalue of this matriz since the sum of all the elements of every row and column is zero. Indeed, the multiplicity of this eigenvalue gives the number of connected components of the graph. (2) As usual, the smallest non-zero eigenvalue of the Laplacian matrix gives the spectral gap, and its second smallest eigenvalue is the algebraic connectivity (or Fiedler value) of the graph, which measures the connectivity degree of the graph, with applications in robustness and synchronizability of networks.

2.1 Structure of Complex Networks

The analysis of the structure of networks is made using a set of relevant parameters that allow the classification of these objects. In the rest of this subsection we introduce some of the most important and frequently used (Fig. 1).

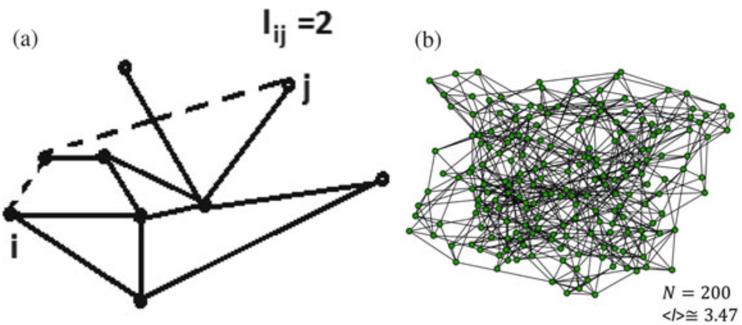


Fig. 1 (**a**) Example of chemical distance. (**b**) Average path length in a SW network with $N = 200$ and $p = 0.3$ and the number of connections to nearest neighbours in ring topology $k = 6$

- **Small-worlds**: relatively short path between any two nodes, defined as the number of edges along the shortest path connecting them. The connectedness can also be measured by means of the diameter of the graph, d, defined as the maximum distance between any pair of its nodes. Networks do not support the definition of a "distance". They are no proper metric space although some attempts to introduce a metric on hidden metric space have been made Serrano et al. (2008). Hence, a chemical distance between two vertices l_{ij} must be introduced, which is defined by the number of steps from one vertex to the other following the shortest path. Associated to this distance emerges the one of the most important measures for describing the structure of a complex network is the so called average path length between connected nodes

$$< l >= \frac{2}{N(N-1)} \sum_{i<j} l_{ij} = \sum_{l} p(l) \tag{4}$$

  Contrary to what happens in a regular lattice -in which $< l >\sim \sqrt{N}$- in most real networks, $< l >$ is a very small quantity (small-world) that scales with the number of nodes as $< l >\sim \log N$.
- **Centrality** Centrality measures the relative importance of the vertices of a network in terms of a real-valued function, where the values produced are expected to provide a ranking which identifies the most important node. Obviously, importance is highly dependent on the context and a criteria of importance must be provided. Several types of centrality have been introduced:
  - Degree centrality: number of connections a node has, $C_D(v) = \deg(v)$. This is a direct measure of how exposed a given vertex v is to whichever it flows through the network.
  - Closeness centrality, defined for vertex i as

$$\tilde{C}_i = \frac{1}{\sum_j d_{ij}} \tag{5}$$

    where d_{ij} is the distance of node i to node j.
  - Path centrality: To go from one vertex to another in the network, following the shortest path, a series of other vertices and edges are visited. The ones visited more frequently will be more central in the network.
  - Betweenness centrality, number of shortest paths that pass through a given node for all the possible paths between two nodes. Measures the "importance" of a node in a network, and it is given by

$$C_B(v) = \frac{\text{Number of shortest paths including } v}{\text{Total number of shortest paths}} = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{6}$$

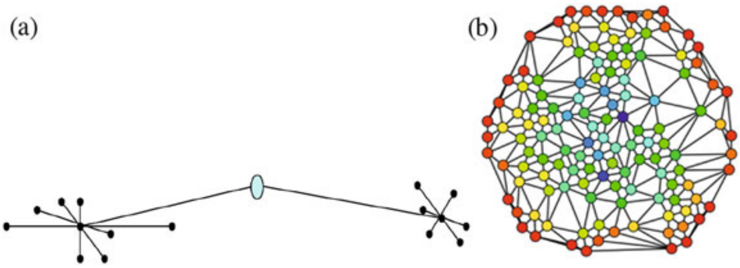


Fig. 2 (**a**) The node in *blue* has the largest betweenness centrality, as does the node in *navy blue* in (**b**) (Source of (**b**): Claudio Rocchini, Wikipedia)

– Eigencentrality: a measure of the influence of a node in a network (Fig. 2).
– Katz centrality: Katz centrality is a generalization of degree centrality that measures the number of all nodes that can be connected through a path, penalizing the contributions of distant nodes.

$$C_{\mathrm{Katz}}(i) = \sum_{k=1}^{\infty}\sum_{j=1}^{n} \alpha^k (A^k)_{ji} \tag{7}$$

where $0 < \alpha < 1$ is an attenuation factor. This centrality measure is related to the very famous PageRank,

$$P_i = \alpha \sum_j A_{ji} \frac{\tilde{C}_j}{L(j)} + \frac{1-\alpha}{N} \tag{8}$$

with $L(i) = \sum_j a_{ij}$ is the number of neighbors of node i, used by Google to rank websites in their search engine results.
– Bonacich centrality, which allows for negative values of the attenuation factor β.

$$C_{BON}(i) = e_i^T \left(\frac{1}{\beta}\sum_{k=1}^{\infty}(\beta A)^k\right)\mathbf{1} = \frac{1}{\beta}\sum_{k=1}^{\infty}\sum_{j=1}^{n} \beta^k (A^k)_{ij} \tag{9}$$

Closely related to those centrality coefficients is the so called core-periphery coefficient, introduced by Holme (2005). This core-periphery coefficient measure if the network is divided in a core that is both distance, and densely connected and central in terms of graph distance and a sparsely connected periphery, and it is defined as

$$c_{cp}(G) = \frac{C[V_{core}(G)]}{C[V(G)]} - \left\langle \frac{C[V_{core}(G')]}{C[V(G')} \right\rangle_{G\in\mathcal{G}(G)} \tag{10}$$

where $\mathscr{G}(G)$ is the ensemble of graphs with the same set of degrees as G. and $V_{core}(G)$ is the k core (i.e. a maximal subgraph with the minimum degree k) with maximal closeness.

- Clustering coefficient: This coefficient provides a quantitative measure of the degree to which links in a network follow a transitive property (i.e. if i is linked to j, and j is linked to k, then i may or may not be connected to k). Such transitive links are very common in social networks, and realistic structures must take this into account. The local clustering coefficient of node i of a set of k_i nodes linked to node i is defined as the ratio between the number E_i of edges that actually exist between these k_i nodes and the total number $C_i = k_i(k_i - 1)/2$. This clustering coefficient provides a measure of the local connectivity structure of the network, i.e. how close the local neighbors of a given node are to being a clique (complete graph). This coefficient was introduced by Watts and Strogatz (1998) to determine whether a graph is a small-world network. The global clustering coefficient is defined as the ratio between the total number of closed triplets and the total number of connected triplets of networks (Fig. 3). The average clustering coefficient

$$< C >= \frac{1}{N} \sum_j C_j \tag{11}$$

can be introduced, as well as the clustering spectrum for vertices of degree k,

$$< C(k) >= \frac{1}{Np(k)} \sum_j \delta_{k_j k} C_j \tag{12}$$

in the usual way.

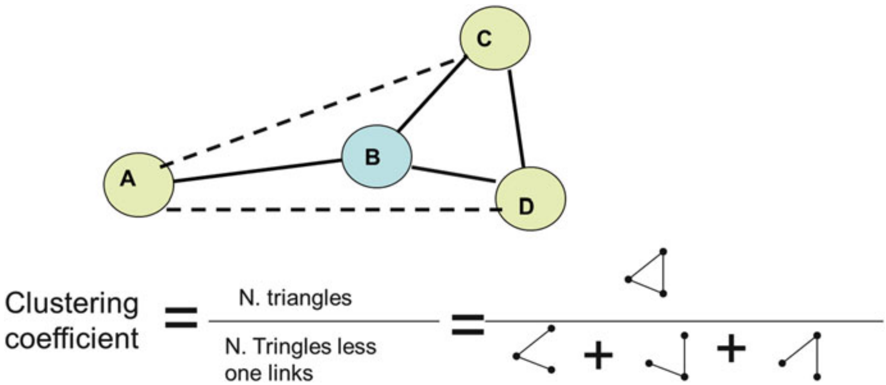


Fig. 3 Calculation of the global clustering coefficient

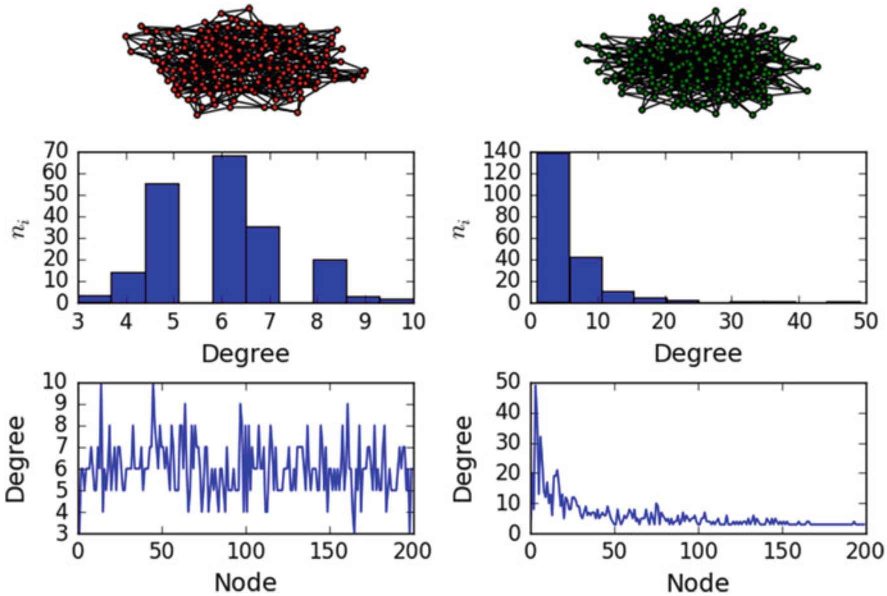


Fig. 4 WS distribution (*left column*) and AB distribution (*right column*)

- Degree distribution: The degree distribution, $p(k_i)$ is the probability that a given node i has a definite number of edges, k_i. In directed networks an in-degree and an out-degree are to be defined, but in undirected network both coincide. These degree distribution can fall in two categories: those within the basin of attraction of the Gaussian distribution (exponential distribution with all their moments finite) and those in the basin of attraction of some other Lévy distribution (subexponential, fat-tailed distributions). The degree distributions associated to the main four types of complex networks introduced above are (Fig. 4 shows those for WS and AB networks):

$$p(k) \sim \begin{cases} \delta(k-k_0) & \text{regular lattice} \\ \binom{N}{k} p^k (1-p)^{N-k} & \text{random network} \\ e^{-\gamma k} & \text{WS networks} \\ k^{-\alpha} & \text{AB networks} \end{cases} \tag{13}$$

Of course, these distributions open the doors to defining the whole set of moments of the degree of nodes. Specifically, we can mention the average degree

$$<k>= \frac{1}{N} \sum_i k_i = \sum_{k'} k' p(k') \tag{14}$$

A network is called sparse if its average degree remains finite when taking the limit $N \gg$. In real (finite) networks, $< k > \ll N$. Of course, real networks are usually correlated, i.e. the degrees of the nodes are not in general independent. This must be described by a conditional probability distribution $p(k'|k)$ which represents the probability of a k-node pointing to a k-node. For an uncorrelated network,

$$p(k'|k) = \frac{k'}{< k >} p(k') \tag{15}$$

independent of k, while for a correlated network this probability depends on the degree of both linked nodes. $p(k)$ and $p(k'|k)$ are not independent, but they are related by a detailed balance condition arising form the fact that the number of nodes of degree k linked to nodes of degree k' equals the number of nodes of degree k' linked to nodes of degree k. On the other hand, in terms of these conditional probability the average degree of the nearest neighbours of vertices of degree k can be calculated as

$$< k_{nn} >= \sum_{k'} k' p(k|k) \tag{16}$$

If $< k_{nn} >$ is an increasing (decreasing) function of the degree k the mixing is assortative (disassortative) meaning that highly connected nodes tend to be linked to other highly (lowly) connected nodes, i.e. there exists a bias in favor of connections between network nodes with similar (dissimilar) characteristics (Pastor-Satorras et al. 2003).

- Nestedness index (Araujo et al. 2010). This concept was originally introduced in the context of ecological studies (Atmar and Patterson 1993). A bipartite network formed by islands and species—linked if the former inhabits the latter—is said to be nested if the species that exist on a few islands tend always to be found also on those islands inhabited by many different species. It indicates the likelihood of a node being linked to the neighbours of the nodes with larger degrees. The mean topological overlap between nodes (Almeida-Neto et al. 2008) has been introduced to quantify nestedness.

3 Dynamics of and on Complex Networks

We have previously mentioned that complex networks are at the root of the structure and dynamics of complex systems. As usual -and in this case probably to a much larger extent than in simple conventional systems- there exists a strong interplay between structure and dynamics, so it is difficult to say if the observed equilibrium structure is a cause or a consequence of the dynamics. Without entering into such philosophical questions, there are different dynamic processes that one can consider in complex networks:

– Synchronization of networks and collective dynamics.

– Evolution of the network itself (evolving networks, rewiring).
– Spreading processes (epidemics, rumors).

Closely associated to dynamics is the concept of robustness, which can be defined as the ability of the network to withstand failures and perturbations. These concepts have been extensively reviewed in the past (see the reviews of Boccaletti et al. 2006 and of Arenas et al. and the references therein Arenas et al. 2008) and the reader is referred to this literature for in-depths treatments.

4 Nonlinear Phenomena on Complex Networks

This section shows some applications of nonlinear dynamics on complex networks for the study of economic and financial problems. The need for new and more fundamental understanding of economic and financial crises has fostered the development of studies on the structure and dynamics of economic and financial networks. In such networks, nodes may represent firms, banks, or even countries, and links represent their mutual interaction, be it cross-ownership, credit-debt, or trade relationships. The network structure allows to focus on the propagation of shocks and on resilience, which show quite different behavior depending on the structure of the network (da Fontoura Costa et al. 2011; Schweitzer et al. 2009a). This section is organized in four subsections: Epidemiological models, Synchronization in networks and financial contagions, Games on networks, and Clustering of network nodes and geographical information.

4.1 Epidemiological Models

Understanding risk, unfolding different sources of risk, preventing failure on large systems, spreading of crises, and large crashes, are among the major tasks to undertake in modern economic and financial literature. Conventional epidemiological models have been used as models of propagation and contagion well beyond the limits of medical epidemiology. They have indeed served as models of any kind of diffusion in social and economic contexts, where large populations with a complex pattern of interaction among agents exist (spreading of rumours or information, diffusion in markets and financial markets, etc.). From this perspective, epidemiological models have also been reconsidered in terms of transmission of financial shocks. For instance, the average lifetime and persistence of viral strains on the Internet is examined in Pastor-Satorras and Vespignani (2001). It so happens that the spreading pattern depends crucially on the properties of the social network and on the diffusion (or contagion) mechanism underlying the complex network (López-Pintado 2008; Barrat et al. 2008). A key factor is the scale-free feature of the network, and a relevant result is the absence of a critical threshold for the spread of computer viruses. That means that any virus is going to propagate—unless properly blocked (Pastor-Satorras and Vespignani 2002). In Barthélemy et al. (2005) simulation studies proved that assortative networks have an early burst, against the late spread in disassortative networks. This also gives recommendations on vaccination policies: blocking the infection at the level of hubs is the optimal choice.

Compartmental models are a special kind of epidemic models that split the population over a discrete set of states (or compartments) depending on their health status with respect to the disease. Individuals can change state as time goes by. Probably the best known of all these models is that by Kermack and McKendrick (1927), a cornerstone in the mathematical modelling of epidemics introduced in order to understand the evolution of infected patients observed in epidemics such as the London (1665–1666) or Bombay (1906) plagues or cholera (London 1865). This deterministic model uses three compartments corresponding to Susceptible (S), Infected (I) and Recovered (R) for classifying the population, whose size is supposed to be fixed (i.e., no births, deaths due to disease, or deaths by natural causes). No incubation period is allowed so the disease is passed instantaneously and a final immunity is acquired by the individuals. The mechanism that the pathogen uses for its propagation is the direct contact between the susceptible and the infected individuals, assuming homogeneous mixing of a completely homogeneous population (no age, spatial, or social structure). This hypothesis states that the per capita rate of acquisition of the disease by the susceptible individuals is directly proportional to the density of infective individuals, and it is a typical mean-field assumption. These assumptions leads to the following set of equations (Kermack and McKendrick 1927):

$$\begin{aligned}\frac{dS(t)}{dt} &= -\beta S(t)I(t)\\ \frac{dI(t)}{dt} &= \beta S(t)I(t) - \gamma I(t)\\ \frac{dR(t)}{dt} &= \gamma I(t)\end{aligned} \tag{17}$$

These compartmental models focus on a specific set of characteristics, and divide network nodes in subsets (*compartments*) that are homogeneous with respect to the specific characteristic. For instance, in a vaccination program, the study of the possible propagation of the infection can be done in accord to the different age ranges. In such context, it is natural to draw network links among the compartments. The relevance of the structure of the network has also been shown on the resilience: assortative networks resist random attacks better than others, but not to targeted attacks (Albert et al. 2000; Chen and Cheng 2015).

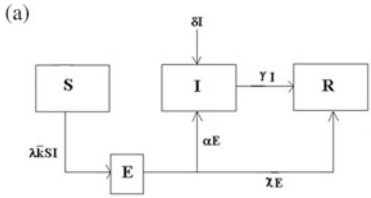
Underlying the epidemic spreading there is always a complex network defined by the contacts of the individuals that support the contagion. The topology of this network drastically determines the propagation of the disease. Essentially, in homogeneous exponentially distributed networks the number of contacts of a given individual can be replaced under certain circumstances by its average value, neglecting fluctuations in the number of contacts (mean-field theory). This is not possible in epidemic diseases where fluctuations in the number of contacts of the individuals are important (e.g. sexually transmitted diseases), but is relatively accurate in diseases like flu. For diseases travelling in homogeneous networks the infection rate β—which controls the epidemic dynamics together with the recovery rate, γ—is given by the probability of infection per contact, λ times the average number of contacts of a given characteristic individual, $\beta = \lambda < k >$. This is, of course, not possible in the case of heterogeneous scale-free networks, in which the average number of contacts is not a good representative of the distribution, or fluctuations are too important to define an "average individual".

One of the main predictions of the SIR model is the existence of an epidemiological threshold, i.e. a cut-off for the infection rate below which the epidemic disease is not capable of infecting a significant fraction of the population and dies out. In the specific case of the Kermack-McKendrick SIR model this is given by the basic reproductive number

$$R_0 = \frac{\beta S}{\gamma} \tag{18}$$

This variable quantifies the number of secondary infections coming from a primary case in a completely susceptible population, and $R_0 = 1$ is the critical value that divides the region where the outbreak peters out ($R_0 < 1, \dot{I} < 0$) from that where the epidemic spreads ($R_0 > 1, \dot{I} > 0$). The specific value $R_0 = 1$ corresponds to an endemic behaviour, with a finite fraction of the population infected in the asymptotic limit, $t \to \infty$.

The description of flu epidemic spreading demands an epidemiological model of the different states the individual goes through during the infection period. Moreover, the network structure must be provided, and since in this case the effect of hubs is negligible, it may be reduced to a mere number, the average connectivity of the network, $\langle k \rangle$, which amounts to a mean-field level of description. Furthermore, a good description of the epidemiology of the disease must be used. In our case, in order to describe the propagation of an epidemic of avian flu in Galicia (Spain) we proposed a compartmental model based on the well-known SEIR model, including an incubation period of exposed individuals $E(t)$ and the possibility that a fraction of those exposed to the virus recover without being effectively infective at any point of the process (see Fig. 5a). Moreover, we allowed for the seed of a fraction of infective individual from outside the region at a rate δ. This is a deterministic model since the factors intervening in the process are controllable and a deterministic differential equation system governs the whole process (Fig. 5b). However, the most natural way to describe an epidemic disease is by means of stochastic models, which describe the process of propagation of the infective agent in a probabilistic way. These models are used when random fluctuations or heterogeneities of the system are important, such as in small or isolated populations. They have several



(b)

$$\frac{dS(t)}{dt} = -\lambda \bar{k} S(t) I(t)$$
$$\frac{dE(t)}{dt} = \lambda \bar{k} S(t) I(t) - \alpha E(t) - \chi E(t)$$
$$\frac{dI(t)}{dt} = \alpha E(t) - \gamma I(t) + \delta I(t)$$
$$\frac{dR(t)}{dt} = \chi E(t) + \gamma I(t)$$

Fig. 5 Deterministic modified SEIR compartmental model (**a**) and its associated system of nonlinear differential equations (**b**)

advantages over the deterministic model, since they study the dynamic of the illness at an individual level instead of on a collective fashion. However, they normally demand numerical simulations. Deterministic models are approximations of the corresponding stochastic models (Nåsell 2002) where the latter are Markov population processes with continuous time and discrete state space, and the former take the form of ordinary differential equations.

The propagation of epidemic diseases on scale-free networks has been studied by Pastor-Satorras and Vespignani (2001), who proved the absence of an epidemic threshold and its associated critical behavior. This is the case when there is a large heterogeneity in the degree of contacts of the individuals, as in the case of sexually transmitted diseases.

The analogy of financial contagion with epidemiological models is well evidenced by attention-capturing titles as "When Belgium sneezes, the world catches a cold",[1] "Why Does The U.S. Sneeze When Europe Gets A Cold?",[2] that echo the sentence of Metternich "When France sneezes, Europe catches a cold".[3] Such sentence was pointing out the relevance of France. Nowadays, such studies point out that the interlinking between national economies is so high that even small countries may trigger a crisis.[4]

The difficulty of a straightforward connection between the structure of the economic network and the propagation of shocks, attacks and crises, resides in the fact that they are weighted fully connected networks (complete graphs). Therefore, the crux of the issue is keeping only the most relevant links. Networks examined in Garas et al. (2010) are the Corporate Ownership Network (CON) and the International Trade Network (ITN). Such networks are nearly fully connected, so thresholds are considered for the detection of non trivial structure (Fagiolo et al. 2007). To identify the uneven roles of different countries in the global economic network, the k-shell decomposition method is used.

[1] http://phys.org/news/2010-11-belgium-world-cold.html.

[2] http://www.npr.org/2011/08/18/139714695/why-does-the-u-s-sneeze-when-europe-gets-a-cold.

[3] http://www.enotes.com/homework-help/metternich-said-when-france-sneezes-europe-catches-120443.

[4] "The Worlds Most Contagious Countries" https://twistedeconotwist.wordpress.com/2010/11/27/the-worlds-most-contagious-countries-heres-the-list/.

In graph theory, the concept of k-shell is related to the colouring number of a graph. It has been applied in many areas, including social networks (Li et al. 2013). By definition, the k-shell of a graph G is a maximal connected subgraph of G in which all vertices have degree at least k. Equivalently, it is one of the connected components of the subgraph of G formed by repeatedly deleting all vertices of degree less than k. The k-core is the k-shell corresponding to the highest k. A nucleus of twelve countries has been found to be quite dangerous in the spread of economic crises. Only six of them are large economies, while the other are medium or small economies. Therefore, the position in the network is quite relevant: while it is clear that large economies and big financial institutions affect the entire network, it turns out that the position in the network makes the potential failure of some small-medium economies and institutions could be as dangerous as the failure of the big ones. Such remarks are in line with the paper "Too interconnected to fail" (Markose et al. 2012), that can be grouped in the same discussion line as the papers "Too central to fail" (Battiston et al. 2012), and "Too big to fail" (White 2014).[5] All such papers consider financial networks. The specific characteristics of Credit Default Swaps are examined in Markose et al. (2012). In Battiston et al. (2012) several perspectives contribute to introduce the *DebtRank*, a centrality measure that states the level of risk of financial institutions. In White (2014) it is shown that the potential loss incurred by saving financial institutions is lower than that of allowing them to fail, due to the nonlinear cascade effects on the network.

Usually, the spread of the economic crisis through a global economic network is examined through the construction of specific models. In Garas et al. (2010) a Susceptible-Infected-Recovered epidemic model is built: the probability of infection depends on the strength of economic relations between the pair of countries, and its strength on the target country. This allows to introduce a ranking of countries in accord to their crises spreading power, that can be used as a further centrality measure. In Chen and Cheng (2015) the credit-debt relation among each couple of financial institutions is specifically modeled.

An interesting phenomenon is the fact that as the average number of connections of each node increases, the probability of default first increases and then drops (Chen and Ghate 2011; Elliott et al. 2014). In fact, the set up of financial links connects the network initially, allowing shocks to propagate; but, as the number of links increases further, organizations are better shielded against other's failures. Beside the specific models for the banks, financial ties among companies are well modeled through the cross-shareholding network. In this context, the integration of a company is the percentage of its shares sold to other companies. It plays a role different from the number of connections: increased dependence on other organizations lowers the sensitivity to own investments (Elliott et al. 2014).

[5] Such ideas and concepts became popular through the book A.R. Sorkin (2009) "Too Big to Fail: Inside the Battle to Save Wall Street", Viking Press; and through the movie "Too big to fail" (2011) HBO movie.

4.2 Synchronization in Networks and Financial Contagions

The theoretical approach to phase synchronization that was used by Huygens in the seventeenth century (Bennett et al. 2002) was only a first step for the study of complete synchronization of coupled chaotic systems (Wu 2007). In this form of synchronization, the state variable of individuals converge towards each other's value. In the last decades, the research of coupled systems started with Fujisaka and Yamada (1983) and was followed by the research on the synchronization on coupled chaotic systems (Pecora and Carroll 2015). Such research lines were applied to the modelization of many natural phenomena. Synchronization has been shown to have potentially catastrophic effects in constructions (for instance: bridge resonance), earthquakes, economics and finance (Arenas et al. 2008). By way of example, the synchronization of many market agents on short position lead to an increase of supply not well balanced by demand, so triggering price drawdowns and, eventually, large financial crashes (Kaizoji 2009; Malevergne and Sornette 2001; Sieczka et al. 2011; Sornette and Andersen 2002). In behavioral finance, herding is addressed as the main cause of synchronization. In fact, herding behavior shows aggregation and imitation. When it concerns financial market investors, it may lead to the same position either buying, thus causing high rises in prices, or selling, thus causing drawdowns and even crashes in the case of Harmon et al. (2011). However, the reaction on buy or sell are far from being symmetric. Another behavioral finance concept, the so called *blindness to small changes*, describes the under-reaction to positive or small news, and the over-reaction to negative or quite alarming news. The coupling of this concept with earthquake models has been used for describing the interaction among worldwide markets, and for showing the propagation of market value cascades and avalanches (Andersen et al. 2011; Bellenzier et al. 2016).

The theme of avalanches has been widely studied also in the Self-Organized Criticality (SOC) models. A key example is the Bak and Sneppen (BS) model. The model was developed for describing the co-evolution of species. in that context the disappearance of the less fit species also causes changes in the fitness value of the species that are directly connected to it. The model shows a critical threshold for avalanches. The distribution of avalanches follows a power-law distribution. This model can be seen as a first approximation to the interaction among market agents for the replication of stylized facts (Rotundo and Ausloos 2007; Petroni et al. 2007). While the original BS model was developed on lattices, several extensions have been studied using methods from nonequilibrium statistical mechanics (Lee and Lee 2009; Moreno and Vazquez 2002; Ausloos and Petroni 2014; Rotundo and Scozzari 2009). As expected, the time span and number of network nodes involved in each avalanche depend on the structure of the network. In the framework of systemic risk and financial fragility, it is straightforward to move from the concept of less fit species to the idea of less fit firm, and to generalize from SOC to a general framework for models of cascade and contagion processes on networks (Lorenz et al. 2009; Rotundo and Scozzari 2009).

4.3 Games on Networks

A paradigmatic model in game theory is the Prisoner's Dilemma (PD): two prisoners (players of the game) have to decide wether to remain silent (cooperate) or betray the other player. Each of them incurs in the worst situation when he cooperates and the other betrays. Nash equilibria appear when each player is maximizing their payoff, conditioned to the decision of the other. Therefore, the information available to each player is of utmost relevance. When the 2-player game is extended to n players, it is reasonable to assume that each player has access to some—yet limited—amount of information on the decisions of the others. Therefore, the topology of the network becomes relevant due to the possibility of each player to establish links and access to the information of the other nodes. Statistical mechanics methods are suitable for detecting the optimal Nash equilibria in random as well as scale free networks (Hauert and Szabó 2005; Szabó and Fath 2007).

Of course, many extensions of the PD model have been studied. For instance, evolutionary game theory is designed to capture the essentials of the characteristic interactions among individuals through the evolution of cooperation. The interplay among network structure, clustering and behavioral strategies constitutes an intriguing link between only apparently unrelated disciplines. They are all represented in the framework of game theory, which links the individual behavior to the collective state of equilibrium. On large networks, typical questions are the detection of the percentage of cooperators versus betrayers. Depending on the parameters of the specific model under examination, critical phase transitions may be found. Models have been developed that fall into the universality class of directed percolation on square lattices and mean-field-type transitions on regular small world networks and random regular graphs (Hauert and Szabó 2005).

Evolutionary game theory is the extension of game theory to evolving populations. The population dynamics may concern any evolving set of individual units, ranging from lifeforms in biology to economic or financial agents in economic and financial models. Evolutionary games are not limited to a sequence of game at each step of the evolution, but also consider the frequency of the changes in the strategy (Perc and Szolnoki 2010; Szabó and Fath 2007).

Lattices serve as structure for defining the connection of the players, and open the way to spatial models. In Guan et al. (2006), Szabó and Tőke (1998), and Szabó et al. (2005) the update of a strategy is based on a random selection of nearest neighbours, and the target is understanding the density of cooperators versus betrayers.

Other network structures have been examined. For instance, in Chen et al. (2007), the PD is explored on community networks. Simulation results show that the average degree plays a universal role in cooperation occurring on random networks, small-world networks, star networks, and scale-free networks (Chen et al. 2007; Du et al. 2009b,c; Pinheiro et al. 2013; Rong et al. 2007; Shang and Wang 2015; Szolnoki et al. 2008; Tang et al. 2006; Wu 2007; Wu et al. 2007). When PD are studied on growing networks of contacts generated via preferential attachment, the correlation among individual increases, and this improves the amount of cooperating players (Santos and Pacheco 2005), while the opposite result occurs when a network becomes assortative mixing by degree (Rong et al. 2007). Nash equilibria have been used also for the optimal design of networks, in relation to the optimal transport and optimal navigation problem (Gulyás et al. 2014). Learning has been shown to be another relevant factor for the emergence of cooperators (Du et al. 2009c,a). In conclusion, (evolutionary) game theory constitutes a still growing field of application for complex networks, statistical mechanics, and optimization.

4.4 Clustering of Network Nodes and Geographical Information

Real life networks—like transport networks and Internet—are quite naturally embedded in a spatial dimension (Gastner and Newman 2006). The information on the geographical location of economic and financial networks has been considered in studies on the Gross Domestic Product (GDP) (Ausloos and Lambiotte 2007; Gligor and Ausloos 2008a,b; Miśkiewicz and Ausloos 2008; Miśkiewicz and Ausloos 2008; Redelico et al. 2009). A technique that is used for both economic and financial data consists in building a network of distances based on the correlation matrix calculated from the time series of the objects under observation. Of course, each of the entries of the empirical estimate of the correlation matrix is non zero, and thus it corresponds to a full network. Thus, techniques for extracting the most relevant information are applied. One of them is the Minimum Spanning Tree (MST). The MST is an acyclic network that connects all the nodes of the networks through the most relevant links. It has been applied on both economic and financial data. For instance, in Miśkiewicz and Ausloos (2008) it is shown that the distance among countries decreases as the year progresses, and this is interpreted as a sign of globalization (Miśkiewicz and Ausloos 2010).

In finance, MST has been used for detecting the backbone of the correlation among financial markets, together with other techniques, like the Maximally Planar Graph and Maximally Filtered Graph (Massara and Matteo 2011; Tumminello et al. 2005, 2007). The combination of tools from statistical physics and network theory allow to test the centrality of economic sector and their evolution during the years (Aste et al. 2010; Pozzi et al. 2008).

Another approach for passing from a complete network to a (usually) complex network consists in keeping only nodes with a connectivity higher than k. This leads to k-shells and k-cores and has been applied to both economic and financial data (Garas et al. 2012). On economic data, it leads to cluster of countries. Another possibility for working on a non complete network is keeping the links that have weights over a specific threshold (Garas et al. 2010).

A different way to explore the economic links among companies, even those belonging to different countries, has been studied in Braha et al. (2011), where a complex networks perspective on interfirm organizational networks is proposed by mapping, analyzing and modeling the spatial structure of a large inter-firm competition network across a variety of sectors and industries within the United States. A probabilistic model for the growth of the network is proposed, and it is shown that it is able to reproduce experimentally observed characteristics.

However, the analysis of ownership and control allows to detect economic super-entities that span over different countries. In fact, trans-national corporations form a giant bow-tie structure and a large portion of control flows to a small tightly-knit core of financial institutions. This raises new important issues both for researchers and policy makers (Vitali et al. 2011a).

Difficulties in the analysis arise from the fact that such networks show strongly connected components, which lead to circular relations in clusters (Schweitzer et al. 2009b). The proper value for the ownership through intermediaries can be disentangled through the calculus of integrated ownership (Rotundo and D'Arcangelis 2010b, 2013; Rotundo 2011; Vitali et al. 2011b). Questions on the maximal concentration in markets give rise to nonlinear optimization problems, that also involve the network structure (Rotundo and D'Arcangelis 2013). Networks of companies can be explored also through the board interlocks (D'Errico et al. 2008; Rotundo and D'Arcangelis 2010a). The coupling of information available from such different networks gives an example of application of *networks of networks* (Dorogovtsev and Mendes 2003).

5 Conclusions

In the present contribution we have summarized the main measures of complex networks with an eye on the combination of these topological objects with nonlinear models as a way to describe the phenomenology of complex systems. Specifically, we have analysed epidemiological compartmental models in combination with complex networks and we have briefly reviewed the main applications these formalisms have found in financial contagions. Moreover, we reviewed on games on networks and also on the usefulness of network nodes for analyzing geographical aspects of economic and financial networks.

Acknowledgements This work has been performed in the framework of COST Action IS1104 "The EU in the new economic complex geography: models, tools and policy evaluation". L.M.V. thanks the financial support of the Dirección Xeral de Saude Pública of the Xunta de Galicia (Spain) and the many fruitful discussions—and personal friendship—of X. Hervada and A. Malvar. L.M.V. wants to explicitly and strongly acknowledge the many contributions and useful discussions with his very close friends and colleagues J. Carrete and C. Ricoy-Riego, without whom this work would surely have not been done and life would be totally different.

References

Albert R, Barabási AL (2002) Statistical mechanics of complex networks. Rev Mod Phys 74:47–97
Albert R, Jeong H, Barabasi A (2000) Error and attack tolerance of complex networks. Nature 406(6794):378–382
Almeida-Neto M, Guimarães P, Guimarães Jr PR, Loyola RD, Ulrich W (2008) A consistent metric for nestedness analysis in ecological systems: reconciling concept and measurement. Oikos 117:1227–1239
Andersen JV, Nowak A, Rotundo G, Parrott L, Martinez S (2011) "price-quakes" shaking the world's stock exchanges. PloS One 6(11):e26472
Araujo AIL, Corso G, Almeida AM, Lewinsohn TM (2010) An analytic approach to the measurement of nestedness in bipartite networks. Physica A 389:1405–1411
Arenas A, Díaz-Guilera A, Kurths J, Moreno Y, Zhou C (2008) Synchronization in complex networks. Phys Rep 469(3):93–153
Aste T, Shaw W, Matteo TD (2010) Correlation structure and dynamics in volatile markets. New J Phys 12(8):085009
Atmar W, Patterson B (1993) The measure of order and disorder in the distribution of species in fragmented habitat. Oecologia 96(3):373–382
Ausloos M, Lambiotte R (2007) Clusters or networks of economies? A macroeconomy study through gross domestic product. Phys A Stat Mech Appl 382(1):16–21
Ausloos M, Petroni F (2014) Threshold model for triggered avalanches on networks. Preprint. arXiv:14014270
Barabási AL, Albert R (1999) Emergence of scaling in random networks. Science 286:509–512
Barrat A, Barthlemy M, Vespignani A (2008) Dynamical processes on complex networks, 1st edn. Cambridge University Press, New York
Barthélemy M, Barrat A, Pastor-Satorras R, Vespignani A (2005) Dynamical patterns of epidemic outbreaks in complex heterogeneous networks. J Theor Biol 235(2):275–288
Battiston S, Puliga M, Kaushik R, Tasca P, Caldarelli G (2012) DebtRank: too central to fail? Financial networks, the FED and systemic risk. Sci Rep 2:541
Bellenzier L, Vitting AJ, Rotundo G (2016) The world's stock exchange seen as network of coupled oscillators. In: Economic modelling (in press)
Bennett M, Schatz MF, Rockwood H, Wiesenfeld K (2002) Huygens's clocks. Proc R Soc Lond A Math Phys Eng Sci 458(2019):563–579
Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang DU (2006) Complex networks: structure and dynamics. Phys Rep 424:175–308
Braha D, Stacey B, Bar-Yam Y (2011) Corporate competition: a self-organized network. Soc Networks 33(3):219–230
Chen PY, Cheng SM (2015) Sequential defense against random and intentional attacks in complex networks. Phys Rev E 91:022805
Chen X, Ghate A (2011) Financial contagion on power law networks. http://dx.doi.org/10.2139/ssrn.1751143
Chen X, Fu F, Wang L (2007) Prisoner's dilemma on community networks. Phys A Stat Mech Appl 378(2):512–518
Chua L (1998) CNN: a paradigm for complexity. World Scientific series on nonlinear science: monographs and treatises. World Scientific, Singapore. https://books.google.es/books?id=G86zMVD3yNYC
Costa LDF, Oliveira ON Jr, Travieso G, Rodrigues FA, Villas Boas PR, Antiqueira L, Correa Rocha LE (2011) Analyzing and modeling real-world phenomena with complex networks: a survey of applications. Adv Phys 60(3):329–412
D'Errico M, Grassi R, Stefani S, Torriero A (2008) Shareholding networks and centrality: an application to the Italian financial market. In: Naimzada A, Stefani S, Torriero A (eds) Network, topology and dynamics. Theory and applications to economics and social systems. Springer, Berlin, pp 215–228

Dorogovtsev SN, Mendes JFF (2003) Evolution of networks: from biological nets to the internet and WWW (Physics). Oxford University Press, Inc., New York
Du WB, Cao XB, Hu MB (2009a) The effect of asymmetric payoff mechanism on evolutionary networked prisoner's dilemma game. Phys A Stat Mech Appl 388(24):5005–5012
Du WB, Cao XB, Hu MB, Yang HX, Zhou H (2009b) Effects of expectation and noise on evolutionary games. Phys A Stat Mech Appl 388(11):2215–2220
Du WB, Cao XB, Zhao L, Hu MB (2009c) Evolutionary games on scale-free networks with a preferential selection mechanism. Phys A Stat Mech Appl 388(20):4509–4514
Elliott M, Golub B, Jackson MO (2014) Financial networks and contagion. Am Econ Rev 104(10):3115–53
Erdös P, Rényi A (1959) On random graphs I. Publ Math 6:290–297
Fagiolo G, Reyes J, Schiavo S (2007) On the topological properties of the world trade web: a weighted network analysis. Papers, arXiv.org. http://EconPapers.repec.org/RePEc:arx:papers:0708.4359
Fujisaka H, Yamada T (1983) Stability theory of synchronized motion in coupled-oscillator systems. Prog Theor Phys 69(1):32–47
Garas A, Argyrakis P, Rozenblat C, Tomassini M, Havlin S (2010) Worldwide spreading of economic crisis. New J Phys 12:113043
Garas A, Schweitzer F, Havlin S (2012) A *k*-shell decomposition method for weighted networks. New J Phys 14:083030
Gastner MT, Newman ME (2006) The spatial structure of networks. Eur Phys J B-Condensed Matter Complex Syst 49(2):247–252
Gligor M, Ausloos M (2008a) Clusters in weighted macroeconomic networks: the eu case. introducing the overlapping index of gdp/capita fluctuation correlations. Eur Phys J B 63(4):533–539
Gligor M, Ausloos M (2008b) Convergence and cluster structures in eu area according to fluctuations in macroeconomic indices. Papers, arXiv.org. http://EconPapers.repec.org/RePEc:arx:papers:0805.3071
Guan JY, Wu ZX, Huang ZG, Xu XJ, Wang YH (2006) Promotion of cooperation induced by nonlinear attractive effect in spatial prisoner's dilemma game. Europhys Lett 76(6):1214
Gulyás A, Bíró J, Körösi A, Rétvári G, Krioukov D (2014) Complex networks as Nash equilibria of navigation games. Preprint arXiv:1412.7229
Harmon D, de Aguiar MA, Chinellato DD, Braha DD, Epstein I, Bar-Yam Y (2011) Predicting economic market crises using measures of collective panic. Preprint. http://ssrn.com/abstract=1829224 or http://dx.doi.org/10.2139/ssrn.1829224
Hauert C, Szabó G (2005) Game theory and physics. Am J Phys 73(5):405–414
Holme P (2005) Core-periphery organization of complex networks. Phys Rev E 72:046111
Kaizoji SDT (2009) Chap Market Bubbles and Crashes
Kermack WO, McKendrick AG (1927) A contribution to the mathematical theory of epidemics. Proc R Soc Lond A Math Phys Eng Sci 115(772):700–721
Lee KE, Lee JW (2009) Avalanches of Bak–Sneppen coevolution model on directed scale–free network. Fractals 17(02):233–237
Li J, Cao L, Wang C, Tan KC, Liu B, Pei J, Tseng VS (2013) Trends and applications in knowledge discovery and data mining: PAKDD 2013 Workshops DMApps, DANTH, QIMIE, BDM, CDA, CloudSD, Golden Coast, QLD. Springer Publishing Company, Incorporated
López-Pintado D (2008) Diffusion in complex social networks. Game Econ Behav 62:573–590
Lorenz J, Battiston S, Schweitzer F (2009) Systemic risk in a unifying framework for cascading processes on networks. Eur Phys J B 71(4):441–460
Malevergne Y, Sornette D (2001) Multi-dimensional rational bubbles and fat tails. Quant Financ 1(5):533–541
Markose S, Giansante S, Shaghaghi AR (2012) "Too interconnected to fail" financial network of {US} {CDS} market: topological fragility and systemic risk. J Econ Behav Organ 83(3):627–646

Massara GP, Di Matteo T (2011) Network filtering for big data: triangulated maximally filtered graph. Preprint. arXiv:150502445
Milgram S (1967) The small-world problem. Psychol Today 1:60–67
Miśkiewicz J, Ausloos M (2008) Correlation measure to detect time series distances, whence economy globalization. Phys A Stat Mech Appl 387(26):6584–6594
Miśkiewicz J, Ausloos M (2008) Correlation measure to detect time series distances, whence economy globalization. Phys A Stat Mech Appl 387(26):6584–6594
Miśkiewicz J, Ausloos M (2010) Has the world economy reached its globalization limit? Phys A Stat Mech Appl 389(4):797–806
Moreno Y, Vazquez A (2002) The Bak-Sneppen model on scale-free networks. Europhys Lett 57(5):765
Newman M (2003) The structure and function of complex networks. SIAM Rev 45:167–256
Newman M, Watts D, Barabási AL (2006) The structure and dynamics of networks. Princeton University Press, Princeton
Nåsell I (2002) Stochastic models of some endemic infections. Math Biosci 179(1):1–19
Pastor-Satorras R, Vespignani A (2001) Epidemic spreading in scale-free networks. Phys Rev Lett 86:3200–3203
Pastor-Satorras R, Vespignani A (2002) Epidemic dynamics in finite size scale-free networks. Phys Rev E 65:035,108
Pastor-Satorras R, Rubi M, Díaz-Guilera A (eds) (2003) Statistical mechanics of complex networks. Springer, Berlin
Pecora LM, Carroll TL (2015) Synchronization of chaotic systems. Chaos 25(9):097611
Perc M, Szolnoki A (2010) Coevolutionary games—a mini review. Biosystems 99(2):109–125
Petroni F, Ausloos M, Rotundo G (2007) Generating synthetic time series from Bak–Sneppen co-evolution model mixtures. Phys A Stat Mech Appl 384(2):359–367
Pinheiro FL, Vasconcelos V, Santos F, Pacheco JM (2013) Self-organized game dynamics in complex networks. ECAL - General Track 61-62
Pozzi F, Aste T, Rotundo G, Di Matteo T (2007) Dynamical correlations in financial systems. In: Microelectronics, MEMS, and Nanotechnology. International Society for Optics and Photonics, p 68021E
Redelico FO, Proto AN, Ausloos M (2009) Hierarchical structures in the gross domestic product per capita fluctuation in Latin American countries. Phys A Stat Mech Appl 388(17):3527–3535
Rong Z, Li X, Wang X (2007) Roles of mixing patterns in cooperation on a scale-free networked game. Phys Rev E 76:027101
Rotundo G (2011) Centrality measures in shareholding networks. In: Use of risk analysis in computer-aided persuasion. NATO science for peace and security series, vol 88. IOS Press, Amsterdam, pp 12–28
Rotundo G, Ausloos M (2007) Microeconomic co-evolution model for financial technical analysis signals. Phys A Stat Mech Appl 373:569–585
Rotundo G, D'Arcangelis AM (2010a) Network analysis of ownership and control structure in the Italian stock market. Adv Appl Stat Sci 2:255–273
Rotundo G, D'Arcangelis AM (2010b) Ownership and control in shareholding networks. J Econ Interaction Coord 5:191–219
Rotundo G, D'Arcangelis AM (2014) Network of companies: an analysis of market concentration in the Italian stock market. Quality Quantity 48(4):1893–1910
Rotundo G, Scozzari A (2009) Co-evolutive models for firms dynamics. In: Networks, topology and dynamics, vol 613. Springer, Berlin, pp 143–158
Santos FC, Pacheco JM (2005) Scale-free networks provide a unifying framework for the emergence of cooperation. Phys Rev Lett 95(9):098104
Schweitzer F, Fagiolo G, Sornette D, Vega-Redondo F, Vespignani A, White DR (2009a) Economic networks: the new challenges. Science 325(5939):422
Schweitzer F, Fagiolo G, Sornette D, Vega-Redondo F, White DR (2009b) Economic networks: what do we know and what do we need to know? Adv Complex Syst 12:407–422

Serrano MA, Krioukov D, Boguñá M (2008) Self-similarity of complex networks and hidden metric spaces. Phys Rev Lett 100:078701

Shang L, Wang XF (2007) Evolutionary minority game on complex networks. Phys A Stat Mech Appl 377(2):616–624

Sieczka P, Sornette D, Holyst JA (2011) The Lehman brothers effect and bankruptcy cascades. Eur Phys J B 82(3–4):257–269

Sornette D, Andersen JV (2002) A nonlinear super-exponential rational model of speculative financial bubbles. Int J Mod Phys C 13(02):171–187

Szabó G, Fath G (2007) Evolutionary games on graphs. Phys Rep 446(4):97–216

Szabó G, Tőke C (1998) Evolutionary prisoner's dilemma game on a square lattice. Phys Rev E 58(1):69

Szabó G, Vukov J, Szolnoki A (2005) Phase diagrams for an evolutionary prisoner's dilemma game on two-dimensional lattices. Phys Rev E 72(4):047107

Szolnoki A, Perc M, Danku Z (2008) Towards effective payoffs in the prisoner's dilemma game on scale-free networks. Phys A Stat Mech Appl 387(8):2075–2082

Tang CL, Wang WX, Wu X, Wang BH (2006) Effects of average degree on cooperation in networked evolutionary game. Eur Phys J B-Condensed Matter Complex Syst 53(3):411–415

Tumminello M, Aste T, Di Matteo T, Mantegna RN (2005) A tool for filtering information in complex systems. Proc Natl Acad Sci USA 102(30):10421–10426

Tumminello M, Di Matteo T, Aste T, Mantegna R (2007) Correlation based networks of equity returns sampled at different time horizons. Eur Phys J B 55(2):209–217

Vitali S, Glattfelder JB, Battiston S (2011a) The network of global corporate control. PLoS One 6:e25995

Vitali S, Glattfelder JB, Battiston S (2011b) The network of global corporate control. PloS One 6(10):e25995

Wang XF (2002) Complex networks: topology, dynamics and synchronization. Int J Bifurcation Chaos 12(05):885–916

Watts D, Strogatz S (1998) Collective dynamics of small-world networks. Nature 393:440–442

White LJ (2014) The basics of'too big to fail'. NYU Working Paper No. 2451/33564. http://papers.ssrn.com/sol3/Papers.cfm?abstract_id=2405433

Wu CW (2007) Synchronization in complex networks of nonlinear dynamical systems, vol 76. World Scientific, Singapore

Wu ZX, Guan JY, Xu XJ, Wang YH (2007) Evolutionary prisoner's dilemma game on Barabási–Albert scale-free networks. Phys A Stat Mech Appl 379(2):672–680

An Overview of Diffusion in Complex Networks

Dunia López-Pintado

Abstract We survey a series of theoretical contributions on diffusion in random networks. We start with a benchmark contagion process, referred in the epidemiology literature as the Susceptible-Infected-Susceptible model, which describes the spread of an infectious disease in a population. To make this model tractable, the interaction structure is considered as a heterogeneous sampling process characterized by the degree distribution. Within this framework, we distinguish between the case of *unbiased-degree* networks and *biased-degree* networks. We focus on the characterization of the *diffusion threshold*; that is, a condition on the primitives of the model that guarantees the spreading of the product to a significant fraction of the population, and its persistence. We also extend the analysis introducing a general diffusion model with features that are more appropriate for describing the diffusion of a new product, idea, behavior, etc.

Keywords Degree distribution • Diffusion threshold • Endemic state • Homophily • Random networks

JEL Classification Numbers: C73, L14, O31, O33

1 Introduction

In this chapter we discuss a number of different models of diffusion, where by *diffusion* (or contagion) we mean the process by which information (or any kind of signal) travels along a population of agents that are influenced by each other in some well defined way. The general objective is to understand how the network

D. López-Pintado (✉)
CORE, Université catholique de Louvain, Louvain-la-Neuve, Belgium

Department of Economics, Universidad Pablo de Olavide, Carretera de Utrera, Km. 1, 41013 Seville, Spain
e-mail: dlopez@upo.es

P. Commendatore et al. (eds.), *Complex Networks and Dynamics*, Lecture Notes in Economics and Mathematical Systems 683,
DOI 10.1007/978-3-319-40803-3_2

structure determines the reach of the process. This question is relevant for many different disciplines, ranging from sociology and economics to molecular biology and neurology.

In economics, technological diffusion has been a central topic of industrial organization and development which has lead to many well-known contributions (e.g., Rogers 1995; Conley and Udry 2001, among others). The issue of diffusion has also been extensively analyzed in the game theoretic literature where we can distinguish between models of learning (Bala and Goyal 1998), opinion formation Golub and Jackson (2012a,b,c) and network games (Morris 2000; Galeotti et al. 2010; Jackson and Yariv 2007). Finally, a direct application of diffusion is the study of disease transmission in a population, an issue which has been addressed widely in the epidemiology literature (e.g., Bailey 1975; Pastor-Satorrás and Vespignani 2001a,b). These last contributions build on the theoretical framework of random networks which provides a natural setup for the study of complex systems (Bollobás 2001; Erdös and Rényi 1959).

The purpose of this chapter is to present a series of models to extend those proposed in epidemiology. In doing so, we aim to understand diffusion not only of an infectious disease in a population, but also of an idea, a product, a cultural fad, or a technology. Our results focus on the characterization of the *diffusion threshold*, a condition on the primitives of the model which guarantees the spreading of the product (to a significant fraction of the population) and its persistence.

In Sect. 1 we study a benchmark contagion process referred in the literature as the Susceptible-Infected-Susceptible (SIS) model. The interaction structure is considered as the realization of a random sampling process characterized by the degree distribution, where the degree of an agent refers to the number of agents sampled by this agent. Within this framework, we distinguish between the case of *unbiased-degree* networks and *biased-degree* networks. In the first case, agents are homogeneous with respect to how much they are observed by others. Thus, heterogeneity in this framework is only related to the number of observations taken by agents before making a choice, but all agents are equally influential. In the second case, however, the number of agents observed by an agent roughly coincides with the number of agents observing such an agent. Thus, this framework can be considered as an approximation of an undirected network, where if agent i is influenced by agent j then j is influenced by i. The results we present shed light on the relevance of the degree distribution on the predictions of the model. In particular, we report how the diffusion threshold changes due to first order stochastic dominance shifts and mean-preserving spreads in the degree distribution.

In Sect. 2, we extend the previous analysis to account for general contagion processes which embody different models, including those based on best-response dynamics of coordination games, imitation dynamics, etc. We find that the diffusion threshold crucially depends on the contagion process. Thus, it becomes a relevant empirical question to determine which models are more appropriate for which applications. For instance, the well-known result that scale-free degree distributions exhibit a zero epidemic threshold for the SIS model is not robust to other contagion processes.[1] In particular, for those contagion processes in which the relative number of adopters (with respect to the size of the sample) is what determines the adoption rate (and not just absolute exposure), degree distributions with intermediate variance might be more appropriate for fostering diffusion.

In Sect. 3, we generalize the model even further to include the issue of homophily (i.e., the tendency of agents to associate with others similar to themselves). To do so, agents are distinguished by their types (e.g., race, gender, age, religion, profession). The interaction patterns are biased by types and different types of individuals might have different proclivities for adoption. In this context, we can analyze how such biases in interactions together with heterogeneity in susceptibility for adopting the new product (idea, disease, etc.) affect the reach of the process. For example, how does the diffusion of a new product that is more attractive to one age group depend on the interaction patterns across age groups? The main result is that homophily actually facilitates diffusion. That is, having a higher rate of homophily allows the diffusion to get started within the more vulnerable type and this can generate the critical mass necessary to diffuse the behavior or infection to the wider society.

2 The SIS Model

Consider a new product, an infectious disease, or an idea spreading in a population. Our objective is to analyze whether diffusion occurs. That is, if we start with an infinitesimal small fraction of initial adopters, would the *product* be adopted by a significant fraction of the population and become endemic? In order to answer this question theoretically we make several crucial assumptions. On the one hand, the contagion process considered is the standard Susceptible-Infected-Susceptible model (SIS hereafter) introduced in the epidemiology literature to study the diffusion of an infectious disease in a population.[2] On the other hand, we introduce a *directed* random sampling process to describe how agents are influenced by each other.

Formally, assume a continuum of agents $N = [0, 1]$. Agents can be in two possible states: *active* (infected) or *passive* (susceptible). A passive agent can become active, and conversely, an active agent can become passive. The SIS model assumes the simplest possible process of contagion characterized by the following parameters. A passive agent becomes active with a probability $\nu > 0$ when interacting with an active agent. Conversely, with a probability $\delta > 0$ an

[1] The existence of a zero epidemic threshold for scale-free networks was first shown by Pastor-Satorrás and Vespignani (2001a).

[2] The so-called SIS model has extensively been studied in the literature (see e.g., Pastor-Satorrás and Vespignani 2001a; Jackson and Rogers 2007, etc.).

active agent can become passive again.[3] The crucial parameter of the model is the (effective) *spreading rate* denoted by $\lambda = \frac{\nu}{\delta}$, which measures how contagious the behavior is. In this setting, the system must always remain in continuous flux since the particular identity of active and passive agents is permanently changing. The objective in this context is to predict the convergence to some population profile where the frequency of active agents remains stable over time. To make the approach tractable, the dynamics is described in continuous time. Thus, the previously defined probabilities ν and δ are instead interpreted as rates. In addition, the stochastic process is approximated by its deterministic counterpart.[4]

Let us consider that individuals observe each other before changing their states. Assume that each agent is characterized by her degree. In particular, an agent has degree d if she samples from the population (and is potentially influenced by) d other agents per unit of time. Observation is typically directed; that is, if an agent observes agent i, this does not imply that j observes i, although an approximation of an undirected network will be considered as well. Let $P(d)$ be the degree distribution; that is, the fraction of the population with degree d. Equivalently, $P(d)$ can be viewed as the probability that a randomly selected node has degree d.

There are several focal degree distributions. For instance, if the population is homogeneous then $P(\overline{d}) = 1$ for some degree $\overline{d} \geq 1$. Moreover, empirical studies have led to the conclusion that many complex networks are characterized by a scale-free degree distribution (i.e., a fat-tailed property). Price (1965) was the first to find such distributions in a network setting (in particular, in citation networks among scientific articles). The scale-free distribution, or power-law distribution, can be expressed as

$$P(d) = bd^{-\gamma}$$

where $2 < \gamma \leq 3$ and b is a positive normalizing constant. The main feature of this distribution is that the relative probabilities of two different degrees $(d, \widehat{d})$ only depend on their ratio $(\frac{d}{\widehat{d}})$ and not on their absolute values. In these distributions, the average degree cannot be conceived as a good estimate of the typical node degree found in the network. In particular, the population has a significant fraction of *hubs*, i.e., nodes with very high degree compared to the average.

Within the context of directed random networks, we consider two paradigmatic cases depending on how agents choose who to observe: *unbiased-degree* (case 1) and *biased-degree* (case 2). In case 1, agents select other agents completely at

[3]Note that in the context of a disease, it is implicitly assumed that there is no full immunization and therefore a recovered person can catch the disease again. An obvious instance is the standard flu.

[4]Benaïm and Weibull (2003) show that the continuous (deterministic) approximation is appropriate when dealing with large populations. In particular, they find that if the deterministic population flow remains forever in some subset of the state space, then the stochastic process will remain in the same subset space for a very long time with a probability arbitrarily close to one, provided the population is large enough.

random and, thus, the probability of choosing an agent with degree d is precisely $P(d)$. In case 2, agents are biased by the degree of others, so that an agent with degree d is sampled d times more often than an agent with degree 1. Therefore, the probability of selecting an agent with degree d is proportional to $dP(d)$. More precisely, let $Q(d)$ be the probability of selecting an agent with degree d. Then

$$Q(d) = P(d) \text{ in case 1}$$

whereas

$$Q(d) = \frac{dP(d)}{\langle d \rangle_P} \text{ in case 2}$$

where $\langle d \rangle_P = \sum_{d \geq 1} dP(d)$ is the average degree.

Note that Case 2 can be considered as an approximation of an undirected network, as the number of the agents is observed by an agent is the same as the (expected) number of times this agent is being observed. Note that, for some applications (e.g., the diffusion of a disease) it is a more accurate description of the reality as personal interaction is required for contagion. We analyze next the two cases separately.

2.1 *The SIS Model and Unbiased-Degree Random Networks*

In this section we focus on the unbiased-degree network case which represents the simplest framework to study random interactions characterized by a degree distribution.

Let us first introduce some notation. Let $\rho_d(t)$ denote the frequency of active agents among those with degree d at time t and $\rho(t)$ be the total frequency of active agents in the population at time t. Thus,

$$\rho(t) = \sum_d P(d)\rho_d(t).$$

The adoption dynamics describes the evolution of $\rho_d(t)$ as a function of the parameters of the SIS model. For each $d \geq 1$ we have the following differential equation:

$$\rho_d'(t) = -\rho_d(t)\delta + (1 - \rho_d(t))\nu d\rho(t),$$

where the first term on the sum $(-\rho_d(t)\delta)$ tracks the transitions from active to passive, whereas the second term tracks the transitions from passive to active $((1-\rho_d(t))\nu d\rho(t))$. To understand this term, note that the expected number of active agents in the sample of an agent with degree d is $d\rho(t)$. Thus, the probability that a passive agent becomes active in the small interval of time from t to $t + dt$ is given by $[1 - (1 - \upsilon dt)^{d\rho(t)}]$ and $\lim_{dt \to 0} \frac{[1-(1-\upsilon dt)^{d\rho(t)}]}{dt} = \nu d\rho(t)$.

The stationary states of this dynamics can be computed by imposing that $\rho'_d(t) = 0$ for all d. Therefore, for each d,

$$\rho_d = \frac{\lambda d\rho}{1 + \lambda d\rho},$$

where $\lambda = \frac{\nu}{\delta}$ is the (effective) *spreading rate*. The following fixed-point equation characterizes the fraction of adopters in the stationary state:

$$\rho = H_{\lambda,P}(\rho), \tag{1}$$

where

$$H_{\lambda,P}(\rho) = \sum_d P(d) \frac{\lambda d\rho}{1 + \lambda d\rho}.$$

Notice that $\rho = 0$ is always a solution of Eq. (1), which implies that the state where all agents are passive is stationary. Thus, in order to spread the "active state" in the population, there must be an initial seed of active agents. To be more precise, let us introduce the following two definitions:

We say that there is *diffusion* in the population if by seeding it randomly with an infinitesimally small initial fraction of active agents, the behavior spreads to a positive fraction of the population and becomes persistent.

We say that λ^* is the *diffusion threshold* if there is diffusion if and only if $\lambda > \lambda^*$.

Theorem 1 (López-Pintado 2012) *Let P be the degree distribution of a (unbiased-degree) random network. The diffusion threshold for the SIS model is:*

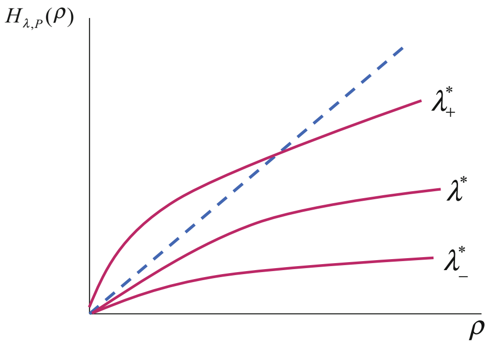
$$\lambda^* = \frac{1}{\langle d \rangle_P}$$

The outline of the proof is the following. In this context, diffusion occurs whenever Eq. (1) has a positive solution. It is straightforward to show that $H_{\lambda,P}(\rho)$ is an increasing and concave function of ρ. Moreover, $H_{\lambda,P}(0) = 0$ and $H_{\lambda,P}(1) < 1$. Therefore, as depicted in Fig. 1, there exists a positive solution of Eq. (1) if and only if $\frac{dH_{\lambda,P}(\rho)}{d\rho}\rfloor_{\rho=0} = \lambda \sum_d P(d)d > 1$.

The diffusion threshold is inversely proportional to the average degree. That is, the higher the average degree the easier it is to foster diffusion. Nevertheless, Theorem 1 does not provide information about the reach of the process whenever there is diffusion. We analyze this issue next.

We say that the adoption dynamics has reached an *endemic state* with a fraction of adopters ρ^* if this fraction remains constant in the upcoming periods.

Fig. 1 Representation of $H_{\lambda,P}$ when (1) λ equals the diffusion threshold λ^* (2) λ is above the diffusion threshold $\lambda = \lambda^*_+$ and (3) λ is below the diffusion threshold $\lambda = \lambda^*_-$



Equation (1) provides a characterization of the endemic states as a function of the degree distribution. Notice that Eq. (1) has one solution ($\rho = 0$) when $\lambda \leq \lambda^*$ and two solutions ($\rho = 0$ and a positive one) when $\lambda > \lambda^*$, as depicted in Fig. 1. Nevertheless, $\rho = 0$ is not a stable solution whenever $\lambda > \lambda^*$. Thus, we define as $\rho^*(P)$ to the (stable) endemic state of the diffusion process.

Two definitions are required before presenting the next result. Consider the degree distributions P and $\widetilde{P}$. We say that $\widetilde{P}$ *first order stochastic dominates* P if

$$\sum_{d=0}^{x} \widetilde{P}(d) \leq \sum_{d=0}^{x} P(d) \text{ for all } x.$$

The intuitive idea is that $\widetilde{P}$ is obtained by shifting mass from P to place it on higher values.

We can also say that $\widetilde{P}$ is a *mean-preserving spread* of P if $\widetilde{P}$ and P have the same mean and

$$\sum_{z=0}^{x} \sum_{d=0}^{z} \widetilde{P}(d) \leq \sum_{z=0}^{x} \sum_{d=0}^{z} P(d) \text{ for all } x.$$

This condition implies that $\widetilde{P}$ has a (weakly) higher variance than P, but it also implies a more structured relationship between the two. In fact the reverse is not true, having a higher variance and the same mean is not sufficient for one distribution to be a mean-preserving spread of another.

Proposition 1 *Let P be the degree distribution of a (unbiased-degree) random network and consider the SIS model. The following holds:*

(1) If $\widetilde{P}$ first order stochastic dominates P then $\rho^*(\widetilde{P}) \geq \rho^*(P)$.
(2) If $\widetilde{P}$ is a mean-preserving spread of P then $\rho^*(\widetilde{P}) \leq \rho^*(P)$.

Before showing this result, let us describe the following two well-known *properties*.

- Property 1: If $\widetilde{P}$ first order stochastic dominates P then, for all non-decreasing functions f,

$$\sum_d f(d)\widetilde{P}(d) \geq \sum_d f(d)P(d).$$

and

- Property 2: If $\widetilde{P}$ is a mean-preserving spread of P then, for all concave functions f,

$$\sum_d f(d)P(d) \geq \sum_d f(d)\widetilde{P}(d).$$

Notice that the endemic state is characterized by Eq. (1) and since ρ_d is nondecreasing and concave (as a function of d), applying properties (1) and (2) we obtain the desired result.

As illustrated in Fig. 2, the diffusion threshold decreases and the endemic state increases with a first order stochastic dominance shift of the degree distribution (see the graph on the left). This is a consequence of the fact that in the SIS contagion process the higher the degree of an agent, the easier it is to become an adopter. We also illustrate how, even though the diffusion threshold does not vary if we shift the degree distribution with a mean-preserving spread, the endemic state decreases with such a shift (see the graph on the right). The intuition for this is that an increase in the degree of an agent increases her adoption rate, but it has a decreasing marginal effect.

In the next section we analyze a biased-degree random network instead and find critical differences with the degree-unbiased case.

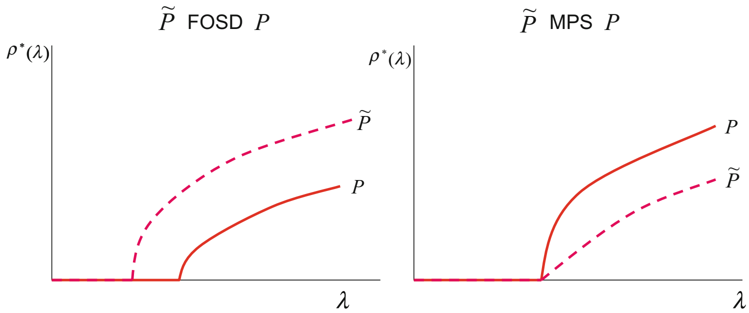


Fig. 2 The graphs represent qualitatively the endemic state (ρ^*) as function of the spreading rate (λ) for the degree distribution P and $\widetilde{P}$, where $\widetilde{P}$ first order stochastic dominates (FOSD) P in the graph on the *left* and $\widetilde{P}$ is a mean-preserving spread of P (MPS) in the graph on the *right*

2.2 The SIS Model and Biased-Degree Random Networks

The methodology applied in the previous section is useful for understanding the predictions on diffusion in a directed network, that is, a network where if i interacts with j this does not imply that j interacts with i. This, of course, is a strong assumption as many socioeconomic interactions among agents are bilateral in nature and thus links in the network are undirected. We therefore can extend the specification of the model proposed above in order to approximate an undirected interaction structure. A tractable attempt to do so is to assume that sampling is not performed uniformly at random but that, instead, it is biased by degree. That is, the probability that an agent samples an agent with degree d is proportional not only to $P(d)$ but also to d. This captures the idea that agents with higher degree are sampled more often.[5] Formally, as already highlighted in Sect. 1, in this case, the probability that an agent samples another agent with degree d is:

$$Q(d) = \frac{dP(d)}{\langle d \rangle_P},$$

where the average degree $\langle d \rangle_P$ is used for normalization purposes.

Some additional notation is required before presenting the dynamics. Let $\theta(t)$ be the probability that an agent samples an active agent at time t. Thus,

$$\theta(t) = \sum_d Q(d)\rho_d(t). \tag{2}$$

This is due to the fact that the probability that at the end of a link an agent has degree d is $Q(d)$ (which in this case is different from $P(d)$). Following analogous steps to those already presented in Sect. 2.1 we find that the adoption dynamics is now described as follows:

$$\rho_d'(t) = -\rho_d(t)\delta + (1 - \rho_d(t))\nu d\theta(t),$$

where the first term of the sum $(-\rho_d(t)\delta)$ stands for the transition from active to passive. The second term stands for the transition from passive to active $((1 - \rho_d(t))\nu d\theta(t))$, where the expected number of adopters in a sample of size d is $d\theta(t)$. The stationary states of this dynamics can be computed by imposing that $\rho_d'(t) = 0$ for all d which leads to:

$$\rho_d = \frac{\lambda d\theta}{1 + \lambda d\theta} \text{ for all } d. \tag{3}$$

[5] Pastor-Satorrás and Vespignani (2001a) used this specification.

Substituting (3) in (2) we find that the fixed-point equation which characterizes the value of θ in the stationary state is

$$\theta = \widetilde{H}_{\lambda,P}(\theta), \tag{4}$$

where

$$\widetilde{H}_{\lambda,P}(\theta) = \frac{1}{\langle d \rangle_P} \sum_d P(d) \frac{\lambda d^2 \theta}{1 + \lambda d \theta}.$$

The next result characterizes the diffusion threshold. To do so let us denote by $\langle d^2 \rangle_P$ to the second order moment of the degree distribution P. That is,

$$\langle d^2 \rangle_P = \sum_d d^2 P(d).$$

Theorem 2 (Pastor-Satorrás and Vespignani 2001a) *Let P be the degree distribution of a (biased-degree) random network. The diffusion threshold for the SIS model is:*

$$\lambda^* = \frac{\langle d \rangle_P}{\langle d^2 \rangle_P}.$$

The outline of the proof is the following. There is diffusion if there exists a positive solution of Eq. (4). In addition, it is straightforward to show that $\widetilde{H}_{\lambda,P}(\theta)$ is an increasing and concave function of θ, $\widetilde{H}_{\lambda,P}(0) = 0$ and $\widetilde{H}_{\lambda,P}(1) < 1$. Therefore, there is diffusion if and only if $\frac{d\widetilde{H}_{\lambda,P}(\theta)}{d\rho}\rfloor_{\theta=0} = \lambda \frac{\langle d^2 \rangle_P}{\langle d \rangle_P} > 1$.

Note that, unlike for the unbiased case, now the variance of the degree distribution also determines the diffusion threshold as $\lambda^* = \frac{\langle d \rangle_P}{\langle d^2 \rangle_P} = \frac{\langle d \rangle_P}{\langle d \rangle_P^2 + var(p)}$, where $var(P)$ denotes the variance of distribution P.

As a consequence of Theorem 2, we provide the following comparative statics results on the degree distribution.

Corollary 1 *Let P be the degree distribution of a (biased-degree) random network and consider the SIS model. The following holds:*

(1) If $\widetilde{P}$ first order stochastic dominates P then $\lambda^*(\widetilde{P}) \leq \lambda^*(P)$.
(2) If $\widetilde{P}$ is a mean-preserving spread of P then $\lambda^*(\widetilde{P}) \leq \lambda^*(P)$.

This proof of this corollary is straightforward given the expression for the diffusion threshold provided by Theorem 2. If we compare these results with those obtained for the unbiased-degree case (Proposition 1), we find that for both cases the diffusion threshold decreases with the density of the network. Nevertheless, the effect of a mean-preserving spread is different as in the biased-degree case diffusion is triggered more easily the higher the variance of the degree distribution. The intuition behind this finding relies on the relevant role that hubs (i.e., high-degree nodes) play for diffusion in the biased-degree case, which is not as important in the unbiased-degree case. In the biased-degree case, agents with high degree not only observe many others, but are also observed by many others. Therefore, they easily become infected and also infect others afterwards. For the unbiased-degree case, however, agents with high degree are observed equally as much as any other agent in the population and, thus, they do not necessary promote diffusion once they become infected.

The study of the endemic state and how it depends on the degree distribution is not straightforward (see Jackson and Rogers 2007). The main reason for this is that the values of ρ (the fraction of adopters in the population) and θ (the probability of sampling an adopter) do not necessarily move in the same direction when there is a shift in P. The next result shows that θ increases with a mean-preserving spread of the degree distribution. Nevertheless, this does not imply that ρ also increases. A piece of notation is needed. Given P, let $\theta^*(P)$ denote the value of θ in the (stable) endemic state of the dynamics. Then the following result holds:

Proposition 2 (Jackson and Rogers 2007) *Let P be the degree distribution of a (biased-degree) random network and consider the SIS model. If $\widetilde{P}$ is a mean-preserving spread of P, then $\theta^*(P) \leq \theta^*(\widetilde{P})$.*

The proof of this result is a direct consequence of Eq. (4), property (2) described in Sect. 2.1, and the fact that $\frac{\lambda d^2 \theta}{1+\lambda d \theta}$ is a convex function of d.[6]

3 Beyond the SIS Model: General Adoption Rules

The SIS model corresponds with a specific contagion process, which is directly imported from epidemiology. In social contexts, however, the diffusion of information, or a behavior, often exhibits features that do not match well those of the epidemic models. For instance, in the SIS formulation of diffusion, the transmission of infection to a healthy agent depends on her total exposure to the disease, i.e., the absolute number of infected neighbors. In the spread of many social phenomena, there is a factor of coordination (or persuasion) involved and therefore relative considerations are important (i.e., the number of infected versus non-infected). Moreover, unlike in the SIS model, the adoption rate does not necessarily need to increase linearly with the number of adopters. For instance, if agents adopt only if a significant fraction of others have adopted (a threshold rule). Finally, in the SIS model, the transition from active to passive occurs at a constant rate and therefore does not depend on the behavior of others, something which seems artificial for diffusion in many socioeconomic contexts. For all the reasons listed above, we now

[6]Jackson and Rogers (2007) were the first to analyze the diffusion proprieties of networks ordered through the stochastic dominance of their degree distributions.

present a general family of contagion models which extend the SIS model in several directions.

We assume that in each period agents are in one of two states: active or passive (as before). The agents' actions are influenced by the actions of others, but in a stochastic manner. A passive agent adopts the behavior at a rate described by an adoption rule $f_d(a)$ where d is her degree and a is the number of sampled agents who have adopted the behavior. Conversely, an active agent becomes passive at a rate given by $g_d(a)$ where, again, d is her degree and a is the number of sampled agents who have adopted the behavior. The adoption rules $f_d(a)$ and $g_d(a)$ are the primitives of the diffusion process and must satisfy the following assumptions:

- $f_d(0) = 0$ for each d. In words, a passive agent cannot become active unless she samples at least one active agent.
- $f_d(a)$ is a non-decreasing function of a. In words, the adoption rate is non-decreasing in the number of adopters in the sample.
- $f_d(1) > 0$ for some d such that $P(d) > 0$. This is a technical condition and it implies that there exists a certain degree such that the rate of adoption for agents with such a degree, when only one agent in the sample is active, is positive.[7]
- $g_d(0) = \delta > 0$ for all d. That is, the transition rate from active to passive, when all agents in the sample are passive, is positive and constant for all degrees.
- $g_d(a)$ is a non-increasing function of a. That is, the transition rate from active to passive is non-increasing in the number of active agents in the sample.

This general approach encompasses a number of different models. Three simple examples are the following.

First, the SIS model presented in the previous section corresponds with the adoption rules $f_d(a) = \nu a$ and $g_d(a) = \delta$.

Second, consider the following *Imitation model*. Every period, a non-smoker considers the possibility of smoking at a rate $\nu > 0$. This agent engages in smoking if he or she happens to sample a smoker among those agents that influence him or her. Conversely, at a rate δ a smoker considers the possibility of quitting smoking. This agent decides to quit if him or her happens to sample among those agents that influence him or her a non-smoker. This diffusion process corresponds with the following specification of the adoption rules: $f_d(a) = \nu \frac{a}{d}$ and $g_d(a) = \delta \frac{d-a}{d}$.

Third, consider the following *Majority threshold model*.[8] We consider again the example of choosing whether to smoke or not. Every period, a non-smoker considers the possibility of smoking at a rate $\nu > 0$. This agent engages in smoking behavior if he or she observes that more than half of the agents that influence him or her are smokers. Conversely, a smoker considers the possibility of quitting smoking at a rate δ. This agent decides to quit if him or her observes that at least half of the agents

[7] For example, a rule where agents adopt only if at least two sampled agents have adopted does not satisfy this assumption.

[8] These threshold models have been extensively analyzed in the literature (see Granovetter 1978, Watts 2002, López-Pintado 2006, and Jackson and Yariv 2007).

that influence him or her are non-smoker. This diffusion process corresponds with the following specification of the adoption rules: $f_d(a) = \nu$ if $\frac{a}{d} > 0.5$ and $f_d(a) = 0$ otherwise. Also, $g_d(a) = \delta$ if $\frac{a}{d} \leq 0.5$ and $g_d(a) = 0$ otherwise.

Notice that both in the Imitation model and the Majority threshold model relative considerations (i.e., $\frac{a}{d}$ instead of a) are important. Moreover, in the Majority threshold model the adoption rules do not depend linearly on a. Finally, the transition from active to passive, in both models, is not constant and crucially depends on the behavior observed by the agent when making such a decision.

The diffusion threshold can also be calculated for these general models as presented in the next result.

Theorem 3 (Jackson and López-Pintado 2013) *Let P be the degree distribution of a random network. The diffusion threshold for the general model is:*

$$\sum_d Q(d)d\frac{f_d(1)}{\delta} > 1$$

where $Q(d) = P(d)$ in an unbiased-degree random network, whereas $Q(d) = \frac{dP(d)}{\langle d \rangle_P}$ in a biased-degree random network.

Notice that Theorem 3 shows that the diffusion threshold depends on the degree distribution and on the values of the adoption rules $f_d(1)$ and $g_d(0)$. The reason why $f_d(a)$ and $g_d(a)$ for $a > 1$ does not appear in the condition is that in the initial periods of the dynamics there is only a small fraction of adopters in the population and, thus, the probability that an agent observes more than one adopter in her sample is negligible. Nevertheless, as briefly explained below, further properties of the adoption rule (e.g., the concavity of the rule) crucially affect other properties of the diffusion process, as for example, the type of transition occurring at the diffusion threshold (i.e., whether it is a second order phase transition or not).

The results obtained for the Imitation model are striking. The diffusion threshold is $\lambda^* = 1$, both for the unbiased-degree and biased-degree random network. The reason is that, in this case, all agents have the same probability of becoming an adopter, independently of their degree. To see this, consider two agents *i* and *j* where *i* has degree d and *j* has degree $2d$. It is straightforward to show that the probability that agent *i* observes a active agents in the sample coincides with the probability that agent *j* observes $2a$. Moreover, the Imitation model assumes that both agents, *i* and *j*, would have the same probability of adopting as $\frac{a}{d} = \frac{2a}{2d}$).

For the Majority threshold model, we obtain that the diffusion threshold is $\lambda^* = 1/p(1)$ for the unbiased random network, and $\lambda^* = \langle d \rangle_P / p(1)$ for the biased random network. Notice that, in this case, only agents with degree 1 that happen to sample an adopter will adopt in the initial periods of the dynamics, which is why the diffusion threshold decreases with respect to $p(1)$. The diffusion threshold is higher for the biased-degree case than for the unbiased-degree case. The reason is that, in the former case, the agents with degree 1 are observed (in expectation) only by 1 agent and thus are less efficient in spreading the behavior than in the later case where an agent with degree 1 is observed by the same number of agents as any other agent in the population (i.e., roughly by $\langle d \rangle_P$ other agents in each unit of time).[9]

To conclude, let us concentrate on *absolute adoption rules*, that is, on rules satisfying that it is the total exposure to the activity what determines the adoption rate. Formally, $f_d(a) = f(a)$ and $g_d(a) = g(a)$ for all d and $0 \leq a \leq d$. We can distinguish three focal absolute adoption rules, which are the following:

(a) $f(a) = \upsilon a$ and $g(a) = \delta$ (i.e., the SIS model), where the marginal impact on adoption of having one more adopter in an agents's sample is constant.
(b) $f(a) = \upsilon \sqrt{a}$ and $g(a) = \delta$, where the marginal impact on adoption of having one more adopter in an agent's sample is decreasing.
(c) $f(a) = \upsilon a^2$ and $g(a) = \delta$, where the marginal impact on adoption of having one more adopter in an agents's sample is increasing.

Corollary 2 *Let P be the degree distribution of a random network. The diffusion threshold for the absolute adoption rules (a), (b) and (c) is*

$$\lambda^* = \frac{1}{\langle d \rangle_P}$$

for an unbiased-degree random network, whereas it is

$$\lambda^* = \frac{\langle d \rangle_P}{\langle d^2 \rangle_P}$$

for a biased-degree random network.

This corollary is a direct application of Theorem 3. Notice that the diffusion threshold coincides for all the absolute adoption rules considered. Nevertheless, as described in López-Pintado (2008), for cases (a) and (b) the endemic state $\rho^*(\lambda)$ exhibits a second order phase transition at the diffusion threshold $\lambda = \lambda^*$. In other words, the endemic state $\rho^*(\lambda)$ is a continuous function of the spreading rate and, therefore, as λ converges to λ^*, $\rho^*(\lambda)$ converges to $\rho(\lambda^*)$. For case (c), however, we obtain a discontinuity in the endemic fraction of adopters at $\lambda = \lambda^*$ (first order phase transition or hysteresis) as illustrated in Fig. 3. This is due to the existence of multiple stationary states of the adoption dynamics, depending on the size of the initial seed of adopters.

[9]The study of how collective outcomes depend on the details of the contagion process has also been highlighted by Young (2009), Galeotti and Goyal (2009), López-Pintado and Watts (2008), etc.

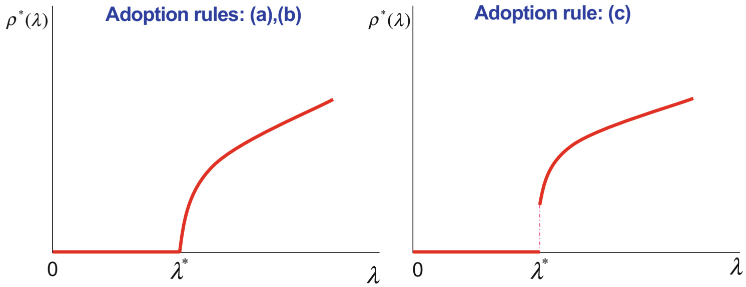


Fig. 3 The graphs represent qualitatively the endemic state (ρ^*) as function of the spreading rate (λ). The graph on the *left* corresponds to concave adoption rules such as (a) and (b), where there is a second order phase transition at λ^*. The graph on the *right* corresponds to adoption rule (c), where there is a first order phase transition at λ^*

4 Homophily

In this section we want to understand the effect that homophily has on diffusion, something which despite its importance has received little attention in the diffusion literature.[10] Homophily is the tendency of agents to associate with others similar to themselves. For example, young children in day care have higher rates of interaction with other young children than with older children. Adults of a certain profession, religion and education are more likely to interact with other adults with similar characteristics. In addition we allow for heterogeneity in preferences regarding the new product or behavior (or different susceptibilities for catching a disease). For instance, children can be more vulnerable to some diseases than adults, a new movie can be more attractive to women than men, etc. In particular, we examine whether or not diffusion occurs in a heterogeneous and homophilous society.

To be consistent with the rest of the paper, we start analyzing homophily in the SIS model, and extend later the analysis to more general models of diffusion in random networks.

4.1 Homophily in the SIS Model

For ease of exposition, we assume that the population is only made of two groups (the young and the elder). Agents in each group have different proclivities for getting infected of a certain disease. In particular, imagine that the elder are more vulnerable to this disease than the young. More precisely, if λ_1 is the spreading rate of the young

[10]There are some exceptions such as Currarini et al. (2009), Golub and Jackson (2012b), among others.

and λ_2 of the elder then $\lambda_1 < \lambda_2$. Let π be the probability that an individual interacts with an individual of the same age range. We also allow for heterogeneity with respect to the degree distribution. In particular, let $P_i(d)$ be the degree distribution of individuals of type i. In this example it is reasonable to assume bilateral interaction and, thus, let us focus on the biased-degree network case. Hence, conditional on sampling an agent of type i, this agent will have degree d with probability $Q_i(d) = \frac{dP_i(d)}{\langle d \rangle_i}$. The result on diffusion is the following:

Proposition 3 (Jackson and López-Pintado 2013) *Let* $\pi_0 = \frac{1 - \widetilde{d}_1 \lambda_1 \widetilde{d}_2 \lambda_2}{\frac{\widetilde{d}_1 \lambda_1 + \widetilde{d}_2 \lambda_2}{2} - \widetilde{d}_1 \lambda_1 \widetilde{d}_2 \lambda_2}$, *where* $\widetilde{d}_i = \frac{\langle d^2 \rangle_i}{\langle d \rangle_i}$. *Diffusion occurs for the SIS model with homophily if and only if one of the following conditions hold:*

1) $\lambda_1 \lambda_2 > \frac{1}{\widetilde{d}_1 \widetilde{d}_2}$ *or*
2) $\lambda_1 \lambda_2 \leq \frac{1}{\widetilde{d}_1 \widetilde{d}_2}$ *and* $\pi > \pi_0$

Recall that the condition for diffusion in the standard (homogenous) SIS model is $\lambda > \frac{1}{\widetilde{d}}$, which is a particular case of the previous result. Note also that if we considered, instead, an unbiased-degree random network, we obtain the same result as in Proposition 3, but with $\widetilde{d}_i = \frac{1}{\langle d \rangle_i}$.

The most interesting scenario turns out to be one where one of the types would foster diffusion if isolated, whereas the other would not (i.e., $\lambda_1 < \frac{1}{\widetilde{d}_1}$ and $\lambda_2 > \frac{1}{\widetilde{d}_2}$). In that scenario, we show that homophily either plays no role (if $\lambda_1 \lambda_2 > \frac{1}{\widetilde{d}_1 \widetilde{d}_2}$) or it actually facilitates diffusion (if $\lambda_1 \lambda_2 < \frac{1}{\widetilde{d}_1 \widetilde{d}_2}$). Note that in the latter case diffusion occurs only if the two types are sufficiently biased in interactions towards their own types (i.e., π is sufficiently large).

4.2 Homophily Beyond the SIS Model

We now generalize the previous analysis beyond the SIS model with two types. Let us assume that all relevant characteristics are captured by a finite set of m types where $m \geq 1$ (e.g., agents in type i are male, aged 30–40, atheist and university professors). Formally, the continuum of agents $N = [0, 1]$ is partitioned by types where n_i denotes the fraction of agents of type i. Thus, $\sum_{i=1}^{m} n_i = 1$. Again, we assume that agents have a degree which measures the number of individuals sampled per unit of time. The distribution of degrees can be different across types (i.e., the elderly might have lower mean and variance in degrees than teenagers) and thus $P_i(d)$ indicates the degree distribution of individuals of type i.

The random meeting process now incorporates biases across types. In particular, the rate at which an agent of type i meets agents of other types is described by the following matrix:

$$\Pi = \begin{pmatrix} \pi_{11} & \dots & \pi_{1m} \\ \vdots & \dots & \vdots \\ \pi_{m1} & \dots & \pi_{mm} \end{pmatrix},$$

where π_{ij} is the probability that an agent of type i meets an agent of type j in any given meeting. Thus, $\sum_{j=1}^{m} \pi_{ij} = 1$. To guarantee that a behavior that starts spreading in one group reaches any other group we must assume that Π is a primitive matrix, that is, $\Pi^t > 0$ for some t.[11]

In any given period, an agent of type i with degree d expects to meet $d\pi_{ij}$ agent of type j, and conditional on meeting an agents of type j, the probability that this agent has degree d is

$$Q_j(d) = P_j(d),$$

in the unbiased-degree network case, and

$$Q_j(d) = \frac{dP_j(d)}{\langle d \rangle_{P_j}},$$

in the biased-degree network case, where $\langle d \rangle_{P_j}$ is the average degree of P_j.[12]

Let $\rho_{i,d}(t)$ denote the frequency of active agents at time t among those of type i with degree d. Thus,

$$\rho_i(t) = \sum_d P_i(d)\rho_{i,d}(t)$$

is the frequency of active agents at time t among those of type i, and

$$\rho(t) = \sum_d n_i \rho_i(t)$$

is the overall fraction of active agents in the population at time t.

[11] Notice that $\Pi^t = \Pi * \Pi * \dots * \Pi$, t times.

[12] In the biased-degree case, certain constraints on the parameters of the model would be required in order to approximate it to an undirected network. For example, the number of interactions from type i to type j should coincide with the number of interactions from type j to type i in a unit of time. That is, $n(i)\, \langle d \rangle_i\, \pi_{ij} = n(j)\, \langle d \rangle_j\, \pi_{ji}$.

Finally, the contagion model is defined with the general adoption rules $f_{i,d}(a)$ and $g_{i,d}(a)$ presented in Sect. 2 but note that these rules can differ across types. This allows us to define $\theta_i(t)$ as the probability that an agent of type i samples an active agent. Note that

$$\theta_i(t) = \sum_j \pi_{ij} \sum_d Q_j(d)\rho_{j,d}(t). \tag{5}$$

Let us now define the rates at which a passive agent becomes active and vice versa. To do so, let $rate_{i,d}^{0\to 1}(t)$ be the rate at which a passive agent of type i and with degree d becomes active, whereas $rate_{i,d}^{1\to 0}(t)$ stands for the reverse transition. We assume that the number of infected agents in a sample follows a binomial distribution with parameters d (number of draws) and $\theta_i(t)$ (probability of each draw being active). That is,

$$rate_{i,d}^{0\to 1}(t) = \sum_{a=0}^{d} f_{i,d}(a) \binom{d}{a} \theta_i(t)^a (1-\theta_i(t))^{(d-a)},$$

$$rate_{i,d}^{1\to 0}(t) = \sum_{a=0}^{d} g_{i,d}(a) \binom{d}{a} \theta_i(t)^a (1-\theta_i(t))^{(d-a)}.$$

The diffusion dynamics is described as follows:

$$\rho'_{i,d}(t) = -\rho_{i,d}(t) rate_{i,d}^{1\to 0}(t) + (1-\rho_{i,d}(t)) rate_{i,d}^{0\to 1}(t) \tag{6}$$

where the right-hand side represents the increase in the level of active agents, whereas the left-hand side represents the decrease in such a level due to the transition of some active agents to passive.

As in a stationary state $\rho'_{i,d}(t) = 0$ then

$$\rho_{i,d} = \frac{rate_{i,d}^{0\to 1}}{rate_{i,d}^{0\to 1} + rate_{i,d}^{1\to 0}}. \tag{7}$$

We substitute Eq. (7) in Eq. (5) and find that the values for θ_i in the steady states are

$$\theta_i = H_i(\theta_1, \theta_2 \ldots \theta_n),$$

where

$$H_i(\theta_1, \theta_2 \ldots \theta_n) = \sum_j \pi_{ij} \sum_d Q_j(d) \frac{rate_{i,d}^{0\to 1}}{rate_{i,d}^{0\to 1} + rate_{i,d}^{1\to 0}}.$$

This system of equations characterizes the steady states for θ_i, but from here we can compute the steady states for the fraction of adopters of each type ρ_i and ultimately the overall fraction of adopters ρ.

The objective is to find conditions for diffusion. Note that $\boldsymbol{\theta} = \mathbf{0}$ (i.e., $(\theta_1, \theta_2 \ldots \theta_n) = (0, 0, \ldots, 0)$) is a steady state of the diffusion dynamics. We must explore the stability of such state. If $\boldsymbol{\theta} = \mathbf{0}$ is not stable, following fixed-point arguments applied to monotone correspondences on lattices, it can be shown that there exists another strictly positive steady state of the dynamics (see Jackson and López-Pintado 2013 for details on this argument). From (5) and (6) we find

$$\theta_i'(t) = \sum_j \pi_{ij} \sum_d Q_j(d) \left[-\rho_{j,d}(t) rate_{j,d}^{1\to 0}(t) + (1 - \rho_{j,d}(t)) rate_{j,d}^{0\to 1}(t) \right].$$

Note that near $\boldsymbol{\theta} = \mathbf{0}$ (and assuming that there is an upper bound on the degree of agents) we have that:

$$rate_{j,d}^{1\to 0}(t) = \delta \text{ and } (1 - \rho_{j,d}(t)) rate_{j,d}^{0\to 1}(t) = df_{jd}(1)\theta_j(t)$$

and thus we can rewrite

$$\theta_i'(t) = \sum_j \pi_{ij} \sum_d Q_j(d) df_{jd}(1) \theta_j(t) - \theta_i(t)\delta,$$

which can be expressed in matricial form as

$$\boldsymbol{\theta}'(\mathbf{t}) = [A\boldsymbol{\theta} - \boldsymbol{\theta}]\,\delta$$

where

$$A = \begin{pmatrix} \pi_{11}x_1 & \ldots & \pi_{1m}x_m \\ \vdots & \ldots & \vdots \\ \pi_{m1}x_1 & \ldots & \pi_{mm}x_m \end{pmatrix},$$

and

$$x_i = \sum_d Q_i(d) d \frac{f_{i,d}(1)}{\delta}.$$

The term x_i can be interpreted as the relative growth of adoption due to type i and adjusted by the relative rates at which agents of type i will be met by other agents.

With this information we can now state the following result.

Theorem 4 (Jackson and López-Pintado 2013) *Diffusion occurs if the largest eigenvalue of A is larger than* 1.

Note that if we only have *one type* in the population then we can drop subindex i and the condition for diffusion is

$$x = \sum_d Q(d)d\frac{f_d(1)}{\delta} > 1$$

which coincides with the condition provided in Theorem 3.

Assume now that there are *two types* with symmetry in how introspective groups are in their meetings. Therefore, $\pi_{11} = \pi_{22} = \pi$. The result in this case is the following:

Proposition 4 (Jackson and López-Pintado 2013) *Let* $\pi_0 = \frac{1-x_1x_2}{x_1+x_2-x_1x_2}$. *Diffusion occurs if and only if one of the following conditions hold:*

1) $x_1x_2 > 1$ *or*
2) $x_1x_2 < 1$ *and* $\pi > \pi_0$.

Note that if diffusion occurs within each type when isolated, it would also occur when there is interaction among the two (in such a case $x_1 > 1$ and $x_2 > 1$ and thus part (1) of the proposition holds). If diffusion does not occur among either type when isolated, then it would not occur if there is interaction between them (note that $\pi > \pi_0$ cannot occur if both x_1 and x_2 are below 1). Finally, if diffusion occurs among only one of the types when isolated, then it would occur among the entire population if homophily is high enough. The intuition behind this result is that having a higher rate of homophily allows the diffusion to get started within the more vulnerable type, and this can generate the critical mass necessary to diffuse the behavior to the wider society.

5 Conclusions

In this chapter we have surveyed a series of stylized models of diffusion in networks. In order to make the analysis tractable, the interaction or influence structure is described by an explicit sampling process where two extreme cases have been considered: the unbiased-degree and the biased-degree case. In the former case out-degree (information level) and expected in-degree (visibility level) of agents are uncorrelated, whereas in the latter case these two measures coincide. López-Pintado (2012) and Jackson and López-Pintado (2013) extend this idea to comprise a wide array of sampling options depending on the level of correlation assumed between agent's in and out degree. The main focus of most of the work surveyed in this chapter, however, is to discuss the hypothesis that more dense and heterogeneous networks always favor diffusion, something which is true for standard epidemiology models but that does not generalize to other models of diffusion based on coordination and imitation behavioral rules.

We have also tried to understand the effect that homophily has on diffusion, concentrating on the concept of the diffusion threshold. That is, the spreading to a significant fraction of the population of a new behavior when starting with a small initial seed. Nevertheless, there are other issues which are not addressed here, but that are relevant. For example, one could evaluate the size of the adoption endemic state as a function of the homophily level.

Acknowledgements I want to thank Juan D. Moreno-Ternero for his helpful comments. Financial support from the Spanish Ministry of Economy and Competitiveness (ECO2014-57413-P) is gratefully acknowledged.

This chapter is based upon work from COST Action ISCH COST Action IS1104 "The EU in the new complex geography of economic systems: models, tools and policy evaluation", supported by COST (European Cooperation in Science and Technology), www.cost.eu.

References

Bailey NTJ (1975) The mathematical theory of infectious diseases. Griffin, London
Bala V, Goyal S (1998) Learning from neighbors. Rev Econ Stud 65:595–621
Benaïm M, Weibull JW (2003) Deterministic approximation of stochastic evolution in games. Econometrica 71:873–903
Bollobás B (2001) Random graphs. Cambridge University Press, Cambridge
Conley TG, Udry CR (2001) Social learning through networks: the adoption of new agricultural technologies in Ghana. Am J Agric Econ 83:668–673
Currarini S, Jackson MO, Pin P (2009) An economic model of friendship: homophyly, minorities, and segregation. Econometrica 77:1003–1045
Erdös P, Rényi A (1959) On random graphs. Publ Math Debr 6:290–297
Galeotti A, Goyal S (2009) A theory of strategic diffusion. Rand J Econ 40:509–532
Galeotti A, Goyal S, Jackson MO, Vega-Redondo F, Yariv L (2010) Network games. Rev Econ Stud 77:218–244
Golub B, Jackson MO (2012a) How homophily affects the speed of learning and best response dynamics. Q J Econ 127:1287–1338
Golub B, Jackson MO (2012b) Does homophily predict consensus times? Testing a model of network structure via a dynamic process. Rev Netw Econ 11:Article 9
Golub B, Jackson MO (2012c) Network structure and the speed of learning: measuring homophily based on its consequences. Ann Econ Stat 107/108:33–35
Granovetter MS (1978) Threshold models of collective behavior. Am J Sociol 83:1420–1443
Jackson MO, López-Pintado D (2013) Diffusion and contagion in networks with heterogeneous agents and homophily. Netw Sci 1:49–67
Jackson MO, Rogers B (2007) Relating network structure to diffusion properties through stochastic dominance. B.E. J Theor Econ (Advances) 7:1–13
Jackson MO, Yariv L (2007) Diffusion of behavior and equilibrium properties in network games. Am Econ Rev (Papers and Proceedings) 97:92–98
López-Pintado D (2006) Contagion and coordination in random networks. Int J Game Theory 34:371–381
López-Pintado D (2008) Diffusion in complex social networks. Games Econ Behav 62:573–90
López-Pintado D (2012) Influence networks. Games Econ Behav 75:776–787

López-Pintado D, Watts DJ (2008) Social influence, binary decisions and collective dynamics. Ration Soc 20:399–443
Morris S (2000) Contagion. Rev Econ Stud 67:57–78
Pastor-Satorrás R, Vespignani A (2001a) Epidemic spreading in scale-free networks. Phys Rev Lett 86:3200–3203
Pastor-Satorrás R, Vespignani A (2001b) Epidemic dynamics and endemic states in complex networks. Phys Rev E 63:066117
Price DJS (1965) Networks of scientific papers. Science 149:510–515
Rogers EM (1995) Diffusion of innovations. Free Press, New York
Watts DJ (2002) A simple model of information cascades on random networks. Proc Natl Acad Sci 99:5766–5771
Young HP (2009) Innovation diffusion in heterogeneous populations: contagion, social influence, and social learning. Am Econ Rev 99:1899–1924

Opinion Dynamics on Networks

Ugo Merlone, Davide Radi, and Angelo Romano

Abstract Sociophysics has devoted a lot of attention to social influence and opinion dynamics. Among the others, the pioneering works by Galam, where agents randomly gather in groups of different size until consensus is reached, have been used to analyze the spreading of rumors. Galam's model however, considers only special kinds of social spaces. In this chapter we survey some of the most recent contributions on opinion dynamics, illustrate Galam's model of rumor diffusion and extend it to consider more general networks.

Keywords opinion dynamics • social networks • Galam's model • rumor diffusions • majority influence

1 Introduction

The study of opinion dynamics and consensus is a fascinating research area studied by several disciplines. For example, social psychology has devoted considerable attention in investigating how majority and minority can influence individuals' opinions (Asch 1956; Moscovici et al. 1969). Consensus has also been analyzed through mathematical modeling, as for example in DeGroot (1974) and Berger

U. Merlone (✉)
Department of Psychology, Center for Cognitive Science, University of Torino, via Verdi 10, I 10124 Torino, Italy
e-mail: ugo.merlone@unito.it

D. Radi
School of Economics and Management, LIUC - Università Cattaneo, Castellanza, Lombardy, Italy
e-mail: davide.radi@unibg.it

A. Romano
Department of Psychology, University of Torino, via Verdi 10, I 10124 Torino, Italy

Department of Social and Organizational Psychology, VU University Amsterdam, Van der Boechorststraat 1, 1081 BT, Amsterdam, The Netherlands
e-mail: a.romano@unito.it

P. Commendatore et al. (eds.), *Complex Networks and Dynamics*, Lecture Notes in Economics and Mathematical Systems 683,
DOI 10.1007/978-3-319-40803-3_3

(1981). These approaches investigate the processes and the factors interacting to have consensus in groups. Furthermore, there are models analyzing social influence and strategic interaction where agents' opinion is influenced by the others. For example in Buechel et al. (2013), an analysis of how cultural factors change individual opinions is provided in an overlapping setting and within a game theoretical framework. According to this stream of research consensus (conformity) on a specific opinion is reached as the results of strategic interactions among agents.

Another effort in the analysis of processes of social influence has been provided by sociophysics. The backbone of this approach consists of the contributions by Galam, see, e.g., Galam (1990, 2003, 2012), which provide a theoretical justification for the frequently observed phenomenon of the diffusion of rumors within a population such that the great majority of its members refuses to believe in it. For example, in Galam (2003), there is a formalization of the spread of the notice in which it was claimed that no plain crashed in the Pentagon on September 11. In Galam's model, individuals have two possible opinions about a rumor. One opinion is '$-$', individuals believe in the rumor, the other opinion is '$+$', individuals do not believe in the rumor. The rumor at stake is the object of the discussion within different social spaces where members—separately in each group—discuss the issue. Within each social space individuals line up to the opinion of the majority. In case no majority can be reached, a bias is assumed in favor of '$-$'. The social structure of the bias and the particular structure of the social space determine the spreading of the rumor despite a small minority of the population believe in its truthfulness.

In this chapter, we embed Galam's model into a network perspective, showing how it becomes a particular case in which networks consist of a collection of connected components that are cliques. Then, we use the networks to describe more complicated social spaces where agents can interact and exchange opinions about the rumor at stake even if seated at different tables of the social space.

The road map of the Chapter is as follows. First, we present a brief review of opinion dynamics on networks. Second, we present the features of Galam's model of rumor spreading. Third, we show how the social spaces considered in Galam's model can be seen as special cases of networks and extend this model to more general ones. Finally, the results of the simulation and implications for future research will be discussed.

2 Networks and Opinion Dynamics on Networks

In the last 50 years complex systems approaches have been widely employed to understand, simulate, and analyze complex systems (Knoke and Kuklinski 1982; Newman 2010). Among the others, network analysis has a particular attention on the patterns of connections in a complex system (Newman 2010). Basically, in network analysis complex systems are modeled as networks and consist of points linked by lines (Newman 2010). In graph theory terminology, points are called *nodes* or vertices and lines are called *edges* (Knoke and Yang 2008). Nodes are the elements and actors of a specific system while edges are connections among them. Some suitable example of networks are the internet, the world wide web, and the social networks (Newman 2010). For example, concerning the world wide web, nodes can be considered the web pages while edges are the hyperlinks to switch from one page to another.

The focus of network analysis is on relational data (Knoke and Yang 2008; Knoke and Kuklinski 1982) because of the analysis of patterns of connection in which it is possible to understand the basics of the structure of complex networks. Network analysis can provide a further understanding of the regularities of these connection patterns as well as the absence of relation among them (Newman 2010). For this reason, the attributes of nodes within a network are not relevant in this kind of analysis. Nevertheless, network analysis allows to understand how relations affect individual behavior and the performance of a system.

Some important concepts of complex networks are *centrality*, *small world effect*, and *communities* (Newman 2010). First, the property of centrality allows us to have a picture of how important are some vertices rather one other. Among the other measures of centrality, the *degree* of a node is a measure OF the number of lines incident with it. Using this approach, it is possible to identify some central elements within a system (Knoke and Yang 2008). Second, the small world effect (Milgram 1967; Travers and Milgram 1969) refers to the fact that some networks exhibit properties such as high clustering path lengths and short average (Watts and Strogatz 1998). Third, another property which can arise from network analysis is the presence of specific subnetworks in a network. The possibility to catch the presence of subclusters, and links between these subclusters provides a further understanding on the presence of cohesive subnetworks and thus on how a system is structured (Newman 2010).

In particular, among the most influential network group ideas, we list *cliques* which can be defined as a maximal complete subgraphs of three or more nodes and can be thought of "as a collection of actors all of whom choose each other, and there is no other actor in the group who also 'chooses' and is 'chosen' by all of the members of the clique" (Wasserman and Faust 1994, p. 254). In the following we will see the role of cliques in a well known model of opinion dynamics.

Network analysis has been widely implemented to analyze opinion dynamics, see for example Valente (1996), Amblard and Deffuant (2004), and Altafini (2012).

In Valente (1996), the Author examines collective phenomena as collective action and diffusion of innovation. In this case, network analysis is employed to understand the presence of opinion leaders who can attract opinion, to change behavioral boundaries of influences and to predict trajectories of diffusion.

As argued in Amblard and Deffuant (2004) one of the most used model of opinion formation is the Ising-like influence dynamics of binary or discrete opinions. In this paper the Authors analyze the influence of extreme beliefs on opinion dynamics. In particular, they investigate the emergence of centrism or extremism in small world networks. Some application of this model on opinion dynamics in social networks can be found in Barrat and Weigt (2000) and Kuperman and Zanette (2002). In both Barrat and Weigt (2000) and Kuperman and Zanette (2002), the Ising model is applied in small-world networks and provides further understanding of opinion dynamics under situation of noise, disorder and imitation.

In Altafini (2012) opinions are analyzed in order to understand how the structure of social networks communities can lead opinions to be divided in two main polarized factions. In this kind of analysis individuals are the network nodes and the relations in the networks are represented by edges. By contrast, in Benczik et al. (2009) the Authors apply network analysis to the study of voter dynamics. In particular, they investigate how relationships can change according to interpersonal relations, how preferences affect the strength of connection within individuals in the network and provide some insight into the analysis of consensus and polarization. A model of the influence of conformity processes on opinion formation is presented also in Buechel et al. (2015). There, the Authors present a model of opinion formation in a social network in which individuals' opinions are affected by weighted average of opinions hold by their neighbours. This model provides a further understanding of how conformity processes influence opinion leadership and the quality of information aggregation. Another contribution which analyzes the influence of different topologies on opinion dynamics is Weisbuch (2004). Moreover, in Lorenz and Urbig (2007) different communication rules and strategies are analyzed in order to understand consensus. Other interesting contributions about the interplay between opinion dynamics and network topologies are Fortunato (2004), and Fortunato (2005). In Stauffer and Meyer-Ortmanns (2004), opinion dynamics are analyzed considering people connected in scale-free networks. For a survey on social networks analysis and opinion dynamics see also Lorenz (2007).

3 Galam's Model

As discussed in the introduction, Galam's model provides a formalization to understand how false information (specifically hoaxes or rumors) can be propagated within a population (Galam 2003). In the Galam's model we have different social gatherings and discussions. At the beginning, at time 0, individuals having one of two possible opinions ('+' and '−') meet in different groups which can be thought as tables with a fixed number of seats. In the course of time, $t = 1, 2, \ldots$ the same individuals will meet again and again randomly in the same cluster configuration of size groups, and change their opinion according to the majority influence exerted within the specific group they belong to. Within each group the discussion is assumed to follow a majority rule dynamics, with a bias in favor of '−' in case of a local doubt. Interestingly, this model shows how an opinion—despite being supported by a minority—can, after several social gatherings, be propagated to the whole population.

Formally, in Galam's model we have a N-person (finite) population having only two opinions, '+' and '−'. At time $t = 0, 1, 2, 3, \ldots$ individuals seat randomly at different tables $\mathscr{T} = \{T_1, T_2, \ldots, T_L\}$ with $L < N$. The discussion takes place within the people seated at the same tables. Two individuals seating at two different tables do not share any idea or opinion about the subject at stake. Each table has a limited number of seats, given by $|T_r|$, and $\sum_{T_r \in \mathcal{T}} |T_r| = N$. The vector $\mathbf{n} = (|T_1|, |T_2|, \dots, |T_L|) \in \mathbb{R}^L$, indicating the number of seats for each table, is known as *social space*, see Merlone and Radi (2014).

The social space does not change over time and an individual has probability

$$a_k = \frac{k}{N} \sum_{T_r \in \mathcal{T}} \delta_{k,|T_r|} \quad \text{where } \delta_{k,|T_r|} \text{ is the Kronecker's delta,} \tag{1}$$

of seating at a table of size $k = 1, 2, \dots, K$, where K is the size of the largest table of the social space. Given the social space $\mathbf{n}$, vector $\mathbf{y} = (y_1, y_2, \dots, y_L)$ indicates a generic seating configuration where $y_r \in [0, |T_r|]$, $r = 1, 2, \dots, L$, is the number of agents with opinion '+' seated at table T_r.

As shown in Merlone and Radi (2014), assuming $y \in S = \{0, 1, \dots, N\}$ agents with opinion '+', the probability of a seating configuration $\mathbf{y}$, can be computed as follows:

$$P_y(\mathbf{y}) = \frac{\binom{|T_1|}{y_1} \cdots \binom{|T_L|}{y_L}}{\binom{N}{y}} \quad \text{where } \binom{\cdot}{\cdot} \text{ are binomial coefficients} \tag{2}$$

and this $\forall \mathbf{y} \in \Omega_y$, where Ω_y is the set of all possible seating configurations of y agents with opinion '+' given the social space $\mathbf{n}$:

$$\Omega_y = \left\{ \mathbf{y} : 0 \le y_r \le \min(|T_r|, y), r = 1, 2, \dots, L \text{ and } \sum_{r=1}^{L} y_r = y \right\}$$

Once seated, at each table of the social space individuals line up with a consensual opinion. In particular, they agree with the majority at the table they seat. As in Galam (2003), when opinion is exactly split at an even size table the outcome is determined assuming a bias in favor of opinion '−'. This change of opinion is described by the majority rule discussion function D

$$D(y_r, |T_r|) = \begin{cases} |T_r| & \text{if } y_r \ge \left\lfloor \frac{|T_r|}{2} \right\rfloor + 1 \\ 0 & \text{if } y_r \le \left\lfloor \frac{|T_r|}{2} \right\rfloor \end{cases} \qquad r = 1, \dots, L$$

Therefore, the probability to move from a state i to a state j, where i and j is the number of agents having opinion '+', is given by

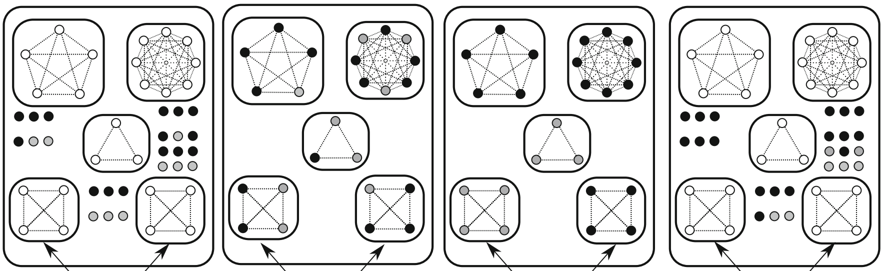


Fig. 1 A one step opinion dynamics. First stage, people sharing the two opinions are moving around. *Gray* have opinion '−' while *black* have opinion '+'. No discussion is occurring with 9 *gray* and 15 *black*. Second stage right, people take place at the tables the social space consists of. Third stage, within each group consensus has been reached. As a result, they are now 7 *gray* and 17 *black*. Last stage, people are again moving around with no discussion

$$p_{i,j} = \sum_{\mathbf{y}\in\omega_{i,j}} P_i(\mathbf{y}) \tag{3}$$

where P_i is defined in (2) and $\omega_{i,j} \subseteq \Omega_i$ is the set of seating configurations such that, starting with i agents with opinion '+', after the discussion j agents end up with opinion '+':

$$\omega_{i,j} = \left\{ \mathbf{y} : \mathbf{y} \in \Omega_i \text{ and } \sum_{r=1}^{L} D\left(y_r, |T_r|\right) = j \right\} \tag{4}$$

Figure 1 provides an example of how Galam's model works with a population of 24 agents and social space $\mathbf{n} = (3, 4, 4, 5, 8)$. According to $\mathbf{n}$ entries, the social space consists of 6 tables, with respective number of seats 3, 4, 4, 5 and 8.

Taking place at each time $t = 1, 2, 3, \ldots$, the described mechanism of diffusion of rumors can be formalized as an homogeneous Markov process $(Y^t)_{t\geq 0}$ where Y^t is a random variable representing the number of agents with opinion '+' at time t. In particular, from the transition probability p_{ij}, $\forall i, j \in S$, it is possible to define the transition matrix $M_{\mathbf{n}} = \left(p_{i,j}\right)_{i,j\in S}$, from which we have that the global dynamics of such a process is described by the following theorem (see Merlone and Radi 2014 for details):

Theorem 1 (Merlone and Radi 2014) *The stochastic process that describes the diffusion of the opinions '+' and '−' in a population with N agents according to Galam's model is an absorbing Markov chain with the property that consensus on opinion '+', state N (all agents of the population believe that the rumor is fake), and consensus on opinion '−', state* 0 *(all agents believe in the rumor), are absorbing states. Another absorbing state, namely N/*2*, exists if and only if the social space* *consists of two identical tables with an odd number of seats. No other absorbing state exists.*

This theorem specifies that, excluding the special case of a social space consisting of two identical tables with odd number of seats, the possible long-run outcome of the Galam's model is consensus either on opinion '$-$' or on opinion '$+$'. The probability to converge to one of such absorbing states depends on the initial distribution of opinions among the members of the population. In this respect, Theorem 1 allows us to use standard technique on absorbing Markov chain, see, e.g. Ibe (2009), to compute the *absorption probability* (the probability that the chain that starts in a transient state will be absorbed in a specific absorbing state) and the *mean time to absorption* (the expected number of transitions until the process reaches an absorbing state, given that it starts from a particular transient state).

Let us consider the example depicted in Fig. 1, for which the absorbing Markov chain is characterized by only two absorbing states: N, consensus on opinion '$+$' and 0, consensus on opinion '$-$'.

Although as long as we deal with the original Galam's model, it is possible to rely on the closed form expression derived in Merlone and Radi (2014), when considering more complex social spaces makes such a calculation non trivial. Therefore, as suggested in Merlone et al. (2015), we use simulation to compute the transition probability matrix. In particular, for each initial state from 0 to N, we randomly assign the agents to the seats and compute the state of the process at the next time. This is replicated 100,000 times in order to provide an accurate estimation of the transition probabilities (3). This way we obtain the transition matrices. For example, when considering the social space introduced in Fig. 1, the transition matrix is reported in (5).

$$M = \begin{bmatrix}
1.00 & 0.00 \\
1.00 & 0.00 \\
0.99 & 0.00 & 0.00 & 0.01 & 0.00 \\
0.96 & 0.00 & 0.00 & 0.03 & 0.00 \\
0.91 & 0.00 & 0.00 & 0.06 & 0.01 & 0.02 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.82 & 0.00 & 0.00 & 0.10 & 0.04 & 0.04 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.71 & 0.00 & 0.00 & 0.14 & 0.07 & 0.07 & 0.00 & 0.00 & 0.01 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.57 & 0.00 & 0.00 & 0.17 & 0.10 & 0.11 & 0.00 & 0.01 & 0.03 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.41 & 0.00 & 0.00 & 0.19 & 0.14 & 0.15 & 0.00 & 0.02 & 0.07 & 0.01 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.25 & 0.00 & 0.00 & 0.20 & 0.16 & 0.17 & 0.00 & 0.05 & 0.13 & 0.03 & 0.00 & 0.01 & 0.01 & 0.01 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.11 & 0.00 & 0.00 & 0.17 & 0.16 & 0.16 & 0.00 & 0.08 & 0.19 & 0.06 & 0.00 & 0.03 & 0.03 & 0.02 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.03 & 0.00 & 0.00 & 0.11 & 0.12 & 0.13 & 0.00 & 0.11 & 0.22 & 0.09 & 0.00 & 0.06 & 0.07 & 0.05 & 0.00 & 0.01 & 0.01 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.05 & 0.06 & 0.07 & 0.00 & 0.12 & 0.21 & 0.12 & 0.00 & 0.09 & 0.14 & 0.08 & 0.00 & 0.02 & 0.03 & 0.01 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.01 & 0.01 & 0.02 & 0.00 & 0.09 & 0.15 & 0.11 & 0.00 & 0.10 & 0.20 & 0.11 & 0.00 & 0.06 & 0.08 & 0.05 & 0.00 & 0.00 & 0.01 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.04 & 0.07 & 0.06 & 0.00 & 0.09 & 0.21 & 0.10 & 0.00 & 0.11 & 0.16 & 0.10 & 0.00 & 0.01 & 0.04 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.01 & 0.02 & 0.02 & 0.00 & 0.05 & 0.17 & 0.07 & 0.00 & 0.11 & 0.21 & 0.13 & 0.00 & 0.04 & 0.14 & 0.03 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.02 & 0.09 & 0.04 & 0.00 & 0.08 & 0.22 & 0.11 & 0.00 & 0.06 & 0.27 & 0.07 & 0.00 & 0.00 & 0.04 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.03 & 0.01 & 0.00 & 0.04 & 0.17 & 0.08 & 0.00 & 0.07 & 0.35 & 0.09 & 0.00 & 0.00 & 0.14 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.01 & 0.00 & 0.00 & 0.01 & 0.10 & 0.04 & 0.00 & 0.06 & 0.37 & 0.10 & 0.00 & 0.00 & 0.31 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.04 & 0.02 & 0.00 & 0.04 & 0.31 & 0.08 & 0.00 & 0.00 & 0.51 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.01 & 0.00 & 0.00 & 0.02 & 0.22 & 0.06 & 0.00 & 0.00 & 0.69 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.01 & 0.12 & 0.03 & 0.00 & 0.00 & 0.84 \\
0.00 & 0.04 & 0.01 & 0.00 & 0.00 & 0.95 \\
0.00 & 1.00 \\
0.00 & 1.00
\end{bmatrix}. \quad (5)$$

From such a matrix, following Kemeny and Snell (1960), it is possible to compute the absorption probabilities and expected time of convergence to consensus as in Merlone and Radi (2014). They are represented graphically together with the

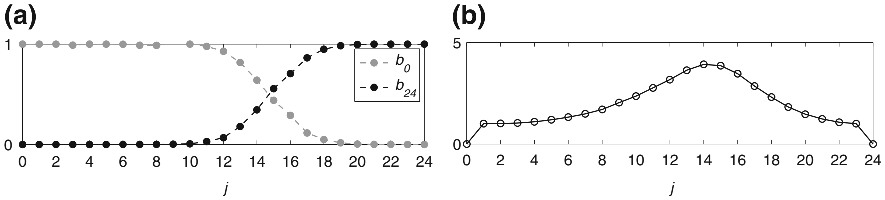


Fig. 2 Figure 1 example: (**a**) absorbing probability to states 0 (*gray circles*) and 24 (*black circles*); (**b**) expected time of convergence to consensus (either 0 or 24)

expected time of convergence to the absorbing states. The left side of these figures represents the probabilities to converge, starting from state $j \in S$, to the respective absorbing states. The right side of the figures represents the expected time, z_j, of convergence to absorbing states for each initial state $j \in S$. As we are considering discrete states the dashed lines are there for illustrative purpose only. In this case, Fig. 2a depicts the absorption probabilities of state $N = 24$ (0) by black dots (gray dots) for each state. We observe that for all states smaller or equal than 9, probability to converge to absorbing states 0 and 24 are 1 and 0 respectively. On the other hand, from states equal or larger than 21 the probability to converge to the absorbing state 0 and 24 are 0 and 1 respectively. As it concerns the remaining transient states the probability to converge to either absorbing state is strictly positive, and there is uncertainty about the final outcome of the process of opinion diffusion. Clearly, the probability to converge to the absorbing state $N = 24$ increases as the initial number of agents with opinion'+' increases and, conversely, the probability to converge to the absorbing state $N = 0$ decreases. By the absorbing probabilities we can identify the so-called *stochastic killing point*, i.e., the larger initial state for which the probability to be absorbed into the state 0 is equal to or larger than the probability to be absorbed into the state N, see, e.g. Galam (2003) and Merlone and Radi (2014). For a given N-person population, the *stochastic killing point* depends on the different configurations of the social space and is used as an indicator of the rumor diffusion. In particular, the larger the killing point, the higher is the possibility that the rumor spreads among the population. The other important information is depicted in Fig. 2b and it is the mean time to absorption, that is the time required to the rumor either to spread to all members of the population or to decay. As shown in Fig. 2b, starting with 9 (or less) agents not believing in the rumor, on the average, it takes not more than two iterations for a rumor to spread; on the contrary, starting with 21 (or more) not believing in the rumor, on the average, no more than two iterations are necessary for the rumor decay. Finally, when the initial number of agents not believing in the rumors is between 9 and 21, the average number of iterations to converge to a final consensus is 5 iterations, although the final consensus cannot be determined a priori. These convergence times are consistent with Galam (2003), which indicate that a rumor either spreads or decays quickly among a population.

4 Extending Galam's Model to Networks

The social space in Galam's model can be considered as a social network in which the nodes represent the seats and the connected components are cliques and represent the tables. This way agents communicate only if seated at nodes belonging to the same clique. Considering the example depicted in Fig. 1, the network consists of 5 cliques with respectively 3, 4, 4, 5 and 8 nodes. We remark that the network implicit in Galam's model does not change over time; by contrast, the assignment of each agent to a single node (seat) is randomly determined at each time period. Specifically, each agent has the same probability to be assigned to any node of the network and this probability is constant over time.

As social networks can be represented using a graph theoretic notation (Wasserman and Faust 1994, p. 95) we use such a notation to formalize our model. In particular, a social network with an undirected dichotomous relation can be modeled by a graph $\mathscr{G}(\mathscr{N},\mathscr{L})$ where $\mathscr{N}$ is a set of nodes—in our case the seats—and a set of lines $\mathscr{L}$ between pairs of nodes. In our model we have a line between two seats if and only if agents seated at those seats can communicate. In order to formalize social spaces we recall some basic definitions.

Definition 1 (Subgraph (Wasserman and Faust 1994, p. 97)) A graph $\mathscr{G}_S$ is a *subgraph* of $\mathscr{G}$ if the set of nodes of $\mathscr{G}_S$ is a subset of the set of nodes of $\mathscr{G}$, and the set of lines in $\mathscr{G}_S$ is a subset of the set of lines in $\mathscr{G}$.

Definition 2 (Complete Graph (Wasserman and Faust 1994, p. 102)) If all lines are present, then all nodes are adjacent, and the graph is said to be *complete*.

Definition 3 (Walk (Wasserman and Faust 1994, p. 105)) A *walk* is a sequence of nodes and lines, starting and ending with nodes in which each node is incident with the lines following and preceding it in the sequence.

Definition 4 (Path (Wasserman and Faust 1994, p. 107)) A *path* is a walk in which all nodes and all lines are distinct.

Definition 5 (Connected Graph (Wasserman and Faust 1994, p. 109)) A graph is *connected* if there is a path between every pair of nodes in the graph.

Definition 6 (Component (Wasserman and Faust 1994, p. 109)) The connected subgraphs in a graphs are called *components*. A component of a graph is a maximal connected subgraph.

Definition 7 (Clique (Wasserman and Faust 1994, p. 254)) A *clique* in a graph is a maximal complete subgraph of three or more nodes.

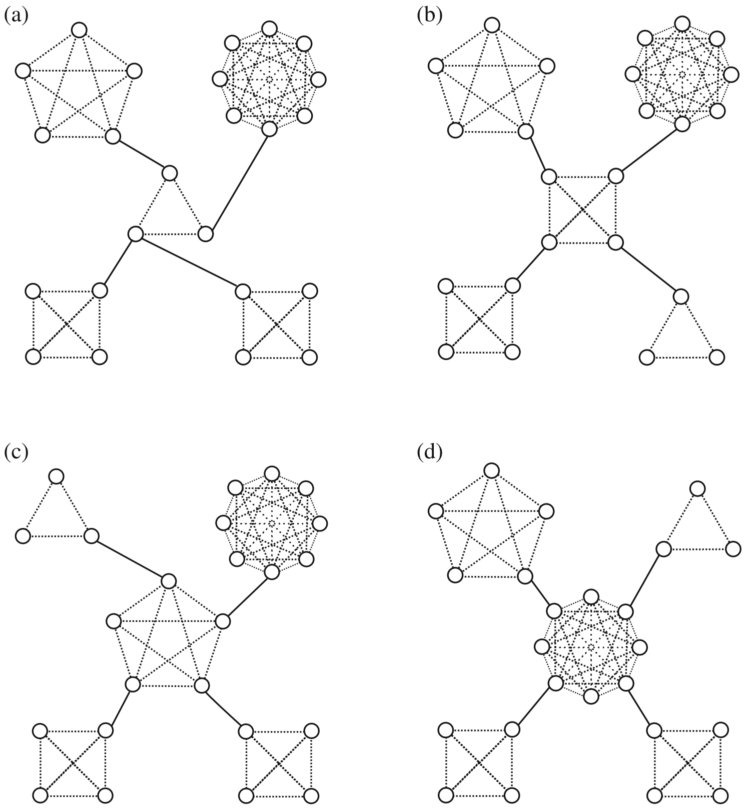


Fig. 3 Some general networks obtained by adding further links to the one considered in Fig. 1. (**a**) All tables connected to a 3-size table. (**b**) All tables connected to the 4-size table. (**c**) All tables connected to the 5-size table. (**d**) All tables connected to the 8-size table

With these definitions it follows

Proposition 1 *Any social spaces considered in Galam (2003), Bischi and Merlone (2010), and Merlone and Radi (2014), can be formalized as a network in which all components are cliques.*

As a consequence of this formalization it is evident that more general networks can be obtained adding further links as illustrated in Fig. 3.

We illustrate how these additional links may affect the dynamics considering a particular opinion configuration in the social space illustrated in Fig. 1. We can assess the role of the additional links by comparing the Figs. 4 and 5. In both cases the initial opinion is 9 agents with opinion '−' (gray) and 15 with opinion '+' (black). When no additional links are considered we end up with 12 agents with each opinion (Fig. 4), on the contrary the mere presence of four links makes the final configuration consisting of respectively 14 gray and 10 black agents (Fig. 5).

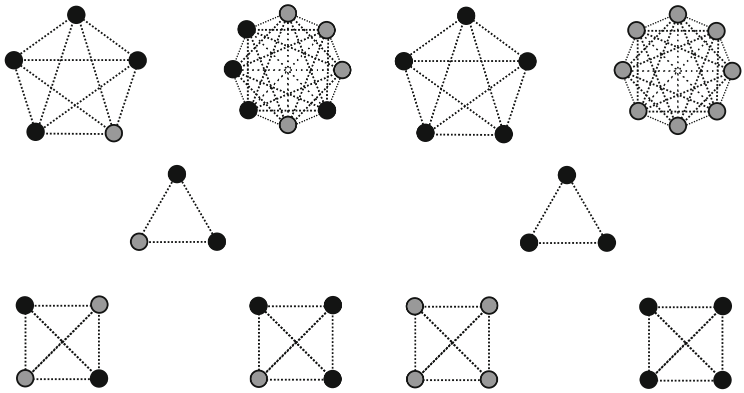


Fig. 4 A one step opinion dynamics starting from 9 agents with opinion '−' (*gray*) and 15 with opinion '+' (*black*) and no additional link between tables

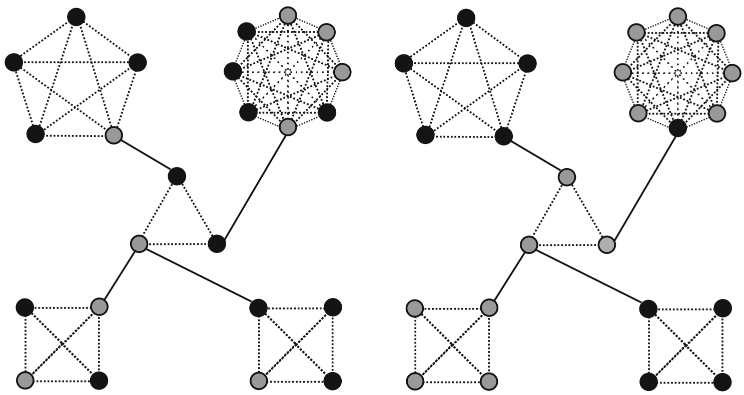


Fig. 5 A one step opinion dynamics starting from 9 agents with opinion '−' (*gray*) and 15 with opinion '+' (*black*) and four additional links between tables

Obviously, these additional links affect the discussion and increase the complexity of the model. Indeed, as illustrated in Fig. 5, after discussion, some agents may end up with an opinion that is different from those of the other members of the clique they would belong to if no additional links were considered. This occurs, because agents with links to other tables may be influenced by agents not belonging to their clique. This is clearly an element of novelty with respect to Galam's original model.

As we have seen the social spaces considered in Galam's original model (Galam 2003) are special cases of networks in which all components are cliques. As we allow influence between agents seating at different tables the network density increases as well the complexity of interaction.

This further complexity influences the opinion dynamics as these additional links may affect the number of situations in which the bias in favor of opinion '−' plays a role. We recall that, following Galam's framework, the bias of the majority rule function is the agents propensity to believe the rumor (opinion '−') when there is not a clear majority within the discussion group that does not believe in the rumor. This bias, together with the configuration of network or discussion groups, affects the killing point and is responsible for the possible diffusion of the rumor despite only a minority of the entire population believe in it.

It is interesting to observe how the killing point and the expected time of convergence of the network considered in Fig. 1 are affected when different additional links are considered as illustrated in Fig. 3. Nevertheless, we do not expect large differences as the diameter and the density of the networks with additional links is the same; furthermore such a density is slightly larger than the one of the original network. As mentioned above, when considering large social spaces we have to resort to simulation. This holds especially when introducing new links as, in this case, the calculation of transition probability is not trivial. Therefore, for the networks considered in Fig. 3 we used simulation. Also in this case we run 100,000 runs in order to obtain estimations of the transition probabilities (3); for the sake of brevity we omit the matrices. The killing point and the expected time of convergence for each social network, are presented in Figs. 6, 7, 8, and 9.

When comparing Fig. 2a to Figs. 6a, 7a, 8a, and 9a, the additional links have some effects on the killing point which moves around 15. By contrast, comparing Fig. 2b to Figs. 6b, 7b, 8b, and 9b, we can notice that the additional links introduced in the networks depicted in Figs. 3a–d do not reduce significantly the expected number of steps to convergence to an absorbing state. As matter of facts, the time to converge to an absorbing state depends on the degree of connectivity of the network; for example, in a network in which all nodes are directly connected convergence would occur in a single step. In this respect, the density of the networks represented in Fig. 3a–d is only slightly larger than the one without additional links. This justifies the small differences in terms of the time to converge to absorbing states.

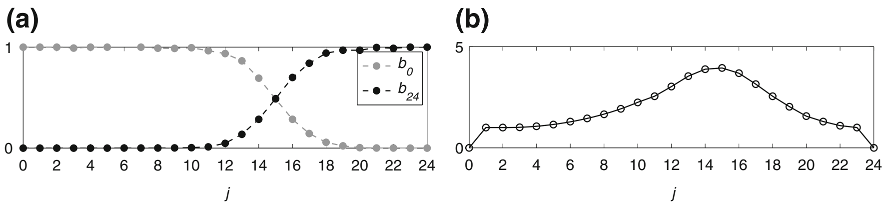


Fig. 6 Case of all tables connected to the 3-size table (Fig. 3a). Depending on the initial state $j \in S$: (**a**) probability to converge to the absorbing states 0 (*gray circles*) and 24 (*black circles*); (**b**) expected time of convergence to consensus (either 0 or 24)

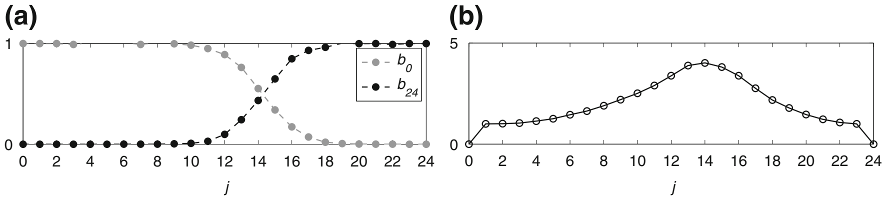


Fig. 7 Case of all tables connected to the 4-size (Fig. 3b). Depending on the initial state $j \in S$: (**a**) probability to converge to the absorbing states 0 (*gray circles*) and 24 (*black circles*); (**b**) expected time of convergence to consensus (either 0 or 24)

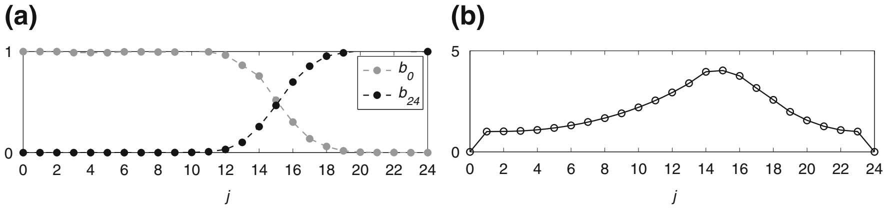


Fig. 8 Case of all tables connected to the 5-size table (Fig. 3c). Depending on the initial state $j \in S$: (**a**) probability to converge to the absorbing states 0 (*gray circles*) and 24 (*black circles*); (**b**) expected time of convergence to consensus (either 0 or 24)

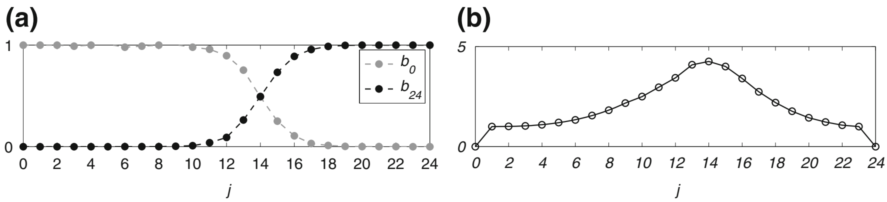


Fig. 9 Case of all tables connected to the 8-size table (Fig. 3d). Depending on the initial state $j \in S$: (**a**) probability to converge to the absorbing states 0 (*gray circles*) and 24 (*black circles*); (**b**) expected time of convergence to consensus (either 0 or 24)

5 Conclusions and Further Research

We extended Galam's opinion spreading model in order to consider more complex social spaces such as networks. The original model can be obtained as particular case in which the connected components of the network are cliques.

With the proposed network representation it is possible to extend the discussion to more general social spaces. This way, Galam's model can be extended to consider opinion diffusion on general networks and it is possible to consider more *small world* like situations (Milgram 1967; Travers and Milgram 1969). Analyzing some examples of possible configurations of social spaces which consider forms of inter table discussion, we provide evidence that this kind of interactions can alter the dynamics of rumor diffusion. Moreover, we show how adding even a small number of links may affect the opinion diffusion.

Our contribution represents a first attempt to generalize the social space as proposed in Galam in order to consider more complex structures. In particular, the use of networks to specify a social space offers the possibility to analyze several discussion groups configuration. A detailed investigation of the effects deriving by considering different network configurations is left to further research. Given the complexity of interactions, writing the closed formulas for the transition probabilities can be cumbersome. Therefore, as illustrated by the examples in this chapter, a strategy to analyze rumors diffusion on such networks is given by simulation.

Acknowledgements This Chapter is based upon work from COST Action ISCH COST Action IS1104 "The EU in the new complex geography of economic systems: models, tools and policy evaluation", supported by COST (European Cooperation in Science and Technology) www.cost.eu.

References

Altafini C (2012) Dynamics of opinion forming in structurally balanced social networks. PloS One 7(6):e38135

Amblard F, Deffuant G (2004) The role of network topology on extremism propagation with the relative agreement opinion dynamics. Phys A Stat Mech Appl 343:725–738

Asch SE (1956) Studies of independence and conformity: I. A minority of one against a unanimous majority. Psychol Monogr Gen Appl 70(9):1–70

Barrat A, Weigt M (2000) On the properties of small-world network models. Eur Phys J B-Condensed Matter Complex Syst 13(3):547–560

Benczik I, Benczik S, Schmittmann B, Zia R (2009) Opinion dynamics on an adaptive random network. Phys Rev E 79(4):046104

Berger LR (1981) A necessary and sufficient condition for reaching a consensus using DeGroot's method. J Am Stat Assoc 76(374):415–418

Bischi GI, Merlone U (2010) Binary choices in small and large groups: a unified model. Phys A Stat Mech Appl 389(4):843–853

Buechel B, Hellmann T, Klößner S (2013) Opinion dynamics under conformity. Working Paper, Institute of Mathematical Economics, Bielefeld University

Buechel B, Hellmann T, Klößner S (2015) Opinion dynamics and wisdom under conformity. J Econ Dyn Control 52:240–257

DeGroot MH (1974) Reaching a consensus. J Am Stat Assoc 69(345):118–121

Fortunato S (2004) Universality of the threshold for complete consensus for the opinion dynamics of Deffuant et al. Int J Mod Phys C 15(09):1301–1307

Fortunato S (2005) On the consensus threshold for the opinion dynamics of Krause–Hegselmann. Int J Mod Phys C 16(02):259–270

Galam S (1990) Social paradoxes of majority rule voting and renormalization group. J Stat Phys 61(3–4):943–951 (1990)

Galam S (2003) Modelling rumors: the no plane Pentagon French hoax case. Phys A Stat Mech Appl 320:571–580

Galam S (2012) Sociophysics, a physicist's modeling of psycho-political phenomena. Understanding complex systems. Springer, Berlin (2012)

Ibe OC (2009) Markov processes for stochastic modeling. Elsevier Academic Press, London

Kemeny JG, Snell JL (1960) Finite Markov chains. Springer, New York
Knoke D, Kuklinski J (1982) Network analysis, 1 edn. Sage Publications, Thousand Oak
Knoke D, Yang S (2008) Social network analysis, vol 154. Sage, Thousand Oaks
Kuperman M, Zanette D (2002) Stochastic resonance in a model of opinion formation on small-world networks. Eur Phys J B-Condensed Matter Complex Syst 26(3):387–391
Lorenz J (2007) Continuous opinion dynamics under bounded confidence: a survey. Int J Mod Phys C 18(12):1819–1838
Lorenz J, Urbig D (2007) About the power to enforce and prevent consensus by manipulating communication rules. Adv Complex Syst 10(02):251–269
Merlone U, Radi D (2014) Reaching consensus on rumors. Phys A Stat Mech Appl 406:260–271. doi:10.1016/j.physa.2014.03.048. http://dx.doi.org/10.1016/j.physa.2014.03.048
Merlone U, Radi D, Romano A (2015) Minority influence in opinion spreading. In: Yilmaz L, Chan WKV, Moon I, Roeder TMK, Macal C, Rosetti M (eds) Proceedings of the 2015 winter simulation conference. IEEE
Milgram S (1967) The small world problem. Psychol Today 1:60–67
Moscovici S, Lage E, Naffrechoux M (1969) Influence of a consistent minority on the responses of a majority in a color perception task. Sociometry 32(4):365–380
Newman MEJ (2010) Network Analysis: an Introduction, 1 edn. Oxford University Press, Oxford
Stauffer D, Meyer-Ortmanns H (2004) Simulation of consensus model of Deffuant et al. on a Barabasi–Albert network. Int J Mod Phys C 15(02):241–246
Travers J, Milgram S (1969) An experimental study of the small world problem. Sociometry 32(4):425–443
Valente TW (1996) Social network thresholds in the diffusion of innovations. Soc Netw 18(1):69–89
Wasserman S, Faust K (1994) Social network analysis. Cambridge University Press, Cambridge
Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393:440–442
Weisbuch G (2004) Bounded confidence and social networks. Eur Phys J B-Condensed Matter Complex Syst 38(2):339–343

Econometric Aspects of Social Networks

Mariano Matilla-García and Jesús Mur

Abstract This chapter aims to show why and how is it possible to use econometric analysis tools to the empirical study of social networks, what are the current technical limitations and the open challenges, and what are the theoretical (micro)economic underpinnings of the (few) recent econometric models that serve to analyze social interactions in form of networks, neighborhoods, and groups. The chapter tries also to devise the links between the recent econometric literature (identification and estimation) on networks and the literature in spatial econometrics.

Keywords Interactions • Neighborhoods • Networks • Spatial econometrics • W matrix

1 Introduction

In the last 30 years, there has been a renaissance of interest among economists in the social determinants of individual behavior and aggregate outcomes. Regional, urban, labor and family economics are specific areas that have made use of networks formed by neighborhoods. Broadly speaking, neighborhoods do not require to be defined by geographical proximity, rather they can rely on the notion "social proximity", from which spatial proximity is a particular form of proximity. Social proximity enriches the scope of economic interaction studies, and it does through the existence and conformation of social networks.

Much of the interest has been on peer effects that are widespread in many economic domains, being education and labor economics two of most prominent fields where peer effects have been studied. Peer group influences are understood to

M. Matilla-García (✉)
Departamento de Economía Aplicada Cuantitativa, Universidad Nacional de Educación a Distancia (UNED), Paseo Senda del Rey 11, 28040 Madrid, Spain
e-mail: mmatilla@cee.uned.es

J. Mur
Department of Economic Analysis, Faculty of Economics, Campus Rio Ebro University of Zaragoza, María de Luna, s/n, 50018 Zaragoza, Spain
e-mail: jmur@unizar.es

P. Commendatore et al. (eds.), *Complex Networks and Dynamics*,
Lecture Notes in Economics and Mathematical Systems 683,
DOI 10.1007/978-3-319-40803-3_4

produce imitative behavior due to an intrinsic (genetic) desire to behave like others in the group or due to an economic optimal behavior in the sense that the cost of an action depends on whether the others do the same.

Peer effects are important in as much as there is some kind of social multiplier, which is one of the main interests of this chapter. Consider the following example.[1] A policy maker wants to improve high-school graduation rates and to this end she is going to (randomly) provide college scholarships. She has two possibilities: either to choose students from across a set of high-schools, or to choose them among one or two (but no more) high-schools. It is evident that there will be a direct effect on those selected students, and if only these direct effects are taking into account when making a decision by the policy maker, then the second alternative makes no meaning. However because students of the same school form networks, there can be a more appealing advantage for the second alternative; namely, concentrating the scholarships in one school facilitates that granted students induce peer effects that affect all students in the school, including those who were not granted, but that are in the network. In other words, there is an amplification of the consequences thanks to the existence of a network. Identifying these multiplier is a central objective of the econometric research in social networks. Social networks have applications in other social and economic contexts, and the relevant literature has shown interest is areas where spillovers occur between industries that share some degree of similarity (industrial, technological or locational) or between individuals.

In this chapter we are particularly concerned with econometric modelization and estimation problems that appear when dealing with social networked data sets. We try to provide a synoptic view that may be insightful to theorists and practitioners. Econometric analysis tools for the empirical study of social networks are generally based on linear models. We explore the main current available techniques, and show some technical limitations and challenges. We also try to devise the links between the recent econometric literature on networks and the fairly well established literature based on statistical tools and methods develop in spatial econometrics. In doing so, the role play by sociomatrices are crucial. These matrices are also important when considering the problem of identifying social interactions in various empirical contexts. Somehow econometric identification in networks is a variation of the classical identification problem of simultaneous equations systems that reflect the specific structure of social interactions models.

One reason social networks matter in social interactions analysis is because they facilitate identification by breaking the reflection problem. Unidentification is not a minor problem because, at the end, it refers to the recovering of the structural parameters from the reduced form parameters. If this is not possible, we can not discriminate between the different mechanisms of social interaction, which is the main objective of this literature; at the best, we will be able to test for the existence of social interaction. It is clear that identification is not an econometric issue, but it depends on the fundamentals of the model and the way the sampling information is

[1] The same example has been used throughout the literature on neighborhoods effects.

collected. In this sense, the question of how individuals interact is a key point, as highlighted by Bramoullé et al. (2009), who demonstrate that it is easier to attain identification in the case that individuals interact in groups, like peers in the same classroom, than in a single network, like friendship relationships. Nonlinearities, heterogeneity, measures for the ties among individuals, ... are other factors usually employed in the literature to attain identification.

A typical social interaction model has to cope with, at least, three basic problems: simultaneity, because of the presence of lags of the endogenous variable in the right hand side of the equation; the existence of unobservable effects in the case of group interactions which, if correlated with the observable x variables, results in the *correlated effect* of Manski (1993), and the selection bias or the endogenous group membership problem related to the way individuals in the network form their own reference group. Some innovative lines, appeared recently in this strand of literature, refer to the use of centrality measures in the networks, reflecting that the nodes (individuals) have different bargaining power, and the introduction of dynamics into the models. Both issues raise econometric technicalities that are under discussion in the literature.

To this end Sect. 2 deals with exploring the long tradition in economics in the study of social interactions and the reasons for the current renaissance under the umbrella of social networks. Social networks models provide a elemental focus on the microeconomic structure of interactions emphasizing heterogeneity in these interactions across individual pairs. Section 3 presents the microeconomic fundamentals for agents to interact through a network and shows how is it possible that economic based decisions give rise to the possibility of using linear econometric models, like the well-known linear-in-means models, among others. The correct identification of the model allows for the estimation of peer effects and other direct effects. These effects are then used to the estimation of social multipliers. As stated in Sect. 4, the '*reflection problem*' of Manski (1993) directly points to the identification issue as one of the main concerns in the field, that must be checked, by the econometrician, case-by-case. Section 5 is devoted to connections between micro-foundations and econometric inference on network models. Section 6 focuses on the specification of econometric models of social networking. To some extent, this literature has evolved paralleling the literature on spatial econometrics because the two deal with similar problems in which interaction is a central point. We review the most recent econometric literature, appeared mainly in the last decade, dealing with all these problems. The results seem to be satisfactory in the sense that the researcher has an adequate toolkit to cope with them. Finally, Section7 returns back, once again, to Manski (1993), to the problem of identifying the reference group for each individual. Often, applied literature assumes that the econometrician has this information but this is a very demanding hypothesis. We introduce the key points of a long standing discussion that is taking place, on the same question, in the spatial econometrics literature: how to specify the adjacency or weighting matrix. The most common and simple approach consist in building this matrix exogenously using prior knowledge from the researcher, which may lead directly to unidentified models if this information is wrong. There are other proposals that suggest estimate the matrix from the data which are, from our perspective, a step forward on a journey, yet long and complex. Another way of approaching the problem focuses, directly, on the endogenous treatment of the adjacency matrix which requires of some additional information, in the form of instrumental variables and the like. Section 8 revises the most popular and, perhaps, useful software in the field. The chapter finishes with a brief section of conclusions and final thoughts.

2 Social Interactions Through Social Networks

Social interactions are central to social sciences and, particularly, for economics. Social networks clearly can help to the understanding and to the study of social interactions. This is so mainly because social interactions studied via social networks seem adequate to solve a relevant problem in the social sciences, namely, the observation of large differences in outcomes in the absence of statistical relevant differences in the fundamentals.

The term '*social network*', as used in this chapter, has to be distinguished from the platforms (social network services or sites) that serve to build social networks or social relations among people who share related interests, activities, backgrounds or, in general, real-life connections. Mostly these platforms are web-based providing means for users to interact, to share information, opinions in different forms. In this regard, social networks platforms or sites or services are, broadly speaking, social media. These social network services provides means for social interaction; therefore, they are also an object of scholarly research as they play a key role allowing people to interact on a regular basis, or even sporadically. These interactions influence our decisions, preferences, information, constrains and expectations.

Of course, social networks have been subject of scholar study before the existence of Internet. Understanding how social networks (interactions) influences economic activity has been central for economics since Adam Smith, particularly in *The Theory of Moral Sentiments* (more clearly than in *The Wealth of Nations*) where Smith focused on explaining economic behavior of economic agents that interact mainly with acquaintances or a reduced number of connected agents. In this framework, social networks facilitate particular forms of externalities in which the decisions or actions of a reference group affect an individual's preferences. The reference group can be relatives, neighbors, friends, peers, and so on, depending on the context where the decisions or actions are taking place. As in Smith, these actions or interactions are not necessarily regulated by price mechanism.

Perhaps the most remarkable initial contribution to the economic literature on social interactions is Schelling's (1971; 1972) formal analysis on the influence of social groups in individual behavior. Of particular interest is how Schelling's models have implications for the sorting of agents (individuals) and activities (firms) across the space. Segregation across space is a possible equilibrium when there are some type of interaction models even for individuals that are eager and content to live in an integrated neighborhood, as long her/his group does not form a small minority. This line of reasoning based on interactions (possible, non-market interactions) also help in explaining cities agglomerations, massive fast social movements and fashion followers. Apart from Schelling, earlier less formal contributors to the literature on social interactions are Veblen's (1934) analysis of conspicuous consumption, Duesenberry (1949) and Leibenstein (1950). If social interactions are central since the origins of economic analysis and some relevant scholars were aware of it, how is it possible that economists did not pay enough attention to its study?

Scholar economists did pay attention to social interaction. The heart of noncooperative game theory encourage economists to understand all interactions as games, being markets special cases of interactions. Once that economics has broadened its scope, it is realistic to treat and analyze non-market economic interactions. This is the case for studying the evolutions of institutions and social norms, employment patterns, schooling outcomes, participation in welfare programs, residential segregation, crime rates, diffusion of ideas, human ecology, organizational studies, industrial agglomerations, and urban design, among others. All these topics have been covered by economic analysis, although not all of them by means of social network analysis.

The study of social interactions via social networks is more recent. The main reason for this delay is the (scarce) availability of data, together with the inherent difficulty of drawing inferences from the data. From this viewpoint social network analysis is part of the network science, which considers and represents individuals by nodes and the connections among them by as links (edges). Network science not only considers complex networks as social networks, but also telecommunications networks, computer networks, biological networks, and semantic-cognitive networks.

Many, but not all, market services and goods are contracted upon via networks and therefore the prices, products and the terms of trade that evolve and appear in such networked markets can crucially depend on who is connected to whom. It makes sense that, for centuries, many economic interactions have been held provided that there was a social network. Currently, depending on the size of the market, this is also true. This has been extensively studied for the job market where networks have the role of transmitting information to workers about the specifics job opportunities, and also the role of transmitting information to firms on the potential fit of different workers. In general, there are transactions that might benefit from placing them within the context of a network of transactions as long as the network might help to circumvent difficulties inherent in a given transaction, for example, those where reputation and repeated relationships are critical to make transactions.

Non-market economic issues like diffusion of opinions, behaviors (criminal, philanthropic actions, charity), technology and, even, diseases, are also clear situations where social networks make a role. The same network role appears in learning processes at any cognitive stage, making education a preferred target of network analysis (Benhabib et al. 2011).

What does make the network important, either in market or non-market activities? The activation of a social multiplier which happens when the marginal utility to one person of undertaking an action increases with the average amount of the action taken by her/his peers. Social networks (its existence and iarchitecture) facilitates, and qualifies, that a change in fundamentals had a direct effect on behavior and an indirect effect of the same sign. That is, the behavior of your peers affects fundamentally your own decisions via indirect effects. The result of these indirect effects is the social multiplier. The size of the social multiplier depends on the social network. Phenomena like market crashes, industrial locations, social norms seem to be characterized by large variations in the endogenous variable relative to a (small) change in the fundamentals.

Economic fundamentals are also able to explain the formation and evolution of social networks. Agents (individuals and firms) choose their relationships based on the payoffs that emerge as a function of the network. Naturally, there are well-established and known ways that explain the formation of networks in general. In this context, individuals choose friendship that make them happy (benefit them), also firms choose other firms with which to make transactions or companies pick out which workers to hire to improve specific purposes (reducing risk, for example). As anticipated in this chapter and also in others chapters of this book, there are other ways that explain the formation and creation of networks, like the random network approach, which try reproduce some observed features of social networks.

An important point is that social network is worthy to be studied as long as we can carry some kind of empirical analysis. That is to say that we want to look for regularities and patterns, and also, if possible, we want to make some inferences for causal and non-causal economic relationships. These objectives are, by no means, straightforward to accomplish

As way of example, Patacchini and Zenou (2010) work with the National Longitudinal Survey of Adolescent Health (AddHealth), analyzing the role of conformism (that is, the strategy of blending with the surrounding and to do nothing to draw attention to oneself) in juvenile crime from a network perspective. The main hypotheses are that conformism and deterrence are key factors to explain teenagers decisions to commit crimes. The Survey offers information on the friendship network for each teenager but there are severe problems of identification and measurement due to simultaneity (the reflection problem) and endogeneity (self-selection and unobserved correlated group effects). Conformity means peer effects and deterrence a collection of other conditioning factors (family ties, residential neighborhood, school environment, etc.) that leads the authors to a mixed autoregressive model or (spatial) lag model. Using the same Survey, AddHealth, Hsieh and Lee (2014) analyze peer effects on academic achievement, GPA index, and smoking behaviour once again through a social network approach. The authors exploit the information about the friendship network of each teenager assuming that young people in small groups tend to obtain higher average GPA and a better (least) smoking behaviour. To a great extent, both assumptions are corroborated in the study, which reinforces the importance of the group of close friends. A correct treatment of this issue implies consider the endogeneity in the friendship network formation and the selection bias.

The works of Patacchini and Zenou and Hsieh and Lee are, from our view point, outstanding examples of the singularities inherent to this delicate research field which requires of very specific analytical techniques. The following section deals with economic based models that make possible that the observed social network data can be used for conducting econometric studies.

3 Micro-Foundations for Econometric Analysis

We consider that social interaction takes place within a network of N agents. A model of social interactions studies the joint behavior of the agents who are members of this network. The primary aim is to probabilistically describe the agent choices, that we will refer by y_i, which are made from a set of possible outcomes $\mathcal{Y}_i$. For each i, we denote with y_{-i} the choices of others in the network, which are potential sources of social interactions.

Each agent in the network is described by a vector of attributes, say, $\mathbf{x}_i^*$. This vector contains observable (to other agents and to the econometrician) and non-observable attributes relative to agent i, so $\mathbf{x}_i^* = [\mathbf{x}_i, \mathbf{z}_i]$, where $\mathbf{z}$ refers to a non-observable set of attributes.

Agent choices $\{y_i; i = 1, 2, .., N\}$ represent the maximization of some payoff function. As we represent economic agents interacting through the network, it is reasonable that a decision may affect the actions of other agents via the three key elements, which provide the micro-foundations to the decision problem: preferences (represented by the payoff function), constraints (captured in the set $\mathcal{Y}_i$), and expectations, that is, some structure on the beliefs each agent possesses about behaviors of others in the network.

The payoff function is an utility function for which agent's utility depends on: individual's (1) own action and (2) own attributes, network member's (3) actions and (4) attributes. A common form of utility function is

$$U_i = v_i\left(\mathbf{x}_i^*, \mathbf{x}_{-i}^*, \cdot\right) y_i - \frac{1}{2} y_i^2 + \beta \cdot f(y_i, y_{-i}, \cdot), \tag{1}$$

with $|\beta| < 1$.

$v_i(\cdot)$ is a private part of the utility function that depends on members' attributes, but also (see below) on the network. The middle term reflects the private costs of action. $f(\cdot)$ is a social component of utility that depends on agent own actions, other agents actions and also (see below) on the network structure.

In the *private part* of the utility function, network effects are known, since Manski (1993), as contextual-effects or exogenous-effects, capturing the direct influence of others' attributes on i's decisions. In the *social part* of the utility function, network influences are known as endogenous-effects (also referred as peers-effects), capturing the influence of the decisions of the members of the network on own actions of agent i.

The network structure is introduced in the utility function through $\mathbf{W}_e$ (endogenous) and $\mathbf{W}_x$ (exogenous) sociomatrices or adjacency matrices. Both matrices can be the same or can be different. These matrices can be redefined to normalize the information of their rows and columns.

A typical and recently appeared formulation using these matrices is Blume et al. (2015):

$$v_i(\mathbf{W}_x, \mathbf{X}) = \mathbf{x}_i^* \gamma^* + \mathbf{W}_{x(i)} \mathbf{X} \delta$$

where $\mathbf{X}$ is a $N \times K$ matrix of networked agents' attributes and $\mathbf{W}_{x(i)}$ denotes the ith row of the matrix. The second term captures contextual (exogenous) effects on i's behavior. The first term indicates that marginal private utility is linear in individual's attributes. Provided that $\mathbf{x}_i^*$ incorporates non-observable (private) characteristics of agent i, it is convenient to rewrite that expression as follows

$$v_i(\mathbf{W}_x, \mathbf{X}) = \mathbf{x}_i \gamma + \mathbf{W}_{x(i)} \mathbf{X} \delta + \mathbf{u}_i$$

where $\mathbf{u}_i$ refers to a vector of random agent-level attributes describing agent i, so accounting for individual heterogeneity unobservable to the econometrician. For simplicity many studies suppose that there is only one observable attribute and also one non-observable characteristic, although it is not compulsory because results are easily generalized. To complete, notice that $\mathbf{W}_{x(i)} \mathbf{X} = \sum_j \omega_{x,ij} X_j$ so the product can be understood as a contextual weighted average of agent peers' attribute vectors.

On the other hand, the term $f(y_i, y_{-i}, \cdot)$ captures endogenous effects $f(y_i, y_{-i}, \mathbf{W}_e)$, and its form depends on whether we model social interactions as emerging from social norms or from strategic complementarities. Particularly, we can consider the average action of agent i's peers, that is

$$\mathbf{W}_e \mathbf{y} = \sum_j \omega_{e,ij} y_j.$$

In this case, agent i can obtain utility of the following type

$$f(y_i, y_{-i}, \mathbf{W}_e) = \mathbf{W}_e \mathbf{y} y_i = \sum_j \omega_{e,ij} y_j \cdot y_i$$

which implies that an agent gets pleasure from jointing or matching with peers. Therefore the utility function (1) will be

$$U_i = \left(\gamma x_i + \delta \mathbf{W}_{x(i)} \mathbf{X} + u_i\right) y_i - \frac{1}{2} y_i^2 + \beta \cdot \mathbf{W}_e \mathbf{y} y_i, \tag{2}$$

where $0 < \beta < 1$.

Symmetrically, the agent can obtain disutility when not conforming with network's members, that is, there is social pressure (social norm) to conform, and therefore the utility function (1) can be characterized as

$$U_i = \left(\gamma x_i + \delta \mathbf{W}_{x(i)}\mathbf{X} + u_i\right) y_i - \frac{1}{2}y_i^2 + \frac{1}{2}\beta \cdot (y_i - \mathbf{W}_e\mathbf{y})^2\,, \tag{3}$$

where $-1 < \beta < 0$. The last term is a distance between agent i's action and the weighted average of networked members.

Utility functions (2) and (3) posit the existence of both endogenous and exogenous social interactions. Particularly, the marginal utility associated with an change in y_i increases with respect to the average action of network members, $\mathbf{W}_e\mathbf{y}$. In the two utility functions of (2) and (3), the β parameter captures the strength of any endogenous social interaction through the network:

$$\frac{\partial U_i}{\partial y_i \partial \mathbf{W}_e\mathbf{y}} = \beta$$

which reflects that own- and peer-effects are complements.

On the other hand, the marginal utility of y_i also varies with member attributes:

$$\frac{\partial U_i}{\partial y_i \partial \mathbf{W}_x\mathbf{X}} = \delta,$$

which indexes the strength of exogenous effects or contextual effects. Notice that this effect is common to the two utility configurations, hinting that both characterizations are, in this sense, equivalent.

Summing up, $U_i(y_i, y_{-i})$, in any of the configurations, shows that an agent's utility depends upon her or his own choice and the choices of others. To look for the best response function of each agent, it is considered that all actions are chosen simultaneously under a setting where information is incomplete. Each agent makes a decision to maximize expected utility given their private information and the public information about others, that is, given $[x, z_i] \in R^{N+1}$. To form beliefs (expectations) there is a priori probability distribution μ of the set of all $[x, z_i]$, $i = 1, \ldots, N$. Then, a conditional distribution of μ given $[x, z_i]$ reflects agent i's belief on others members attributes (see Blume et al. 2011 for more details on expectation formations).

To find the best response in this simultaneously incomplete information game, the Bayes-Nash equilibrium concept is required. An equilibrium action is such that no agent can increase her or his utility by changing her action given the actions of all other agents in the network, assuming that beliefs are correct.

Interestingly for the aim of this chapter, the first order condition for optimal action with (1), given the strategy profiles of other agents, generates the following best response function

$$y_i = \frac{\gamma}{1+\beta}x_i + \frac{\delta}{1+\beta}\mathbf{W}_x\mathbf{X} + \frac{\beta}{1+\beta}\mathbf{W}_e\mathbb{E}\left(\mathbf{y}\,|x\right) + \frac{1}{1+\beta}u_i. \tag{4}$$

The econometric model implied by (4) leads, after configuring or restricting sociomatrices, to the well-known linear in means model introduced by Manski (1993) and followed by other scholars. Particularly, linear-in-means models have the following (shared) sociomatrix for each member of the network:

$$w_{ij} = \begin{cases} (N-1)^{-1} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}. \tag{5}$$

Notice that the econometrician have a dataset formed from a random sample of networks, so it is possible that agents belong to one network but not others, and therefore i's agent could be non-networked with j's agent. Given this sociomatrix, model (4) reduces to

$$y_i = \frac{\gamma}{1+\beta} x_i + \frac{\delta}{1+\beta} \bar{x} + \frac{\beta}{1+\beta} \mathbb{E}(\bar{y}\,|x) + \frac{1}{1+\beta} u_i, \tag{6}$$

where agents actions vary with the average action of those whom s(he) is directly connected, $\bar{y}$, her or his own observed attributes, x_i, the average attributes of direct peers, $\bar{x}$, and unobserved attributes, u_i.

Interestingly, neighborhoods can also serve as a mean to analyze social effects (see Durlauf 2004). Neighborhood's based models generate sociomatrices from agents' relative locations. Unlike linear-in-means model agents, those models are not partitioned into non-overlapping structures, but an *i's* agent can have connections with more than one neighborhood and hence her influence can be more widespread than other agent with less connections. For this kind of models the sociomatrix is of the following form

$$w_{ij} = \begin{cases} (n_h)^{-1} & \text{if } j \in n_h \\ 0 & \text{if } j \notin n_h \end{cases},$$

where n_h is the number of agents in neighborhood h, which are agents to whom a given agent, say i, is connected with.

Linear-in-mean models (either via neighborhood or other channels) share the facts that social interactions are generated through level of group-specific averages and that there is only one type of connection between agents. Sociologists have insisted in that is worthy considering much richer social networks so that the network allows for heterogeneity of interactions across pairs of agents. Different kinds of connections in a social network have an impact on network potential outcomes.

Sociomatrices can cope with this demand. The basic linear-in-means model assumes that $\mathbf{W}$ is symmetric binary, that the edges of the network are bidirectional, and that there is transitivity in the structure. To this end w_{ij} are either 1 or 0. For each pair of agents i and j, $w_{ij} = 1$ if these agents are socially connected and 0 otherwise.

However, $\mathbf{W}$ can express richer specifications of social relations by allowing the elements of the sociomatrix to be arbitrary real numbers. In such models, the degree of influence j has on i is measured by the magnitude and sign of the real number w_{ij}.

For example, connections between agents can be separate in two groups, say, strong and weak. Agent i can have a given number of weak connections, n_{iW}, and strong connections n_{iW}. Then, the sociomatrix has the following form

$$w_{ij} = \begin{cases} (n_{iW} + \theta n_{iS})^{-1} & \text{if } j \text{ is weakly connected to } i, \\ \theta (n_{iW} + \theta n_{iS})^{-1} & \text{if } j \text{ is strongly connected to } i, \\ 0 & \text{otherwise} \end{cases}$$

where parameter θ is a ratio of the strength of strong to weak connections.

The behavioral equation of (4) provides also an intuitive bridge between social interaction and spatial interaction. Such equation can be interpreted as a SAR model with one spatial lag. Indeed, sociomatrix $\mathbf{W}_e$ could capture the strength of the social interactions between two agents, but also the effect of any kind of social or geographical distance. This relationship has been defended in Lee (2007), Bramoullé et al. (2009) and Lee et al. (2010), allowing for a social structure also in the error term. In particular, spatial-type autocorrelations in the errors assume that the vector of shocks, for a given agent of a given group, consists of the sum of group-specific fixed affects and a stochastic group-specific component. Such group relations are characterize by another sociomatrix, let us say $\mathbf{W}_u$, coincident or not with $\mathbf{W}_x$ and/or $\mathbf{W}_e$.

3.1 Economic Identification

It has been said that the Bayes-Nash equilibrium equation (4) embraces a rich variety of network structures, which in fact indicates that linear models, like those is studied and applied in the literature, are economically well-founded. Although micro-foundation is important, researchers have focused more their attention in analyzing under what circumstances (conditions) the models implied by (4) are identified.

Identification is central in these types of models, so it is worth to extend a bit more its treatment before going into the econometric details. In short, identification means the ability to recover (γ, δ, β) from the set of observables. In this micro-motivated setting, it implies to identify the parameters of the utility function. Let us write the model of (4) in matrix form as

$$\mathbf{y} = \gamma^* \mathbf{X} + \delta^* \mathbf{W}_x \mathbf{X} + \beta^* \mathbf{W}_e \mathbb{E}(\mathbf{y} | \mathbf{x}) + \tilde{\mathbf{u}}.$$

Provided that $\mathbf{y}$ are observed, the econometrician faces the following sampling analog model

$$\mathbf{y} = \gamma^* \mathbf{X} + \delta^* \mathbf{W}_x \mathbf{X} + \beta^* \mathbf{W}_e \mathbf{y} + \tilde{\mathbf{u}},$$

which can be solved for $\mathbf{y}$

$$\mathbf{y} = \left(I - \beta^* \mathbf{W}_e\right)^{-1} \left(\gamma^* \mathbf{X} + \delta^* \mathbf{W}_x \mathbf{X}\right) + \left(I - \beta^* \mathbf{W}_e\right)^{-1} \tilde{\mathbf{u}}.$$

In continuation, let us assume, for simplicity, that sociomatrices are the same and that there are only endogenous effects in the model. The last equation therefore simplifies to

$$\mathbf{y} = \left(I - \beta^* \mathbf{W}\right)^{-1} \left(\gamma^* \mathbf{X}\right) + \left(I - \beta^* \mathbf{W}\right)^{-1} \tilde{\mathbf{u}}. \tag{7}$$

Taking into account the series expansion

$$\left(I - \beta^* \mathbf{W}\right)^{-1} = \sum_{k=0}^{\infty} \beta^{*k} \mathbf{W}^k,$$

Eq. (7) is

$$\mathbf{y} = \sum_{k=0}^{\infty} \beta^{*k} \mathbf{W}^k \mathbf{X} \gamma^* + \sum_{k=0}^{\infty} \beta^{*k} \mathbf{W}^k \tilde{\mathbf{u}}$$

After simple algebraic manipulations, the last equation can be written as

$$\mathbf{y} = \gamma^* \mathbf{X} + \left[\left(\gamma^* \beta^*\right) \sum_{k=0}^{\infty} \beta^{*k} \mathbf{W}^{k+1} \right] \mathbf{X} + \left[\sum_{k=0}^{\infty} \beta^{*k} \mathbf{W}^k \right] \tilde{\mathbf{u}}.$$

The last expression displays the *social multiplier effect*, introduced in the example of Sect. 1 and motivates the importance of networks effects. Consider now a policy (public or private) which increases by d the ith agent's value of x_i. The change affects, firstly the outcome of agent i by $d\gamma^*$. Then, agent $i's$ friends will best reply to the increase in agent $i's$ outcome, and therefore all friends of agent i will experience outcome variation, particularly $d\gamma^* \beta^* \mathbf{W}$. As a consequence of this second round of reactions, agent $i's$ friends' friends change (as best response) their outcomes by $d\gamma^* \beta^{*2} \mathbf{W}^2$. In general, in the k-round, the change will be $d\gamma^* \beta^{*k-1} \mathbf{W}^{k-1}$. The long-run or full effect of the initial change in x_i on the optimal response will be

$$d\gamma^* \left(\mathbf{I} - \beta^* \mathbf{W}\right)^{-1},$$

which is understood as the *social multiplier*.

Notice that according to the notion of social multiplier, policy-makers can take advantage of the structure of the social network **W** to efficiently target interventions; they can do it at low cost, as long as the cost of perturbing x_i does not vary with i. For this result to be operational, it is needed that endogenous (but also exogenous) effects can be identified and distinguished from other sources of heterogeneity. Put it differently, outcomes of members in a common network tend to covary. Isolating (identifying) the true nature of this covariance, either from spillovers or from heterogeneity is useful. The distribution of outcomes can just be modified by adding or deleting networks' links, being then a potential powerful political tool.

It is now evident the important role of sociomatrices. As in the context of spatial econometrics and spatial statistics, sociomatrices have been considered central pieces of analysis and, generally, as known (given) for the researcher. It is possible that sociomatrices are chosen for theoretical reasons, but also is possible that they are just empirical constructions. It is therefore a common assumption to consider the **W** matrix as given.

The fact that the vast majority of applied and theoretical studies have considered that $\mathbf{W}_e$ and $\mathbf{W}_x$ are the same is a limitation, because is not well motivated why these weights must coincide. Moreover, it seems more feasible to obtain information about exogenous-effect matrix, $\mathbf{W}_x$, than for $\mathbf{W}_e$ which captures a primitive proclivity to act similarly to others. Sometimes the $\mathbf{W}_x$ can be thought of formed from some type of (social or even physical) distance. Recent theoretical research on how to identify the model consists on looking for the restrictions that allow for identification when there is only knowledge of $\mathbf{W}_x$, but not from $\mathbf{W}_e$.

The distinction between $\mathbf{W}_e$ and $\mathbf{W}_x$ is well established in a spatial econometrics context and, more important, this literature has dealt recently with the issue of unknown sociomatrices, pointing that they do not necessarily have to be known (Sect. 7 abounds on this issue). In practice, the observer-researcher has information about connections between individuals or firms, but not the sociomatrices themselves. In this context, the researcher needs to interpret the data on direct connections in terms of sociomatrices. Consider, for example, the case of a dataset formed from a survey where respondents indicate to whom they are connected, but not the weights corresponding to each relation. Once again this fact highlights the link between network econometric literature and spatial econometric literature.

4 The Reflection Problem and the Lack of Identification

As indicated, in the last decades there has been a revival of the interest in social interaction, neighborhood effects and social dynamics, which has spurred the research on the subject. At the same time, applied literature has receive a strong impulse especially after the seminal works of Manski (1993), Brock and Durlauf (2001) and Moffitt (2001). The first of them coined the term '*reflection problem*' that appears when, observing the behavior of a population, we try to infer whether the behavior in a group influences the decisions of the individuals that is part of the same group. It is difficult to disentangle if the collective behavior of the group is causing the decisions of the individuals or the former is a mere reflection of the last ones. This problem appears explicitly in Eq. (4), where $E(y|x)$ integrates y_i, and has been the reason of a long-standing econometric discussion.

A consequence of the reflection is that renders the model unidentified, unless the researches introduces additional information. For example in the case of the simple model in (6), it is immediate to obtain that $E(y|x) = (\gamma + \delta)E(x)$. The reduced form follows from the sample analog of $E(x)$; then we obtain:

$$y_i = \frac{\gamma}{1+\beta}x_i + \frac{\delta\beta + (\delta + \beta)}{1+\beta}\bar{x} + \frac{1}{1+\beta}u_i \tag{8}$$

It is possible to test for the presence of social effects in the equation, given that the composite parameter $\frac{\delta\beta+(\delta+\beta)}{1+\beta}$ will be different from zero whenever δ or β, or both, are different from zero; however is not possible separate the effects of contextual and endogenous factors. The variance of the error term can not help us to attain identification.

The social interaction model in Manski (1993) is a bit more elaborated but the conclusions remain the same: identification is problematic in these kind of models and must be checked case-by-case. For instance, the problem with the example of (6) is that the conditional mean of y is proportional to the mean characteristics of the agents in the network, $E(y|x) = (\gamma + \delta)E(x)$. In order to break with this constraint we would need more heterogeneity in the form, for example, of a sample made of several networks, a sequence of sociomatrices with weights different from that of (5), nonlinearities (Brock and Durlauf 1995) or directly resorting to experimentalism (Keane 2010).

The identification problem is closely related to how individuals interact. We can distinguish two main approaches: through social networks (a structure made of nodes, individuals, that are tied by some specific type of dependence such as friendship, scientific collaboration, spatial location, etc) or in groups (there is a partition of the population in subsets so that each individual is affected by all others in their group but none outside it; classical examples are mates in a classroom or segregation patterns). As shown in Bramoullé et al. (2009), identification is easier in the case of group interaction where, in general, it suffices that at least two groups have different size (of course, there must be social effects in the model); this is their Proposition 2, where identification arises due to the impact of group sizes on reduced-form coefficients within each group. In the case of interaction through social networks and according to Proposition 3 in Bramoullé et al. (2009), the conditions are the existence of social interactions, as before, plus the restriction that the endogenous sociomatrix, $\mathbf{W}_e$, its square, $\mathbf{W}_e^2$ and the identity matrix, I, be linearly independent. In this case, identification arises from the use of instrumental variables built as powers of the sociomatrix that multiply the x's variables. A sufficient condition is the existence of *intransitive triads* in the network which imply that the friends' friends are not necessarily my friends or that the neighbors' neighbors are not necessarily my neighbors.

5 Model Specification

The model of Manski (1993) has become an obliged reference for the applied literature due of its simplicity, consistency and flexibility:

$$y_i = \theta_0 + \theta_1 E(y|g) + E(x|g)' \theta_2 + x_i' \theta_3 + u_i \tag{9}$$

where y_i is the outcome of individual i (qualifications, productivity of the factory, housing value, etc.), x_i are a set of k factors that directly affect the outcome of individual i (ability, formation, appliances in the house, etc.) and g attributes characterizing the reference group for i. There is social interaction because the behavior of i varies with the collective behavior of the group, given that $\theta_1 \neq 0$, and also because the exogenous factors corresponding to the group have an impact on the behavior of i, given that $\theta_2 \neq 0$. Of course, the econometrician observes (y, x, g) and then builds the sample (linear) analog of (9) as:

$$y_i = \theta_0 + \theta_1 \mathbf{W}_{ei}\mathbf{y} + \mathbf{W}_{xi}\mathbf{X}\theta_2 + x_i' \theta_3 + u_i \tag{10}$$

In the following, to simplify the notation, we assume that the two adjacency matrices are the same $\mathbf{W}_e = \mathbf{W}_x = \mathbf{W}$. The sample size is N. In the case of social networks, $\mathbf{W}$ is a general $(N \times N)$ matrix with zeros in the diagonal (individual i is not considered as being part of its own reference group; the exception is Manski 1993), with nonnegative weights, not necessarily symmetric and typically subject to some normalization (row/column sum to one, maximum eigenvalue, etc.; Corrado and Fingleton 2012, for details). In the case of group interaction, the size of each group is N_j; $j = 1, 2, \ldots, G$ and the adjacency matrix is block-diagonal where each matrix in the diagonal is $\left(N_j \times N_j\right)$.

The treatment of an equation like that of (10) presents some difficulties. The first is simultaneity, a problem very well-known in the spatial econometrics literature and connected to the presence of the lag of the outcome variable in the right hand side of the equation, $\mathbf{Wy}$. It is clear that if factory $A's$ decisions affect factory $B's$ actions and vice-versa, this generates a conventional simultaneous equations problem. In short, we cannot regress directly factory $A's$ actions on factory $B's$. The econometric toolkit contains a great variety of estimation algorithms suitable for this case (IV, GMM, QML, etc.).

A second question, more subtle, is the correlation between unobservables which arises in the case of group interaction. If there are unobserved effects in a behavioral equation, that is structured in groups, it is natural to think that these unobservables vary across groups. Let us call the unobserved effects as μ_j; $j = 1, 2, \ldots, G$. Moreover, it is very likely that the unobservables are also correlated with the observed characteristics of the individuals (with the x variables), so that $E\left(\mu_j; x_j\right) \neq 0$. This is the *correlated effect* introduced by Manski (1993) as $E\left(\mu_j \left| x_j; g\right.\right) = g'\lambda$ being λ a conformable vector of linear parameters. The unobservables arise from a variety of sources. For example, we may think in the case or unobserved preferences encouraging that individuals tend to locate where there are other individuals of the same type, like in the models of social segregation (Schelling 1978) or the tendency for the companies to cluster in space to benefit from different types of agglomeration economies (Fujita and Thisse 2002). The two cases are endogenous group membership models, and it is very likely that the unobservables are correlated with the characteristics of the individuals in the groups. For the case of exogenous group membership models, the unobservables may represent contextual, or environmental, influences such as school ethos, sense of security in the neighborhood or confidence in the political institutions of a region. The meaning of all these unobservables is rather clear in the respective model, but they are hardly measurable by the econometrician.

The consequences of the correlated effect are well-known in the econometric literature, implying that the LS estimation of Eq. (10) yields inconsistent estimates of all the $\theta's$. In particular the LS estimate of θ_2 and θ_3 will be severely upward biased as the covariance between the $x's$ and the unobserved effect, μ_j, increases. Note that part of the regressors represent the average across individuals in a group, which likely increments the correlation between the unobservable and the observable terms. Thus, endogeneity due to correlated unobservables is a new risk factor for identification.

Consistent estimation of (10) requires breaking the correlation between the $x's$ and the unobservables. Moffitt (2001) suggests a randomization assignment of individuals to groups, which can be useful if the source of such correlation is a endogenous group membership. More simple is the solution of Bramoullé et al. (2009) which, similar to the within transformation in the panel data literature, consists in taking *local differences* to eliminate the unobservables:

$$(\mathbf{I}-\mathbf{W})\mathbf{y} = (\mathbf{I}-\mathbf{W})\theta_0 + \theta_1(\mathbf{I}-\mathbf{W})\mathbf{W}\mathbf{y} + (\mathbf{I}-\mathbf{W})\mathbf{W}\mathbf{X}\theta_2 + (\mathbf{I}-\mathbf{W})\mathbf{X}\theta_3 + (\mathbf{I}-\mathbf{W})\mathbf{u} \tag{11}$$

This transformation is useful in the case that the adjacency matrix has been row-standardized. However, there is a price to pay for this transformation: bits of information have been lost to compensate for the presence of correlated effects, which makes identification more difficult. Now the identity and the powers of the adjacency matrix, up to order *three*, must be linearly independent in order to ensure identification in model (11); simply, the *diameter* of the network should be greater than or equal to three (where diameter means the maximal distance, in terms of connections, between two nodes in the network). Bramoullé et al. (2009) show that if a model is identified after local differences,will be also identified when taking *global differences* (the deviations are obtained with respect to network means, not with respect to the group means).

The endogenous group membership is a different question which leads directly to inconsistent LS estimates. In this case, the model is made of two structural equations. One is the behavioral equation of (9) whereas the second describes how the groups are formed. It is clear that each individual decides to join a given group based on personal utility maximization considerations. This function may depend on several factors, among which should appear the individual and group average characteristics and mean structural errors from (9). The consequence is that, ignoring the endogenous group membership pattern, leads to a model that is, once again, unidentified. The LS estimates will be inconsistent because the $x's$ regressors of (9) are not independent from the error terms of the equation. This is know as the *selection bias* in the econometrics literature. The consequences are identical to the correlated unobservables issue; in fact, group membership selection equation gives specific form to this correlation. Section 7 resumes this discussion.

6 Estimation Issues

The technicalities for estimating social interaction models have evolved somewhat in parallel to the spatial econometrics literature; in fact, often is difficult distinguish between both kind of models. Starting from (10), it is usual to capture the group unobservables through a fixed effect in the equation so that (Lee 2010):

$$\mathbf{y}_{\mathrm{j}} = \theta_1 \mathbf{W}_{\mathrm{j}} \mathbf{y}_{\mathrm{j}} + \mathbf{W}_{\mathrm{j}} \mathbf{X}_{2\mathrm{j}} \theta_2 + \mathbf{X}_{1\mathrm{j}} \theta_3 + \boldsymbol{\iota}_j \alpha_j + \mathbf{u}_j \tag{12}$$

where $\mathbf{y}_{\mathrm{j}}$, $\mathbf{X}_{1\mathrm{j}}$ and $\mathbf{X}_{2j}$ are the vector and matrices of the N_j observations in group j ($\mathbf{X}_{2j}$ contains the contextual variables and $\mathbf{X}_{1j}$ the personal attributes of the individuals forming the group; both sets of variables can coincide); α_j is the fixed effect of the j-th group and $\boldsymbol{\iota}_j$ a $\left(N_j \times 1\right)$ vector of ones. The error term is assumed to be $i.i.d.N(0, \sigma^2)$ and the adjacency matrix corresponding to this group, $\mathbf{W}_{\mathrm{j}}$, is exogenous, known by the econometrician and conforms to usual rules (Kelejian and Prucha 2001). In the peer effects model this matrix has a specific form: $\mathbf{W}_{\mathrm{j}} = \frac{1}{N_j}\left(\boldsymbol{\iota}_j \boldsymbol{\iota}_j' - I_j\right)$.

Lee (2007) points that model (12) is not identified in the presence of unobservables and suggests to use the within equation, equivalent to the local differences of Bramoullé et al. (2009). Identification in this case is attained when different groups have different numbers of members. Identification can be weak, especially if the size of the groups is very large. Note that the within transformation is useful only for the peer effects model. In a more general case, where the adjacency matrix reflects other interaction mechanisms, Lee and Yu (2010) develop the so-called *orthonormal transformation* using the eigenvectors of the matrix $\mathbf{W}_{\mathrm{j}} = \frac{1}{N_j}\left(\boldsymbol{\iota}_j \boldsymbol{\iota}_j' - I_j\right)$. Multiplying the model by the eigenvectors matrix eliminates the unobservables from the equation, at the price of disturbing some of the basic properties of the (transformed) adjacency matrix such as zero diagonal terms and row (column) normalization.

Under quite general conditions, Lee (2007) obtains the conditional ML estimates, CMLE, corresponding to the within equation of (12), conditional on the means equation. CMLE have optimal asymptotic properties: identification uniqueness, consistency and asymptotic normality. The asymptotic is solved under two scenarios: small group interaction, where the average size of the groups ($\overline{N}$) is bounded, and large group interaction, where both the size and the number of groups increases (as a regularity condition, the growth rate of the number of groups should be higher than that of the average size). The convergence rate of the CMLE of the endogenous interaction effect, $\hat{\theta}_1$, is the usual $\sqrt{N}$ in the case of small group interaction, but it is scaled down in the case of large group interaction $\sqrt{N}/\overline{N}$. However, the rate of convergence of the CMLE for the contextual effects, $\hat{\theta}_2$, and direct effects, $\hat{\theta}_3$, remain unaffected: $\sqrt{N}/\overline{N}$ for the contextual effects and $\sqrt{N}$ for the direct effects. This means that is more difficult to obtain accurate estimates of the contextual or endogenous effects (in the case of large group interaction) than for the direct effects.

The IV estimation algorithm reproduces this general picture. Lee (2007) shows that the within equation can be estimated by IV and that there is a best IV vector, whose variables are not collinear if the groups have different size. The use of LS to estimate contextual and direct effects from the within equation would only be justified in the large interaction case when $\overline{N}/G \to \infty$. In this situation the interaction is distorted by the increasing density of each group which dominates the converge. Moreover, the rate of convergence of the LS estimates for the direct effects would be $\sqrt{\overline{N}}$, smaller than the usual $\sqrt{N}$.

The model of (12) has received several extensions. Lee et al. (2010) introduce a pattern of dependence between the disturbances of the individuals connected in the network, with the purpose of capturing the selection bias effect due to the endogenous group formation; that is:

$$\mathbf{u_j} = \eta \mathbf{M_j} \mathbf{u_j} + \boldsymbol{\varepsilon}_j \tag{13}$$

$\mathbf{M}_\mathrm{j}$ is a new adjacency matrix for group j (it can be the same, $\mathbf{W}_\mathrm{j} = \mathbf{M}_\mathrm{j}$) and $\boldsymbol{\varepsilon}_\mathrm{j}$ is a vector of error terms not necessarily normal, $i.i.d.(0, \sigma^2)$. The quasi ML estimates, QMLE, have good properties for the small-world case (N increases because the number of groups, G, increases but the average size of the groups remains bounded). The parameters are identified even if the groups have the same size, based on the mean regression function and the correlation structure of the dependent variable. However, care should be taken in the case of using a unique adjacency matrix ($\mathbf{W}_\mathrm{j} = \mathbf{M}_\mathrm{j}$) because if $\theta_1 \boldsymbol{\theta}_\mathbf{3} + \theta_2 = 0$, the contextual effects, θ_2, are indistinguishable from the error correlation coefficient, η. In fact, this is the well-known problem of Common Factors (Davidson 2000) which implies that the model of (12) plus the common factors restriction is observationally equivalent to the model $\mathbf{y}_\mathrm{j}^\circ = \mathbf{X}_{1\mathrm{j}}^\circ \theta_3 + \mathbf{u}_j^\circ; \mathbf{u}_j^\circ = \eta \mathbf{W}_j^\circ \mathbf{u}_j^\circ + \boldsymbol{\varepsilon}_j^\circ$ ('$\circ$' means '*after orthonormal transformation*'). The QMLE are consistent and, in the small-word scenario, asymptotically normally distributed at the usual rate of $\sqrt{N}$; they are efficient relative to the Generalized Two-Stages Least Squares estimator, G2SLSE, based on IV estimates (Kelejian and Prucha 1999).

Liu and Lee (2010) warn against the routinely row(column)-normalization of the adjacency matrix because this practice removes valuable information in terms of centrality of the individuals. Bonacich (1987), develops a simple centrality index based in the sum of the rows in the adjacency matrix, the so-called *indegress*. He summarizes the situation: higher value of the index means more centrality in the network and more bargaining power. It is evident that if we row-standardize the matrix, this information would be lost. Using the original, not row-normalized adjacency matrix, the parameter space for the endogenous and correlation effects coefficients must be adjusted to the spectral radius of the matrix (Kelejian and Prucha 2001).

Another difficulty appears in the case of combining no row-normalization and unobserved fixed effects. The problem is that the partial likelihood approach, that factorizes the join distribution function into the product of the likelihood function of the transformed data and the conditional likelihood function of the sufficient statistic for the transformation, is not feasible. Lee et al. (2010) develop 2SLS and GMM estimation algorithms to compensate the gap; both are adjusted to account for a $K/\sqrt{N}$ order bias due to the increasing number of IVs that appears for the mean regression equation (K is related to the number of groups). Adjusted 2SLS and GMM estimators are $\sqrt{N}$consistent and asymptotically normally distributed.

All the models discussed up to now are static, assuming contemporaneous effects; however, social networks are dynamic by nature. Dynamics is a relatively new strand of development in the literature of social networks where there are not many contributions so far. Manski (1993) admitted the utility of dynamics specifications such as:

$$y_{it} = \theta_0 + \theta_1 E_{t-1}(y|g) + E_{t-1}(x|g)'\theta_2 + x_{it}'\theta_3 + u_{it};\ E_t(u_{it}; x) \neq 0 \quad (14)$$

where E_t and E_{t-1} denote expectations taken at periods t and $t-1$. That is, non social forces, like the personal attributes of the agent, operates contemporaneously but there is a delay with respect to the impact of social forces, endogenous and/or exogenous. The unobservables are permanent characteristics so that the correlation with the group features is different from zero whenever the expectation is taken. The sample analog of the last equation leads to a recursive model:

$$\mathbf{y}_{\mathrm{jt}} = \theta_1\mathbf{W}_{\mathrm{j}}\mathbf{y}_{\mathrm{jt-1}} + \mathbf{W}_{\mathrm{j}}\mathbf{X}_{\mathrm{2jt-1}}\theta_2 + \mathbf{X}_{\mathrm{1jt}}\theta_3 + \boldsymbol{\iota}_j\alpha_j + \mathbf{u}_{jt};\ j = 1, \ldots, G;\ t = 1, \ldots, T \quad (15)$$

For which we need of panel series for each individual. Now the sample size is $NT = G \times \bar{N} \times T$. The problem of simultaneity disappears which facilitates inference. In fact, this equation can be estimated using standard methods such as LS (if the time span is large) conditional ML, etc.

Lee and Yu (2012) consider a fully dynamic spatial panel data model that is suitable for social network models:

$$\mathbf{y}_{\mathrm{t}} = \gamma_0 \mathbf{W}_{\mathrm{t}} \mathbf{y}_{\mathrm{t}} + \gamma_1 \mathbf{W}_{\mathrm{t-1}} \mathbf{y}_{\mathrm{t-1}} + \gamma_2 \mathbf{y}_{\mathrm{t-1}} + \mathbf{X}_{1\mathrm{t}} \pi_1 + \mathbf{W}_{\mathrm{t}} \mathbf{X}_{2\mathrm{t}} \pi_2 + \boldsymbol{\alpha}_0 + \iota c_t + \mathbf{u}_t;\ t = 1, \ldots, T \tag{16}$$

The network, of size N, does not contain groups. There is a simultaneous and a lag endogenous interaction effect, measured by γ_0 and γ_1, a purely autoregressive impact, γ_2, to capture short-term dynamics in the personal scores, a contemporaneous contextual effect, π_2, (which can be lagged) and a direct effect, π_1. There are unobservables related to time (c_t is the unobserved term for period t) and also to individuals ($\boldsymbol{\alpha}_0$ is a $(N \times 1)$ vector); finally $u_{it} \sim i.i.d.(0, \sigma^2)$. Moreover, Lee and Yu (2012) allow that the adjacency matrix changes from period to period, $\mathbf{W}_t;\ t = 1, \ldots, T$, (assuming that it is exogenous and known by the econometrician). These matrices are all row-normalized.

The unobserved temporal effects are eliminated using a orthogonal transformation to each cross-section, and the personal unobservables after concentrating out the $\boldsymbol{\alpha}_{i0}$ $(i = 1, \ldots, N)$ terms from the log-likelihood function using first-order conditions. This facilitates the obtention of the QMLE for the parameters of interest, $\gamma' s$ and $\pi' s$.

Lee and Yu (2012) show that the dynamic network model of (9) is identified and that the QMLE are consistent, under usual regularity conditions and a large interaction cross-section (N is an increasing function of T and T goes to infinity). The QMLE are asymptotically normally distributed. The rate of convergence depends on the relation between N and T, for which three situations can be distinguished: (1)- if $N/T \to 0$, then the QMLE converge to a centered normal distribution with a finite covariance matrix, at a rate $\sqrt{(N-1)\,T}$; (2) if $N/T \to \kappa < \infty$, the incidental parameter problem emerges (together with the question of initial conditions) in the form of a bias of magnitude $O(T^{-1})$ that disappears as T goes to infinity; the asymptotic distribution is normally non-centered; (3) if $N/T \to \infty$, the distribution function degenerates.

Social networks are entities in permanent transformation, whose behavior, in the long run, is dynamic. In fact a network consists of ties between individuals that change over time. In general, these connections are not to be seen as transient events, such as telephone calls or email traffic at a given point in time, but should be regarded as states with a tendency to endure over time such as friendship, commercial relations or migration flows, all of which are very stable but exhibit gradual changes over time (in fact, they can be interpreted as Markovian processes). Snijders et al. (2010) work out a general approach to model the dynamic evolution of a social network based on the interdependency between ties and behaviors. Between the simple recursive model of (15) and the simultaneous dynamic model of (16) or the fully endogenous model of Snijders et al. (2010) there is a large variety of candidates to introduce dynamics into the social interaction equations. This strand of literature seems very promising and fruitful.

7 Identifying Reference Groups

This is the title of one of the sections in the seminal article of Manski (1993). The concern of Manski is that, commonly, it is assumed that the researcher knows the reference group for each individual, same as the individuals themselves who should correctly perceive all endogenous and contextual effects. This is, indeed, the case with most of the applications in social networks where the dyadic information on the nodes emerges naturally such as peer effects at schooling. However there are situations where the connections are not so clear, as in the case of friendships networks or real estate markets. Moreover the group boundaries are often arbitrary, partly reflecting limitations arising from the availability of disaggregated data, as in local crime rates (Calvó-Armengol et al. 2009) or technology diffusion models (Comin and Mestieri 2014). In these and related cases, the question of specifying the adjacency matrix is in the center of the discussion, although the issue has remained at a low level in the literature on modelling social networks. However, the issue has received quite attention from a spatial econometrics perspective. We think that this discussion may have some interest here.

Overall, we can identify in this literature two general approaches to the**W** question: (1) specifying the matrix exogenously; (2) estimating the matrix from the data. The exogenous approach is by far the most popular but requires an important prerequisite, the individuals must be located in a certain support with a well-defined metric (Harris et al. 2011). In this case, we can use, for example, the k-nearest neighbors, kernel functions based on some measure of distance between the individuals, a pure binary criteria based on some notion of proximity or nearness between individuals, etc.

The second approach considers the physical characteristics of the network together with the nature of the data, and takes many forms. Kooijman (1976), for example, suggests build the weights of the adjacency matrix so as to maximize the value of some cross-sectional correlation coefficient (Moran's I, for example). Griffith (1996) proposes to find a **W** that absorbs all the cross-sectional effects from the data (arising either from the endogenous effect or from the endogenous group formation process); Fernández et al. (2009), in the same vein, propose a specification of **W** based on a measure of entropy, while the LSM (local statistical model) of Getis and Aldstadt (2004) tries to find, for each individual, a critical distance beyond which there is no direct interaction with other mates in the network. The weights, exceeding the threshold, are set to zero but inside the threshold are fixed proportional to a certain interaction measure.

This approach also includes the proposals where**W** is directly extracted from the data. This is a complex problem because of the large number of parameters that should be estimated if the weights of the adjacency matrix, ω_{ij}, are not restricted. Meen (1996) sets out the problem in a SUR framework where the interaction between the individuals occurs in the error terms. The **W** matrix is unknown but can be estimated from the SUR residuals. Bhattacharjee and Holly (2013) and Bhattacharjee and Jensen-Butler (2013) improve Meen's algorithm, based on the restrictions of symmetry, zero diagonal terms and row-normalization of the matrix. Beenstock and Felsenstein (2012) extend the discussion to a cross-sectional autoregressive model.

The algorithms assume panel data, that the model is correctly specified, a finite N and the asymptotics is with T. Under these circumstances, Bhattacharjee and Jensen-Butler (2013) show that their optimization procedure identifies a unique adjacency matrix, which is consistently estimated as T goes to infinity. The estimates of the adjacency matrix support a CLT which offers an adequate framework for testing for drivers of social interaction. Finally Ahrens and Bhattacharjee (2015) develop a two step Lasso estimator that, in a panel data framework, mimics two-stage least squares (2SLS) to account for endogeneity of the spatial lag. The condition for identifying **W** requires that each unit/node in the sample is affected by only a limited number of other units which they called *approximate sparsity* of the spatial weights matrix; also strict exogeneity of the regressors (others than the spatial lag) across space is required, as they become the source of instruments to tackle with the endogeneity problem. Roughly, the procedure is $\sqrt{T}$ consistent.

Another relevant problem related to**W** is that, in many circumstances, the assumption of exogeneity is not reasonable. For example, Waldinger (2011), in the case of research productivity, points out that sorting of individuals affects the estimation of peer effects, as highly productive scientists often choose to co-locate: staff members self-select into departments with peers of similar level and departments tend to appoint new staff compatible with the existing staff; all of which inflates the pure peers effect. In a different context, Conley and Topa (2002), for the case of the spatial pattern of unemployment in Chicago, use several weighting matrices based on generalized measures of socio-economic distance between the Census tracts that involve physical distance, travel time, as well as differences in ethnic and occupational distribution.

The endogeneity of the adjacency matrix has severe consequences on inference given that usual 2SLS and GMM algorithms will produce non valid instruments; neither the QML methods will be appropriate, as shown in Lee et al. (2010). To our knowledge, there are few papers attempting to cope with this problem, all of which coincide, one way or another, in the use of instruments to safeguard inference.

Kelejian and Piras (2014) specify a network interaction model with endogenous effects (no contextual variables are included), unobservables and endogenous regressors:

$$\mathbf{y}_\mathrm{t} = \gamma_0 \mathbf{W}_\mathrm{t}\mathbf{y}_\mathrm{t} + \mathbf{Y}_\mathrm{t}\varpi + \mathbf{X}_{1\mathrm{t}}\pi_1 + \boldsymbol{\alpha}_0 + \mathbf{u}_t;\ t = 1, \ldots, T \tag{17}$$

where $\mathbf{Y}_\mathrm{t}$ is a $N\times q$ matrix of observations on q endogenous variables at time t and ϖ the associated vector of parameters. The error term is not necessarily $i.i.d.(0, \sigma^2)$. The weights in the adjacency matrix, $\omega_{ij,t}$, are assumed to be endogenous so that their mean exists and depends on a set of r variables; r_1 of them, in the $p_{ij,t}$ $(r_1 \times 1)$ vector, are observable whereas r_2 are unobservable, in the $q_{ij,t}$ $(r_2 \times 1)$ vector, so that:

$$\mathrm{E}\left[\omega^{\bullet}_{ij,t}\right] = f\left(p_{ij,t}; q_{ij,t}\right); \; i,j = 1, \ldots, N; \; t = 1, \ldots, T \tag{18}$$

f is an unknown function and the dot denotes that the corresponding weight is different from zero. The adjacency matrix is allowed to change from period to period. Kelejian and Piras (2014) develop the IV estimators for the case of (17)–(18), where T is finite and N increases. The instruments are a combination of all the exogenous elements in the system (that is, the p and the x variables plus a feasible set of m variables do not included in Eq. (17) but related with the $\mathbf{Y}$ variables). The IV estimates thus obtained are consistent and asymptotically normally distributed at a rate $\sqrt{NT}$.

Qu and Lee (2015) generalize the work of Kelejian and Piras above. As before, the weights of the matrix are not predetermined but depend on a set of observable random variables, $\omega_{ij,t} = h\left(Z; \delta_{ij}\right)$ where h is a bounded unknown function, δ_{ij} a measure of distance in the corresponding physical or social support of the network between individuals i and j, Z another set of k_3 interaction factors, linearly dependent on p_2 exogenous variables in D, which is $(k_3 \times p_2)$:

$$Z = D\Gamma + \epsilon \tag{19}$$

Γ contains the corresponding set of parameters. For example, in international economics, Z may refer to economic distance between two economies while D are import-export flows; or, in migration models, Z accounts for information flows and D accumulated migrants in the destination country. Qu and Lee (2015) use a simple autoregressive model with only one cross-section ($T = 1$), and no endogenous regressors nor unobservables in the right hand side:

$$\mathbf{y} = \gamma_0 \mathbf{W}\mathbf{y} + \mathbf{X}_1 \pi_1 + \mathbf{u} \tag{20}$$

The endogeneity of the weights arises because the error terms of (19) and (20) have a joint distribution, $\left(\mathrm{u}_i; \boldsymbol{\epsilon}_i'\right)' \sim i.i.d.\,(0; \boldsymbol{\Sigma}_{u,\boldsymbol{\epsilon}})$ where $\Sigma_{u,\boldsymbol{\epsilon}} = \begin{pmatrix} \sigma_u^2 & \sigma_{u,\boldsymbol{\epsilon}}' \\ \sigma_{u,\boldsymbol{\epsilon}} & \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \end{pmatrix}$ being $\sigma_{u,\boldsymbol{\epsilon}}$ a covariance vector and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ a $(k_3 \times k_3)$ covariance matrix, σ_u^2 is a scalar. Under this setting, Qu and Lee (2015) obtain the 2SIV estimators, using control function approach to account for the endogeneity of the weights (there is a second source of endogeneity in (20), which is $\mathbf{Wy}$), QMLE and GMM, using the strong exogeneity of the x variables. The three estimators are consistent and asymptotically normally distributed at a rate $\sqrt{N}$. 2SIV and GMM avoid the calculus of the Jacobian term, which can be computationally demanding for the large sample case. As it is well-known, QMLE are asymptotically efficient under normality but is not longer efficient in the absence of normality. The 2SIV uses only mean linear moments; it is less efficient than the QMLE. The GMM adds proper quadratic moment conditions and can be asymptotically as efficient as the QMLE under normality; it is more efficient than the 2SIV.

The framework of (19)–(20), with the same endogeneity structure, appears in Hsieh and Lee (2014) where the authors present a model of endogenous friendship formation, and possible selection bias. The endogeneity in the adjacency matrix arises because the weights are parameterized through a logit model on a set of exogenous (observed) and latent (unobserved) variables that describe the (unknown) position of each individual in the social space of the group. The link decisions may remain independent of each other decisions in the network, which requires that the utility function behind the binary variable equation be separable. However, relations in a social network are characterized by homophily, transitivity and clustering which point to dependency in the individual decisions. The resulting specification, called *Selection Corrected SAR* model (SCSAR), breaks with the correlation between the error terms ϵ and $\mathbf{u}$ but is highly nonlinear. Hsieh and Lee (2014) introduce Bayesian MCMC procedures for the two cases of dependent and independent decision, that seem to work properly.

8 Computational Software

One characteristic of social network analysis is that most of the available software is free and/or open sourced. A second, also important characteristic, is the value given to visualization. Visual representation of social networks is vital to understand the network data and convey the result of the analysis. This explains why many of the analytic available software requires to have modules for network visualization. In general, this software aims the user to be able to interact with the representation, manipulate the structures, shapes and colors to reveal hidden patterns. The goal is to help data analysts to make hypothesis, intuitively discover patterns, isolate singularities or faults during data sourcing.

Generally, network analysis software consists of either packages based on graphical user interfaces (GUIs), or packages built for scripting/programming languages. Widely used and well-documented GUI packages include freeware software like NetMiner, UCINet, Pajek; opensource software like GUESS, ORA, Cytoscape, Gephi and muxViz; and also private GUI packages (mainly directed at business customers) like: Orgnet, which provides training on the use of its software, Polinode, Keyhubs, KeyLines, KXEN and Keynetiq.

Commonly used and well-documented scripting tools used for network analysis include: NetMiner with Python scripting engine, the StatNet suite of packages for the R statistical programming language, igraph, which has packages for R and Python, muxViz (based on R statistical programming language and GNU Octave) for the analysis and visualization of multilayer networks.[2]

[2] Readers interested in software aspects of social networks analysis are recommended to visit the NetWiki web page where they can find updated information about current available software. The content of this section has clearly benefited from the information contained in that webpage.

9 Concluding Remarks

One of the main points that this chapter highlights is that the econometric treatment of social network (interaction) analysis is closely similar to what is known by spatial econometric methods. In ample terms, the very nature of social interaction models and spatial economics have a common origin, namely, the social multiplier. Distance can be understood in physical terms, but not only. Social distances, as a term inclusive of notions like political, linguistic, cognitive, and genetic distances—among others—are the common epycenters of spatial econometrics and social network analysis. It is therefore not a surprise that the evolution of both strings, although through different avenues, has lead to share the main technical issues, as for example the role of the sociomatrices **W-type**, and the identification concerns. It is also expected the solutions eventually provided might serve for both (connected) lines of research.

Acknowledgements This chapter is based upon work from COST Action ISCH COST Action IS1104 "The EU in the new complex geography of economic systems: models, tools and policy evaluation", supported by COST (European Cooperation in Science and Technology), www.cost.eu. Mariano Matilla-García and Jesús Mur are grateful to the following research grants, respectively: ECO2015-65637-P and ECO2015-65758-P from the Ministerio de Economia y Competitividad de España and FEDER funds.

References

Ahrens A, Bhattacharjee A (2015) A two-step lasso estimation of the spatial weights matrix. Econometrics 3:128–155

Beenstock M, Felsenstein D (2012) Nonparametric estimation of the spatial connectivity matrix using spatial panel data. Geogr Anal 44:386–397

Benhabib J, Bisin A, Jackson M (eds) (2011) Handbook of social economics, vol 1. Elsevier, North-Holland

Bhattacharjee A, Holly S (2013) Understanding interactions in social networks and committees. Spat Econ Anal 8:23–53

Bhattacharjee A, Jensen-Butler C (2013) Estimation of the spatial weights matrix under structural constraints. Reg Sci Urban Econ 43:617–634

Blume LE, Brock WA, Durlauf SN, Ioannides YM (2011) Identification of social interactions. In: Benhabib J, Bisin A, Jackson MO (eds) Handbook of social economics, vol 1. North-Holland, pp 853–964

Blume LE, Brock WA, Durlauf SN, Jayaraman R (2015) Linear social interactions models.J Polit Econ 123:444–496

Bonacich P (1987) Power and centrality: a family of measures. Am J Sociol 92:1170–1182

Bramoullé Y, Djebbari H, Fortin B (2009) Identification of peer effects through social networks. J Econ 150:41–55

Brock W, Durlauf S (1995) Discrete choice with social interactions i: theory. Working paper 5291. Madison: Social Science Research Institute

Brock WA, Durlauf SN (2001) Interactions-base models. In: Heckman JJ, Leamer EE (eds) Handbook of econometrics. North-Holland, Amsterdam, pp 3297–3380

Calvó-Armengol A, Patacchini E, Zenou Y (2009) Peer effects and social networks in education. Rev Econ Stud 76:1239–1267
Comin D, Mestieri M (2014) Technology diffusion: measurement, causes and consequences. In: Aghion P, Durlauf S (eds) Handbook of economic growth, vol 2. Elsevier, Amsterdam, pp 565–622
Conley TG, Topa G (2002) Socio-economic distance and spatial patterns in unemployment. J Appl Econ 17:303–332
Corrado L, Fingleton B (2012) Where is the economics in spatial econometrics? J Reg Sci 52:210–239
Davidson J (2000) Econometric theory. Blackwell Publishers, Oxford
Durlauf S (2004) Neighborhood interactions. In: Henderson JV, Thisse J-F (eds) Handbook of regional and urban economics, vol 4. North Holland, Amsterdam
Duesenberry JS (1949) Income, saving, and the theory of consumer behavior. Cambridge, Harvard University Press
Fernández E, Mayor M, Rodríguez J (2009) Estimating spatial au-toregressive models by GME-GCE techniques. Int Reg Sci Rev 32:148–172
Fujita M, Thisse JF (2002) Economics of agglomeration. Cities, industrial location and regional growth. Cambridge University Press, Cambridge
Getis A, Aldstadt J (2004) Constructing the spatial weight matrix using a local statistic. Geogr Anal 36:90–104
Griffith D (1996) Some guidelines for specifying the geographic weights matrix contained in spatial statistical models. In: Arlinghaus S (ed) Practical handbook of spatial statistics. CRC, Boca Raton, pp 104–132
Harris R, Moffat J, Kravtsova V (2011) In search of W. Spat Econ Anal 6:249–270
Hsieh C-S, Lee LF (2014) A social interactions model with endogenous friendship formation and selectivity. J Appl Econ. doi:10.1002/jae.2426
Keane MP (2010) Structural vs. atheoretic approaches to econometrics. J Econ 156:3–20
Kelejian HK, Piras G (2014) Estimation of spatial models with endogenous weighting matrices, and an application to a demand model for cigarettes. Reg Sci Urban Econ 46:140–149
Kelejian HH, Prucha IR (1999) A generalized moments estimator for the autoregressive parameter in a spatial model. Int Econ Rev 40:509–533
Kelejian HK, Prucha IR (2001) On the asymptotic distribution of the Moran I test statistic with applications. J Econ 104:219–257
Kooijman S (1976) Some remarks on the statistical analysis of grids especially with respect to ecology. Ann Syst Res 5:25–48
Lee L (2007) Identification and estimation of econometric models with group interactions, contextual factors and fixed effects. J Econ 140:333–374
Lee LF (2010) Identification and estimation of econometric models with group interactions, contextual factors and fixed effects. Econ J 140:333–374
Lee LF, Yu J (2010) A dynamic panel estimator with both fixed and spatial effects. Econ Theory 26:564–597
Lee LF, Yu J (2012) QML estimation of spatial dynamic panel data models with time varying spatial weights matrices. Spat Econ Anal 7:31–74
Lee LF, Liu X, Lin X (2010) Specification and estimation of social interaction models with network structures. Econ J 13:145–176
Leibenstein J (1950) Bandwagon, snob, and veblen effects in the theory of consumer's behavior. Q J Econ DXIV(2):183–207
Liu X, Lee LF (2010) GMM estimation of social interaction models with centrality. J Econ 159:99–115
Manski C (1993) Identification of endogenous social effects: the reflection problem. Rev Econ Stud 60:531–542
Meen G (1996) Spatial aggregation, spatial dependence and predictability in the UK housing market. Hous Stud 11:345–372

Moffitt RA (2001) Policy interventions, low-level equilibria, and social interactions. In: Durlauf S, Young HP (eds) Soc Dyn. MIT, Cambridge, pp 45–82

Patacchini E, Zenou Y (2010) Juvenile delinquency and conformism. J Law Econ Organ 28:1–31

Qu X, Lee LF (2015) Estimating a spatial autoregressive model with an endogenous spatial weight matrix. J Econ 184:209–232

Schelling T (1971) Dynamic models of segregation. J Math Soc 1:143–86

Schelling T (1972) A process of residential segregation: neighborhood tipping. In: Pascal A (ed) Racial discrimination in economic life. Heath, Lexington

Schelling T (1978) Micromotives and macromotives. Norton, New York

Snijders T, van de Bunt G, Steglich C (2010) Introduction to stochastic actor-based models for network dynamics. Soc Netw 32:44–60

Veblen T (1934) The theory of leisure class: an economic study of institutions. Modern Library, New York

Waldinger F (2011) Peer effects in science: evidence from the dismissal of scientists in nazi Germany. Rev Econ Stud 20:1–24

An Overview of the Measurement of Segregation: Classical Approaches and Social Network Analysis

Antonio Rodriguez-Moral and Marc Vorsatz

Abstract We present a comprehensive overview of the literature on the measurement on segregation. With a focus on the evenness and exposure dimensions—two of the five dimensions of segregation in the multi-dimensional framework defined by Massey and Denton (Soc Forces 67(2):281–315, 1988)—we introduce some of the most relevant segregation measures developed under the classical statistical approach and under the social networks analysis framework. We also briefly describe two different approaches for the definition of segregation measures when using social networks, namely the use of descriptive graph statistics and the use of spectral graph theory.

Keywords Assortative mixing • Graph statistics • Homophily • Network • Segregation • Spectral graph theory

JEL-Numbers: C0, D85, Z13

1 Introduction

The term segregation is often used to refer to the "unequal" distribution of two or more groups of people according to some characteristic like race, religion, gender, income, or wealth, across different social units or positions. Hutchens (2001) refers

A. Rodriguez-Moral
Universidad Nacional de Educación a Distancia (UNED), Paseo Senda del Rey 11, 28040 Madrid, Spain
e-mail: arodrigue2431@alumno.uned.es

M. Vorsatz (✉)
Departamento de Análisis Económico II, Universidad Nacional de Educación a Distancia (UNED), Paseo Senda del Rey 11, 28040 Madrid, Spain

Fundación de Estudios de Economía Aplicada (FEDEA), Calle Jorge Juan 46, 28001 Madrid, Spain
e-mail: mvorsatz@cee.uned.es

P. Commendatore et al. (eds.), *Complex Networks and Dynamics*, Lecture Notes in Economics and Mathematical Systems 683,
DOI 10.1007/978-3-319-40803-3_5

to segregation as a shortcut for the more cumbersome "inequality in the distribution of people across groups", while noting that "it should, however, be clear that they are one and the same." Following this perspective, ethnic residential segregation would address the unequal distribution of people belonging to different ethnic groups across neighborhoods, while occupational segregation would address the way in which some groups (e.g. women or minorities) are unequally represented in different occupational classes. In an alternative but related view, segregation refers to restrictions on the interactions between people, or to restrictions on the access of people or organizations to specific resources or physical space. Along this line, some authors—for example, Berry (1958) and Hunt and Walker (1974)—define segregation as a form of isolation which places limits upon contact, communication, and social relations, while van der Zanden (1972) defined segregation as a process or state whereby people are separated or set apart.

Segregation has been, and continues to be, a source of great concern; see, among many others, Benabou (1993)—who finds that underemployment is more extensive the easier it is for high-skill workers to isolate themselves from others—Borjas (1995)—who finds that ethnic residential segregation influence inter-generational mobility—Poterba (1997) and Goldin and Katz (1999)—who show how ethnic fragmentation results in less spending on education—Easterly and Levine (1997) and Alesina and La Ferrara (2005)—who show how ethnic fragmentation reduces growth—Edin et al. (2003)—who estimate the causal effect on labor market outcomes of living in ethnic "enclaves" within metropolitan areas—Card and Rothstein (2007)—who study the effects of school and neighborhood segregation on the relative SAT scores of black students across different metropolitan areas—Kling et al. (2007)—who, although found no consistent evidence of neighborhood effects on adult earnings or welfare participation in the U.S. Moving to Opportunity (MTO) program, did find some positive effects on teenagers—and Zenou (2009)—who studies the connection between ethnic identity and the labor market outcome of immigrants in different European countries.

It is then clear that empirical work on segregation requires to transform perceptions of segregation into something measurable, that can be used either in a normative way, such as when striving for low segregation in a population, or in a positive way, with the goal of comparing specific outcomes across different groups, settings, or time points. For example, to determine whether the distribution of women across occupations has become less segregated (or more equal) over time, or to compare different districts of a city, or several cities, in terms of the ethnic segregation of neighborhoods, we require measures of segregation that can work with these variables in cross-sectional and inter-temporal analytical settings. Developing accurate measures of segregation is also a necessary first step in assessing the causes of separation between groups and the social and the economic consequences of these patterns, and in formulating appropriate policies for them.

The objective of this work is thus to present a comprehensive overview of the literature on the measurement on segregation. First, we introduce the "classical" statistical approach to the measurement of segregation, and the multi-dimensional framework proposed by Massey and Denton (1988). We briefly describe some of the most relevant segregation measures found in the literature along two of the dimensions in that framework (evenness and exposure). We then introduce the analysis of segregation under the social networks framework, and briefly describe some of the most relevant measures of segregation in this area, which build either on descriptive graph statistics or on spectral graph theory. Finally, we conclude.

2 The Classical Setting

Consider a set $N = \{1, 2, \ldots, n\}$ of n individuals that, for simplicity, we will call a society. These individuals live in a spatial region C (e.g. a city), which is composed of a finite set $M = \{1, 2, \ldots, m\}$ of m subareas or units (e.g. neighborhoods or census tracts). Also, there is a set G of k mutually exclusive groups that forms a partition of the society. One can think of a group $g \in G$ as a subset of members of the society that share a particular attribute such as religion or ethnicity. Let $n_{g,i}$ be the number of individuals of group $g \in G$ who live in subarea $i \in M$, n_i the total number of individuals who live in subarea $i \in M$, and n_g the total population of group $g \in G$. Note that $n_g = \sum_{i \in M} n_{g,i}$, $n_i = \sum_{g \in G} n_{g,i}$, and $n = \sum_{i \in M} n_i = \sum_{g \in G} n_g$. Thus, the distribution of individuals belonging to each group across subareas can be expressed by a $g \times m$ matrix $\mathbf{N} = (n_{g,i})_{g \in G, i \in M}$. Let $\mathcal{N}$ denote the set of all those matrices. Given matrix $\mathbf{N}$, we can define the concentration and density column vectors of group g as $\mathbf{c_g}(\mathbf{N}) = (n_{g,i}/n_i)_{i \in M}$ and $\mathbf{d_g}(\mathbf{N}) = (n_{g,i}/n_g)_{i \in M}$, respectively. Whenever there is no room for confusion we will drop the matrix $\mathbf{N}$ to which all these operators refer to. Let $c_{g,i}$ and $d_{g,i}$ be the i–th elements of $\mathbf{c_g}$ and $\mathbf{d_g}$, respectively. The proportion of group g in the total population is defined as $c_g = n_g/n$.

For example, if there are only two groups X and Y in a spatial region C_1 with four subareas, we could have

$$\mathbf{N_1} = \begin{pmatrix} 8 & 4 & 0 & 4 \\ 16 & 16 & 4 & 12 \end{pmatrix}.$$

So, the density and concentration vectors for group X would be $\mathbf{c_X}(\mathbf{N_1})^{\mathrm{T}} = (1/3, 1/5, 0, 1/4)$ and $\mathbf{d_X}(\mathbf{N_1})^{\mathrm{T}} = (1/2, 1/4, 0, 1/4)$. An alternative distribution for region C_1 could be

$$\mathbf{N_1'} = \begin{pmatrix} 2 & 8 & 4 & 2 \\ 9 & 18 & 15 & 6 \end{pmatrix}.$$

Let us also consider a second region C_2, with the following distribution of individuals of groups X and Y, in this case across five subareas:

$$\mathbf{N_2} = \begin{pmatrix} 4 & 2 & 3 & 0 & 3 \\ 6 & 6 & 12 & 3 & 9 \end{pmatrix}.$$

Given this information, one would like to know if the segregation in region C_1 is reduced when the distribution moves from $\mathbf{N_1}$ to $\mathbf{N_1'}$. Similarly, is region C_2, under distribution $\mathbf{N_2}$, more or less segregated than region C_1 under distribution $\mathbf{N_1}$?

The main type of mathematical device developed to measure segregation are segregation indices. Formally, a normalized segregation index is a function $S : \mathcal{N} \rightarrow [0, 1]$ that maps any distribution $\mathbf{N}$ into the unit interval. Reardon and Firebaugh (2002) introduce various approaches to multi-group segregation indices. In what follows, we consider the so-called disproportionality approach according to which one assesses the variability of the values of $c_{g,i}$ across units. Let $\mathbf{R} = (c_{g,i}/c_g)_{g \in G, i \in M}$ be the $g \times m$ *disproportionality matrix*. The ratio $r_{g,i}$ reflects the extent to which group g is disproportionately represented in subarea $i \in M$ with respect to its overall ratio c_g. Then, let $W(\mathbf{N})$ be the weighted average of all deviations of $r_{g,i}$, measured by a *disproportionality function f*:

$$W(\mathbf{N}) = \sum_{g \in G} c_g \sum_{i \in M} \frac{n_i}{n} f(r_{g,i}).$$

There are many possible disproportionality functions, which will lead to different segregation indices; what is required from an appropriate disproportionality function f is that (a) $f(1) = 0$, so a group g that is neither under- nor over-represented in subarea i does not contribute to segregation, and (b) the segregation measure $W(\mathbf{N})$ satisfies the disproportionality axiom for segregation: segregation is zero only when $r_{g,i} = 1$ for all g and i, otherwise segregation is greater than zero.

It is important to note that all the functions defined for the segregation indices included in this section are such that the maximum value $W(\mathbf{N})$ is obtained when each unit contains individuals from a single group—see Reardon and Firebaugh (2002):

$$\begin{aligned} \max\{W(\mathbf{N})\} &= \sum_{g \in G} c_g \left[\sum_{i \in M: r_{g,i}=0} \frac{n_i}{n} \cdot f(0) + \sum_{i \in M: r_{g,i}=1/c_g} \frac{n_i}{n} \cdot f(1/c_g) \right] \\ &= \sum_{g \in G} c_g [(1 - c_g) f(0) + c_g \cdot f(1/c_g)]. \end{aligned}$$

The *normalized segregation index* $S(\mathbf{N})$ is then equal to the ratio between the weighted average disproportionality and its maximum possible value:

$$S(\mathbf{N}) = \frac{W(\mathbf{N})}{\max\{W(\mathbf{N})\}}.$$

The probably most widely used segregation measure is the *dissimilarity index*, defined originally in its two-group version by Jahn et al. (1947) and Duncan and Duncan (1955). In its generalized, multi-group version (Morgan 1975 and Sakoda (1981)), D is obtained from disproportionality function $f(r_{g,i}) = \frac{1}{2} \left| r_{g,i} - 1 \right|$. In this case, the maximum of $W(\mathbf{N})$ is $\max\{W(\mathbf{N})\} = \sum_{g\in G} c_g(1-c_g)$, also known as the Simpson's interaction index $I(\mathbf{N})$. The dissimilarity index then becomes:

$$D(\mathbf{N}) = \frac{1}{2I(\mathbf{N})}\sum_{g\in G} c_g \sum_{i\in M}\frac{n_i}{n}\left|r_{g,i}-1\right| = \frac{1}{2I(\mathbf{N})}\sum_{g\in G}\sum_{i\in M}\frac{n_i}{n}\left|c_{g,i}-c_g\right|.$$

In its two-group version, D reduces to

$$D(\mathbf{N}) = \frac{1}{2\,n\,c_X\,(1-c_X)}\sum_{i\in M} n_i|c_{X,i}-c_X|.$$

Some algebraic manipulation gives an alternative way of computing D from the deviations of the density of both groups in each subarea:

$$\begin{aligned}
D(\mathbf{N}) &= \tfrac{1}{2\,n\,c_X\,(1-c_X)}\textstyle\sum_{i\in M} n_i|c_{X,i}-c_{X,i}\,c_X-c_X+c_{X,i}\,c_X| \\
&= \tfrac{1}{2\,n\,c_X\,(1-c_X)}\textstyle\sum_{i\in M} |n_i\,c_{X,i}\,(1-c_X)-n_i\,c_X\,(1-c_{X,i})| \\
&= \tfrac{1}{2}\left|\tfrac{\sum_{i\in M}\;n_i\,c_{X,i}}{n\,c_X}-\tfrac{\sum_{i\in M}\;n_i\,(1-c_{X,i})}{n\,(1-c_X)}\right| \\
&= \tfrac{1}{2}\sum_{i\in M}\left|\tfrac{n_{X,i}}{n_X}-\tfrac{n_{Y,i}}{n_Y}\right| \\
&= \tfrac{1}{2}\sum_{i\in M}\;|d_{X,i}-d_{Y,i}|.
\end{aligned}$$

One interesting feature of the dissimilarity index is that it summarizes the degree to which subareas mirror the demographic balance of the larger area. If integration is completely even throughout the considered area, then each subarea must contain the same share of group X as in the whole region. To see this, let $n^*_{X,i}$ be the population of group X associated with this complete integration on subarea $i\in M$. Then, it must be true that $n^*_{X,i}/n_X - n_{Y,i}/n_Y = 0$. Actual levels of $n_{X,i}$ deviate from $n^*_{X,i}$ by some amount, say $n^\epsilon_{X,i}$, so that $n_{X,i} = n^*_{X,i} + n^\epsilon_{X,i}$. It follows that

$$D(\mathbf{N}) = \tfrac{1}{2}\sum_{i\in M}\left|\tfrac{n_{X,i}}{n_X}-\tfrac{n_{Y,i}}{n_Y}\right| = \tfrac{1}{2}\sum_{i\in M}\left|\tfrac{n^*_{X,i}+n^\epsilon_{X,i}}{n_X}-\tfrac{n_{Y,i}}{n_Y}\right| = \tfrac{1}{2}\sum_{i\in M}\left|\tfrac{n^\epsilon_{X,i}}{n_X}\right|.$$

Thus, D is the fraction of the group X population that would have to move in order to achieve complete integration. Note that the $1/2$ in the formula indicates that in order to achieve this balance, at most one-half of group X individuals would actually have to move; only the individuals living in subareas where they are relatively overrepresented have to move to subareas where they are relatively underrepresented. For distribution $\mathbf{N_1'}$, we obtain that $D(\mathbf{N_1'}) = \frac{1}{8}$, so 2 individuals ($1/8$ of the total group population of 16) from group X would have to move in order to obtain complete integration, going from subarea 2 to subareas 1 and 3, ending up with the following distribution $\mathbf{N_1''}$, for which D equals 0:

$$\mathbf{N_1''} = \begin{pmatrix} 3 & 6 & 5 & 2 \\ 9 & 18 & 15 & 6 \end{pmatrix}.$$

Duncan and Duncan (1955) was probably the first systematic attempt to provide an extensive analysis of measures of residential segregation. Their main conclusion was that there is little information provided by the various segregation indices introduced until then that is not captured by the dissimilarity index. This conclusion, shared by Taeuber and Taeuber (1965), contributed to the popularity of the dissimilarity index, making it the standard index of residential segregation for over 20 years. However, the consensus built around the dissimilarity index was challenged by Cortese et al. (1976), which opened another period of debate until the mid 1980s.[1] At that point an important step forward was made by James and Taeuber (1985), who—drawing on Schwartz and Winship (1980) work on inequality measurement—developed a set of four criteria, based on how segregation indices must react to changes in the groups distribution between subareas. Shortly later a crucial advance was made by Massey and Denton (1988), not only with respect to how to define segregation but also with respect to how it should be measured. In the context of racial segregation of city neighborhoods Massey and Denton (1988) defined segregation as *the degree to which two or more groups live separately from one another.* In an effort to bring some order to the field, they undertook a systematic analysis of nineteen segregation indices they identified from a review of the extant literature. By looking at residential segregation as the result of various social phenomena, they argued that segregation is not a unidimensional concept, but encompasses the following five distinct dimensions of variation.

1. *Evenness* refers to the distribution differences of social groups across subareas.
2. *Exposure* addresses the degree of potential contact of a group with other groups.
3. *Concentration* is the amount of relative physical space occupied by a group in a given geographical area.
4. *Centralization* reflects the degree to which a group is spatially distributed close to the center of an urban area.
5. *Clustering* refers to the degree of agglomeration, i.e., the extent to which areas inhabited by members from a social group adjoin one another in space.

Figure 1 illustrates the differences among the five dimensions of segregation by means of a simple example: the distribution of households within a city with a square layout, where each square is occupied by either a minority family (black squares) or by a majority family (white squares). In each case, the city is subdivided in 16 geographical subareas, each comprising a 3×3 square. We can observe in the first column that, when considering the evenness dimension, the upper pattern displays a high level of segregation, since each subarea is occupied by individuals from a single group (either all white or all black), while the lower pattern displays a low level of segregation, since in each subarea the proportion of black squares is roughly

[1]This debate can be recreated through Cohen et al. (1976), Cortese et al. (1976), Taeuber and Taeuber (1976), Winship (1977, 1978), Falk et al. (1978), Kestembaum (1980), Coleman et al. (1982), Lieberson and Carter (1982), White (1983), James and Taeuber (1985), and the many references therein.

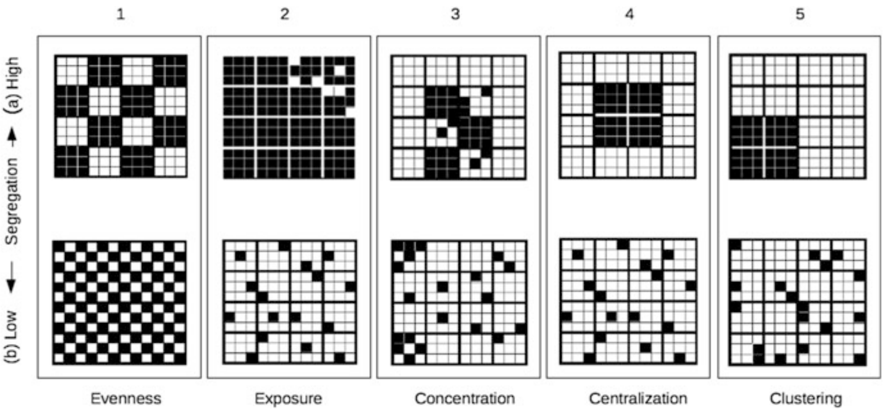


Fig. 1 The dimensions of segregation

the same as in the overall region (1/2). Column 3 shows that segregation in the concentration dimension arises when the minority group occupies a small proportion of land within the city, regardless of whether these small areas are contiguous or they are not. On the other hand, column 5 shows that a high degree of clustering arises from a residential structure where minority areas are arranged contiguously, creating one large agglomeration, whereas a low level of clustering means that minority areas are widely scattered around the urban environment.

At this point it is important to note that the concrete measurement of the five dimensions requires different kinds of data. Measures of evenness and exposure only require information about the population of the different groups in every subarea, so they can defined on the set of matrices $\mathcal{N}$. On the other hand, measures of concentration, centralization, and clustering need some sort of geographical or spatial data. Consequently, we concentrate for the moment on the dimensions of evenness and exposure, and will come back formally to the other dimensions in the next section. Also, in a historical context, note that most of the empirical analysis of segregation in economics has focused until lately on the study of evenness and exposure. This could be due to the fact that, from a conceptual perspective, it is easier to define and derive indices for these two dimensions.

2.1 Evenness

Massey and Denton (1988) recommend using the dissimilarity index in most cases. Two commonly mentioned benefits of the dissimilarity index are that it is easy to compute and interpret and that it can be consistently computed with available census data over long periods of time. Some of its main drawbacks are that (a) it is strongly affected by random factors when the number of minority members is small relative to the number of units, (b) it is insensitive to the redistribution of minority members among units with minority proportions above or below the city's minority proportion, and (c) it fails to satisfy the adapted Pigou-Dalton "transfer principle" from the literature on the measurement of income inequality to the measurement of segregation (James and Taeuber 1985). Alternatives to the dissimilarity index are the *Gini index* (Gini 1921), the *entropy* or *information index* (Theil 1972 and Theil and Finizza 1971), or the *Atkinson index* (Atkinson 1970).

From a formal point of view the multi-group Gini index $G(\mathbf{N})$ is obtained if $f(r_{g,i}) = 1/2 \cdot \sum_{j\in M} n_j/n \cdot |r_{g,i} - r_{g,j}|$, which takes into account absolute deviations of all pairwise comparisons in the region. This results in a measure of average disproportionality

$$W(\mathbf{N}) = \sum_{g\in G} c_g \sum_{i\in M} \frac{n_i}{n} f(r_{g,i}) = \frac{1}{2} \sum_{g\in G} c_g \sum_{i\in M} \sum_{j\in M} \frac{n_i n_j}{n^2} \left| r_{g,i} - r_{g,j} \right| .$$

Its maximum value is, as for the dissimilarity index, the Simpson's interaction index $I(\mathbf{N})$. The resulting generalized Gini index is defined as

$$G(\mathbf{N}) = \frac{1}{2I(\mathbf{N})} \sum_{g\in G} c_g \sum_{i\in M} \sum_{j\in M} \frac{n_i n_j}{n^2} \left| r_{g,i} - r_{g,j} \right| = \frac{1}{2I(\mathbf{N})} \sum_{g\in G} \sum_{i\in M} \sum_{j\in M} \frac{n_i n_j}{n^2} \left| c_{g,i} - c_{g,j} \right|.$$

The multi-group index reduces to the standard expression of the Gini index in the case of two groups:

$$G(\mathbf{N}) = \frac{1}{2\, n^2\, c_X\, (1 - c_X)} \sum_{i\in M} \sum_{j\in M} n_i\, n_j\, |c_{X,i} - c_{X,j}|.$$

The entropy or Theil information theory index[2] $H(\mathbf{N})$ measures the weighted average deviation of each subarea from the region-wide "entropy" or diversity. Although Theil originally derived H from information theory, Reardon and Firebaugh (2002) prove that it corresponds to the disproportionality function $f(r_{g,i}) = r_{g,i} \ln(r_{g,i})$. Then,

$$W(\mathbf{N}) = \sum_{g\in G} c_g \sum_{i\in M} \frac{n_i}{n} f(r_{g,i}) = \sum_{g\in G} c_g \sum_{i\in M} \frac{n_i}{n} r_{g,i} \ln(r_{g,i}).$$

[2]Often referred to as the Shannon index, after the related work on information theory (Shannon 1948; Shannon and Weaver 1949; see also Khinchin 1957). Its application to the analysis of segregation was introduced by Theil and Finizza (1971) in an analysis of racial entropy in the Chicago public schools.

Its maximum value is $E(\mathbf{N}) \equiv \max\{W(\mathbf{N})\} = \sum_{g \in G} c_g \ln(1/c_g)$, which is known as Theil's entropy index. The resulting generalized Theil index is defined as

$$H(\mathbf{N}) = \frac{1}{E(\mathbf{N})} \sum_{g \in G} c_g \sum_{i \in M} \frac{n_i}{n} r_{g,i} \ln(r_{g,i}) = \frac{1}{E(\mathbf{N})} \sum_{g \in G} \sum_{i \in M} \frac{n_i}{n} c_{g,i} \ln\left(\frac{c_{g,i}}{c_g}\right).$$

In the two-group case the entropy E of the whole region and the entropy E_i of a given subarea $i \in M$ reduce to

$$E(\mathbf{N}) = c_X \ln\left(\frac{1}{c_X}\right) + (1 - c_X) \ln\left(\frac{1}{1 - c_X}\right)$$

and

$$E_i(\mathbf{N}) = c_{X,i} \ln\left(\frac{1}{c_{X,i}}\right) + (1 - c_{X,i}) \ln\left(\frac{1}{1 - c_{X,i}}\right).$$

The information index H is then equal to

$$H(\mathbf{N}) = \sum_{i \in M} \left(\frac{n_i}{n}\right) \left(\frac{E - E_i}{E}\right).$$

The family of Atkinson segregation indices A were introduced by James and Taeuber (1985) for the two-group case. They are based on the Atkinson inequality indices (Atkinson 1970), and are defined as

$$A_b(\mathbf{N}) = \left(\frac{1}{n\, c_X\,(1 - c_X)}\right) \left| \sum_{i \in M} \left(1 - c_{X,i}\right)^{1-b} \cdot c_{X,i}^{b} \cdot n_i \right|^{\frac{1}{1-b}}.$$

For a given distribution $\mathbf{N}$ the specific Atkinson index depends on the value of the shape parameter b. For small values of the shape parameter, $0 < b < 0.5$, subareas where the proportion of the group is smaller than the region's average contribute more to the segregation index; for large values of the shape parameter, $0.5 < b < 1.0$, the reverse is true. When the shape parameter is $b = 0.5$ all subareas are weighted equally. The Atkinson segregation index has been generalized to the multi-group case by Frankel and Volij (2008), who define the generalized Atkinson index as

$$A_w(\mathbf{N}) = 1 - \sum_{i \in M} \prod_{g \in G} \left(c_{g,i}\right)^{w_g},$$

where $\mathbf{w} = (w_1, \ldots, w_k)$ is a vector of k fixed non-negative weights that sum to one. When all weights are equal, we obtain the symmetric Atkinson index, which reduces to an increasing transformation of the original index defined by James and Taeuber (1985) in the two-group case, preserving the same properties.

An alternative concept and associated mathematical device for measuring segregation in the evenness dimension is that of *segregation curves*. Being aware of the similarities between the large body of research on income inequality and the research that had been developed in parallel on occupational segregation—already noticed by Duncan and Duncan (1955) and Winship (1978)—Hutchens (1991) took the analytic framework used to analyze income inequality—Lorenz curves—and translated it to the problem of analyzing occupational segregation in the two-group case. de la Vega and Volij (2014) adapt a similar mathematical device, borrowed from the literature on the value of information to extend the Lorenz criterion to the multi-group case. This allows them to define a partial ordering of cities (distributions), according to the informativeness of their neighborhoods (units). Given a city, the location of a randomly selected individual is seen as a signal that provides information about the group he belongs to. In this sense, the collection of distributions of the various groups across locations can be seen as an experiment in the sense of Blackwell (1951, 1953), one in which locations play the role of signals and groups play the role of states of nature. A city whose locations are more informative than another city's locations is then considered more segregated than the latter. This correspondence between informativeness and segregation allows to construct a partial order of cities. de la Vega and Volij (2014) show that any partial segregation order of cities that satisfies four basic axioms that go back to James and Taeuber (1985) must be consistent with the partial segregation order induced by the informativeness of their neighborhoods, which then can be regarded as a generalization of the standard order based on segregation curves. When restricted to the two-group case, this partial ordering coincides with the Lorenz partial ordering derived from segregation curves.

2.2 Exposure

Rather than measuring segregation as a departure from some abstract ideal of evenness, exposure indices try to measure the experience of segregation from the viewpoint of the average individual. Although measurement indices for exposure and evenness are usually correlated empirically, they are conceptually distinct. Members of a minority group can be evenly distributed among residential areas of a city, but at the same time experience little exposure to majority members if they comprise a relatively large share of the city. Conversely, if the areas they occupy comprise a small proportion of the city, minority members tend to experience high levels of exposure to majority members, no matter what the level of evenness.

Early research on exposure measurement was done by Bell (1954). However, after the 1955 consensus around the dissimilarity index, exposure was largely forgotten until Lieberson and Carter (1982) reintroduced the *isolation index* and its counterpart, the *interaction index*. More precisely, the isolation index $P^*_{g,g}(\mathbf{N})$ equals the probability that a randomly drawn member of group g shares a neighborhood with a member from the same group; that is,

$$P^*_{g,g}(\mathbf{N}) = \sum_{i \in M} \left(\frac{n_{g,i}}{n_g}\right)\left(\frac{n_{g,i}}{n_i}\right) = \sum_{i \in M} d_{g,i}\, c_{g,i} = \mathbf{d}_g^T \mathbf{c}_g.$$

Related to the interaction index is the isolation index, which reflects the probability that a member of group g interacts with a member of a distinct group h in her neighborhood:

$$P^*_{g,h}(\mathbf{N}) = \sum_{i \in M} \left(\frac{n_{g,i}}{n_g}\right)\left(\frac{n_{h,i}}{n_i}\right) = \sum_{i \in M} d_{g,i}\, c_{h,i} = \mathbf{d}_g^T \mathbf{c}_h.$$

The main objections to the isolation index are that it confounds population composition with segregation patterns in such a way that the resulting index value heavily reflects the former and that it creates asymmetrical measures of segregation when it is more desirable to have a single measure for both groups. However, the isolation index can be normalized to control for population composition and eliminate the asymmetry mentioned above. Then, one obtains the well-known *correlation ratio* η^2 (White 1986):

$$\eta_g^2(\mathbf{N}) = \frac{P^*_{g,g}(\mathbf{N}) - c_g}{1 - c_g} = \frac{\sum_{i \in M} n_i \cdot c_{g,i}^2}{n\, c_g\, (1 - c_g)^2} - \frac{c_g}{1 - c_g}.$$

Stearns and Logan (1986) argue that η_g^2 constitutes an independent dimension of segregation, but Massey and Denton (1988) hold that it actually encompasses both the exposure and evenness dimensions. In fact, being derived from $P^*_{g,g}(\mathbf{N})$, η_g^2 displays some properties associated with an exposure measure, but normalization also gives it the qualities of an evenness index. Massey and Denton (1988) demonstrated this duality empirically and argued that it is better to use D and P^* as separate measures of evenness and exposure. Nonetheless, Jargowsky (1996) has shown that one version of η_g^2 yields a better and more concise measure of segregation when one wishes to measure segregation between multiple groups simultaneously.

3 Segregation in Social Networks

A natural way to capture the dimensions of concentration, centralization, and clustering, and, to some extent, the spatial aspects of segregation, is to introduce a graph-theoretical approach.[3]

[3] Massey and Denton (1988) propose some indices as the most appropriate for these three dimensions. A large body of research in the literature relates to the definition of spatial measures, see Reardon and O'Sullivan (2004) and Cohn and Jackman (2011) for a detailed account.

Formally, a graph is a pair (V, E), where V denotes its set of v nodes (vertices) and E denotes its set of e edges (links or ties). Each edge in E is a pair of vertices, with the edge connecting distinct nodes i and j written as (i, j). We assume that the graph connectivity is defined through a binary, irreflexive, and symmetric relation on $V \times V$. This relation induces a $v \times v$ adjacency matrix $\mathbf{A}$ such that $a_{i,j} = 1$ if i and j are connected and $a_{i,j} = 0$ otherwise. Observe that this framework can be straightforwardly extended to asymmetric connections, for example, when i is connected to j but j is not connected to i. We do not allow for self-loops (hence $a_{i,i} = 0$) or overlapping links; i.e., there cannot be more than one link between i and j. Such a graph is known as a simple graph. The degree of a node $i \in V$ is $\eta_i = \sum_{j \in V} a_{i,j}$.

Nodes have attributes, which model some properties of interest. Enumerative or discrete attributes are those that lack any particular ordering and represent properties like ethnicity, gender, *etc.* Scalar attributes, on the other hand, are those that can be represented by a scalar variable (typically by integers or real numbers) and model properties like age, income, *etc.* For simplicity, we will only consider a single node attribute t that, depending on the context, can be an enumerative or a scalar attribute. Let $\mathbf{t} = (t_i)_{i \in V}$ be the column vector that represents the attributes of all nodes. We consider the situation when V is equal to the set of all individuals N. An alternative but very much related approach would be to assume $V = M$. Hence, for us, a network is a tuple $R = \langle N, \mathbf{t}, \mathbf{A} \rangle$.

In social networks, individuals have a disproportionately strong tendency to associate with others who are similar to them in characteristics such as age, nationality, language, socioeconomic status, educational level, political beliefs, *etc.* This property, called *homophily* by Lazarsfeld and Merton (1954), is also known in the broader context of generic networks as *assortative mixing* (see, Newman 2003). One consequence of homophily is that social networks show a large degree of homogeneity with regard to many sociodemographic, behavioral, and intrapersonal characteristics. As it happens with segregation, homophily limits people's social worlds in a way that has some powerful implications for the information they receive, the attitudes they form, and the interactions they experience.[4]

From a statistical perspective, assortative mixing in a network simply means that the attributes of nodes correlate across edges; that is, given the link $(i, j) \in E$ the values t_i and t_j tend to be more similar than those in a randomly chosen unconnected pair. If attributes are enumerative, a simple way to quantify the degree of assortative mixing in a graph consists of measuring the increase in the fraction of edges that join nodes of the same type with respect to the fraction we would expect to find if edges

[4]Empirical studies have found strong evidence of homophily with respect to age (e.g. Verbrugge 1977, Marsden 1988, and Burt 1991), education (e.g. Marsden 1987 and Kalmijn 2006), income (e.g. Laumann 1966 and Laumann 1973), ethnicity (e.g. Baerveldt et al. 2004 and Ibarra 1995) or geographical distance (e.g. Campbell 1990 and Wellman 1996). Homophily has been found to matter in a wide range of social interactions such as friendship and marriage, job market outcomes, speed of information diffusion, speed or learning, consensus reaching and even social mobility (see Currarini et al. 2009, Brammoullie and Kranton 2007, and Golub and Jackson 2012a,b).

were formed at random in the network. This measure is known as the *modularity Q* of the network (see Newman 2003). Formally, given two nodes $i, j \in N$, let $\delta(t_i, t_j)$ denote the Kronecker delta function; that is, $\delta(t_i, t_j) = 1$ whenever $t_i = t_j$ and $\delta(t_i, t_j) = 0$ otherwise. Then, the total number of edges that join nodes of the same type is equal to $\sum_{(i,j)\in E} \delta(t_i, t_j) = \frac{1}{2}\sum_{i,j\in N} a_{i,j}\,\delta(t_i, t_j)$.

Calculating the expected number of edges that would form between nodes of the same type at random requires that we clearly specify what "at random" means. Different models of random network formation have been studied, for example the one known as *Erdös-Rènyi random graphs*, in which edges form independently with the same probability p. However, this simple random network model does not work well as benchmark for measuring assortativity, since if some nodes have very high degrees, then their attribute value will show up in more connected pairs, and this will skew the fraction of same-type edges when compared to the one expected in the Erdös-Rènyi model. The solution is to use a different random network model known as the *configuration model*, in which edges are formed at random while preserving the nodes degrees. In this model the probability that two nodes $i, j \in N$ are connected by a link is given by $p_{i,j} = (\eta_i \cdot \eta_j)/(2 \cdot e)$. Hence, the configuration model implies *preferential attachment* (Barabási and Albert 1999)—the higher the degrees of two nodes, the more likely they are to connect. Consequently, under the configuration model, the expected number of edges connecting nodes with the same attribute value is $\frac{1}{2}\sum_{i,j\in N} p_{i,j}\,\delta(t_i, t_j) = \frac{1}{2}\sum_{i,j\in N}\frac{\eta_i\cdot\eta_j}{2e}\,\delta(t_i, t_j)$. Then, the modularity Q of network R with respect to attribute t is given by

$$Q(R) = \frac{1}{2e}\sum_{i,j\in N}\left(a_{i,j} - \frac{\eta_i\eta_j}{2e}\right)\delta(t_i, t_j).$$

As noted by Newman (2010), this measure is strictly less than 1, takes positive values if the number of links between nodes of the same type exceeds what we would expect by chance (in which case we say that the network displays assortative mixing), and negative ones otherwise (in which case we say that the network displays disassortative mixing). Although the value of $Q(R)$ is always less than 1, in general it does not achieve a value of $Q(R) = 1$ even for a perfectly mixed network, defined as one in which all the links run only between same-type nodes. In that case, $\sum_{i,j\in N} a_{i,j}\,\delta(t_i, t_j) = 2e$, and then

$$\max\{Q(R)\} = \frac{1}{2e}\left(2e - \sum_{i,j\in N}\frac{\eta_i\eta_j}{2e}\,\delta(t_i, t_j)\right) = 1 - \sum_{i,j\in N}\frac{\eta_i\eta_j}{(2e)^2}\,\delta(t_i, t_j).$$

Then we can define the *normalized modularity* as

$$\hat{Q}(R) = \frac{Q(R)}{\max\{Q(R)\}} = \frac{\sum\limits_{i,j\in N}\left(a_{i,j} - \frac{\eta_i\eta_j}{2e}\right)\delta(t_i, t_j)}{2e - \sum\limits_{i,j\in N}\left(\frac{\eta_i\eta_j}{2e}\right)\delta(t_i, t_j)}.$$

If the node attribute under consideration is a scalar quantity, then there is additional information to be taken into account, and the appropriate measure becomes a covariance measure $C(R)$ which can be seen as a network-based generalization of the Pearson correlation coefficient. Let $\mu(\mathbf{t})$ be the mean value of t observed at either end of a link:

$$\mu(\mathbf{t}) = \frac{\sum_{i,jN} a_{i,j} t_i}{\sum_{i,j\in N} a_{i,j}} = \frac{\sum_{i\in N} \eta_i t_i}{\sum_{i\in N} \eta_i} = \frac{1}{2e} \sum_{i\in N} \eta_i t_i.$$

Similar, the variance of t is:

$$Var(\mathbf{t}) = \frac{1}{2e} \sum_{i,j\in N} a_{i,j} (t_i - \mu(\mathbf{t}))^2 = \frac{1}{2e} \sum_{i\in N} \eta_i (t_i - \mu(\mathbf{t}))^2.$$

The covariance is then

$$C(R) = \frac{\sum_{i,j\in N} a_{i,j} (t_i - \mu(\mathbf{t}))(t_j - \mu(\mathbf{t}))}{\sum_{i,j\in N} a_{i,j}} = \frac{1}{2e} \sum_{i,j\in N} a_{i,j} t_i t_j - \mu(\mathbf{t})^2 = \frac{1}{2e} \sum_{i,j\in N} \left(a_{i,j} - \frac{\eta_i \eta_j}{2e} \right) t_i t_j.$$

As in the case of the modularity measure, it is useful to normalize this measure by its maximum value so that it ranges from −1 to 1. This results in a normalized covariance measure that has the following functional form:

$$\hat{C}(R) = \frac{C(R)}{\mathrm{Var}(\mathbf{t})} = \frac{\sum_{i,j\in N} \left(a_{i,j} - \frac{\eta_i \eta_j}{2e} \right) t_i t_j}{\sum_{i,j\in N} \left(\eta_i \delta(i,j) - \frac{\eta_i \eta_j}{2e} \right) t_i t_j}.$$

A simple example illustrates the measurement of assortative mixing by a enumerative or scalar attribute. Consider the network depicted in Fig. 2. Each node color represents the node's ethnicity, while each node's label represents the node's age. A quick visual inspection of the graph suggests that this network displays assortative mixing by node color (since nodes link mostly to other nodes of the same color), but disassortative mixing by age (since nodes with higher age values

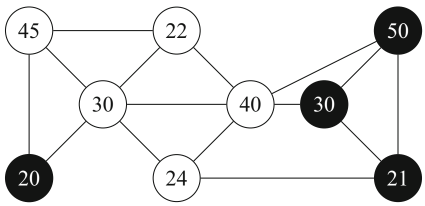


Fig. 2 Measurement of assortative mixing

link mostly to nodes with lower values). This is confirmed by the values of $\hat{Q}(R)$ and $\hat{C}(R)$, which result in 0.31 and −0.21, respectively.

The measures of assortative mixing introduced above provide some general insight of whether individuals with the same characteristics stick together and, in terms of the dimensions of Massey and Denton (1988), they are clearly connected to the dimension of exposure. Yet, several other measures, which can be classified into two approaches, have been suggested. The *descriptive graph statistics* approach uses mathematical devices similar to those introduced for the measurement of modularity in graphs. Different segregation problems can then be studied depending on the type of node attribute that is considered in the analysis; for example, the analysis of occupational segregation by gender, or the analysis of residential segregation by ethnicity can be modeled as a problem of assortative mixing by an enumerative attribute, while the analysis of residential segregation by income may be modeled as a case of assortative mixing by a scalar attribute. The *spectral graph theory* approach, on the other hand, uses a different way of representing and analyzing the graph associated with a network, by using a set of matrices associated with the graph, rather than the matrix that represents the graph of vertices and edges itself. It then proceeds by computing some characteristic values of those associated matrices, such as their eigenvalues and eigenvectors, which reveal some structural properties of the original graph.

3.1 Measures Based on Descriptive Graph Statistics

In what follows, we consider the enumerative attribute t that can take values from the set of groups G. That is, the notation $t_i = g$ indicates that individual $i \in N$ belongs to group $g \in G$. With this notation at hand, we obtain the $n \times k$ type indicator matrix $\mathbf{T}$, where $t_{i,g} = 1$ if individual $i \in N$ is a member of group g, or $t_{i,g} = 0$ otherwise. Given the set $\mathcal{R}$ of networks with the generic element $R = \langle N, \mathbf{T}, \mathbf{A} \rangle$, we can define segregation measures or indices at three levels: (1) A network-level segregation index $S : \mathcal{R} \to \mathbb{R}$, (2) a group-level segregation index $S_g : \mathcal{R} \to \mathbb{R}^k$, and (3) a node-level segregation index $S^i : \mathcal{R} \to \mathbb{R}^n$.

Following Bojanowski and Corten (2014), let the *mixing matrix* $\mathbf{M} = (m_{ghy})_{k \times k \times 2}$ associated to the adjacency matrix $\mathbf{A}$ be a three-dimensional distribution of all the dyads (pairs of nodes) in the graph, based on the following characteristics: (1) The group g to which the first node in the dyad belongs, (2) the group h to which the second node in the dyad belongs, and (3) whether the two nodes in the dyad are connected in the analyzed network. In particular, m_{gh1} denotes the number of connections between individuals of group g and group h. Similarly, m_{gh0} denotes the number of links between type g nodes and type h nodes that are not present in the network. One can look at the mixing matrix as a single matrix composed of two different layers, the "contact layer" mixing matrix $\bar{\mathbf{M}} = \left(m_{gh1}\right)_{g,h \in G}$ and the "non-contact layer" mixing matrix $\tilde{\mathbf{M}} = \left(m_{gh0}\right)_{g,h \in G}$. Both of these matrices can be expressed in terms of the Kronecker delta function:

$$m_{gh1} \equiv \sum_{i,j \in N: t_i = g, t_j = h} a_{i,j} = \sum_{i,j \in N} a_{i,j}\, \delta(t_i, g)\, \delta(t_j, h)$$

and

$$m_{gh0} \equiv \sum_{i,j \in N: t_i = g, t_j = h} (1 - a_{i,j}) = \sum_{i,j \in N} (1 - a_{i,j})\, \delta(t_i, g)\, \delta(t_j, h).$$

The values of the contact layer mixing matrix $\bar{\mathbf{M}}$ can be conveniently calculated as $\bar{\mathbf{M}} = \mathbf{T}^{\mathrm{T}}\mathbf{AT}$. We also use the + sign to denote summation over a particular subscript when dealing with marginal distributions of the mixing matrix. Some of these marginal distributions represent useful descriptive statistics of the network graph:

- $m_{gh+} = m_{gh1} + m_{gh0}$ denotes the total number of potential links between nodes of type g and nodes of type h. Thus, $m_{gh+} = n_g \cdot n_h$ whenever $g \neq h$, while $m_{gg+} = n_g \cdot (n_g - 1)$.
- $m_{g+1} = \sum_{h \in G} m_{gh1}$ denotes the total number of observed links between nodes of type g and all other nodes in the network (including also within-group links for group g). Thus, $m_{g+1} = \sum_{i \in N: t_i = g} \eta_i$.
- $m_{g++} = \sum_{h \in G} (m_{gh1} + m_{gh0}) = m_{g+1} + m_{g+0}$ denotes the total number of potential links between nodes of type g and all other nodes in the network.
- $m_{++1} = \sum_{g \in G} \sum_{h \in G} m_{gh1}$ denotes the total number of links actually formed in the network.

A straightforward way of measuring segregation in social networks consists of collecting and analyzing some descriptive statistics of the network graph, usually related to proportions and ratios of between-group and within-group links across the graph. For example, for group g the total number of within-group (segregative) links found in the network is m_{gg1}, and the total number of between-group (integrative) links is $(m_{g+1} - m_{gg1})$. Then, let $W(R) = \sum_{g \in G} m_{gg1} = Tr(\bar{\mathbf{M}})$ be the total number of within-group links and $B(R) = \sum_{g \in G}(m_{g+1} - m_{gg1}) = m_{++1} - Tr(\bar{\mathbf{M}})$ be the total number of between-group links in the network. Given a group $g \in G$, the density of within-group and between-group links, denoted by $w_g(R)$ and $b_g(R)$, respectively, are equal to

$$w_g(R) = \frac{m_{gg1}}{m_{gg+}}$$

and

$$b_g(R) = \frac{m_{g+1} - m_{gg1}}{m_{g++} - m_{gg+}}.$$

Segregation indices can then be constructed by comparing some of these statistics, as observed in the network, either among themselves or against some benchmark that represents null segregation. Additionally, a segregation index may consider only information on connected nodes, in which case it will be derived exclusively from the contact layer mixing matrix $\bar{\mathbf{M}}$, or it may also consider information about isolated nodes, in which case it will include statistics derived from the non-contact layer mixing matrix $\tilde{\mathbf{M}}$. There are supporting arguments for both approaches. One could argue that isolated nodes in a network should not play any role in segregation, since they do not contribute any "relational" information. On the other hand, one could argue that disconnected nodes may be a source of opportunities for creating links, and thus adding isolated nodes from a given (minority) group creates additional opportunities for integration (reducing segregation) in the form of between-group links. This decision is specially relevant when analyzing segregation in a dynamic context, i.e. when modeling social dynamics that may result in (or be affected by) segregation.

We introduce below four of the most relevant network segregation indices based on descriptive network statistics, and provide some detail on them. Of these, the *GAM index*, and the *segregation matrix index* do not use an equivalent random network as a benchmark for baseline segregation. The *assortativity coefficient* and the *Coleman's homophily index* do. None of these indices use information about isolation nodes. For other relevant indices we refer the reader to Bojanowski and Corten (2014), who also study whether all these indices satisfy a series of properties that can be considered desirable, and provide some empirical examples of their application.

1. **The Assortativity Coefficient**
   The assortativity coefficient was proposed by Newman and Girvan (2003) and Newman (2003), in the context of analyzing mixing patterns in networks of sexual contacts and marriage matching. First, we define a proportions matrix $\mathbf{P}$ derived from the mixing matrix $\mathbf{M}$:

$$(p_{gh})_{g,h\in G} = \left(\frac{m_{gh1}}{m_{gh+}}\right)_{g,h\in G}$$

   The matrix $\mathbf{P}$ normalizes the contact layer of the mixing matrix $\mathbf{M}$ by computing, for each dyad of node types $\langle g, h\rangle$, the proportion of links actually formed in the network (m_{gh1}) over the total number of possible links that could be formed between them (m_{gh+}). The assortativity coefficient is then defined as

$$AC(R) = \frac{\sum_{g\in G} p_{gg} - \sum_{g\in G} p_{g+}p_{+g}}{1 - \sum_{g\in G} p_{g+}p_{+g}}.$$

The assortativity coefficient is, in fact, the normalized modularity measure $\hat{Q}(R)$ of assortative mixing by group membership of the graph, expressed in the mixing matrix terminology. The index reaches its maximum value of 1 when all the links are within-group, in which case the diagonal entries of **P** sum up to 1. The minimum value of AC does not necessarily take the value -1, but rather depends on the relative number of links in each group. It is equal to

$$\min\{AC(R)\} = \frac{-\sum\limits_{g\in G} p_{g+}\, p_{+g}}{1-\sum\limits_{g\in G} p_{g+}\, p_{+g}}.$$

Finally, the index assumes the value of 0 when $p_{gh} = p_{g+}\, p_{+h}$, that is, when group memberships in connected dyads are stochastically independent.

2. **The GAM Index**
Gupta et al. (1989) define a multi-group segregation index (which we will refer to as the *GAM* index) in order to assess the "within-group mixing" in a population. Let the "proportions matrix" **Q** associated to the mixing matrix **M** be such that

$$(q_{gh})_{g,h\in G} = \left(\frac{m_{gh1}}{m_{g+1}}\right)_{g,h\in G}.$$

So, q_{gh} is the proportion of the number of links from nodes in group g to nodes in group h over the total number of links formed from nodes in group g. The *GAM* index is then

$$GAM(R) = \frac{\sum\limits_{g\in G} q_{gg} - 1}{k-1}.$$

Values of the *GAM* index vary between $-1/(k-1)$ when $q_{gg} = 0$ for all $g \in G$ and 1 when $\sum_{g\in G} q_{gg} = k$ for all $g \in G$. As reported by Newman (2003), the *GAM* index may give misleading results in those networks with perfect integration except for a group that, representing a small share of the total number nodes, has perfect assortative mixing. In these cases such a small group can have a disproportionately large effect on the value of the GAM index. Then, the value of the *GAM* index will signal that the network has very strong assortative mixing, when in fact it does not. The reason for this lies in the fact that the *GAM* index gives equal weight to each group, as opposed to giving it to each edge, as the *AC* index does.

3. **The Segregation Matrix Index**
The segregation matrix index proposed by Fershtman (1997), based on the relative number and intensity of inward to outward interactions, is intended to serve as a measure of the cohesiveness of a group. Fershtman (1997) defines a cohesive group as a social group of actors who prefer to interact with one another more than with others and reveal a highly self-preference segregative attitude.

The *SMI* index is defined for directed graphs at the group level. Bojanowski and Corten (2014) generalize the original measure for two-groups to any number of groups by properly formulating the densities of within-group and between-group links in the network. Formally, the multi-group index *SMI* for group g is

$$SMI_g(R) = \frac{w_g(R) - b_g(R)}{w_g(R) + b_g(R)}.$$

Values of the *SMI* index for group g range from -1 for minimum segregation (when type-g nodes only form between-group links, so $w_g(R) = 0$) to +1 for maximum segregation (when type-g nodes only form within-group links, so $b_g(R) = 0$)).

The *SMI* index suffers from two main limitations. First, the index expresses cohesiveness properly provided there are equal probabilities for each node to participate in edge formation. In large networks in which that probability may be significantly small, or in networks in which situational limitations obstruct free social interactions, the *SMI* index is not appropriate for describing cohesiveness. Second, networks may exist in which some hierarchical structures (formed by a node that links only to a cluster of nodes from the same group but the reverse links are not formed in the network) are not really cohesive groups, and additional criteria are needed to exclude them from being defined as such.

4. **Coleman's Homophily Index**

Coleman (1958) introduces a group-level segregation measure for directed networks that represents the propensity of an individual from group g to create a link to someone else from the same group, as opposed to choosing randomly. Formally, let

$$m^*_{gg1} = \sum_{i \in N: t_i = g} \eta_i p_g = \sum_{i \in N: t_i = g} \eta_i \frac{n_g - 1}{n - 1}$$

be the expected number of within-group links for group g in a random network R^*, where $p_g = (n_g - 1)/(n - 1)$ is the probability for a node of type g to link to another node from the same group in network R^* (and which can be approximated by n_g/n for large n and large n_g). Coleman's Homophily Index for group $g \in G$ is then computed as

$$CHI_g(R) = \begin{cases} \frac{m_{gg1} - m^*_{gg1}}{\sum_{t_i = g} \eta_i - m^*_{gg1}} & \text{if } m_{gg1} \geqslant m^*_{gg1} \\ \frac{m_{gg1} - m^*_{gg1}}{m^*_{gg1}} & \text{if } m_{gg1} < m^*_{gg1} \end{cases}.$$

$CHI_g(R)$ varies between -1 (individuals are perfectly avoiding their own group) and 1 (perfect segregation). The index takes the value 0 if and only if the observed number of within-group links is exactly equal to the expected number under random choice, given the total degree of a group. The corresponding network-level index $CHI(R)$ has been introduced in Bojanowski and Corten (2014):

$$CHI(R) = \begin{cases} \frac{W(R)-W^*(R)}{\sum\limits_{i \in N} \eta_i - W^*(R)} & \text{if } W(R) \geqslant W^*(R) \\ \frac{W(R)-W^*(R)}{W^*(R)} & \text{if } W(R) < W^*(R) \end{cases},$$

where $W^*(R) = \sum_{g \in G} \sum_{t_i = g} (\eta_i (n_g - 1))/(n-1)$ is the expected value of $W(R)$ in the equivalent random network R^*.

3.2 Measures Based on Spectral Graph Theory

All segregation measures defined in the previous section result from analyzing a social network through the graph that directly represents it as a structured collection of vertices and edges, and then deriving properties on segregation (or assortative mixing) from that graph. Spectral graph theory is the study and exploration of graphs through some characteristic values naturally associated with a graph, such as its eigenvalues and eigenvectors. Spectral graph theory starts with a different way of representing and analyzing the graph associated with a network, by using a set of matrices associated with the graph, rather than the adjacency matrix itself. It then proceeds by computing some characteristic values of those associated matrices, such as their eigenvalues and eigenvectors, which then reveal some properties and structure of the original graph.

1. **Spectral Segregation Index**
   Echenique and Fryer Jr. (2007) were the first to propose a segregation index based on social interactions. Their spectral segregation index *SSI* is based upon two premises: (1) a measure of segregation should disaggregate to the level of individuals and (2) an individual is more segregated the more segregated are the individuals to whom he interacts with. They also recover important empirical results on segregation patterns that formerly were only obtained under the more restrictive dissimilarity index. Although defined at the group level, the *SSI* can be easily defined down to the node level giving segregation values for individual nodes. It can also be aggregated flexibly to provide segregation scores for network components and for the network as a whole.
   The procedure to obtain these indices is as follows. Given the row-stochastic matrix $\mathbf{P}$ that describes the percentage of time individuals spend with each other—in particular, $p_{i,j}$ is the fraction of time i spends with j—the group-level index $SSI_g(R)$ is computed by determining first the subgraph of R that contains only interactions between members of g, denoted by $\mathbf{P^g}$. All individuals belonging to a different group are eliminated from the network and, consequently, from the matrix $\mathbf{P}$. Such a subgraph $\mathbf{P^g}$ generally contains more than one strongly

connected component, so the next step is to calculate the *SSI* for each component separately. Let γ be a (strongly connected) component of $\mathbf{P^g}$. The *SSI* for group g in γ, denoted by $SSI_g(\mathbf{P}^{g,\gamma})$ is equal to the spectral radius of the sub-stochastic matrix $\mathbf{P}^{g,\gamma}$, which is the largest absolute value of the eigenvalues of $\mathbf{P}^{g,\gamma}$. The segregation of a node (individual) i in this component is denoted by $SSI^i(\mathbf{P}^{g,\gamma})$, and is the i-th entry of the principal eigenvector of the matrix $\mathbf{P}^{g,\gamma}$, normalized so that the vector average is $SSI_g(\mathbf{P}^{g,\gamma})$. The segregation measure of group g is a weighted average over the segregation measures of group g in all components of $\mathbf{P^g}$. In particular,

$$SSI_g(\mathbf{P}^g) = \sum_{\gamma} d_g^{\gamma}\, SSI(\mathbf{P}^{g,\gamma}),$$

where $d_g^{\gamma} = \frac{1}{n_g} \sum_{\gamma} \sum_{i \in \mathbf{P}^{g,\gamma}} SSI^i(\mathbf{P}^{g,\gamma})$ is the share of individuals of group g found in the component γ. Alternatively, the *SSI* of group g can also be expressed as the average over the individual segregation levels:

$$SSI_g(\mathbf{P}^g) = \frac{1}{n_g} \sum_{\gamma} \sum_{i \in \mathbf{P}^{g,\gamma}} SSI^i(\mathbf{P}^{g,\gamma}).$$

To see how the *SSI* is applied consider Fig. 3, which corresponds to the motivating example in Echenique and Fryer Jr. (2007). The society is composed of two groups, blacks and whites. Each dot represents one individual. It is also assumed that individuals only interact with their horizontal and vertical neighbors. So, individual (A,1) spends 50 % of her time with each (A,2) and (B,1) each. The subgraph $\mathbf{P}^{blacks}$ consists therefore of two connected components $\mathbf{P}^{blacks,1}$ and $\mathbf{P}^{blacks,2}$.

The *SSI* for the black group is determined by taking a weighted average over the spectral radii of the two black connected components. Since $SSI(\mathbf{P}^{blacks,1}) = 0.72$, $SSI(\mathbf{P}^{blacks,2}) = 0.25$, and 80 % of the blacks reside in component 1 and 20 % in component 2, the segregation of the black group is $SSI(\mathbf{P}^{blacks}) = 0.8 \cdot 0.72 + 0.2 \cdot 0.25 = 0.63$.

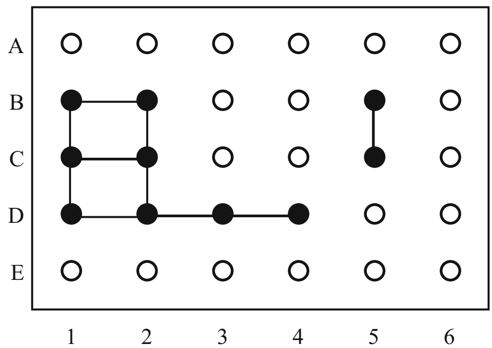


Fig. 3 Calculation of the *SSI*

According to Echenique and Fryer Jr. (2007), the following four main features of *SSI* give this segregation index some important advantages over their alternatives. First, it is invariant to arbitrary partitions of a city. Second, it allows to investigate how segregated multiple minority groups are within and between cities. Third, it allows for analysis of the full distribution of segregation, allowing researchers to move beyond aggregate statistics, which can be misleading. And finally, there are inherent multiplicative effects captured by *SSI* which other indices omit given the fact that it is built based on a social network framework.

2. **Spectral Homophily Index**
In the context of analyzing how homophily—through the segregation patterns it induces—affects the speed of a learning process in a network, Golub and Jackson (2012a) introduce the spectral homophily index *SHI*, a measure of homophily in networks.
The definition of the *SHI* starts with considering a matrix associated to the network graph, the so-called interaction matrix $\mathbf{F}$. To define $\mathbf{F}$, let $p_{g,h}$ be the probability of a type-g node forming a link with a type-h node. Then $q_{g,h} = n_g n_h p_{g,h}$ is the total contribution to the degrees of nodes of type g from links to nodes of type h. Also, let $\eta_g\left(q_{g,h}\right) = \sum_{h \in G} q_{g,h}$ be the total degree of nodes of type g. The interaction matrix $\mathbf{F}$ indicates the expected fraction of their links that nodes of type g will have with nodes of type h. Each entry (g, h) in $\mathbf{F}$ is computed as $f_{g,h} = \frac{q_{g,h}}{\eta_g\left(q_{g,h}\right)}$.
Note that, using the mixing matrix terminology, $q_{g,h} = m_{gh1}$, $\eta_g\left(q_{g,h}\right) = m_{g+1}$, $f_{g,h} = m_{gh1}/m_{g+1}$, and therefore $\mathbf{F}$ is the same matrix as the proportions matrix $\mathbf{Q}$ used to define the *GAM* index. The spectral homophily index *SHI* is then set equal to the second-largest eigenvalue of the interaction matrix $\mathbf{F}$, $SHI(R) = \lambda_2\,[\mathbf{F}]$.
The *SHI* index is therefore based on first simplifying the overall interaction matrix to that of the expected interaction across groups, and then looking at a particular part of the spectrum of that matrix. Golub and Jackson (2012a) note how, on an intuitive level, homophily makes it possible to draw a boundary in the group structure of a network, separating it into two blocks so that there are relatively few links across the boundary and relatively many links not crossing the boundary (staying inside each of the two blocks that result from the split). Since the second-largest eigenvalue of a matrix captures the extent to which it can be broken into two blocks with relatively little interaction across the blocks, the *SHI* index picks up fault lines in the group structure, which should reveal segregation patterns.

3. **Random walk-based segregation index**
Ballester and Vorsatz (2014) propose a segregation index that builds on the following random-walk process on the network graph. Pick any two individuals from a given group $g \in G$ at random and suppose that the first of the two individuals moves over the network in such a way that in the first period she advances from her area of residence to some neighboring area (observe that $V = M$). In each subsequent period, she either moves from her current position

in the network to some adjacent node (this event happens with probability $0 \leq \alpha < 1$) or the process stops (this event happens with probability $1 - \alpha$). Hence, the parameter α can be interpreted as the degree of spatial mobility. The (normalized) random-walk based segregation index for group g is then defined as the probability that the two randomly chosen individuals meet when the random walk of the first individual terminates.

From a formal point of view let $\mathbf{P}$ be the $m \times m$ row-stochastic matrix that collects the exogenous transition probabilities between nodes; that is, $p_{i,j}$ is the probability that an individual currently at node i moves to node j. This probability is linked to the adjacency matrix $\mathbf{A}$ and, for example, one could assume that $p_{i,j} = a_{i,j}/\eta_i$. Also, let $\mathbf{Q}$ be the $m \times m$ matrix such that $q_{i,j}$ is the probability that a walk ends in node j, conditional on that it started in node i. Since it can be shown that $\mathbf{Q} = (1-\alpha)(\mathbf{I} - \alpha\mathbf{P})^{-1}\mathbf{P}$, the normalized measure at the group level can be written as

$$\begin{aligned}
\hat{\sigma}_g(R, \alpha) &= \left(\tfrac{n_g}{n}\right)^{-1} \cdot \textstyle\sum_{i \in M} d_{g,i} \sum_{j \in M} q_{i,j}\, c_{g,j} \\
&= \left(\tfrac{n_g}{n}\right)^{-1} \cdot \mathbf{d}_g^T\, \mathbf{Q}\, \mathbf{c}_g \\
&= \left(\tfrac{n_g}{n}\right)^{-1} \cdot \mathbf{d}_g^T\, (1-\alpha)(\mathbf{I} - \alpha\mathbf{P})^{-1}\mathbf{P}\, \mathbf{c}_g.
\end{aligned}$$

A simple example clarifies the application of the index. Consider the network depicted in Fig. 4.

There are two ethnic groups, blacks and whites. Two whites reside in neighborhoods 1 and 3, three whites in neighborhood 4, and one white in neighborhood 2. Moreover, three blacks live in neighborhood 3 and one black in neighborhoods 1, 2 and 4 each. The transition matrix $\mathbf{P}$ is such that all links (self-links included) are equally likely to be chosen. Since $\mathbf{d}^{\mathsf{T}}_{blacks} = (1/6, 1/6, 1/2, 1/6)$, $\mathbf{d}^{\mathsf{T}}_{whites} = (1/4, 1/8, 1/4, 3/8)$, $\mathbf{c}^{\mathsf{T}}_{blacks} = (1/3, 1/2, 3/5, 1/4)$, and $\mathbf{c}^{\mathsf{T}}_{whites} = (2/3, 1/2, 2/5, 3/4)$, it is easy to verify that $\hat{\sigma}_{blacks}(R, 0.85) = 1.05$ and $\hat{\sigma}_{whites}(R, 0.85) = 0.97$, respectively.

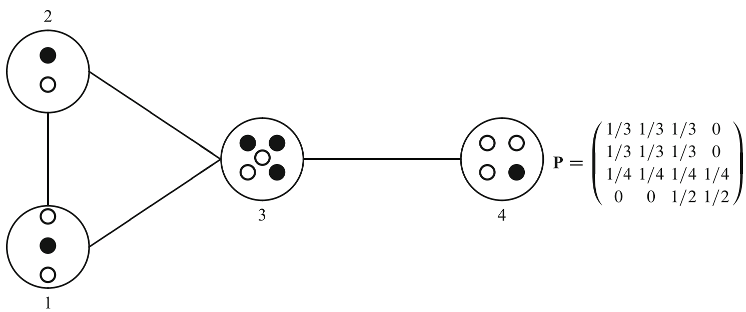


Fig. 4 Calculation of the random-walk based segregation index

This index has several favorable attributes. First, it is a natural spatial generalization of the isolation index to networks. In fact, σ_g reduces to the isolation index of group g if the sociogeographical network is empty and individuals interact only in the neighborhoods they reside in. Second, if $\alpha = 0$, it captures the clustering dimension of Massey and Denton (1988) as the measure reduces to the homophily index introduced by Currarini et al. 2009 and only direct connections count. Finally, it also captures directly the dimension of centralization because the index can be shown to be equal to

$$\hat{\sigma}_g(R, \alpha) = \left(\frac{n_g}{n}\right)^{-1} \mathbf{w}_g^T \, \mathbf{P} \, \mathbf{c}_g,$$

where w_g is the principal eigenvector of $(1 - \alpha)\mathbf{d}_g \mathbf{1}^T + \alpha \mathbf{P}^T$ that essentially underlies Google's PageRank index (Brin and Page (1998)). That is, the vector $\mathbf{w}_g$ assesses the centrality of the different nodes in the network for group g.

4 Conclusion

Most of the classical statistical methods introduced in the literature on the measurement of segregation were aimed to assess the heterogeneity in population compositions among areas. To a very large extent, these measures were derived around the data provided by the public authorities. More recently, sociologists and economists have tried to be explicit about the processes through which segregation arises; namely, through social interactions. Building upon the theoretical tool of graph theory, two main roads can here be identified. First, the use of descriptive graph statistics, such as ratios of within-group and between-group links in a network, which can be expressed in a compact way through the mixing matrix associated to a graph. Second, the use of spectral graph theory, which provides a set of general theoretical tools for a mean-field approach to the analysis of networks, built upon characteristic values of matrices associated to the network graph.

One drawback of the literature is that the employed measures are silent about the particular processes that may lead to the formation of the observed network structure that is characteristic of segregation. According to Ackland (2013) and Wimmer and Lewis (2010) there are three main reasons why segregation forms. First, it may be due to the presence of homophily, which can in principle operate with respect to any attribute. However, when the attribute in question can be modified by the individual (as for example with cultural preferences), it becomes difficult to distinguish the direction of causality, i.e. whether attributes and preferences are influencing link formation or whether they are influenced by existing links. Second, there may be opportunity structures that influence link formation. For example, a reduced group size and social or spatial proximity can be factors influencing whether two individuals form a social tie. The smaller a group the more likely is that its members will form links outside of the group, and proximity between two individuals, either in their spatial location or in their social or institutional environments, can also influence the probability of them forming a link (this is known as the *propinquity mechanism*). Finally, there are other mechanisms that are not related to the attributes of actors in a dyad, but influence link formation and therefore the level of segregation in a network. Among these mechanisms we can mention some balance mechanisms (or network effects) like *reciprocity* and *transitivity* that have been empirically documented. Consequently, it seems desirable to develop measures that can "unpack" the contributions of each of these before-mentioned mechanisms to the level of segregation in a given social network.

Acknowledgements The authors are very grateful to Coralio Ballester for many fruitful discussions.

The author "Marc Vorsatz" gratefully acknowledges financial support from the Fundación Ramón Areces and the Spanish Ministry of Economy and Competitiveness (ECO2012-31985).

References

Ackland R (2013) Web social science: concepts, data and tools for social scientists in the digital age. SAGE, London

Alesina A, La Ferrara E (2005) Ethnic diversity and economic performance. J Econ Lit 43(3):762–800

Atkinson AB (1970) On the measurement of inequality. J Econ Theory 2:244–263

Baerveldt C, Van Duijn MA, Vermeij L, Van Hemert DA (2004) Ethnic boundaries and personal choice. Assessing the influence of individual inclinations to choose intra-ethnic relationships on pupils' networks. Soc Netw 26(1):55–74

Ballester C, Vorsatz M (2014) Random walk-based segregation measures. Rev Econ Stat 96(3):402–417

Barabási A-L, Albert R (1999) Emergence of scaling in random networks. Science 286:509–512

Bell W (1954) A probability model for the measurement of ecological segregation. Soc Forces 32(4):357–364

Benabou R (1993) Workings of a city: location, education, and production. Q J Econ 108:619–652

Berry B (1958) Race and ethnic relations. Houghton Mifflin, Boston

Blackwell D (1951) Comparison of experiments. In: Second berkeley symposium on mathematical statistics and probability

Blackwell D (1953) Equivalent comparisons of experiments. Ann Math Stat 24(2):265–272

Bojanowski M, Corten R (2014) Measuring segregation in social networks. Soc Netw 39:14–32

Borjas GJ (1995) Ethnicity, neighborhoods, and human-capital externalities. Am Econ Rev 85:365–390

Brammoullie Y, Kranton R (2007) Risk sharing across communities. Am Econ Rev 97(2):70–74

Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. In: Computer networks and ISDN systems, vol 33, pp 107–117

Burt RS (1991) Measuring age as a structural concept. Soc Netw 13(1):1–34

Campbell KE (1990) Networks past: a 1939 Bloomington neighborhood. Soc Forces 69(1):139–155

Card D, Rothstein J (2007) Racial segregation and the black-white test score gap. J Public Econ 91(11–12):2158–2184

Cohen JK, Falk RF, Cortese CF (1976) Reply to Taeuber and Taeuber. Am Sociol Rev 41(5):889–893

Cohn MJ, Jackman SP (2011) A comparison of aspatial and spatial measures of segregation. Trans GIS 15(s1):47–66

Coleman J (1958) Relational analysis: the study of social organizations with survey methods. Hum Organ 17(4):28–36
Coleman J, Hoffer T, Kilgore S (1982) Achievement and segregation in secondary-schools - a further look at public and private school differences. Sociol Educ 55(2):162–182
Cortese CF, Falk RF, Cohen JK (1976) Further considerations on the methodological analysis of segregation indices. Am Sociol Rev 41(4):630–637
Currarini S, Jackson MO, Pin P (2009) An economic model of friendship: homophily, minorities, and segregation. Econometrica 77(4):1003–1045
de la Vega CL, Volij O (2014) Segregation, informativeness and lorenz dominance. Soc Choice Welf 43(3):547–564
Duncan OD, Duncan B (1955) A methodological analysis of segregation indexes. Am Sociol Rev 20(2):210–217
Easterly W, Levine R (1997) Africa's growth tragedy: policies and ethnic divisions. Q J Econ 112(4):1203–1250
Echenique F, Fryer RG Jr (2007) A measure of segregation based on social interactions. Q J Econ 122(2):441–486
Edin PA, Fredriksson P, Åslund O (2003) Ethnic enclaves and the economic success of immigrants - evidence from a natural experiment. Q J Econ 118:329–357
Falk RF, Cortese CF, Cohen J (1978) Utilizing standardized indices of residential segregation: comment on winship. Soc Forces 57(2):713–716
Fershtman M (1997) Cohesive group detection in a social network by the segregation matrix index. Soc Netw 19:193–207
Frankel DM, Volij O (2008) Scale-invariant measures of segregation. Technical Report 0814, Department of Economics Ben-Gurion University of the Negev
Gini C (1921) Measurement of inequality of incomes. Econ J 31:124–126
Goldin C, Katz LF (1999) Human capital and social capital: the rise of secondary schooling in america, 1910–1940. J Interdiscip Hist 29(4):683–723
Golub B, Jackson MO (2012a) How homophily affects the speed of learning and best-response dynamics. Q J Econ 127(3):1287–1338
Golub B, Jackson MO (2012b) Network structure and the speed of learning measuring homophily based on its consequences. Ann Econ Stat 108(107):33–48
Gupta S, Anderson, MO, May RM (1989) Networks of sexual contacts: implications for the pattern of spread of HIV. AIDS (London, England) 3(12):807–817
Hunt CL, Walker L (1974) Ethnic dynamics: patterns of intergroup relations in various societies. Dorsey Press
Hutchens RM (1991) Segregation curves, lorenz curves, and inequality in the distribution of people across occupations. Math Soc Sci 21(1):31–51
Hutchens R (2001) Numerical measures of segregation: desirable properties and their implications. Math Soc Sci 42:13–29
Ibarra H (1995) Race, opportunity, and diversity of social circles in managerial networks. Acad Manag J 38(3):673–703
Jahn JA, Schmidt CF, Schrag C (1947) The measurement of ecological segregation. Am Sociol Rev 12:293–303
James DR, Taeuber KE (1985) Measures of segregation. Sociol Methodol 15:1–32
Jargowsky PA (1996) Take the money and run: economic segregation in U.S. metropolitan areas. Am Sociol Rev 61:984
Kalmijn M (2006) Educational inequality and family relationships: influences on contact and proximity. Eur Sociol Rev 22(1):1–16
Kestembaum B (1980) Notes on the index of dissimilarity: a research note. Soc Forces 59(1):275–280
Khinchin A (1957) Mathematical foundations of information theory. Dover Publications, New York
Kling JR, Liebman JB, Katz LF (2007) Experimental analysis of neighborhood effects. Econometrica 75(1):83–119

Laumann EO (1966) Prestige and association in an urban community: an analysis of an urban stratification system. The Bobbs-Merrill Company, Indianapolis
Laumann EO (1973) Bonds of pluralism: form and substance of urban social networks. Wiley, New York
Lazarsfeld PF, Merton RK (1954) Friendship as a social process: a substantive and methodological analysis. In: Berger M (ed) Freedom and control in modern society. Van Nostrand, New York, pp. 18–66
Lieberson S, Carter DK (1982) Temporal changes and urban differences in residential segregation: a reconsideration. Am J Sociol 88(2):296
Marsden PV (1987) Core discussion networks of Americans. Am Sociol Rev 52(1):122–131
Marsden PV (1988) Homogeneity in confiding relations. Soc Netw 10:57–76
Massey DS, Denton NA (1988) The dimensions of residential segregation. Soc Forces 67(2):281–315
Morgan BS (1975) The segregation of socioeconomic groups in urban areas: a comparative analysis. Urban Stud 12:47–60
Newman M (2003) Mixing patterns in networks. Phys Rev E, Stat Nonlinear, Soft Matter Phys 67(2 Pt 2):026126
Newman M (2010) Networks: an introduction. Oxford University Press, Oxford
Newman M, Girvan M (2003) Mixing patterns and community structure in networks. Stat Mech Complex Netw 625:66–87
Poterba JM (1997) Demographic structure and the political economy of public education. J Policy Anal Manage 16(1):48–66
Reardon SF, Firebaugh G (2002) Measures of multigroup segregation. Sociol Methodol 32:33–67
Reardon SF, O'Sullivan D (2004) Measures of spatial segregation. Sociol Methodol 34(814):121–162
Sakoda JM (1981) A generalized index of dissimilarity. Demography 18:245–250
Schwartz J, Winship C (1980) The welfare approach to measuring inequality. Sociol Methodol 11:1–36
Shannon CE (1948) A mathematical theory of communication. Bell Syst Technol J 27(3):379–423
Shannon CE, Weaver W (1949) The mathematical theory of communication. University of Illinois Press, Champaign
Stearns LB, Logan JR (1986) Measuring trends in segregation: three dimensions, three measures. Urban Aff Rev 22(1):124–150
Taeuber KE, Taeuber AF (1965) Negroes in cities: residential segregation and neighborhood change. Aldine Publishing, Chicago
Taeuber KE, Taeuber AF (1976) A practitioner's perspective on the index of dissimilarity. Am Sociol Rev 41(5):884–889
Theil H (1972) Statistical decomposition analysis. North-Holland, Amsterdam
Theil H, Finizza AJ (1971) A note on the measurement of racial integration in schools. J Math Sociol 1(2):187–1193
van der Zanden JW (1972) American minority relations. Ronald Press, New York
Verbrugge LM (1977) Structure of adult friendship choices. Soc Forces 56(2):576–597
Wellman B (1996) Are personal communities local? a dumptarian reconsideration. Soc Netw 18(4):347–354
White MJ (1983) The measurement of spatial segregation. Am J Sociol 88(5):1008
White MJ (1986) Segregation and diversity measures in population distribution. Popul Index 52(2):198–221
Wimmer A, Lewis K (2010) Beyond and below racial homophily: ERG models of a friendship network documented on facebook. Am J Sociol 116(2):583–642
Winship C (1977) A revaluation of indexes of residential segregation. Soc Forces 55(4):1058–1066
Winship C (1978) The desirability of using the index of dissimilarity or any adjustment of it for measuring segregation: reply to Falk, Cortese, and Cohen. Soc Forces 57(2):717–720
Zenou Y (2009) How common is integration policy in Europe? In: Gustavsson S, Oxelheim L, Pehrson L (eds) How unified is the European union? Springer, Heidelberg

Part II
Complex Network Analysis Applied to Economic Theoretical and Empirical Issues

An Investigation of Interregional Trade Network Structures

Roberto Basile, Pasquale Commendatore, Luca De Benedictis, and Ingrid Kubin

Abstract We provide empirical evidence on the network structure of trade flows between European regions and discuss the theoretical underpinning of such a structure. First, we analyze EU regional trade data using Social Network Analysis. We describe the topology of this network and compute local and global centrality measures. Finally, we consider the distribution of higher order statistics, through the analysis of local clustering and main triadic structures in the triad census of interregional trade flows. In the theoretical part, we explore the relationship between trade costs and trade links. As shown by Behrens (J Urban Econ 55(1):68–92, 2004), Behrens (Reg Sci Urban Econ 35(5):471–492, 2005a) and Behrens (J Urban Econ 58(1):24–44, 2005b) in a two-region linear new economic geography (NEG) model, trade costs and the local market size determine, even with finite trade costs, unconditional autarky and unilateral trade, that is, a one-directional flow from one region to the other. Following these contributions and guided by the empirical evidence, we clarify the relationship between market competition, trade costs and the patterns of trade in a three-region NEG model. We identify a larger set of trade

R. Basile
Department of Economics, Second University of Naples, Caserta, Campania, Italy
e-mail: roberto.basile@unina2.it

P. Commendatore (✉)
Department of Law, University of Naples 'Federico II', Via Mezzocannone 16, Naples, Italy
e-mail: commenda@unina.it

L. De Benedictis
Department of Law and Economics, University of Macerata, Macerata, Italy
e-mail: luca.debenedictis@unimc.it

I. Kubin
Department of Economics, Institute for International Economics and Development, WU Vienna University of Economics and Business, Vienna, Austria
e-mail: Ingrid.Kubin@wu.ac.at

P. Commendatore et al. (eds.), *Complex Networks and Dynamics*,
Lecture Notes in Economics and Mathematical Systems 683,
DOI 10.1007/978-3-319-40803-3_6

network configurations other the three elementary ones that occur at the dyadic level between two regions (no trade, one-way trade, reciprocated two-way trade), and relate the model with the triad census.

Keywords EU regional trade data • European trade flows • Social network analysis • Three-region new economic geography model • Triad census

1 Introduction

In this paper we provide some empirical evidence on the network structure of trade flows at the regional level in Europe and we discuss the possible theoretical underpinning of such a structure. In the empirical part of the paper, we look at the EU regional trade data recently produced by the PBL Netherlands Environmental Assessment Agency (Thissen et al. 2013a,b, 2015), and we analyze it using Social Network Analysis tools (Wasserman and Faust 1994). We take advantage of both the binary structure of the European regional trade network (analyzing the presence and absence of regional trade flows) and of its weighted counterpart (making use of the distribution of the value of trade flows, measured in millions of Euros). We use the latter to construct a meaningful threshold to restrict the density of the binary structure, and, following De Benedictis and Tajoli (2011) and De Benedictis et al. (2014), we visualize the trade network at different levels of the threshold, define and describe the topology of the network and produce some of the main local and global centrality measures for the different European regions. Finally,

> ... since the most interesting and basic questions of social structure arise with regard to triads (Hanneman and Riddle 2005),

we account for the distribution of higher order statistics of the network, through the analysis of local clustering and the main triadic structures in the *triad census* of interregional trade flows.

Given the explicit assumption that trade costs, together with regional markets size, are as for the gravity model of international trade (De Benedictis and Tajoli 2011; Anderson 2011; Head and Mayer 2014) among the main determinants of inter-regional trade flows, the network analysis of regional trade flows in Europe informs the main topological properties of the data that must be reflected in the modeling of such trade flows.

From the theoretical point of view we explore how changes in crucial parameters—especially a reduction in trade costs—may favor the creation of trade links. The theoretical framework we adopt is a three-region linear new economic geography (NEG) model. We have chosen the linear version of a NEG model to overcome a crucial weakness of the standard approach as developed in the literature beginning from Krugman (1991). Indeed, in the standard NEG model all regions trade with each other as long as trade costs are finite. This follows from the isoelastic demand function—because of the specific assumption on consumer's CES preferences—and the ad valorem, proportional to price, *iceberg* trade costs. In the linear version of the NEG model, as developed by Ottaviano et al. (2002), this is not necessarily true. As shown by Behrens (2004, 2011, 2005a) for a two-region linear NEG model, trade costs and the dimension of the local market may determine *unconditional autarky* even in the presence of finite trade costs and *asymmetric patterns of trade*, that is a one-directional flow from one region to the other. Of crucial importance is the size and density of the industrial sector, that even in the presence of symmetric bilateral trade costs, may induce differences in local prices—with a lower price in the larger market. As stated by Behrens:

> price competition and trade costs endogenously create interregional asymmetries in market access and give rise to one-way trade in differentiated products (Behrens 2005a, p. 473).

The same result is obtained by Okubo et al. (2014, OPT). These authors remark that while the NEG and trade literature stress

> the importance of trade barriers for the intensity of competition and the spatial pattern of the global economy,

it pays much less attention to the reverse relationship:

> the impact of competition on the nature and intensity of trade as well as on the location of economic activities.

OPT capture the intensity of competition in domestic markets within a linear NEG model by assuming two regions with asymmetric population sizes.

Following the above mentioned contributions, our theoretical analysis aims to clarify the relationship between the intensity of competition, trade costs and the patterns of trade. Differently from the previous literature, and guided by the empirical evidence, we will consider a three-region model.[1] This allows us to identify a larger set of trade network configurations other than three elementary ones that occur at the dyadic level between two regions (no trade, one-way trade from one region to the other, reciprocated two-way trade), and relate the model with characteristics of the triad census. We also elaborate on how the structure of a trade network can be modified as trade costs vary. In order to focus on the properties of the short-run equilibrium and on the emergence of network structures in this time framework, we exclude factor migration.

The paper is structured as follows: Sect. 2 presents some stylized facts on the dominant interregional trade patters in Europe. Section 3 presents the basic economic framework of the theoretical model, i.e. a three-region NEG linear model with asymmetric trade costs and all possible configurations of trade flows between two regions. In Sect. 3 we derive the short-run equilibrium and determine the trade costs thresholds that determine all network configurations. Section 4 reports a brief discussion and concludes.

[1] In the NEG literature, Ago et al. (2006), Melitz and Ottaviano (2008) and Behrens (2011) put also forward three-region linear models à la Ottaviano et al. (2002). However, they limit their analysis to specific trade cost structures.

2 Interregional Trade Network in Europe: Some Stylized Facts

In this section we perform an empirical analysis of the interregional trade network in Europe in 2010 using a new dataset on constructed trade flows between the European NUTS-2 regions. The aim is to provide some stylized facts about interregional trade patterns in Europe. In particular, through a triad census analysis, we want to identify the dominant triadic types, that is the frequency of each possible triadic structure, in the directed binary European regional trade network.

Interregional trade data are noticeably missing from European regional databases. The only database on interregional trade in goods and services at the NUTS-2 territorial aggregation level, fully consistent with international trade data between the Member States and with the rest of the world, is produced by the PBL Netherlands Environmental Assessment Agency (Thissen et al. 2013a,b, 2015). These data are estimated by essentially breaking down international trade flows and national Supply and Use Tables to the regional level (see Thissen et al. 2015, for an overview of the methodology used in the construction of the data). Importantly, the methodology used is a parameter-free approach and therefore deviates from earlier methods based on the gravity model that suffer from analytical inconsistencies. Unlike a gravity model estimation, the methodology stays as close as possible to observed data without imposing any geographical trade patterns. The resulting data can therefore be used as such in our trade network analysis.

Nevertheless, as pointed out by Thissen et al. (2013a), one has to keep in mind that the constructed interregional trade data are inferred from other data sources and are not measured as a flow from one region to another. Given the compatibility constraints with macro variables, some bias in the trade flows between regions inside a country or between regions of different countries, might result from the weighting procedure used in the construction of the data. In particular the number of positive trade flows is extraordinary high with respect to other international trade data at the national level, such as the Comtrade UN database. The number of zeros in the regional trade matrix is minimal, meaning that the resulting trade network is almost a fully connected one. Therefore, we opted to use the information contained in the data to distinguish the main regional trade flows from the flows being lower than a chosen threshold w. Then, we exploit only the resulting binary structure of the truncated regional trade matrix, focusing on the relative dimension of trade links rather than on the individual absolute value of trade flows between pair of regions.

The version of the *bi-regional trade database* used in our empirical work comprises $267 \times 267 = 71{,}289$ observations of intra- or interregional trade flows among European regions for the year 2010 (Thissen et al. 2015). Export and import flows (both priced free on board) are measured in values (million of euros) and divided into six product categories (aggregates AB, CDE, F, GHI, JK and LMNOP of the NACE rev. 1.1). For our purposes, we only use the aggregate CDE, which includes “Mining and quarrying” (Section C), “Manufacturing” (Section D) and

"Electricity, gas and water supply" (Section E). The countries covered by the data are the countries of the EU-27 (Croatia is not included).

We explore this new dataset through network analysis. As in De Benedictis et al. (2014), we visualize the trade network, define and describe the topology of the binary network and produce some of the main network's statistics (i.e. local and global centrality measures). We also calculate higher order statistics, enriching the analysis with the reports on local clustering and the triad census of inter-regional trade links.

In general terms, the fundamental unit of analysis necessary to study regional trade flows is at the dyadic level rs: if between region r—the exporting region—and region s—the importing region—trade takes place, two levels of information are recorded. The first one is about the *existence* of a trade link, and it is a binary measure that takes the value one if a trade link exists and zero otherwise. The second is about the *intensity* of the trade relation between the two regions r and s, and is a continuous measure that is conditional on the associated binary measure: if the binary measure is zero, the only possible value that the intensity of the relation can take is zero; if the binary variable is one, the intensity of the relation takes real positive values. Since trade flows are directional, it is not in general correct to impose any kind of symmetry, and the value of trade flows between r and s will not be equivalent to the value of trade flows between s and r.

Even if the fundamental unit of analysis is at the dyadic level rs, the decision of agents from region r to trade with agents of region s is not taken in isolation, but it must consider the (best) possible option of trading with region k as an alternative. This is true for both r and s: dyadic trade flows do not occur in isolation. This motivates the use of network analysis in studying the relation between r and s, extending to the nth level the logic behind the study of the so-called *third-region* (k) effect.

More formally, a trade network $N = (V, L, P, W)$ consists of a graph $G = (V, L)$, with $V = 1, 2, \ldots, n$ being a set of *nodes* (the regions, labeled with the respective NUTS-2 code) and L a set of *links* between pairs of vertices (e.g., trade partnership), plus P, the additional information on the vertices, and W the additional information on the links of the graph. The additional information included in the line value function W captures the intensity of trade between r and s (in million of euros). The information on the vertices (P) assembles different properties or characteristics of the regions (regions' labels, GDP, population, and so on). As mentioned, the trade graph is a *directed graph* in nature, since $l_{rs} \in \{0, 1\}$ indicates the existence or not of some exports from region r to region s, and $l_{rs} \neq l_{sr}$.

The graph associated to the EU regional trade network, $G = (V, L)$, has an average dimension of 267 vertices ($V = 1, \ldots, 267$) and 70,898 trade links ($L = 1, \ldots, 70{,}898$) out of 71,022 possible links (i.e. there are only 124 zeros, 0.27 %). Indeed, as the data is constructed, the EU regional trade network is *strongly* connected, that is almost every vertex r is reachable from every s by a direct walk. However, for many dyadic observations the amount of exports is negligible. Thus, in order to visualize the network and to compute local and global centrality measures we use the information associated with the intensity of the links to define

Table 1 Trade network properties for different thresholds

	Full	w>25 mln.	w>500 mln.	w>1000 mln.	w>2500 mln.
% of intra-European trade	100.0	91.0	24.4	10.0	1.3
Number of nodes (regions)	267	266	217	100	13
Number of links	70,898	20,086	834	200	12
% of zeros	0.27	71.72	98.82	99.71	99.98
Density	0.998	0.283	0.018	0.003	0.001
Degree centralization	0.002	0.717	0.988	0.997	0.999
Degree SD	0.006	0.181	0.026	0.030	0.115
Eigenvector centralization	0.001	0.572	0.894	0.894	0.583
Eigenvector SD	0.006	0.297	0.162	0.204	0.322
Clustering	0.999	0.690	0.227	0.138	0.000

The density of a graph is the frequency of realized edges relative to potential edges. The clustering coefficient measures the proportion of vertex triples that form triangles (transitivity)

an appropriate threshold for the selection of links, and then exploit the binary information of the resulting network.

As an example, excluding all dyadic observations lower than 25 million of euros, the remaining flows almost cover the 90 % of the total intra-Europe inter-regional trade in the aggregate sector CDE (i.e. imports+exports, which amount to almost 3,000 billions of Euros) (see Table 1). Adopting a threshold of $w > 25$ the number of edges (and so the density) is substantially reduced, from 70,898 to 20,086 (from 0.998 to 0.285), and one region (PT15) appears as an isolate. The density of the truncated network indicates that the inter-regional trade network is not regular and is far from being complete, or in other terms if the heterogeneity in the strength of links is used to select their presence, this heterogeneity is reflected in the connectivity of the network.[2]

The choice of the threshold of $w > 25$, simply based on the criterion to cover the 90 % of the total intra-Europe inter-regional trade, may certainly appear arbitrary. A valid assessment of the robustness of our analysis to alternative choices of the threshold therefore requires a preliminary exploration of the distribution of interregional export values to ascertain whether any discontinuity takes place in the neighborhood of 25. To perform this check, we report in Fig. 1 the estimated kernel density of interregional export values. The visual inspection of the graph shows a reasonably smooth distribution and does not reveal any relevant jump at the cut off of 25. This evidence supports our choice. As a "litmus test", the results of the network analysis turned out to be robust to any alternative choice of the threshold just around 25 (for example 20 or 30 millions). With these results in hand, we can safely proceed by considering the interregional trade network resulting from the application of the threshold of $w > 25$ as our *benchmark*. Moreover, we assess the

[2] The sub-network including links with $w > 25$ is still weakly connected, but not strongly connected, that is not every vertex r is reachable from every s by a directed walk.

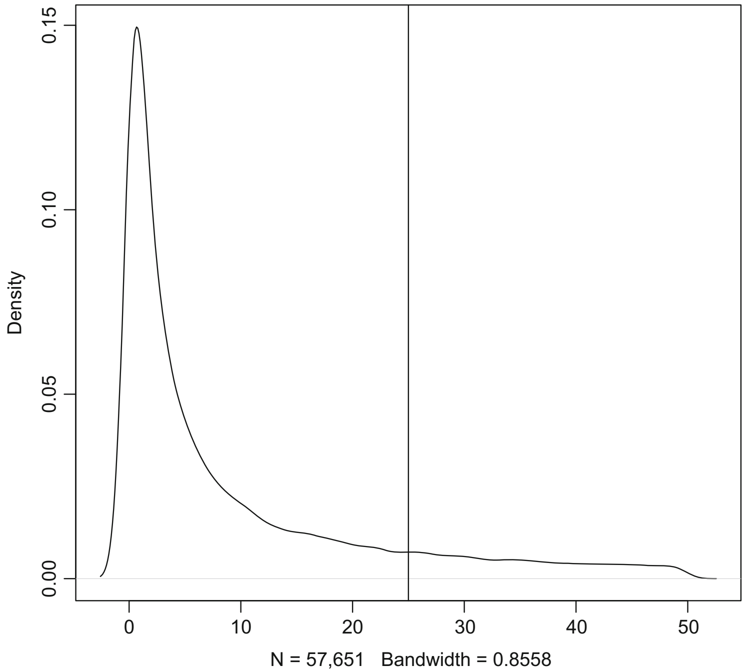


Fig. 1 Density of interregional trade flows

sensitivity of our analysis to alternative thresholds much further from 25 (namely for $w > 500$, $w > 1000$ and $w > 2500$, accounting for about 25, 10 and 1 % of the total intra-Europe trade) (see Table 1).

With a threshold of $w > 500$ or higher, the percentage of inter-regional European trade covered by data used to define the trade network gets substantially reduced, and with a threshold of $w > 1000$ the majority of European regions are excluded from the analysis and appear as isolates (see also Fig. 2), so that the number of zeros become exorbitant. This can be easily visualized using a *sociogram* for the different levels of threshold.

In Fig. 2, each European region is represented by a node in the *topological space*. The application of a so called force-directed algorithm on the regional trade data with valued links makes regions which are strongly connected close to each others, while regions which are not connected tend to be located far apart (Freeman 1979). However, the position of each region does not depend only on its bilateral links but also on the indirect effect of others: the trade partners of its trade partners will contribute to determine the region's position in the network. The role of the *third-region* effect clearly emerges from the visualizations in Fig. 2.

Figure 2 represents the directed network of European trade partners at the regional level in 2010. Nodes are European regions identified by NUTS-2 codes, (P =AT11, ..., UKN0), while links are weighted by the strength of trade flows

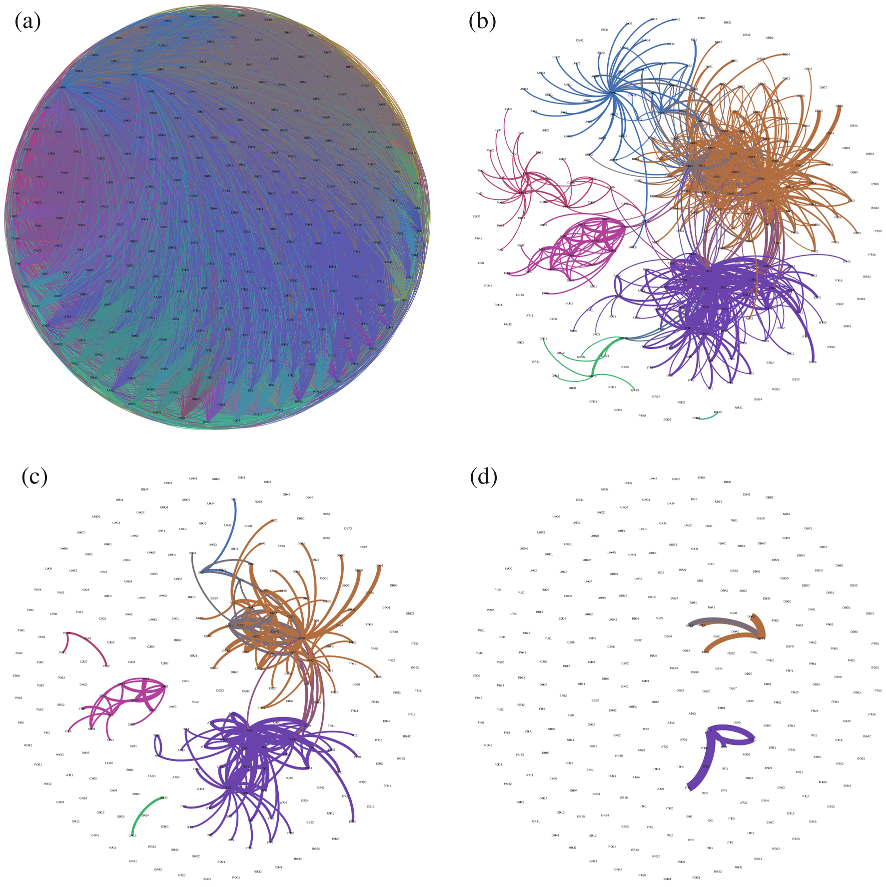


Fig. 2 Regional Trade in Europe: 2010. The figure represents the network of European trade partners at the regional level in 2010. Panel (**a**) visualizes all trade links with weight (exports) $w \geq 25$ (million of euros); panel (**b**) visualizes trade links with $w \geq 500$; panel (**c**) visualizes trade links with $w \geq 1000$; and panel (**d**) visualizes trade links with $w \geq 2500$. Regions (nodes) are identified by their NUTS-2 codes. *Colors* indicate homogeneous clusters defined according to modularity (with 12 clusters) (Color figure online)

(values in million of euros) ($W = \{\text{FR71}, \text{UKD3} = 251.69\}, \ldots, \{\text{ITC4}, \text{ITG1} = 5349.30\}$). Panels (a), (b), (c) and (d) visualize all trade links with weight $w \geq 25; 500; 1000; 2500$, respectively.

As shown in panel (d) the main European trade flows are between Italian regions (ITG1, ITC4, ITE4) and Spain (ES51), on one side, and German landers (DE11, DE12, DE21, DEA1) and France (FR10), on the other. Considering links of lesser strength (panel(c)) reinforces the impression that intra-national trade constitutes a substantial part of the structure of the European regional trade. New communities emerge (e.g. the Nordic countries, Poland, Greece) and the previous ones get reinforced by the inclusion of new links: the Italy-Spain community now includes some Portuguese, Austrian, Hungarian and Slovakian regions; and the Germany-France community is now enlarged to regions in Belgium, UK and the Netherlands. Colors indicate homogeneous cluster/community/modules of regions defined according to modularity (Newman 2006).[3]

Regions like ITC4, DE12, DE21 are at the center of the network, while regions like FI20, PL43, EE00, at the extreme left of the visualizations in Fig. 2, or PT30, BG32, PT15, at the extreme right, are at the boundaries of the network structure. The EU regional trade network displays a core-periphery structure, with the more active regions (i.e. with higher w) at the core. Also other regions are at the center of the visualizations in Fig. 2: UKE3, CZ01, NL12 are visually central, but their level of w is not among the highest percentiles of the distribution of w's. Their position depends on their respective links with their major trade partners. With respect to peripheral regions they are in fact preferentially linked with central regions. As pointed out above, indeed, centrality depends on direct links but can also depend on the centrality of regional trade partners. To clarify these issues, we will report the evidence of different centrality measures.

The simplest measure of centrality of V_r is the number of its neighbors (the number of direct trade connections region r has), namely its *degree*. The standardized *degree centrality* of a vertex is its degree divided by the maximum possible degree (Wasserman and Faust 1994; Newman 2003; Jackson 2008):

$$C_r^d = \frac{d_r}{n-1} = \frac{\sum_{s \neq r}^{n} l_{rs}}{n-1} \tag{1}$$

Since, in simple directed graphs like the one depicted in Fig. 2, a region can be both an exporter (a sender) and an importer (a receiver), we can compute both the *in-degree* of a region, $d_r = \sum_{s \neq r}^{n} d_{sr}$, as the number of incoming links (imports) to region r, and the *out-degree*, $d_r = \sum_{s \neq r}^{n} d_{rs}$, as the number of out-going links (exports) from region r towards its trade partners.

Imposing the condition $w > 25$ as a reasonable threshold that maintain the characteristics of the full network without assuming too much homogeneity between the different European regions (as shown in Table 1), standardized in-degree and out-degree distributions for the European interregional trade network in 2010 are shown in Fig. 3. In both cases a strongly asymmetric and bimodal distribution emerges, suggesting that there are two distinct dominant groups of regions with low and medium standardized degrees, while a small fraction of vertices has a high in-degree (out-degree).

More specifically, the first ten central regions in terms of *in-degree* are Ìle de France (FR10), Lombardia (ITC4), Oberbayern (DE21), Stuttgart (DE11), Dusseldorf (DEA1), Arnsberg (DEA5), Koln (DEA2), Giessen (DE71), Cataluna

[3]Networks with high modularity have dense connections between the nodes within modules but sparse connections between nodes in different modules. We use modularity to detect the community structure of the EU regional trade network. See Newman (2006) on this issue.

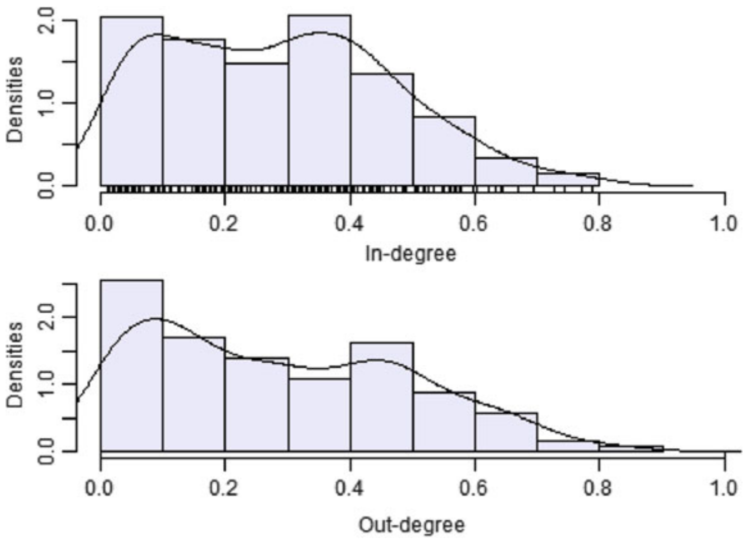


Fig. 3 In-degree and out-degree distribution. Histogram and density plots

(ES51) and Karlsruhe (DE12), with a level of $C_r^d > 0.63$ (the regions are inner linked with a little bit more than 63 % of possible regional partners, with a strength of $w > 25$), with German landers at a core of European markets in terms of imports. If we look at regions as exporters, the first ten regions are, respectively, ITC4, DE21, DE11, DE71, DEA5, DE12, NL33, DEA1, DE25, DE13, with a level of $C_r^d > 0.66$ (the regions are outer linked with a little bit more than 66 % of possible regional partners, with a strength of $w > 25$), with the role of Germany even more prominent. Seven out of ten regions are in the both top-tens, the more connected importing regions are also the most connected exporting regions. More broadly, in-degree and out-degree are positively correlated with a Pearson coefficient of 0.9. There are however some notable exceptions: UK12 is in the top-twenty as an importer, ranking 95th as an exporter.

The scatter plot of in-degree versus out-degree (both derived using Eq. (1)) is depicted in Fig. 4. The French region of Ìle de France (FR10), as previously mentioned, is the EU region with highest In-degree, a characteristic strictly associated with its level of regional GDP, highlighted by the dimension of the dot representing the position of the region in the degree-space. The German regions of Stuttgart (DE11) and Oberbayern (DE21) and the Italian region of Lombardia (ITC4) lead the EU regions in terms of out-degree. The scatter plot clearly confirms the positive correlation between In-degree and Out-degree, but also shows the level of dispersion of the EU regional trade centralities. A notable case is the one of the Great Britain regions of Inner London (UKI1) and Outer London (UKI2) showing a level of in-degree much higher than the level of out-degree, depending on the lesser importance of regional manufacturing with respect to service production and trade.

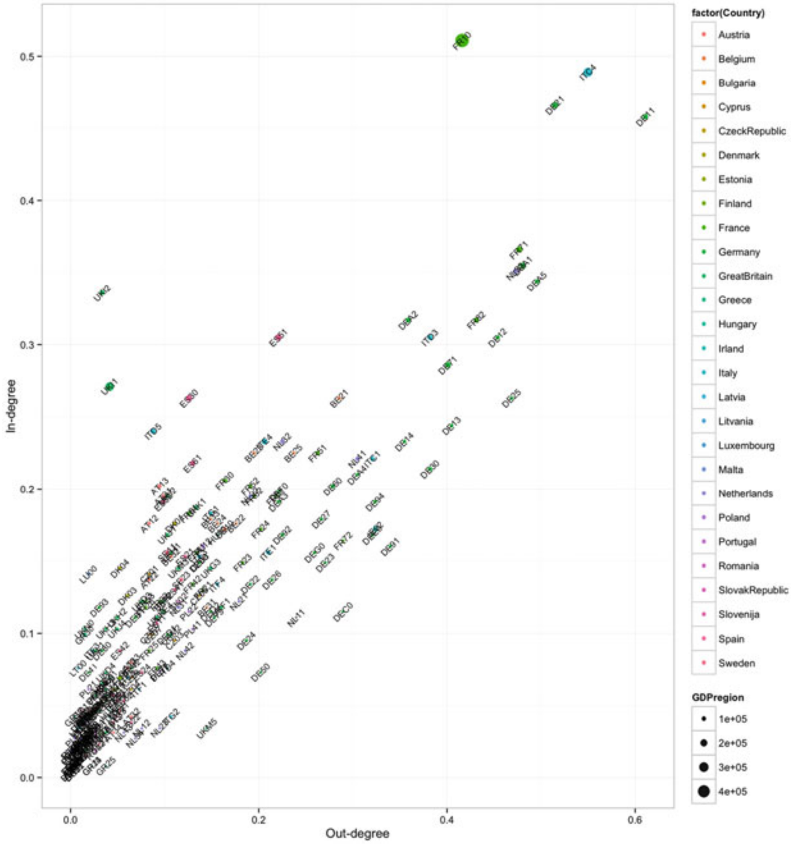


Fig. 4 Scatterplot of in-degree and out-degree centralities. The scatterplot confronts the level of in-degree with the level of out-degree for each EU region. Regions of the same country share the same *color*. The size of the *dots* is proportional to regional GDP. *Dots* are labeled according to NUT-2 codes (Color figure online)

Beyond the degree distribution itself, it is interesting to understand the manner in which regions of different degrees are linked with each other. To this end, we plot the average neighbor degree versus vertex degree (Fig. 5). This plot suggests that, while there is a tendency for regions of higher degrees to link with similar regions, nodes of intermediate degree tend to link with regions of both intermediate and higher degrees. This issue can be better analyzed using a global measure of centrality, namely the eigenvector centrality.[4]

[4]The degree centrality (C_r^d) is classified as a *local measure of centrality* since it takes into consideration only the direct links of a node, its nearest neighborhood, regardless of the position of the node in the network's structure. Contrary to the local measures, *global measures of centrality* uncover the effect of others at a higher level of connection, including the direct and the *indirect*

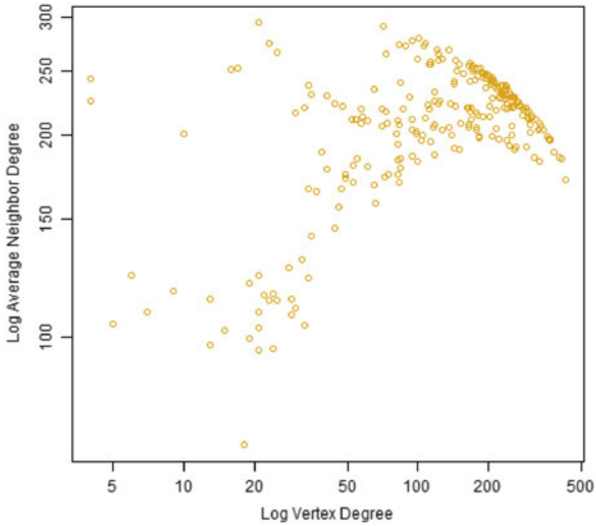


Fig. 5 Average neighbor degree versus vertex degree (log-log scale)

Figure 6 shows the Eigenvector centrality distribution for the European interregional trade network. Again a right skewed bimodal distribution emerges. The ten most central regions (in terms of Eigenvector centrality) are, in order, Ìle de France (FR10), Oberbayern (DE21), Lombardia (ITC4), Stuttgart (DE11), Dusseldorf (DEA1), Arnsberg (DEA5), Koln (DEA2), Karlsruhe (DE12), Darmstadt (DE71) and Rhone-Alpes (FR71).

Last but not least, we explore the new dataset through network analysis to evaluate to what extent two regions that both trade with a third region are likely to trade with each other as well. This notion corresponds to the social network concept of *transitivity* and can be captured numerically through an enumeration of the proportion of vertex triples that form triangles (i.e., all three vertex pairs are connected by edges), typically summarized in a so-called *clustering* coefficient. Table 1 shows that, with $w > 25$, about 70 % of the connected triples close to form triangles. With higher thresholds of exports ($w > 500$, $w > 1000$ and $w > 2500$), this fraction steeply decreases.

More deeply, the role of the third region can be studied (and properties can be tested) using the *Triad Census* (the count of the various type of triads in the network) as a tool (Wasserman and Faust 1994). Classical triad census analysis applies to single, directed and binary network, and we will follow the tradition in this respect.

effect of potentially all nodes in the network. In particular, the *eigenvector centrality* captures the idea that the more central the neighbors of a vertex are, the more central that vertex itself is. In other words, eigenvector centrality gives greater weight to a node the more it is connected to other highly connected nodes. Thus, it is often interpreted as measuring a node's network importance.

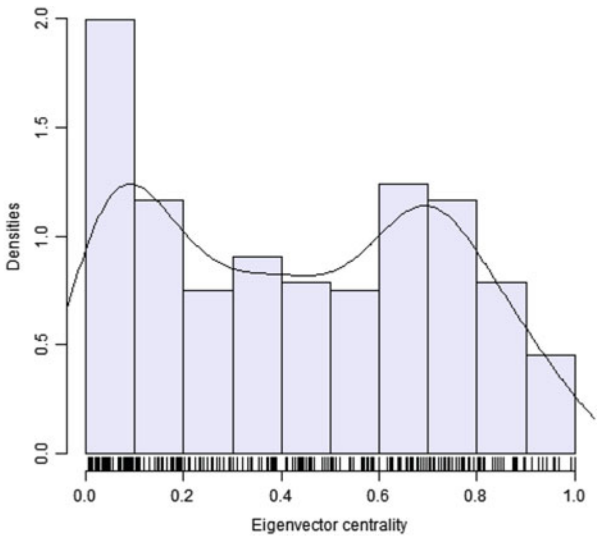


Fig. 6 Eigenvector centrality distribution

Taking the three nodes V_s, V_r and V_k, where $s \neq r \neq k$, we can call them a *triple*, if we also consider the presence or absence of links between the different nodes we have a *triad*. T_{srk} is the triad involving V_s, V_r and V_k. If the network is composed of n nodes, there are $\binom{n}{3} = \frac{n(n-1)(n-2)}{6}$ triads. In the EU regional trade network there are, therefore, 3,136,805 triads.

As far as possible realizations of triads, since there are three nodes in a triad, and each node can be connected to two other nodes, this give rise to six possible links. Since each link can be present or absent, there are $2^6 = 64$ possible realizations of the triads. Excluding isomorphic cases (e.g. if V_s, V_r and V_k are not linked, T_{srk}, T_{rks} and T_{ksr} are isomorphic), we remain with 16 isomorphism classes for 64 different triad states. These classes, represented in Fig. 7, range from the null subgraph to the subgraph in which all three dyads formed by the vertices in the triad have mutual directed links. The figure is from Wasserman and Faust (1994, p. 566) as reproduced in De Nooy et al. (2011). The different classes are labeled with as many as four characters, according to the M-A-N labeling scheme of Holland and Leinhardt (1970), where the first character gives the number of Mutual dyads in the triad, the second the Asymmetric ones, the third the number of Null dyads, and lastly, the forth one, if present, is used to distinguish further among the types (e.g. the two 030 triads—panel 9 and 10 in Fig. 7—can be distinguished by the transitivity of the dyad 9 and the cyclic links of dyad 10). The four letters in the forth character are "U" (for up), "D" (for down), "T" (for transitive) and "C" (for cyclic).

For every network it is possible to calculate the frequencies of the 16 classes. In Table 2 we report the triad census for the European regional trade network at

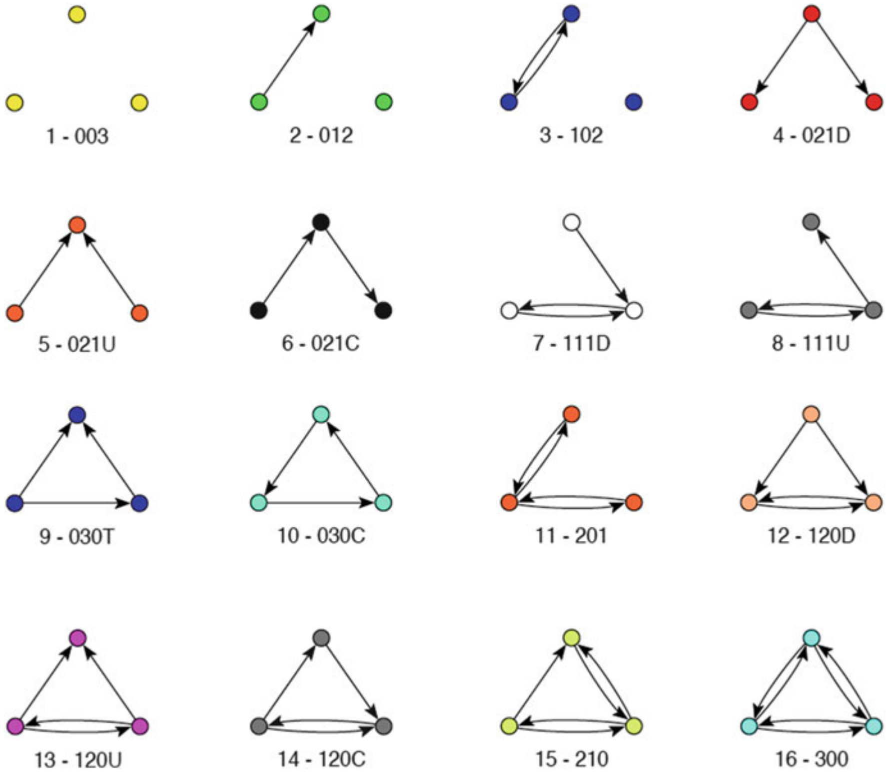
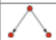


Fig. 7 Triads in a digraph. The figure is from Wasserman and Faust (1994, p. 566) as reproduced in De Nooy et al. (2011). The triples of directional relations are called *triads*. Among the numbers at the bottom of each panel, the first one is progressive from 1 to 16 and indicates all the possible cases of triads, while the second is the M-A-N labeling scheme of Holland and Leinhardt (1970): the first character gives the number of Mutual dyads in the triad, the second the Asymmetric ones, the third the number of Null dyads, and lastly, the forth one, if present, is used to distinguish further among the types (e.g. 003 triad as 0 Mutuals, 0 Asymmetrics and 3 Nulls)

different levels of threshold w. Every triadic census, reported in columns 4 to 8, is calculated excluding isolated regions and zero valued edges, e.g., in the last column of Table 2 the triad census is calculated for the 13 regions and 12 edges of the trade network with $w > 2500$. Here we discuss only our preferred structure (i.e., $w > 25$).

The fifth column in Table 2 shows that the large majority of triads in the EU regional trade is represented by empty-graph structures (corresponding to the 003 MAN code, row 1 in Table 2); followed by single and mutual edges (012 and 102 MAN codes, rows 2 and 3 in Table 2); and then by stars (in-stars 021U, out-stars 021D, and especially mutual stars, 201 MAN code, rows 5, 4 and 11 in Table 2). One noteworthy characteristic of the EU regional trade network is the prevalence of mutual edge + (double) Out structures (111U and 120U MAN codes, rows 8 and 13 in Table 2) over mutual edge + (double) In structures (111D and 120D MAN codes, rows 7 and 12 in Table 2). It seems that when two regions establish a mutual trade

Table 2 Triad census

	MAN code	Figure	Class	Full	w>25 mln.	w>500 mln.	w>1000 mln.	w>2500 mln.
1	003		Empty graph	14	1,040,546	1,549,950	147,072	186
2	012		Single edge	0	487,976	88,052	10,440	73
3	102		Mutual edge	182	737,954	30,906	2907	6
4	021D		Out-star	1	32,297	2665	327	3
5	021U		In-star	0	20,621	806	175	10
6	021C		Line	0	29,254	1407	179	3
7	111D		Mutual edge + In	0	87,483	1314	164	1
8	111U		Mutual edge + Out	0	152,936	2562	276	4
9	030T		Transitive	1	13,299	151	11	0
10	030C		Cycle	0	664	7	0	0
11	201		Mutual-star	7808	157,167	965	95	0
12	120D		Mutual edge + double In	121	17,185	63	2	0
13	120U		Mutual edge + double Out	248	36,178	158	19	0
14	120C		Mutual edge + cycle	21	16,694	131	4	0
15	210		Almost complete graph	15,645	125,402	338	23	0
16	300		Complete graph	3,112,764	145,904	105	6	0[3*pt*]

Every triadic census (columns 4–8) is calculated excluding isolated regions and zero valued edges relationship this fosters them to export to, more that to import from, a third region. This tendency persists at different level of threshold w (as can be seen in columns 6–8 in Table 2). This aspect of the EU regional trade network will be discussed and theoretically motivated in the subsequent sections.

Overall, the evidence emerging from our analysis suggests interesting insights on the interregional trade network in Europe, which also inform our theoretical model described in the following sections. First, it emerges that the interregional trade network (and, thus, the European economic integration) is far from being complete since most regions do not trade (or trade with a very low intensity) with all other regions, but they rather select their partners. This first stylized fact clearly emerges once we neglect bilateral trade flows lower than 25 millions of euros. Second, interregional trade flows, partners and links in Europe are strongly heterogeneous, with a relatively small number of regions playing a central role in the network structure, both in terms of number of links and amount of intra-Europe trade flows accounted. In particular, our findings clearly show that distance matters also in trade between regions and that the emerging clusters are characterized by geographic proximity. Specifically, the national homogeneity of clusters gives evidence that national borders are relevant for regional analysis of EU trade flows. Finally, from the triadic census analysis it emerges that the tendency to reciprocate trade links (i.e., closing triangles) is limited and mutual edge patterns tend to prevail.

3 General Framework

In this section, we build a three-region linear new economic geography (NEG) model along the lines of the evidence emerged from the empirical analysis of the EU regional trade network.

3.1 Basic Assumptions

The economy is composed of three regions (labeled suitably r, s and k). There are two sectors, Agriculture (A-sector) and Manufacturing (M-sector). Workers and entrepreneurs are the two types of agents operating in the economy, each of them is endowed with a factor of production, unskilled labor (L-factor) and human capital (E-factor).[5] L can be used in both sectors; whereas E is specific to M.

[5] A crucial difference between human and, for example, knowledge capital is that the former is embodied into the owner, whereas the second is separated. In a NEG model, this difference enters into play only when factor migration is allowed. When human capital is considered, changes in real incomes alters the migration choice also via the so-called "price index effect" so that changes in local prices may affect the long-run distribution of the industrial sector. Instead, when knowledge capital is concerned, factor movements are only driven by regional nominal profit differentials.

Production in the A-sector involves a homogeneous good, whereas in the M-sector the output consists of N differentiated varieties. The three regions are symmetric—they have the same endowment of L and are characterized by the same production technology and consumption preferences—except for their distance. This translates into regional differences in trade costs.

3.2 Production

The A-sector is characterized by perfect competition and constant returns to scale. The production of 1 unit of the homogeneous good requires only unskilled workers as an input. Without loss of generality, we assume that 1 unit of labor gives 1 unit of output. The A-good is also chosen as numéraire. In the M-sector, instead, (Dixit-Stiglitz) monopolistic competition and increasing returns prevail. In this sector, identical firms produce differentiated varieties with the same technology involving a fixed component, 1 entrepreneur, and a variable component, η units of unskilled labor for each unit of the differentiated variety. Total cost TC for a firm i producing q_i output units corresponds to:

$$TC(q_i) = \pi_i + w\eta q_i$$

where w is the wage rate, π_i represents the remuneration of the entrepreneur and the operating profit. Given consumers' preference for variety (see below) and increasing returns, each firm will always produce a variety different from those produced by the other firms (no economies of scope are allowed). Moreover, since one entrepreneur is required for each manufacturing firm, the total number of firms/varieties, N, always equates the total number of entrepreneurs, $E = N$. Denoting by λ_r the share of entrepreneurs located in region r the number of regional varieties produced in that region is:

$$n_r = \lambda_r N = \lambda_r E.$$

A new economic geography (NEG) model in which the mobile factor is human capital (or, alternatively, skilled labor or entrepreneurship) is known as Footloose Entrepreneur (FE) model (developed originally in Forslid and Ottaviano 2003); a NEG model in which the mobile factor is separated from the owner (such as physical or knowledge capital) is labeled Footloose Capital (FC) model (developed firstly in Martin and Rogers 1995). As mentioned above, this distinction becomes relevant moving from the short to the long-run when factor migration is allowed. Even if the basic structure of the model is equivalent for FC and FE models, we consider the factor specific to the M-sector an entrepreneur.

3.3 Utility Function

Following Ottaviano et al. (2002), to represent individual preferences we adopt a quasi-linear utility function. As we shall see, a crucial difference of this modeling strategy with respect to the standard CES approach is a CIF price (price at destination) which falls as the number of local competing firms rises. This implies a stronger dispersion force (given by the competition effect). Another difference is that it allows to highlight alternative patterns of trade (autarky, one-way, two-way trade) as shown by Behrens (2004, 2005b, 2011) and by Okubo et al. (2014) for the case of two-region economies.

The utility function is composed of a quadratic part defining the preferences across the M goods and a linear component for the consumption of the A-good:

$$U = \alpha \sum_{i=1}^{N} c_i - \left(\frac{\beta - \delta}{2}\right) \sum_{i=1}^{N} c_i^2 - \frac{\delta}{2} \left(\sum_{i=1}^{N} c_i\right)^2 + C_A \tag{2}$$

where c_i is the consumption choice concerning the variety i; α represents the intensity of preferences for the manufactured varieties, with $\alpha > 0$; δ represents the degree of substitutability across those varieties, $\delta > 0$; and where the taste for variety is measured by the (positive) difference $\beta - \delta > 0$.

The representative consumer's budget constraint is:

$$\sum_{i=1}^{N} p_i c_i + C_A = y + \overline{C}_A \tag{3}$$

where $\overline{C}_A$ is the individual endowment of the agricultural good which is assumed sufficiently large to allow for positive consumption of this good in equilibrium; p_i is the price of variety i inclusive of transport costs and y is the income of the individual agent (unskilled worker or entrepreneur).

3.4 Trade Costs

The three regions constitute the nodes of a network economy in which the links are the flows of commodities (esp. those produced in the M-sector). The existence, direction and magnitude of these flows depend on the size of trade costs. There are three types of flows/links between two regions—labeled let's say r and s—a one directional link from r to s, where r is the exporting region and s the importing region; a one directional link from s to r, where s is the exporting region and r the importing region; and a bidirectional link between r and s, where both regions export to and import from the other. Including also the possibility of no links, the maximum number of possible network structures involving three regions is 64; excluding the "isomorphic" cases this number reduces to 16 (see the discussion on triad census above).

Differently from Ago et al. (2006), according to whom the three regions are equally spaced along a line, and from Behrens (2011), according to whom two regions in a trade bloc are at the same distance from a third outside region, we assume that the distance between regions is not necessarily the same. Moreover, we assume identical bilateral trade costs, so that the cost of trading industrial commodities from r to s and from s to r is identical, that is, $T_{rs} = T_{sr}$; and no cost of trading goods within a region, that is, $T_{rr} = 0$. Moreover, we do not assume an a priori specific trade costs configuration,[6] that is, $T_{rs} \neq T_{rk}$ and/or $T_{rk} \neq T_{sk}$ and/or $T_{rs} \neq T_{sk}$.

Letting r, s and k be the three regions under consideration, the trade cost matrix can be written as:

$$\begin{pmatrix} 0 & T_{rs} & T_{rk} \\ T_{rs} & 0 & T_{sk} \\ T_{rk} & T_{sk} & 0 \end{pmatrix}$$

4 Short-Run Equilibrium

We limit our analysis to the equilibrium that emerges in the short run, contingent on given regional shares of entrepreneurs $(\lambda_r, \lambda_s, \lambda_k)$, leaving for future work the analysis of the entrepreneurial migration processes that characterize the long run.

4.1 Equilibrium Determination

Solving for C_A the budget constraint (3), substituting into the utility function (2) and then differentiating with respect to c_i, we obtain the following first-order conditions $(i = 1, \ldots, N)$:

$$\frac{\partial U}{\partial c_i} = \alpha - (\beta - \delta)\, c_i - \delta \sum_{i=1}^{N} c_i - p_i = 0$$

[6]A specific configuration could emerge after empirical analysis. We could have, for example, a "hub and spoke" structure by letting: $T_{rs} \leq T_{rk} \leq T_{sk}$, with $T_{rs} \neq T_{rk}$ and/or $T_{rk} \neq T_{sk}$. This would stress the locational advantage of region r (the "hub") with respect to s and k (the "spokes").

from which

$$p_i = \alpha - (\beta - \delta)\, c_i - \delta \sum_{i=1}^{N} c_i\,.$$

The linear demand function is

$$c_i(p_1, \ldots, p_N) = \frac{\alpha}{(N-1)\delta + \beta} - \frac{1}{\beta - \delta} p_i + \frac{\delta}{(\beta - \delta)[(N-1)\delta + \beta]} \sum_{i=1}^{N} p_i$$
$$= a - (b + cN)p_i + cP$$

where $P = \sum_{i=1}^{N} p_i$, $0 \leq p_i \leq \tilde{p} \equiv \frac{a+cP}{b+cN}$ and

$$a \equiv \frac{\alpha}{(N-1)\delta + \beta}, \quad b \equiv \frac{1}{(N-1)\delta + \beta}, \quad c \equiv \frac{\delta}{(\beta - \delta)[(N-1)\delta + \beta]}.$$

The indirect utility is given by:

$$V = S + y + \overline{C}_A$$

where S corresponds to the consumer's surplus:

$$S = U(c(p_i), i \in [0, N]) - \sum_{i=1}^{N} p_i c_i(p_i) - C_A$$
$$= \frac{a^2 N}{2b} + \frac{b + cN}{2} \sum_{i=1}^{N} p_i^2 - aP - \frac{c}{2} P^2.$$

The consumer's demand originating from region s—but it could be region r or region k, after a suitable change in the subscripts—for a good produced in region r—but it could be region s or k—is

$$c_{rs} = a - (b + cN)p_{rs} + cP_s$$

where c_{rs} is the demand of a consumer living in region s for a good produced in region r; p_{rs} is the price of a good produced in region r and consumed in region s; and P_s is the price index in region s, with

$$P_s = n_r p_{rs} + n_s p_{ss} + n_k p_{ks.}$$

Notice that, following from the assumption of symmetric behavior of firms, prices differ across regions—segmenting markets—only because of trade costs.

Short-run equilibrium requires that in each segmented market demand equals supply:

$$c_{rs} = q_{rs}$$

where q_{rs} is the output produced in region r—but it could be region s or k—that is brought to a market in region s—but it could be region r or k.

In order to derive the short-run solutions, we only consider region r (but the same reasoning applies to region s or k after a suitable change in the subscripts). The operating profit of a representative firm in region r is:

$$\begin{aligned}\pi_r = & \ (p_{rr} - \eta)q_{rr}(L_r + \lambda_r E) \\ + & \ (p_{rs} - \eta - T_{rs})q_{rs}(L_s + \lambda_s E) \\ + & \ (p_{rk} - \eta - T_{rk})q_{rk}(L_k + \lambda_k E).\end{aligned}$$

From the profit maximization procedure and market segmentation, considering further than $N = E$, the first-order conditions follow:

$$\frac{\partial \pi_r}{\partial p_{rr}} = [a + \eta(b + cE) + cP_r - 2p_{rr}(b + cE)](L_r + \lambda_r E) = 0$$

$$\frac{\partial \pi_r}{\partial p_{rs}} = [a + (\eta + T_{rs})(b + cE) + cP_s - 2p_{rs}(b + cE)](L_s + \lambda_s E) = 0$$

$$\frac{\partial \pi_r}{\partial p_{rk}} = [a + (\eta + T_{rk})(b + cE) + cP_k - 2p_{rk}(b + cE)](L_k + \lambda_k E) = 0$$

Taking into account trade costs and letting $\tilde{p}_r = \frac{a+cP_r}{b+cN} > \eta$, profit-maximizing prices correspond to

$$p_{rr} = \frac{a + cP_r + \eta(b + cE)}{2(b + cE)} = \frac{\tilde{p}_r}{2} + \frac{\eta}{2} \tag{4}$$

$$p_{rs} = \begin{cases} \frac{a+cP_s+(\eta+T_{rs})(b+cE)}{2(b+cE)} = \frac{\tilde{p}_s}{2} + \frac{\eta}{2} + \frac{T_{rs}}{2} & \text{if } T_{rs} \le \tilde{p}_s - \eta \\ \tilde{p}_s & \text{if } T_{rs} > \tilde{p}_s - \eta \end{cases} \tag{5}$$

$$p_{rk} = \begin{cases} \frac{a+cP_k+(\eta+T_{rk})(b+cE)}{2(b+cE)} = \frac{\tilde{p}_k}{2} + \frac{\eta}{2} + \frac{T_{rk}}{2} & \text{if } T_{rk} \le \tilde{p}_k - \eta \\ \tilde{p}_k & \text{if } T_{rk} > \tilde{p}_k - \eta \end{cases} \tag{6}$$

where p_{rr} is the price that a firm located in r sets in its own market, p_{rs} the price that such a firm sets in market s, p_{rk} the price set in market k, $\tilde{p}_s = \frac{a+cP_s}{b+cE}$ the reservation price of a consumer living in region s and $\tilde{p}_k = \frac{a+cP_k}{b+cE}$ that of a consumer living in region k.

Using the demand and price functions, we can write:

$$q_{rr} = (b + cE)(p_{rr} - \eta) \tag{7}$$

$$q_{rs} = \begin{cases} (b + cE)(p_{rs} - \eta - T_{rs}) & if\ T_{rs} \leq \tilde{p}_s - \eta \\ 0 & if\ T_{rs} > \tilde{p}_s - \eta \end{cases} \tag{8}$$

$$q_{rk} = \begin{cases} (b + cE)(p_{rk} - \eta - T_{rk}) & if\ T_{rk} \leq \tilde{p}_k - \eta \\ 0 & if\ T_{rk} > \tilde{p}_k - \eta \end{cases} . \tag{9}$$

According to expressions (5)–(6) and (8)–(9), if a firm located in r quotes in market s (or in market k) a price larger than the reservation price for consumers resident in s (or in k), the export from region r to region s (or k) is zero.[7] The boundary condition for trade as reported in these expression is crucial for the following analysis.

The indirect utility for r is given by

$$V_r = S_r + y + \overline{C}_A$$

where S_r corresponds to the consumer's surplus:

$$S_r = \frac{a^2 E}{2b} + \frac{b + cE}{2}\left(\lambda_r p_{rr}^2 + \lambda_s p_{sr}^2 + \lambda_k p_{kr}^2\right) E - aP_r - \frac{c}{2} P_r^2 .$$

We group all the 16 possible network structures created by the trade flows between the regions into four cases: (1) no trade occurs between all the regions; (2) one-way or two-way trade occurs between region r and s and region k is in autarky; (3) one-way or two-way trade occurs between regions r and s and r and k, but regions s and k do not trade with each other (what is called in the triad census terminology a "star" structure); (4) one-way or two-way trade occurs between any two-regions in the economy. For all network structures we derive the relevant conditions as determined by the relationship between trade costs and the distribution of the industrial activity. Some specific cases will be developed in some detail to understand the effects on the three regions of the creation of a new link. As we shall see, there is a one-to-one correspondence between the trade network structures and the triad census taxonomy so that we can match each trade network configuration to a triad as reported in Fig. 7.

[7] See Behrens (2004, 2005b, 2011). On the empirical relevance of the zero in the trade flow matrix, see Melitz (2003).

4.2 Case (1) All Autarkic Regions

First we consider the case in which all the regions are in autarky. This corresponds to triad 1 in Fig. 7. Due to the isomorphic properties, we focus only on region r.

Region r is in autarky when conditions $T_{rs} > \tilde{p}_s - \eta$ and $T_{rk} > \tilde{p}_k - \eta$ apply. Using the Eqs. (4), (5) and (6) and the expression for the reservation price, these "no-trade" conditions can be written as:

$$T_{rs} > \frac{2(a-\eta b)}{2b + c\lambda_s E} = \widetilde{T}_s \quad \text{and} \quad T_{rk} > \frac{2(a-\eta b)}{2b + c\lambda_k E} = \widetilde{T}_k$$

where $\frac{\partial \widetilde{T}_s}{\partial \lambda_s} = -\frac{2(a-\eta b)}{(2b+c\lambda_s E)^2} cE < 0$ and $\frac{\partial \widetilde{T}_k}{\partial \lambda_k} = -\frac{2(a-\eta b)}{(2b+c\lambda_k E)^2} cE < 0$.

These results have two important implications: (1) a lower degree of local competition increases the likelihood of interregional trade—making more permeable the local market in s (or in k) for firms located in r; (2) reducing distance—i.e. trade costs—between regions r and s (or between r and k) has a similar impact. That is, closer regions have a more accessible market.

If trade costs are too high, then no trade occurs among the three regions and firms only sell in the local market. From (4), (5) and (6), considering the linear demand non-negativity constraint, we obtain the price index for region r:

$$P_r = \frac{a(2-\lambda_r) + \eta\lambda_r(b+cE)}{2b + c\lambda_r E} E$$

and the equilibrium prices and quantities:[8]

$$p_{rr} = \frac{a + \eta(b + c\lambda_r E)}{2b + c\lambda_r E} \quad q_{rr} = (b + cE)(p_{rr} - \eta)$$
$$p_{rs} = 2p_{ss} - \eta \quad q_{rs} = 0$$
$$p_{rk} = 2p_{kk} - \eta \quad q_{rk} = 0.$$

As shown by these expressions, the equilibrium prices and quantities depend negatively on the number of local firms λ_r. This is a manifestation of the so-called local competition effect: the larger is the number of firms competing in the local market, the lower the price firms are able to set (and the smaller the output they are able to sell).

Taking into account our symmetry assumption, according to which the regions are endowed with the same number of L, the equilibrium short-run profit for a firm

[8]Note that analogous expressions can be obtained for s or for k by simple switching r and s or r and k.

located in region r, which sells only to the local market, corresponds to:[9]

$$\pi_r = (p_{rr} - \eta)^2(b + cE)\left(\frac{L}{3} + \lambda_r E\right) = (b + cE)\left(\frac{a - \eta b}{2b + c\lambda_r E}\right)^2\left(\frac{L}{3} + \lambda_r E\right).$$

Finally, we obtain the indirect utility of an entrepreneur resident in r:

$$V_r = S_r + \pi_r + \overline{C}_A = \frac{(a - \eta b)^2(b + cE)[3\lambda_r E(3b + c\lambda_r E) + 2bL]}{6b(2b + c\lambda_r E)^2} + \overline{C}_A\,.$$

4.3 Case (2) Trade Only Occurs Between Two Regions

We now consider the case where (one-way or two-way) trade occurs but only between two regions. This implies that, because of high trade costs, the third region has no trade links with the other two. We focus on the possible links between regions r and s; the region in autarky is, therefore, k. Due to the isomorphic properties, the same analysis applies for the links between regions r and k (with s in autarky) or regions s and k (with r in autarky).

There are two subcases: (2.A) one-way trade from region r to region s (isomorphic to the link going in the opposite direction from s to r) corresponding to triad 2 in Fig. 7; and (2.B) two-way trade between r and s, corresponding to triad 3.

Considering the subcase (2.A), from Eqs. (4), (5) and (6), we deduce that one-way trade from region r to region s occurs as long as:

$$\frac{2(a - \eta b)}{2b + c\lambda_r E} = \widetilde{T}_r < T_{rs} \le \widetilde{T}_s = \frac{2(a - \eta b)}{2b + c\lambda_s E}. \tag{10}$$

That is, the region with the larger share of entrepreneurs has a higher chance to be the exporting region, taking advantage of the lower competition in the destination market.

When only one-way trade from region r to region s is allowed, then firms in region r are able to sell both in their local market and in the outside market s. Firms in region s, as before, only produce for the local market but now they have to compete not only with each other but also with firms located in r.

[9]Notice that the assumption of identical workers population has no significant impact on the short-run analysis and it can be easily removed; whereas in the long run, it determines the regional entrepreneurial shares and, via these shares, the trade flows.

We also derive the conditions of "no trade" between regions r and k and between regions s and k:

$$\max\left(\frac{2(a-\eta b)}{2b+c\lambda_r E},\frac{2(a-\eta b)}{2b+c\lambda_k E}\right)=\max\left(\widetilde{T}_r,\widetilde{T}_k\right)<T_{rk} \quad (11)$$

$$\max\left(\frac{2(a-\eta b)+cE\lambda_r T_{rs}}{2b+c\left(\lambda_r+\lambda_s\right)E},\frac{2(a-\eta b)}{2b+c\lambda_k E}\right)=\max\left(\widetilde{T}_{ks},\widetilde{T}_k\right)<T_{sk}. \quad (12)$$

Looking at the expression (12), we note that the boundary condition for a one-directional link from region k to region s is more restrictive than before, $\widetilde{T}_{ks}<\widetilde{T}_s$ since $T_{rs}<\widetilde{T}_s$, taking into account the competition coming from the third region r. Moreover, from these conditions, it follows that the distance from r to s is shorter than the distance from s to k ($T_{rs}<T_{sk}$). This can be proven considering that 1) $T_{rs}\leq\widetilde{T}_{ks}$ from the condition $T_{rs}\leq\widetilde{T}_s$ in (10) and 2) $\widetilde{T}_{ks}<T_{sk}$ from (12). Putting together 1) and 2), we have $T_{rs}<T_{sk}$.

The price indexes when only one-way trade from r to s occurs are:

$$P_r=\frac{a(2-\lambda_r)+\eta\lambda_r(b+cE)}{2b+c\lambda_r E}E$$

$$P_s=\frac{a[2-(\lambda_r+\lambda_s)]+[\eta(\lambda_r+\lambda_s)+T_{rs}\lambda_r](b+cE)}{2b+c(\lambda_r+\lambda_s)E}E$$

$$P_k=\frac{a(2-\lambda_k)+\eta\lambda_k(b+cE)}{2b+c\lambda_k E}E$$

and the equilibrium prices and quantities:

$$p_{rr}=\frac{a+\eta(b+c\lambda_r E)}{2b+c\lambda_r E} \qquad q_{rr}=(b+cE)(p_{rr}-\eta)$$

$$p_{ss}=\frac{a+\eta[b+c(\lambda_r+\lambda_s)E]+\frac{T_{rs}}{2}c\lambda_r E}{2b+c(\lambda_r+\lambda_s)E} \qquad q_{ss}=(b+cE)(p_{ss}-\eta)$$

$$p_{kk}=\frac{a+\eta(b+c\lambda_k E)}{2b+c\lambda_k E} \qquad q_{kk}=(b+cE)(p_{kk}-\eta)$$

$$p_{rs}=p_{ss}+\frac{T_{rs}}{2} \qquad q_{rs}=(b+cE)(p_{rs}-\eta-T_{rs})$$

$$p_{rk}=2p_{kk}-\eta \qquad q_{rk}=0$$

$$p_{sr}=2p_{rr}-\eta \qquad q_{sr}=0$$

$$p_{sk}=2p_{kk}-\eta \qquad q_{sk}=0$$

$$p_{kr}=2p_{rr}-\eta \qquad q_{kr}=0$$

$$p_{ks}=2p_{ss}-\eta \qquad q_{ks}=0.$$

As these equations show, the creation of a one-directional trade link from region r to region s affects negatively the price applied by local firms in region s. Moreover, the competition effect is reinforced since now they have to compete not only with the other local firms but also with the firms located in region r.

Due to lack of space, from now on we do not derive explicit results for profits and indirect utilities—given that, especially the second ones become more and more complicated as the number of links increases. Brief comments on the short-run effects of a link creation on all the regions will be provided for some of the trade network configurations. Indeed, as we shall see, the creation of a link may have an effect not only on the two regions involved but also on the third. Analogously, the presence of a third region may alter the effect of the new link on the two regions directly involved. This is a version of the so-called "third-region" effect.

The equilibrium profits for the case of one-way trade from r to s are:

$$\pi_r = (b + cE)\left[(p_{rr} - \eta)^2\left(\frac{L}{3} + \lambda_r E\right) + (p_{rs} - \eta - T_{rs})^2\left(\frac{L}{3} + \lambda_s E\right)\right]$$

$$\pi_s = (b + cE)\left[(p_{ss} - \eta)^2\left(\frac{L}{3} + \lambda_s E\right)\right]$$

$$\pi_k = (b + cE)(p_{kk} - \eta)^2\left(\frac{L}{3} + \lambda_k E\right).$$

The effect of the creation of a one-directional link (i.e. one-way trade or exports) from region r to region s has a positive effect on the welfare of region r (the exporting region). Indeed, the opening of the new market (the one in region s)—and the generation of the profit accruing from it—causes an increase in the overall profit of the r-firms (i.e. the firms located in region r). The impact on the welfare of region s (the importing region), instead, is ambiguous due to the counterbalancing of two opposite effects: the one on the consumer's surplus—that we call "the surplus effect"—which is positive and the one on profits—that we call "the profit effect"—which is negative. The first is induced by the larger availability of manufactured goods traded in the local market; whereas the second by the stronger competition that the s-firms have to suffer in the local market coming from the r-firms. The "surplus effect", which is smaller at the beginning of the integration process, may overcome the "profit effect", with further trade liberalization, depending on parameter values and on the regional distribution of the industrial sector.

Moving on to the subcase (2.B), when trade costs are reduced enough, two-way trade between region r and s is allowed, implying that with respect to the previous situation, now also firms in s are able to compete in both markets. Two-way trade between region r and region s occurs as long as

$$T_{rs} \leq \min\left(\frac{2(a - \eta b)}{2b + c\lambda_r E}, \frac{2(a - \eta b)}{2b + c\lambda_s E}\right) = \min\left(\widetilde{T}_r, \widetilde{T}_s\right). \tag{13}$$

For future reference, we also derive the conditions for "no trade" between region r and k and between region s and k:

$$\max\left(\frac{2(a-\eta b)+cE\lambda_s T_{rs}}{2b+c\left(\lambda_r+\lambda_s\right)E},\frac{2(a-\eta b)}{2b+c\lambda_k E}\right)=\max\left(\widetilde{T}_{kr},\widetilde{T}_k\right)<T_{rk} \qquad (14)$$

$$\max\left(\frac{2(a-\eta b)+cE\lambda_r T_{rs}}{2b+c\left(\lambda_r+\lambda_s\right)E},\frac{2(a-\eta b)}{2b+c\lambda_k E}\right)=\max\left(\widetilde{T}_{ks},\widetilde{T}_k\right)<T_{sk}. \qquad (15)$$

First of all we notice that, due to additional competition coming from the s-firms operating in the market in region r, also the condition for a one directional link from region k to region r is more restrictive. Second, from the above conditions, it follows that the trade distance between r and s is the shortest, that is, $T_{rs} < \min(T_{rk}, T_{sk})$. This can be proven considering that, as shown before, $T_{rs} < T_{sk}$. Moreover, consider that (1) $T_{rs} < \widetilde{T}_{kr}$ from the condition $T_{rs} \leq \widetilde{T}_r$ in (13); and (2) $\widetilde{T}_{kr} < T_{rk}$ from (14). Putting together (1) and (2), we have that $T_{rs} < T_{rk}$ as well.

The price indexes when two-way trade from r to s occurs are:

$$P_r=\frac{a[2-(\lambda_r+\lambda_s)]+[\eta(\lambda_r+\lambda_s)+T_{sr}\lambda_s](b+cE)}{2b+c(\lambda_r+\lambda_s)E}E$$

$$P_s=\frac{a[2-(\lambda_r+\lambda_s)]+[\eta(\lambda_r+\lambda_s)+T_{sr}\lambda_r](b+cE)}{2b+c(\lambda_r+\lambda_s)E}E$$

$$P_k=\frac{a(2-\lambda_k)+\eta\lambda_k(b+cE)}{2b+c\lambda_k E}E$$

and the equilibrium prices and quantities:

$$p_{rr}=\frac{a+\eta[b+c(\lambda_r+\lambda_s)E]+\frac{T_{rs}}{2}c\lambda_s E}{2b+c(\lambda_r+\lambda_s)E} \qquad q_{rr}=(b+cE)(p_{rr}-\eta)$$

$$p_{ss}=\frac{a+\eta[b+c(\lambda_r+\lambda_s)E]+\frac{T_{rs}}{2}c\lambda_r E}{2b+c(\lambda_r+\lambda_s)E} \qquad q_{ss}=(b+cE)(p_{ss}-\eta)$$

$$p_{kk}=\frac{a+\eta(b+c\lambda_k E)}{2b+c\lambda_k E} \qquad q_{kk}=(b+cE)(p_{kk}-\eta)$$

$$p_{rs}=p_{ss}+\frac{T_{rs}}{2} \qquad q_{rs}=(b+cE)(p_{rs}-\eta-T_{rs})$$

$$p_{rk}=2p_{kk}-\eta \qquad q_{rk}=0$$

$$p_{sr}=p_{rr}+\frac{T_{rs}}{2} \qquad q_{sr}=(b+cE)(p_{sr}-\eta-T_{rs})$$

$$p_{sk}=2p_{kk}-\eta \qquad q_{sk}=0$$

$$p_{kr}=2p_{rr}-\eta \qquad q_{kr}=0$$

$$p_{ks}=2p_{ss}-\eta \qquad q_{ks}=0$$

As these equations show, the effect of the link creation from s to r affects p_{rr}: compared with the previous case, a firm located in r applies a lower price in the local market and faces competition not only from local firms but also from those located in region s.

The equilibrium profits for the case of two-way trade between r and s are:

$$\pi_r = (b + cE)\left[(p_{rr} - \eta)^2 \left(\frac{L}{3} + \lambda_r E\right) + (p_{rs} - \eta - T_{rs})^2 \left(\frac{L}{3} + \lambda_s E\right)\right]$$

$$\pi_s = (b + cE)\left[(p_{ss} - \eta)^2 \left(\frac{L}{3} + \lambda_s E\right) + (p_{sr} - \eta - T_{rs})^2 \left(\frac{L}{3} + \lambda_r E\right)\right]$$

$$\pi_k = (b + cE)(p_{kk} - \eta)^2 \left(\frac{L}{3} + \lambda_k E\right).$$

The effect of the creation of a "link back" from region s to region r, generating two-way trade between the two regions, has an ambiguous impact on region r due the counterbalancing of the two effects mentioned above: The first effect, on profit, is negative, due to the additional competition in the local market coming from the s-firms; whereas the second, on the consumer's surplus, is positive due to the larger availability of manufactured commodities for the r-consumers (the consumers living in region r). The impact on region s, instead, is positive due to the additional profits accruing to the s-firms from the market located in r.

4.4 Case (3) One of the Regions Trade with the Other Two but the Other Two Do Not Trade with Each Other

This third case includes six possible network structures collected in two groups. To the first group, composed of three cases, belong those structures characterized by the existence of one-directional links only, that is, by one-way trade flows only; to the second group belong those structures characterized by one or two bidirectional links. In what follows, we assume that region r always trades with the other two, but these, regions s and k, do not trade with each other. The conditions determining the first group of network structures are listed below:

3.A.1 One-way trade from region r to region s and from region r to region k—corresponding to triad 4 in Fig. 7—occurs as long as:

$$\frac{2(a - \eta b)}{2b + c\lambda_r E} = \widetilde{T}_r < T_{rs} \leq \widetilde{T}_s = \frac{2(a - \eta b)}{2b + c\lambda_s E}$$

$$\frac{2(a - \eta b)}{2b + c\lambda_r E} = \widetilde{T}_r < T_{rk} \leq \widetilde{T}_k = \frac{2(a - \eta b)}{2b + c\lambda_k E}.$$

3.A.2 One-way trade from region r to region s and from region k to region r—corresponding to triad 6 in Fig. 7—occurs as long as:

$$\frac{2(a-\eta b)}{2b+c\lambda_r E}=\widetilde{T}_r<T_{rs}\leq\widetilde{T}_s=\frac{2(a-\eta b)}{2b+c\lambda_s E}$$
$$\frac{2(a-\eta b)}{2b+c\lambda_k E}=\widetilde{T}_k<T_{rk}\leq\widetilde{T}_r=\frac{2(a-\eta b)}{2b+c\lambda_r E}.$$

3.A.3. One-way trade from region s to region r and from region k to region r—corresponding to triad 5 in Fig. 7—occurs as long as:

$$\frac{2(a-\eta b)}{2b+c\lambda_s E}=\widetilde{T}_s<T_{rs}\leq\widetilde{T}_{sr}=\frac{2(a-\eta b)+cE\lambda_k T_{rk}}{2b+c\,(\lambda_r+\lambda_k)\,E}$$
$$\frac{2(a-\eta b)}{2b+c\lambda_k E}=\widetilde{T}_k<T_{rk}\leq\widetilde{T}_{kr}=\frac{2(a-\eta b)+cE\lambda_s T_{rs}}{2b+c\,(\lambda_r+\lambda_s)\,E}.$$

As before, the inequalities holding for cases 3.A.1–3.A.2 confirm that the region with the smaller (larger) share of entrepreneurs has a higher chance to be the importing (exporting) region. Looking at case 3.A.3, we note that the boundary condition for a one-directional link from region s (k) to region r takes into account the competition coming from the third region k (s).

An important remark is useful at this stage: there are many ways in which new links can be added to an existing network structure, taking into account all the isomorphic configurations. Therefore the way we are proceeding (the order we are using) in our analysis—adding one link after the other and looking at the consequences (for example on prices, profits and short-run welfare) of a new link—is not the only possible.

In what follows we study in detail the case 3.A.1 (leaving the analysis of cases 3.A.2 and 3.A.3 to another contribution).

The price indexes when one-way trade from r to s and from to r to k occurs are:

$$P_r=\frac{a(2-\lambda_r)+\eta\lambda_r(b+cE)}{2b+c\lambda_r E}E$$
$$P_s=\frac{a[2-(\lambda_r+\lambda_s)]+[\eta(\lambda_r+\lambda_s)+T_{rs}\lambda_r](b+cE)}{2b+c(\lambda_r+\lambda_s)E}E$$
$$P_k=\frac{a[2-(\lambda_r+\lambda_k)]+[\eta(\lambda_r+\lambda_k)+T_{rk}\lambda_r](b+cE)}{2b+c(\lambda_r+\lambda_k)E}E$$

and the equilibrium prices and quantities:

$$p_{rr} = \frac{a + \eta(b + c\lambda_r E)}{2b + c\lambda_r E} \qquad q_{rr} = (b + cE)(p_{rr} - \eta)$$

$$p_{ss} = \frac{a + \eta[b + c(\lambda_r + \lambda_s)E] + \frac{T_{rs}}{2} c\lambda_r E}{2b + c(\lambda_r + \lambda_s)E} \qquad q_{ss} = (b + cE)(p_{ss} - \eta)$$

$$p_{kk} = \frac{a + \eta[b + c(\lambda_r + \lambda_k)E] + \frac{T_{rk}}{2} c\lambda_r E}{2b + c(\lambda_r + \lambda_k)E} \qquad q_{kk} = (b + cE)(p_{kk} - \eta)$$

$$p_{rs} = p_{ss} + \frac{T_{rs}}{2} \qquad q_{rs} = (b + cE)(p_{rs} - \eta - T_{rs})$$

$$p_{rk} = p_{kk} + \frac{T_{rk}}{2} \qquad q_{rk} = (b + cE)(p_{rk} - \eta - T_{rk})$$

$$p_{sr} = 2p_{rr} - \eta \qquad q_{sr} = 0$$

$$p_{sk} = 2p_{kk} - \eta \qquad q_{sk} = 0$$

$$p_{kr} = 2p_{rr} - \eta \qquad q_{kr} = 0$$

$$p_{ks} = 2p_{ss} - \eta \qquad q_{ks} = 0.$$

As we have seen before, looking at the price applied by firms located in region k, the creation of a one directional link from r to s reduces p_{kk} because now local firms located in region k have to face additional competition from the firms located in r.

The equilibrium profits for the case of one-way trade from r to s and from to r to k are:

$$\begin{aligned}
\pi_r = \quad & (p_{rr} - \eta)^2 \left(\frac{L}{3} + \lambda_r E\right)(b + cE) \\
+ \quad & (p_{rs} - \eta - T_{rs})^2 \left(\frac{L}{3} + \lambda_s E\right)(b + cE) \\
+ \quad & (p_{rk} - \eta - T_{rk})^2 \left(\frac{L}{3} + \lambda_k E\right)(b + cE) \\
\pi_s = \; & (b + cE)(p_{ss} - \eta)^2 \left(\frac{L}{3} + \lambda_s E\right) \\
\pi_k = \; & (b + cE)(p_{kk} - \eta)^2 \left(\frac{L}{3} + \lambda_k E\right).
\end{aligned}$$

Concerning the welfare analysis, adding a one-directional link from r to k has a positive effect on r with an increase in profits for firms and in the surplus for consumers located in that region. As before, the effect of trade liberalization on k is ambiguous with an initial reduction in welfare (with respect to the case of autarky).

With further trade liberalization welfare increases and it may rise above the autarky level depending on parameter values and on the distribution of entrepreneurs.

Turning to the case in which one or two bidirectional links exist, three are the possible configurations. The corresponding "trade conditions" are reported below:

3.B.1. Two-way trade between r and s and one-way trade from r to k—corresponding to triad 8 in Fig. 7—occurs as long as:

$$T_{rs} \leq \min\left(\frac{2(a-\eta b)}{2b+c\lambda_r E}, \frac{2(a-\eta b)}{2b+c\lambda_s E}\right) = \min\left(\widetilde{T}_r, \widetilde{T}_s\right) \tag{16}$$

$$\frac{2(a-\eta b)+cE\lambda_s T_{rs}}{2b+c\left(\lambda_r+\lambda_s\right)E} = \widetilde{T}_{kr} < T_{rk} \leq \widetilde{T}_k = \frac{2(a-\eta b)}{2b+c\lambda_k E}. \tag{17}$$

Moreover, for future reference, we add the condition of "no trade" between s and k:

$$T_{sk} > \max\left(\frac{2(a-\eta b)+cE\lambda_r T_{rk}}{2b+c\left(\lambda_r+\lambda_k\right)E}, \frac{2(a-\eta b)+cE\lambda_r T_{rs}}{2b+c\left(\lambda_r+\lambda_s\right)E}\right) = \max\left(\widetilde{T}_{sk}, \widetilde{T}_{ks}\right). \tag{18}$$

3.B.2. Two-way trade between r and s and one-way trade from k to r—corresponding to triad 7 in Fig. 7—occurs as long as:

$$T_{rs} \leq \min\left(\frac{2(a-\eta b)}{2b+c\lambda_r E}, \frac{2(a-\eta b)+cE\lambda_k T_{rk}}{2b+c\left(\lambda_r+\lambda_k\right)E}\right) = \min\left(\widetilde{T}_r, \widetilde{T}_{sr}\right) \tag{19}$$

$$\frac{2(a-\eta b)}{2b+c\lambda_k E} = \widetilde{T}_k < T_{rk} \leq \widetilde{T}_{kr} = \frac{2(a-\eta b)+cE\lambda_s T_{rs}}{2b+c\left(\lambda_r+\lambda_s\right)E}. \tag{20}$$

Moreover, for future reference, we add the condition of "no trade" between s and k:

$$T_{sk} > \max\left(\frac{2(a-\eta b)}{2b+c\lambda_k E}, \frac{2(a-\eta b)+cE\lambda_r T_{rs}}{2b+c\left(\lambda_r+\lambda_s\right)E}\right) = \max\left(\widetilde{T}_k, \widetilde{T}_{ks}\right). \tag{21}$$

3.B.3. Two-way trade between r and s and between r and k—corresponding to triad 11 in Fig. 7—occurs as long as:

$$T_{rs} \leq \min\left(\frac{2(a-\eta b)}{2b+c\lambda_r E}, \frac{2(a-\eta b)+cE\lambda_k T_{rk}}{2b+c\left(\lambda_r+\lambda_k\right)E}\right) = \min\left(\widetilde{T}_r, \widetilde{T}_{sr}\right)$$

$$T_{rk} \leq \min\left(\frac{2(a-\eta b)+cE\lambda_s T_{rs}}{2b+c\left(\lambda_r+\lambda_s\right)E}, \frac{2(a-\eta b)}{2b+c\lambda_k E}\right) = \min\left(\widetilde{T}_{rk}, \widetilde{T}_k\right).$$

We study in some detail cases 3.B.1 and 3.B.3. Considering case 3.B.1, the price indexes when two-way trade between r to s and one-way trade from r to k occurs are:

$$P_r = \frac{a[2-(\lambda_r+\lambda_s)] + [\eta(\lambda_r+\lambda_s) + T_{rs}\lambda_s](b+cE)}{2b+c(\lambda_r+\lambda_s)E}E$$

$$P_s = \frac{a[2-(\lambda_r+\lambda_s)] + [\eta(\lambda_r+\lambda_s) + T_{rs}\lambda_r](b+cE)}{2b+c(\lambda_r+\lambda_s)E}E$$

$$P_k = \frac{a[2-(\lambda_r+\lambda_k)] + [\eta(\lambda_r+\lambda_k) + T_{rk}\lambda_r](b+cE)}{2b+c(\lambda_r+\lambda_k)E}E$$

and the equilibrium prices and quantities:

$$p_{rr} = \frac{a+\eta[b+c(\lambda_r+\lambda_s)E] + \frac{T_{rs}}{2}c\lambda_s E}{2b+c(\lambda_r+\lambda_s)E} \qquad q_{rr} = (b+cE)(p_{rr}-\eta)$$

$$p_{ss} = \frac{a+\eta[b+c(\lambda_r+\lambda_s)E] + \frac{T_{rs}}{2}c\lambda_r E}{2b+c(\lambda_r+\lambda_s)E} \qquad q_{ss} = (b+cE)(p_{ss}-\eta)$$

$$p_{kk} = \frac{a+\eta[b+c(\lambda_r+\lambda_k)E] + \frac{T_{rk}}{2}c\lambda_r E}{2b+c(\lambda_r+\lambda_k)E} \qquad q_{kk} = (b+cE)(p_{kk}-\eta)$$

$$p_{rs} = p_{ss} + \frac{T_{rs}}{2} \qquad q_{rs} = (b+cE)(p_{rs}-\eta-T_{rs})$$

$$p_{rk} = p_{kk} + \frac{T_{rk}}{2} \qquad q_{rk} = (b+cE)(p_{rk}-\eta-T_{rk})$$

$$p_{sr} = p_{rr} + \frac{T_{rs}}{2} \qquad q_{sr} = (b+cE)(p_{sr}-\eta-T_{rs})$$

$$p_{sk} = 2p_{kk} - \eta \qquad q_{sk} = 0$$

$$p_{kr} = 2p_{rr} - \eta \qquad q_{kr} = 0$$

$$p_{ks} = 2p_{ss} - \eta \qquad q_{ks} = 0$$

The equilibrium profits for the case of two-way trade from r to s and one-way trade from r to k are:

$$\begin{aligned} \pi_r = &\quad (p_{rr}-\eta)^2\left(\frac{L}{3}+\lambda_r E\right)(b+cE) \\ + &\quad (p_{rs}-\eta-T_{rs})^2\left(\frac{L}{3}+\lambda_s E\right)(b+cE) \end{aligned}$$

$$+ \quad (p_{rk} - \eta - T_{rk})^2 \left(\frac{L}{3} + \lambda_k E\right)(b + cE)$$

$$\pi_s = (b + cE)\left[(p_{ss} - \eta)^2 \left(\frac{L}{3} + \lambda_s E\right) + (p_{sr} - \eta - T_{rs})^2 \left(\frac{L}{3} + \lambda_r E\right)\right]$$

$$\pi_k = (b + cE)(p_{kk} - \eta)^2 \left(\frac{L}{3} + \lambda_k E\right).$$

The effect of creating a link from r (the exporting region) to k (the importing region) on these two regions—as trade costs are reduced below the threshold—are analogous to those highlighted previously. It is positive on r and ambiguous on k depending on the distribution of entrepreneurs and on parameter values; whereas the effect on the welfare of the third region—at least in the short-run—is nil.

Moving on to case 3.B.3, the price indexes when two-way trade between r and s and from r to k occurs are:

$$P_r = \frac{a + [\eta + T_{rs}\lambda_s + T_{rk}\lambda_k](b + cE)}{2b + cE}E$$

$$P_s = \frac{a[2 - (\lambda_r + \lambda_s)] + [\eta(\lambda_r + \lambda_s) + T_{rs}\lambda_r](b + cE)}{2b + c(\lambda_r + \lambda_s)E}E$$

$$P_k = \frac{a[2 - (\lambda_r + \lambda_k)] + [\eta(\lambda_r + \lambda_k) + T_{rk}\lambda_r](b + cE)}{2b + c(\lambda_r + \lambda_k)E}E$$

and the equilibrium prices and quantities:

$$p_{rr} = \frac{a + \eta(b + cE) + \left(\frac{T_{rs}}{2}\lambda_s + \frac{T_{rk}}{2}\lambda_k\right)cE}{2b + cE} \qquad q_{rr} = (b + cE)(p_{rr} - \eta)$$

$$p_{ss} = \frac{a + \eta[b + c(\lambda_r + \lambda_s)E] + \frac{T_{rs}}{2}c\lambda_r E}{2b + c(\lambda_r + \lambda_s)E} \qquad q_{ss} = (b + cE)(p_{ss} - \eta)$$

$$p_{kk} = \frac{a + \eta[b + c(\lambda_r + \lambda_k)E] + \frac{T_{rk}}{2}c\lambda_r E}{2b + c(\lambda_r + \lambda_k)E} \qquad q_{kk} = (b + cE)(p_{kk} - \eta)$$

$$p_{rs} = p_{ss} + \frac{T_{rs}}{2} \qquad q_{rs} = (b + cE)(p_{rs} - \eta - T_{rs}) \ldots$$

$$p_{rk} = p_{kk} + \frac{T_{rk}}{2} \qquad q_{rk} = (b + cE)(p_{rk} - \eta - T_{rk})$$

$$p_{sr} = p_{rr} + \frac{T_{rs}}{2} \qquad q_{sr} = (b + cE)(p_{sr} - \eta - T_{rs})$$

$$p_{sk} = 2p_{kk} - \eta \qquad q_{sk} = 0$$

$$p_{kr} = p_{rr} + \frac{T_{rk}}{2} \qquad q_{kr} = (b + cE)(p_{kr} - \eta - T_{rk})$$

$$p_{ks} = 2p_{ss} - \eta \qquad q_{ks} = 0.$$

The equilibrium profits for the case of two-way trade between r and s and from r to k are:

$$\begin{aligned}
\pi_r = \quad & (p_{rr} - \eta)^2 \left(\frac{L}{3} + \lambda_r E\right)(b + cE) \\
+ \quad & (p_{rs} - \eta - T_{rs})^2 \left(\frac{L}{3} + \lambda_s E\right)(b + cE) \\
+ \quad & (p_{rk} - \eta - T_{rk})^2 \left(\frac{L}{3} + \lambda_k E\right)(b + cE)
\end{aligned}$$

$$\pi_s = (b + cE)\left[(p_{ss} - \eta)^2 \left(\frac{L}{3} + \lambda_s E\right) + (p_{sr} - \eta - T_{rs})^2 \left(\frac{L}{3} + \lambda_r E\right)\right]$$

$$\pi_k = (b + cE)\left[(p_{kk} - \eta)^2 \left(\frac{L}{3} + \lambda_k E\right) + (p_{sk} - \eta - T_{rk})^2 \left(\frac{L}{3} + \lambda_k E\right)\right].$$

Comparing the first expression with the previous case, the effect of adding a link from k to r on the profits of region r is negative; whereas the effect on the overall welfare for region r is difficult to assess due to the counterbalancing of the negative effect on profits and the positive effect on the consumer's surplus. The effect on region s' welfare is negative: due to the negative impact on profits; whereas the effect on consumer's surplus is zero for that region. The opposite holds for region k, the overall effect on welfare is positive: this is due to the positive impact on profit; whereas also for this region the impact on the consumer's surplus is nil.

4.5 Case (4) One-Way Trade or Two-Way Trade is Present Between Any Two Region

This fourth case includes seven possible network structures that can be divided into four groups: (4.A) the first group is composed of two structures characterized by the existence of only one-directional links; (4.B) the second of three structures characterized by the existence of one bidirectional link and two one-directional links; (4.C) the third of a single structure characterized by the existence of two bidirectional links and one directional link; (4.D) and the fourth of the single structure characterized by all bidirectional links. The conditions determining these network structures are reported below:

4.A. All regions are involved in one-way trade:
4.A.1 One way trade from r to s, from r to k and from s to k—corresponding to triad 9 in Fig. 7—occurs as long as:

$$\frac{2(a-\eta b)}{2b+c\lambda_r E} = \widetilde{T}_r < T_{rs} \leq \widetilde{T}_s = \frac{2(a-\eta b)}{2b+c\lambda_s E}$$
$$\frac{2(a-\eta b)}{2b+c\lambda_r E} = \widetilde{T}_r < T_{rk} \leq \widetilde{T}_{rk} = \frac{2(a-\eta b)+cE\lambda_s T_{rs}}{2b+c(\lambda_s+\lambda_k)E}$$
$$\frac{2(a-\eta b)}{2b+c\lambda_s E} = \widetilde{T}_s < T_{sk} \leq \widetilde{T}_{sk} = \frac{2(a-\eta b)+cE\lambda_r T_{rs}}{2b+c(\lambda_r+\lambda_k)E}.$$

4.A.2 One way trade from r to s, from k to r and from s to k—corresponding to triad 10 in Fig. 7—occurs as long as:

$$\frac{2(a-\eta b)}{2b+c\lambda_r E} = \widetilde{T}_r < T_{rs} \leq \widetilde{T}_s = \frac{2(a-\eta b)}{2b+c\lambda_s E}$$
$$\frac{2(a-\eta b)}{2b+c\lambda_k E} = \widetilde{T}_k < T_{rk} \leq \widetilde{T}_r = \frac{2(a-\eta b)}{2b+c\lambda_r E}$$
$$\frac{2(a-\eta b)}{2b+c\lambda_s E} = \widetilde{T}_s < T_{sk} \leq \widetilde{T}_k = \frac{2(a-\eta b)}{2b+c\lambda_k E}.$$

4.B. Two regions are involved in two-way trade with each other and in one-way trade with the third region:
4.B.1 Two-way trade between r and s, one-way trade from r to k and from s to k—corresponding to triad 13 in Fig. 7—occurs as long as:

$$T_{rs} < \min\left(\frac{2(a-\eta b)}{2b+c\lambda_r E}, \frac{2(a-\eta b)}{2b+c\lambda_s E}\right) = \min\left(\widetilde{T}_r, \widetilde{T}_s\right)$$

$$\frac{2(a-\eta b)}{2b+c\lambda_r E} = \widetilde{T}_r < T_{rk} \leq \widetilde{T}_{rk} = \frac{2(a-\eta b)+cE\lambda_s T_{rs}}{2b+c(\lambda_s+\lambda_k)E}$$
$$\frac{2(a-\eta b)}{2b+c\lambda_s E} = \widetilde{T}_s < T_{sk} \leq \widetilde{T}_{sk} = \frac{2(a-\eta b)+cE\lambda_r T_{rs}}{2b+c(\lambda_r+\lambda_k)E}.$$

4.B.2 Region r and s are involved in two-way trade between each other; one way trade from region k to region r and from region s to region k. This structure—corresponding to triad 14 in Fig. 7—occurs as long as:

$$T_{rs} \leq \min\left(\frac{2(a-\eta b)+cE\lambda_k T_{rk}}{2b+c(\lambda_r+\lambda_k)E}, \frac{2(a-\eta b)}{2b+c\lambda_s E}\right) = \min\left(\widetilde{T_{sr}}, \widetilde{T}_s\right)$$

$$\frac{2(a-\eta b)+cE\lambda_s T_{rs}}{2b+c(\lambda_s+\lambda_k)E} = \widetilde{T}_{rk} < T_{rk} \leq \widetilde{T}_{kr} = \frac{2(a-\eta b)+cE\lambda_s T_{rk}}{2b+c(\lambda_r+\lambda_s)E}$$

$$\frac{2(a-\eta b)+cE\lambda_r T_{rs}}{2b+c(\lambda_r+\lambda_s)E} = \widetilde{T}_{ks} \leq T_{sk} < \widetilde{T}_k = \frac{2(a-\eta b)}{2b+c\lambda_k E}.$$

4.B.3 Regions r and s are involved in two-way trade between each other; one way trade from region k to region r and from region k to region s. This structure—corresponding to triad 12 in Fig. 7—occurs as long as:

$$T_{rs} \leq \min\left(\frac{2(a-\eta b)+cE\lambda_k T_{rk}}{2b+c(\lambda_r+\lambda_k)E}, \frac{2(a-\eta b)+cE\lambda_k T_{sk}}{2b+c(\lambda_s+\lambda_k)E}\right) = \min\left(\widetilde{T}_{sr}, \widetilde{T}_{rs}\right)$$

$$\frac{2(a-\eta b)}{2b+c\lambda_k E} = \widetilde{T}_k < T_{rk} \leq \widetilde{T}_{kr} = \frac{2(a-\eta b)+cE\lambda_s T_{rs}}{2b+c(\lambda_r+\lambda_s)E}$$

$$\frac{2(a-\eta b)}{2b+c\lambda_k E} = \widetilde{T}_k < T_{rk} \leq \widetilde{T}_{ks} = \frac{2(a-\eta b)+cE\lambda_r T_{rs}}{2b+c(\lambda_r+\lambda_s)E}.$$

4.C. One-way trade between two regions that are both involved in two-way trade with the third region: Two way trade between r and s and r and k and one way trade from s to k—corresponding to triad 15 in Fig. 7—occurs as long as:

$$T_{rs} < \min\left(\frac{2(a-\eta b)+cE\lambda_k T_{rk}}{2b+c(\lambda_r+\lambda_k)E}, \frac{2(a-\eta b)}{2b+c\lambda_s E}\right) = \min\left(\widetilde{T}_{sr}, \widetilde{T}_s\right)$$

$$T_{rk} < \min\left(\frac{2(a-\eta b)+cE\lambda_s T_{rs}}{2b+c(\lambda_r+\lambda_s)E}, \frac{2(a-\eta b)+cE\lambda_s T_{rs}}{2b+c(\lambda_s+\lambda_k)E}\right) = \min\left(\widetilde{T}_{kr}, \widetilde{T}_{rk}\right)$$

$$\frac{2(a-\eta b)+cE\lambda_r T_{rs}}{2b+c(\lambda_r+\lambda_s)E} = \widetilde{T}_{ks} \leq T_{sk} < \widetilde{T}_{sk} = \frac{2(a-\eta b)+cE\lambda_r T_{rs}}{2b+c(\lambda_r+\lambda_k)E}.$$

4.D. All regions are involved in two-way trade: This structure—corresponding to triad 16 in Fig. 7—occurs as long as:

$$T_{rs} \leq \min\left(\frac{2(a-\eta b)+cE\lambda_k T_{rk}}{2b+c(\lambda_r+\lambda_k)E}, \frac{2(a-\eta b)+cE\lambda_k T_{sk}}{2b+c(\lambda_s+\lambda_k)E}\right) = \min\left(\widetilde{T}_{sr}, \widetilde{T}_{rs}\right)$$

$$T_{rk} \leq \min\left(\frac{2(a-\eta b)+cE\lambda_s T_{rs}}{2b+c(\lambda_r+\lambda_s)E}, \frac{2(a-\eta b)+cE\lambda_s T_{rs}}{2b+c(\lambda_s+\lambda_k)E}\right) = \min\left(\widetilde{T}_{kr}, \widetilde{T}_{rk}\right)$$

$$T_{sk} \leq \min\left(\frac{2(a-\eta b)+cE\lambda_r T_{rs}}{2b+c(\lambda_r+\lambda_s)E}, \frac{2(a-\eta b)+cE\lambda_r T_{rs}}{2b+c(\lambda_r+\lambda_k)E}\right) = \min\left(\widetilde{T}_{ks}, \widetilde{T}_{sk}\right).$$

Considering case 4.A.1, the price indexes are:

$$P_r = \frac{a(2-\lambda_r)+\eta\lambda_r(b+cE)}{2b+c\lambda_r E}E$$

$$P_s = \frac{a[2-(\lambda_r+\lambda_s)]+[\eta(\lambda_r+\lambda_s)+T_{rs}\lambda_r](b+cE)}{2b+c(\lambda_r+\lambda_s)E}E$$

$$P_k = \frac{a+(\eta+\lambda_r T_{rk}+\lambda_s T_{sk})(b+cE)}{2b+cE}E$$

and the equilibrium prices and quantities:

$$p_{rr} = \frac{a+\eta(b+c\lambda_r E)}{2b+c\lambda_r E} \qquad q_{rr} = (b+cE)(p_{rr}-\eta)$$

$$p_{ss} = \frac{a+\eta[b+c(\lambda_r+\lambda_s)E]+\frac{T_{rs}}{2}c\lambda_r E}{2b+c(\lambda_r+\lambda_s)E} \qquad q_{ss} = (b+cE)(p_{ss}-\eta)$$

$$p_{kk} = \frac{a+\eta(b+cE)+\left(\frac{T_{rk}}{2}\lambda_r+\frac{T_{sk}}{2}\lambda_s\right)cE}{2b+cE} \qquad q_{kk} = (b+cE)(p_{kk}-\eta)$$

$$p_{rs} = p_{ss}+\frac{T_{rs}}{2} \qquad q_{rs} = (b+cE)(p_{rs}-\eta-T_{rs})$$

$$p_{rk} = p_{kk}+\frac{T_{rk}}{2} \qquad q_{rk} = (b+cE)(p_{rk}-\eta-T_{rk})$$

$$p_{sr} = 2p_{rr}-\eta \qquad q_{sr} = 0$$

$$p_{sk} = p_{kk}+\frac{T_{sk}}{2} \qquad q_{sk} = (b+cE)(p_{sk}-\eta-T_{sk})$$

$$p_{kr} = 2p_{rr}-\eta \qquad q_{kr} = 0$$

$$p_{ks} = 2p_{ss}-\eta \qquad q_{ks} = 0.$$

The equilibrium profits for the case 4.A.1 are:

$$\pi_r = (p_{rr} - \eta)^2 \left(\frac{L}{3} + \lambda_r E\right)(b + cE)$$
$$+ (p_{rs} - \eta - T_{rs})^2 \left(\frac{L}{3} + \lambda_s E\right)(b + cE)$$
$$+ (p_{rk} - \eta - T_{rk})^2 \left(\frac{L}{3} + \lambda_k E\right)(b + cE)$$
$$\pi_s = (b + cE)\left[(p_{ss} - \eta)^2 \left(\frac{L}{3} + \lambda_s E\right) + (p_{sr} - \eta - T_{rs})^2 \left(\frac{L}{3} + \lambda_r E\right)\right]$$
$$\pi_k = (b + cE)(p_{kk} - \eta)^2 \left(\frac{L}{3} + \lambda_k E\right)$$

According to this configuration, a creation of a link from s to k has a negative effect on profits and on welfare for r (with no effect on consumer's surplus in this region). It has a positive effect on region s profits and on this region welfare (with no effect on consumer's surplus). Finally, the effect on k is ambiguous, since the negative effect on profit is counterbalanced by the positive effect on the consumer's surplus.

Considering case 4.B.1, the price indexes are:

$$P_r = \frac{a[2 - (\lambda_r + \lambda_s)] + [\eta(\lambda_r + \lambda_s) + T_{rs}\lambda_s](b + cE)}{2b + c(\lambda_r + \lambda_s)E} E$$
$$P_s = \frac{a[2 - (\lambda_r + \lambda_s)] + [\eta(\lambda_r + \lambda_s) + T_{rs}\lambda_r](b + cE)}{2b + c(\lambda_r + \lambda_s)E} E$$
$$P_k = \frac{a + (\eta + \lambda_r T_{rk} + \lambda_s T_{sk})(b + cE)}{2b + cE} E$$

and the equilibrium prices and quantities:

$$p_{rr} = \frac{a + \eta[b + c(\lambda_r + \lambda_s)E] + \frac{T_{rs}}{2} c\lambda_s E}{2b + c(\lambda_r + \lambda_s)E} \qquad q_{rr} = (b + cE)(p_{rr} - \eta)$$
$$p_{ss} = \frac{a + \eta[b + c(\lambda_r + \lambda_s)E] + \frac{T_{rs}}{2} c\lambda_r E}{2b + c(\lambda_r + \lambda_s)E} \qquad q_{ss} = (b + cE)(p_{ss} - \eta)$$
$$p_{kk} = \frac{a + \eta(b + cE) + \left(\frac{T_{rk}}{2}\lambda_r + \frac{T_{sk}}{2}\lambda_s\right) cE}{2b + cE} \qquad q_{kk} = (b + cE)(p_{kk} - \eta)$$
$$p_{rs} = p_{ss} + \frac{T_{rs}}{2} \qquad q_{rs} = (b + cE)(p_{rs} - \eta - T_{rs})$$

$$p_{rk} = p_{kk} + \frac{T_{rk}}{2} \qquad q_{rk} = (b + cE)(p_{rk} - \eta - T_{rk})$$

$$p_{sr} = p_{rr} + \frac{T_{rs}}{2} \qquad q_{sr} = (b + cE)(p_{sr} - \eta - T_{rs})$$

$$p_{sk} = p_{kk} + \frac{T_{sk}}{2} \qquad q_{sk} = (b + cE)(p_{sk} - \eta - T_{sk})$$

$$p_{kr} = 2p_{rr} - \eta \qquad q_{kr} = 0$$

$$p_{ks} = 2p_{ss} - \eta \qquad q_{ks} = 0.$$

The equilibrium profits for the case 4.B.1 are:

$$\begin{aligned}
\pi_r = &\quad (p_{rr} - \eta)^2 \left(\frac{L}{3} + \lambda_r E\right)(b + cE) \\
+ &\quad (p_{rs} - \eta - T_{rs})^2 \left(\frac{L}{3} + \lambda_s E\right)(b + cE) \\
+ &\quad (p_{rk} - \eta - T_{rk})^2 \left(\frac{L}{3} + \lambda_k E\right)(b + cE) \\
\pi_s = &\quad (p_{ss} - \eta)^2 \left(\frac{L}{3} + \lambda_s E\right)(b + cE) \\
+ &\quad (p_{sr} - \eta - T_{rs})^2 \left(\frac{L}{3} + \lambda_r E\right)(b + cE) \\
+ &\quad (p_{sk} - \eta - T_{sk})^2 \left(\frac{L}{3} + \lambda_k E\right)(b + cE)
\end{aligned}$$

$$\pi_k = (b + cE)(p_{kk} - \eta)^2 \left(\frac{L}{3} + \lambda_k E\right).$$

In this configuration, the creation of a link from s to r has a negative effect on profits for firms located in region r and a positive effect on the surplus of consumers located in that region, with an ambiguous overall effect on welfare. The effect on welfare of region s is positive due to the increase in profits, whereas the effect on the consumer' surplus is nil. The creation of such a link has no short-run effect on region k.

Considering case 4.C, the price indexes are:

$$P_r = \frac{a + (\eta + \lambda_s T_{rs} + \lambda_k T_{rk})(b + cE)}{2b + cE} E$$

$$P_s = \frac{a[2 - (\lambda_r + \lambda_s)] + [\eta(\lambda_r + \lambda_s) + T_{rs}\lambda_r](b + cE)}{2b + c(\lambda_r + \lambda_s)E} E$$

$$P_k = \frac{a + (\eta + \lambda_r T_{rk} + \lambda_s T_{sk})(b + cE)}{2b + cE} E$$

and the equilibrium prices and quantities:

$$p_{rr} = \frac{a + \eta(b + cE) + \left(\frac{T_{rs}}{2}\lambda_s + \frac{T_{rk}}{2}\lambda_k\right) cE}{2b + cE} \qquad q_{rr} = (b + cE)(p_{rr} - \eta)$$

$$p_{ss} = \frac{a + \eta[b + c(\lambda_r + \lambda_s)E] + \frac{T_{rs}}{2}c\lambda_r E}{2b + c(\lambda_r + \lambda_s)E} \qquad q_{ss} = (b + cE)(p_{ss} - \eta)$$

$$p_{kk} = \frac{a + \eta(b + cE) + \left(\frac{T_{rk}}{2}\lambda_r + \frac{T_{sk}}{2}\lambda_s\right) cE}{2b + cE} \qquad q_{kk} = (b + cE)(p_{kk} - \eta)$$

$$p_{rs} = p_{ss} + \frac{T_{rs}}{2} \qquad q_{rs} = (b + cE)(p_{rs} - \eta - T_{rs})$$

$$p_{rk} = p_{kk} + \frac{T_{rk}}{2} \qquad q_{rk} = (b + cE)(p_{rk} - \eta - T_{rk})$$

$$p_{sr} = p_{rr} + \frac{T_{rs}}{2} \qquad q_{sr} = (b + cE)(p_{sr} - \eta - T_{rs})$$

$$p_{sk} = p_{kk} + \frac{T_{sk}}{2} \qquad q_{sk} = (b + cE)(p_{sk} - \eta - T_{sk})$$

$$p_{kr} = p_{rr} + \frac{T_{rk}}{2} \qquad q_{kr} = (b + cE)(p_{kr} - \eta - T_{rk})$$

$$p_{ks} = 2p_{ss} - \eta \qquad q_{ks} = 0.$$

The equilibrium profits for the case 4.C are:

$$\begin{aligned}
\pi_r = \quad & (p_{rr} - \eta)^2 \left(\frac{L}{3} + \lambda_r E\right)(b + cE) \\
+ \quad & (p_{rs} - \eta - T_{rs})^2 \left(\frac{L}{3} + \lambda_s E\right)(b + cE) \\
+ \quad & (p_{rk} - \eta - T_{rk})^2 \left(\frac{L}{3} + \lambda_k E\right)(b + cE) \\
\pi_s = \quad & (p_{sr} - \eta - T_{rs})^2 \left(\frac{L}{3} + \lambda_r E\right)(b + cE) \\
+ \quad & (p_{ss} - \eta)^2 \left(\frac{L}{3} + \lambda_s E\right)(b + cE) \\
+ \quad & (p_{sk} - \eta - T_{sk})^2 \left(\frac{L}{3} + \lambda_k E\right)(b + cE)
\end{aligned}$$

$$\pi_k = \left[(p_{kr} - \eta - T_{rk})^2 \left(\frac{L}{3} + \lambda_r E\right) + (p_{kk} - \eta)^2 \left(\frac{L}{3} + \lambda_k E\right)\right](b + cE).$$

According to this configuration, compared to the case 4.B.1, the creation of a link from k to r determines for region r a reduction in profits; however, since the consumer's surplus in this region is increased, the overall effect on region r welfare is ambiguous. The effect on region s's welfare is negative due to the impact of stronger competition in r on the profits of s-firms accruing from the market in r. Finally, the effect on k is positive due to the profits accruing to k-firms from the market in r.

Considering case 4.D, the price indexes are:

$$P_r = \frac{a + (\eta + \lambda_s T_{rs} + \lambda_k T_{rk})(b + cE)}{2b + cE} E$$

$$P_s = \frac{a + (\eta + \lambda_r T_{rs} + \lambda_k T_{sk})(b + cE)}{2b + cE} E$$

$$P_k = \frac{a + (\eta + \lambda_r T_{rk} + \lambda_s T_{sk})(b + cE)}{2b + cE} E$$

and the equilibrium prices and quantities:

$$p_{rr} = \frac{a + \eta(b + cE) + \left(\frac{T_{rs}}{2}\lambda_s + \frac{T_{rk}}{2}\lambda_k\right) cE}{2b + cE} \qquad q_{rr} = (b + cE)(p_{rr} - \eta)$$

$$p_{ss} = \frac{a + \eta(b + cE) + \left(\frac{T_{rs}}{2}\lambda_r + \frac{T_{sk}}{2}\lambda_k\right) cE}{2b + cE} \qquad q_{ss} = (b + cE)(p_{ss} - \eta)$$

$$p_{kk} = \frac{a + \eta(b + cE) + \left(\frac{T_{rk}}{2}\lambda_r + \frac{T_{sk}}{2}\lambda_s\right) cE}{2b + cE} \qquad q_{kk} = (b + cE)(p_{kk} - \eta)$$

$$p_{rs} = p_{ss} + \frac{T_{rs}}{2} \qquad q_{rs} = (b + cE)(p_{rs} - \eta - T_{rs})$$

$$p_{rk} = p_{kk} + \frac{T_{rk}}{2} \qquad q_{rk} = (b + cE)(p_{rk} - \eta - T_{rk})$$

$$p_{sr} = p_{rr} + \frac{T_{rs}}{2} \qquad q_{sr} = (b + cE)(p_{sr} - \eta - T_{rs})$$

$$p_{sk} = p_{kk} + \frac{T_{sk}}{2} \qquad q_{sk} = (b + cE)(p_{sk} - \eta - T_{sk})$$

$$p_{kr} = p_{rr} + \frac{T_{rk}}{2} \qquad q_{kr} = (b + cE)(p_{kr} - \eta - T_{rk})$$

$$p_{ks} = p_{ss} + \frac{T_{sk}}{2} \qquad q_{ks} = (b + cE)(p_{ks} - \eta - T_{sk}).$$

The equilibrium profits for the case 4.D are:

$$
\begin{aligned}
\pi_r = \; & (p_{rr} - \eta)^2 \left(\frac{L}{3} + \lambda_r E \right) (b + cE) \\
+ \; & (p_{rs} - \eta - T_{rs})^2 \left(\frac{L}{3} + \lambda_s E \right) (b + cE) \\
+ \; & (p_{rk} - \eta - T_{rk})^2 \left(\frac{L}{3} + \lambda_k E \right) (b + cE) \\
\pi_s = \; & (p_{sr} - \eta - T_{rs})^2 \left(\frac{L}{3} + \lambda_r E \right) (b + cE) \\
+ \; & (p_{ss} - \eta)^2 \left(\frac{L}{3} + \lambda_s E \right) (b + cE) \\
+ \; & (p_{sk} - \eta - T_{sk})^2 \left(\frac{L}{3} + \lambda_k E \right) (b + cE) \\
\pi_k = \; & (p_{kr} - \eta - T_{rk})^2 \left(\frac{L}{3} + \lambda_r E \right) (b + cE) \\
+ \; & (p_{ks} - \eta - T_{sk})^2 \left(\frac{L}{3} + \lambda_s E \right) (b + cE) \\
+ \; & (p_{kk} - \eta)^2 \left(\frac{L}{3} + \lambda_k E \right) (b + cE).
\end{aligned}
$$

According to this configuration, the creation of a link from k to s has the following short-run effect on the three regions' welfare: The effect on region r's welfare is negative since the impact of stronger competition in s reduces the profits that are accruing to r-firms from the market in s. The effect on s is ambiguous: indeed the negative effect on profits is counterbalanced by the increase in the surplus of consumers living in that region. Finally, the effect on k is positive given to the profits accruing to the k-firms originating from the market in s.

4.6 From Theory Back to Triad Census: A Special Case as an Example

In the previous section, we derived conditions on the trade costs for the occurrence of different triad patterns. In this section, we illustrate how these conditions can be used to corroborate the empirically found regularities in the trade network as summarized in Table 2.

One striking result of the triad census was that triad 3 (Mutual edge) and triad 8 (Mutual edge + Out) are much more often found than triad 7 (Mutual edge + In), which means that when two regions establish a mutual trade relationship this fosters them to export to, more than to import from, a third region. Theory suggests that more trade links come into existence with lower trade costs. This leads to the following interpretation: If in a triplet of regions all trade costs are very high, no trade occurs. If in a triplet of regions, bilateral trade costs between regions r and s are lower, then mutual trade occurs between those two regions. If in a triplet of regions, trade costs between another pair of regions—say between r and k—are lower as well, then unilateral trade from region r to region k is much more often found (resulting in triad 8) than trade in the opposite direction (which would lead to triad 7). There is an economic rationale for this difference that relies on third country effects, i.e. on the network structure: exporting from region r to region k is comparatively easy, because in region k no third country competition is yet present (there is no trade from s to k). Trade in the opposite direction—i.e. exporting from region k to region r—is more difficult, since in region r also competing firms from the third country, i.e. from region s, are already present.

In order to focus on these network effects, we assume as simplification a uniform distribution of firms, i.e. $\lambda_r = \lambda_s = \lambda_k = \frac{1}{3}$. Then, the conditions for the occurrence of triad 8 (16), (17) and (18) reduce to

$$T_{rs} \leq \min\left(\frac{2(a-\eta b)}{2b+c\frac{1}{3}E}, \frac{2(a-\eta b)}{2b+c\frac{1}{3}E}\right)$$

$$\frac{2(a-\eta b)+cE\frac{1}{3}T_{rs}}{2b+c\frac{2}{3}E} < T_{rk} \leq \frac{2(a-\eta b)}{2b+c\frac{1}{3}E}$$

$$T_{sk} > \max\left(\frac{2(a-\eta b)+cE\frac{1}{3}T_{rk}}{2b+c\frac{2}{3}E}, \frac{2(a-\eta b)+cE\frac{1}{3}T_{rs}}{2b+c\frac{2}{3}E}\right).$$

The second equation can only hold if

$$\frac{2(a-\eta b)+cE\frac{1}{3}T_{rs}}{2b+c\frac{2}{3}E} < \frac{2(a-\eta b)}{2b+c\frac{1}{3}E}$$

which can be easily transformed into the first condition, that is also the condition (13) holding for triad 3, i.e. for bilateral trade. Therefore, if bilateral trade between region r and region s exists and if T_{rk} sufficiently falls, than exports from region r to region k may start—triad 8 may come into existence.

Instead, Triad 7 is characterized by the conditions (19), (20) and (21), that by setting $\lambda_r = \lambda_s = \lambda_k = \frac{1}{3}$ simplify to :

$$T_{rs} \leq \min\left(\frac{2(a-\eta b)}{2b+c\frac{1}{3}E}, \frac{2(a-\eta b)+cE\frac{1}{3}T_{rk}}{2b+c\frac{2}{3}E}\right)$$

$$\frac{2(a-\eta b)}{2b+c\frac{1}{3}E} = \widetilde{T}_k < T_{rk} \leq \widetilde{T}_{kr} = \frac{2(a-\eta b)+cE\frac{1}{3}T_{rs}}{2b+c\frac{2}{3}E}$$

$$T_{sk} > \max\left(\frac{2(a-\eta b)}{2b+c\frac{1}{3}E}, \frac{2(a-\eta b)+cE\frac{1}{3}T_{rs}}{2b+c\frac{2}{3}E}\right).$$

The second equation can only hold if

$$\frac{2(a-\eta b)}{2b+c\frac{1}{3}E} < \frac{2(a-\eta b)+cE\frac{1}{3}T_{rs}}{2b+c\frac{2}{3}E}$$

which can be transformed into

$$\frac{2(a-\eta b)}{\left(2b+c\frac{1}{3}E\right)} < T_{rs}$$

which contradicts the first equation. Therefore, triad 7 is not a possible outcome of a reduction in T_{rk}.

Summing up, for an equal distribution of firms, we can show that a reduction in T_{rk} may lead to triad 8, while triad 7 is not possible. The result might change for unequal distributions, but this analysis is left for further studies.

5 Final Remarks

Some stylized facts emerging from the network analysis of interregional trade flows have confirmed that the European economic integration is still largely incomplete: most regions do not trade with any other region, but they rather select their partners, and a relatively small number of regions play a central role in the network structure. Moreover, the triad census analysis has revealed that the large majority of triads in the EU regional trade is represented by empty-graph structures (autarky), followed by single and mutual edges. This suggests that regions select their partners engaging mostly in bilateral trade. In addition, when two regions, say *r* and *s*, establish a mutual trade relationship, a third region, say *k*, is more likely to participate as an importer rather than as an exporter.

In order to shed more light to the processes that shape the specific network structure, we used a three-region footloose entrepreneur model. This model stresses the fact that an integrated market, for example the one composed of regions *r* and *s*, is more difficult to access than a non-integrated market (for example that represented by region *k*) due to stronger competition. Therefore, for region *r* or *s* is easier to export towards *k* (with an outward link) than the other way round (for *k* to export towards *r* or *s*, with an inward link). Using the model, we derived explicit conditions on the bilateral trade cost for the occurrence of each of the 16 possible triads. For a special case, we exemplified how these conditions can be used to draw interferences on the network structure. Based on this evidence, it has been possible to envisage a specific sequence of links generation. This sequence starts from the case of full autarky, with no links; it proceeds to the creation of a one-directional link, for example from *r* to *s*; next, a bidirectional link between *r* and *s* comes into existence, then a further link towards a third region *k* and, finally, a second bidirectional link, for example between *r* and *k* is created. Implicit in this sequence, there is a corresponding reduction of trade costs, which characterizes the process of European integration. However, this can be assessed only by looking at time series data and we leave this to future work.

Acknowledgements This Chapter is based upon work from COST Action ISCH COST Action IS1104 "The EU in the new complex geography of economic systems: models, tools and policy evaluation", supported by COST (European Cooperation in Science and Technology) www.cost.eu. We also sincerely thank the IPTS (Institute for Prospective Technological Studies) for providing the data on interregional trade in Europe, and an anonymous referee for providing constructive comments and help in improving the contents of this chapter.

Appendix

See Table 3.

Table 3 Regional NUTS-2 classification

AT11	Burgenland	FI19	Lansi-Suomi	PL43	Lubuskie
AT12	Niederösterreich	FI1A	Pohjois-Suomi	PL51	Dolnoslaskie
AT13	Wien	FI20	Aland	PL52	Opolskie
AT21	Kärnten	FR10	Ile de France	PL61	Kujawsko-Pomorskie
AT22	Steiermark	FR21	Champagne-Ardenne	PL62	Warminsko-Mazurskie
AT31	Oberösterreich	FR22	Picardie	PL63	Pomorskie
AT32	Salzburg	FR23	Haute-Normandie	PT11	Norte
AT33	Tirol	FR24	Centre	PT15	Algarve
AT34	Vorarlberg	FR25	Basse-Normandie	PT16	Centro (PT)
BE10	Region de Bruxelles	FR26	Bourgogne	PT17	Lisboa
BE21	Prov. Antwerpen	FR30	Nord - Pas-de-Calais	PT18	Alentejo
BE22	Prov. Limburg (B)	FR41	Lorraine	PT20	Regio Autnoma dos Aores
BE23	Prov. Oost-Vlaanderen	FR42	Alsace	PT30	Regio Autnoma da Madeira
BE24	Prov. Vlaams Brabant	FR43	Franche-Comte	SE11	Stockholm
BE25	Prov. West-Vlaanderen	FR51	Pays de la Loire	SE12	ostra Mellansverige
BE31	Prov. Brabant Wallon	FR52	Bretagne	SE21	Sydsverige
BE32	Prov. Hainaut	FR53	Poitou-Charentes	SE22	Norra Mellansverige
BE33	Prov. Liege	FR61	Aquitaine	SE23	Mellersta Norrland
BE34	Prov. Luxembourg (B)	FR62	Midi-Pyrenees	SE31	ovre Norrland
BE35	Prov. Namur	FR63	Limousin	SE32	Smland med oarna
CZ01	Praha	FR71	Rhone-Alpes	SE33	Vstsverige
CZ02	Stredni Cechy	FR72	Auvergne	SK01	Bratislavsk kraj
CZ03	Jihozapad	FR81	Languedoc-Roussillon	SK02	Zapadne Slovensko
CZ04	Severozapad	FR82	Provence-Alpes-Cote d Azur	SK03	Stredne Slovensko
CZ05	Severovychod	FR83	Corse	SK04	Vchodne Slovensko
CZ06	Jihovychod	GR11	Anatoliki Makedonia Thraki	UKC1	Tees Valley and Durham

CZ07	Stredni Morava	GR12	Kentriki Makedonia	UKC2	Northumberland Tyne and Wear
CZ08	Moravskoslezko	GR13	Dytiki Makedonia	UKD1	Cumbria
DE11	Stuttgart	GR14	Thessalia	UKD2	Cheshire
DE12	Karlsruhe	GR21	Ipeiros	UKD3	Greater Manchester
DE13	Freiburg	GR22	Ionia Nisia	UKD4	Lancashire
DE14	Tübingen	GR23	Dytiki Ellada	UKD5	Merseyside
DE21	Oberbayern	GR24	Sterea Ellada	UKE1	East Riding, North Lincolnshire
DE22	Niederbayern	GR25	Peloponnisos	UKE2	North Yorkshire
DE23	Oberpfalz	GR30	Attiki	UKE3	South Yorkshire
DE24	Oberfranken	GR41	Voreio Aigaio	UKE4	West Yorkshire
DE25	Mittelfranken	GR42	Notio Aigaio	UKF1	Derbyshire and Nottinghamshire
DE26	Unterfranken	GR43	Kriti	UKF2	Leicestershire Rutland
DE27	Schwaben	HU10	Kozep-Magyarorszag	UKF3	Lincolnshire
DE30	Berlin	HU21	Kozep-Dunantul	UKG1	Herefordshire Worcestershire
DE41	Brandenburg - Nordost	HU22	Nyugat-Dunantul	UKG2	Shropshire and Staffordshire
DE42	Brandenburg - Südwest	HU23	Del-Dunantul	UKG3	West Midlands
DE50	Bremen	HU31	eszak-Magyarorszag	UKH1	East Anglia
DE60	Hamburg	HU32	eszak-Alfold	UKH2	Bedfordshire Hertfordshire
DE71	Darmstadt	HU33	Del-Alfold	UKH3	Essex
DE72	Giessen	IE01	Border Midlands	UKI1	Inner London
DE73	Kassel	IE02	Southern and Eastern	UKI2	Outer London
DE80	Mecklenburg-Vorpommern	ITC1	Piemonte	UKJ1	Berkshire Bucks Oxfordshire
DE91	Braunschweig	ITC2	Valle dAosta Vallee dAoste	UKJ2	Surrey East and West Sussex
DE92	Hannover	ITC3	Liguria	UKJ3	Hampshire and Isle of Wight
DE93	Lüneburg	ITC4	Lombardia	UKJ4	Kent
DE94	Weser-Ems	ITD1	Bolzano-Bozen	UKK1	Gloucestershire Wiltshire

(continued)

Table 3 (continued)

DEA1	Düsseldorf	ITD2	Provincia Autonoma Trento	UKK2	Dorset and Somerset
DEA2	Köln	ITD3	Veneto	UKK3	Cornwall and Isles of Scilly
DEA3	Münster	ITD4	Friuli-Venezia Giulia	UKK4	Devon
DEA4	Detmold	ITD5	Emilia-Romagna	UKL1	West Wales and The Valleys
DEA5	Arnsberg	ITE1	Toscana	UKL2	East Wales
DEB1	Koblenz	ITE2	Umbria	UKM2	North Eastern Scotland
DEB2	Trier	ITE3	Marche	UKM3	Eastern Scotland
DEB3	Rheinhessen-Pfalz	ITE4	Lazio	UKM5	South Western Scotland
DEC0	Saarland	ITF1	Abruzzo	UKM6	Highlands and Islands
DED1	Chemnitz	ITF2	Molise	UKN0	Northern Ireland
DED2	Dresden	ITF3	Campania	BG31	Severozapaden
DED3	Leipzig	ITF4	Puglia	BG32	Severen tsentralen
DEE0	Sachsen-Anhalt	ITF5	Basilicata	BG33	Severoiztochen
DEF0	Schleswig-Holstein	ITF6	Calabria	BG34	Yugoiztochen
DEG0	Thüringen	ITG1	Sicilia	BG41	Yugozapaden
DK01	Hovedstadsreg	ITG2	Sardegna	BG42	Yuzhen tsentralen
DK02	Ost for Storeblt	LT00	Lietuva	CY00	Kypros
DK03	Syddanmark	LU00	Luxembourg (Grand-D)	SI01	Vzhodna Slovenija
DK04	Midtjylland	LV00	Latvija	SI02	Zahodna Slovenija
DK05	Nordjylland	MT00	Malta	RO11	Nord-Vest
EE00	Eesti	NL11	Groningen	RO12	Centru
ES11	Galicia	NL12	Friesland	RO21	Nord-Est
ES12	Principado de Asturias	NL13	Drenthe	RO22	Sud-Est
ES13	Cantabria	NL21	Overijssel	RO31	Sud Muntenia
ES21	Pais Vasco	NL22	Gelderland	RO32	Bucureşti Ilfov
ES22	Foral de Navarra	NL23	Flevoland	RO41	Sud-Vest Oltenia

ES23	La Rioja	NL31	Utrecht	RO42	Vest
ES24	Aragon	NL32	Noord-Holland		
ES30	Comunidad de Madrid	NL33	Zuid-Holland		
ES41	Castilla y Leon	NL34	Zeeland		
ES42	Castilla-la Mancha	NL41	Noord-Brabant		
ES43	Extremadura	NL42	Limburg (NL)		
ES51	Cataluna	PL11	Ldzkie		
ES52	Comunidad Valenciana	PL12	Mazowieckie		
ES53	Illes Balears	PL21	Malopolskie		
ES61	Andalucia	PL22	Slaskie		
ES62	Region de Murcia	PL31	Lubelskie		
ES63	Ceuta (ES)	PL32	Podkarpackie		
ES64	Melilla (ES)	PL33	Swietokrzyskie		
ES70	Canarias (ES)	PL34	Podlaskie		
FI13	Ita-Suomi	PL41	Wielkopolskie		
FI18	Etela-Suomi	PL42	Zachodniopomorskie		

References

Ago T, Isono I, Tabuchi T (2006) Locational disadvantage of the hub. Ann Reg Sci 40(4):819–848

Anderson JE (2011) The gravity model. Ann Rev Econ 3(1):133–160

Behrens K (2004) Agglomeration without trade: how non-traded goods shape the space-economy. J Urban Econ 55(1):68–92

Behrens K (2005a) How endogenous asymmetries in interregional market access trigger regional divergence. Reg Sci Urban Econ 35(5):471–492

Behrens K (2005b) Market size and industry location: traded vs non-traded goods. J Urban Econ 58(1):24–44

Behrens K (2011) International integration and regional inequalities: how important is national infrastructure? Manch Sch 79(5):952–971

De Benedictis L, Tajoli L (2011) The world trade network. World Econ 34(8):1417–1454

De Benedictis L, Nenci S, Santoni G, Tajoli L, Vicarelli C (2014) Network analysis of world trade using the BACI-CEPII dataset. Glob Econ J 14(3–4):287–343

De Nooy W, Mrvar A, Batagelj V (2011) Exploratory social network analysis with Pajek. Cambridge University Press, Cambridge

Forslid R, Ottaviano GI (2003) An analytically solvable core-periphery model. J Econ Geogr 3(3):229–240

Freeman LC (1979) Centrality in social networks conceptual clarification. Soc Networks 1(3):215–239

Hanneman R, Riddle M (2005) Introduction to social network methods. University of California, Riverside, CA. http://faculty.ucr.edu/hanneman/nettext

Head K, Mayer T (2014) Gravity equations: workhorse, toolkit, and cookbook. In: Gopinath G, Helpman E, Rogoff K (eds) Handbook of international economics, vol 4. Elsevier, Amsterdam, pp 131–195

Holland PW, Leinhardt S (1970) A method for detecting structure in sociometric data. Am J Sociol 76:492–513

Jackson MO (2008) Social and economic networks. Princeton University Press, Princeton

Krugman PR (1991) Geography and trade. MIT Press, Cambridge, MA

Martin P, Rogers CA (1995) Industrial location and public infrastructure. J Int Econ 39(3):335–351

Melitz MJ (2003) The impact of trade on intra-industry reallocations and aggregate industry productivity. Econometrica 71(6):1695–1725

Melitz MJ, Ottaviano GI (2008) Market size, trade, and productivity. Rev Econ Stud 75(1):295–316

Newman ME (2003) The structure and function of complex networks. SIAM Rev 45(2):167–256

Newman ME (2006) Modularity and community structure in networks. Proc Natl Acad Sci 103(23):8577–8582

Okubo T, Picard PM, Thisse J-F (2014) On the impact of competition on trade and firm location. J Reg Sci 54(5):731–754

Ottaviano GI, Tabuchi T, Thisse J-F (2002) Agglomeration and trade revisited. Int Econ Rev 43:409–436

Thissen M, Diodato D, Van Oort F (2013a) Integrated regional Europe: European regional trade flows in 2000. PBL Netherlands Environmental Assessment Agency, The Hague

Thissen M, Van Oort F, Diodato D (2013b) Integration and convergence in regional Europe: European regional trade flows from 2000 to 2010. PBL Netherlands Environmental Assessment Agency, The Hague

Thissen M, Lankhuizen M, Jonkeren O (2015) Multi-regional trade data for Europe in 2010. Project for the Institute for Prospective Technological Studies. PBL Netherlands Environmental Assessment Agency, The Hague

Wasserman S, Faust K (1994) Social network analysis: methods and applications. Cambridge University Press, Cambridge

The Empirics of Macroeconomic Networks: A Critical Review

Giorgio Fagiolo

Abstract This chapter critically surveys the recent empirical literature applying complex-network techniques to the study of macroeconomic dynamics. We focus on three important macroeconomic networks: international trade, finance and migration/mobility. We discuss both the empirical evidence on the topological properties of these networks and econometric works that identify the impact of network properties on macroeconomic dynamics. Results indicate that a detailed knowledge of macroeconomic networks is necessary to better understand the dynamics of country income, growth and productivity, as well as the diffusion of crises.

Keywords Complex networks • Diffusion of economic shocks • International finance • International migration • International trade • Macroeconomic networks

1 Introduction

In the last two decades, the empirical and theoretical research on economic networks has boomed.[1] Economists have indeed become increasingly aware that the dynamics of economic systems may be strongly influenced by the patterns of interactions among their constituent units (e.g., firms, consumers, institutions, industries, countries). Understanding how the structure of social and economic interactions is shaped and evolves across time, and how it affects—and it is influenced by—economic dynamics, becomes therefore crucial in order to describe, predict and control fundamental economic phenomena such as, among others, country growth, economic development, and the diffusion of global crises.

[1] See Schweitzer et al. (2009), De Martí and Zenou (2009), Jackson (2010), Easley and Kleinberg (2010), and Jackson and Zenou (2015).

G. Fagiolo (✉)
Istituto di Economia, Scuola Superiore Sant'Anna, Piazza Martiri della Libertà, 33, 56127 Pisa, Italy
e-mail: giorgio.fagiolo@sssup.it

P. Commendatore et al. (eds.), *Complex Networks and Dynamics*, Lecture Notes in Economics and Mathematical Systems 683,
DOI 10.1007/978-3-319-40803-3_7

Using an admittedly coarse-grained criterion, existing studies addressing from a complex-network perspective the study of economic and social interactions can be classified in three main classes: (1) micro; (2) meso; and (3) macro. Micro studies address economic networks where nodes (i.e. vertices in the graph) are microeconomic agents, such as firms, banks, financial institutions and consumers.[2] Links in microeconomic networks may represent, depending on the context, buying/selling or borrowing/lending relationships, knowledge and information exchanges, and so on. Meso-economic networks deal instead with interactions among economic entities (nodes) located in-between the micro and the macro layer of the economy. These can be products, technologies and industries (Hidalgo et al. 2007; Acemoglu et al. 2012), connected by links assessing, e.g., their technological similarity or their input-output relations, both within countries and at the level of global value chains (Cerina et al. 2015).

Macroeconomic networks—the topic of this chapter—focus instead on interactions among world countries, which play the role of nodes in the graph. Links in macro-networks describe the ways in which world countries may interact. These range from international trade, financial/banking relations, and foreign direct investment (i.e. mergers and acquisitions or green-field investment) all the way to permanent cross-border human migration and temporary international mobility.[3]

The starting point of this literature is that the study of macroeconomic linkages from a complex-network perspective is important to understand macroeconomic dynamics. For example, macroeconomic linkages may be responsible in transmitting internationally economic fluctuations and other types of shocks occurring at the country level (Galvão et al. 2007). In presence of non-linear transmission mechanisms, understanding the topology of these interaction structure becomes crucial to predict how a shock hitting a certain country may be amplified and diffused to other regions of the world economy. Furthermore, the position of a country in the macroeconomic network at a certain point in time may impact the trajectories of its subsequent growth and development. For instance, the relative centrality and embeddedness of a country in the network of international financial relationships may act either as a shield against (or an amplifier of) shocks transmitted from other countries, thus influencing its subsequent economic performance.

In this chapter we shall critically survey some of the recent literature on macroeconomic networks. In particular, we will focus on three classes of bilateral international linkages, i.e. trade, finance and human migration/mobility. We will organize the discussion in such a way to answer two main questions, namely: (1) How do macroeconomic networks look like? (2) Can we employ the knowledge of the topology of macroeconomic networks to better understand and predict

[2] Cf., among others, May et al. (2008), Guerrero and Axtell (2013), and Saito et al. (2007).

[3] A parallel line of research, which we will not discuss here, has explored infrastructure networks connecting world countries, which facilitate how people and goods move across space and borders (e.g., air, cargo and maritime transportation networks; cf. Barrat et al. 2004; Hua and Zhu 2009; Kaluza et al. 2011; Woolley-Meza et al. 2011).

macroeconomic dynamics? In other words, we are not only interested in empirically characterizing the shape of macroeconomic networks, but also to use this information as predictor for the behavior of world countries in the macroeconomy.

This rest of this chapter is organized as follows. Section 2 sets the stage and formally defines macroeconomic networks using complex-network concepts. In Sect. 3, we discuss the existing empirical evidence on the structure of macroeconomic networks. Section 4 presents some examples dealing with the impact of network structure on macroeconomic dynamics. Finally, Sect. 5 concludes and sketches out some topics for future research.

2 Macroeconomic Networks

Macroeconomic networks are graph-based descriptions of bilateral linkages among pairs of world countries. More formally, at any given point in time t (e.g., a year), consider the graph where nodes are the elements of the set $C^t = \{1, 2, \ldots, N^t\}$ of world countries[4] and links between any pair of countries (i,j), $i,j \in C^t$, and $i \neq j$, represent an existing "type of interaction" between them (e.g., international trade, finance, migration/mobility). Links may be directed if one can in principle differentiate between the effect of i on j and that of j on i; and weighted if directed or undirected links may be associated to their intensity. In the most general terms, a macroeconomic network for a given interaction type ι is defined as a sequence of network snapshots:

$$MN(\iota)^t = \{C^t, W(\iota)^t\}, \quad t = 1, \ldots, T \tag{1}$$

where $W(\iota)^t$ is a weighted, possibly asymmetric, $N^t \times N^t$ matrix fully representing the structure of weighted (directed) links in place among world countries at time t for the interaction type ι, and T is the number of time periods which we have information about.[5]

This chapter mainly discusses three interaction types, which we will shortly describe in the following sections.

[4]Of course, the number of world countries to be considered in the analysis at time t may depend not only on those actually existing at t but also on data availability at that period of time.

[5]Due to data availability, misreporting and the presence of zero flows, C^t may also depend, in principle, on the type of interaction ι. In what follows, we will focus on cases when one restricts the analysis on the minimal set of countries present in all interaction layers, so that $C^t(\iota) = C^t$.

2.1 The International-Trade Network

In the second half of the last century, the volume and value generated by the exchange of goods and services across international borders (aka international trade) have boomed. During the "second wave of globalization"[6] the share of world trade to GDP has more than doubled, increasing from 25 to 60 %.[7] Such a spectacular trend has been achieved not only *intensively* (i.e., through increases of trade flows between countries already trading in the past), but also *extensively* (i.e., via newly created trade relationships). Indeed, according to the estimates in Felbermayr and Kohler (2006), about 40 % of world trade growth after 1950 came from newly-established bilateral trading relationships.

This process has generated an intricate web of trade linkages, which currently connects the great majority of world countries, channels a huge economic value, and facilitates cross-border technological diffusion (Keller 2004) and global human mobility (Egger et al. 2012). It is therefore of a paramount importance to understand its structure, as well as its socio-economic, political and geographical determinants.

Research addressing the properties of international trade from a complex-network perspective has flourished in the last years (Fagiolo et al. 2009).[8]

The object of analysis is the International-Trade Network (ITN), aka World Trade Web (WTW) or World Trade Network (WTN), which is the graph representation of bilateral trade flows among world countries across the years. In its simplest form, the ITN is binary and undirected, that is a link represent the existence of a positive trade relationship (import and/or export) between any two countries. Differentiating between the existence of import vs export relationships makes the graph directed. If any existing (directed or undirected) link is associated to the (deflated) value expressed in a common currency (e.g., USD) the ITN becomes a weighted graph. In the directed case, it is customary to weight each directed link with the value of exports (or imports). If the graph is undirected, links between countries (i, j) may typically represent total trade (i.e. the sum of imports and exports).

Data to study how the ITN is shaped and evolves are easily available, both at the aggregate level and at the commodity-specific one.[9] This chapter mostly deals with

---

[6]By "second wave of globalization" we mean the period from 1945 onwards, as opposed to the "first wave of globalization" (1800–1914). The two waves are separated by a slump in international trade occurred between the two world Wars (Baldwin and Martin 1999).

[7]World Development Indicators Online (WDI) database, see http://data.worldbank.org/data-catalog/world-development-indicators.

[8]Quantitative approaches to the study of trade in terms of networks have been pioneered in sociology and political sciences. For instance, the seminal paper by Snyder and Kick (1979) triggered a fruitful literature mostly aimed at testing some flavor of "world system" or "dependency" theories using social-network analysis techniques, see e.g. Nemeth and Smith (1985), Sacks et al. (2001), Breiger (1981), Smith and White (1992), Kim and Shin (2002) and Mahutga (2006), and the discussion in Fagiolo et al. (2010).

[9]See for example COMTRADE (comtrade.un.org), the BACI dataset at CEPII (cepii.fr) or data on trade in goods and services at UNCTAD (unctadstat.unctad.org). Additional data are available

aggregate representations of the ITN, i.e. where linkages describe aggregate export and import flows. However, data allow one to construct as many interaction layers of the network as data on trade about specific commodities become available. This permits to correlate the properties of different product-specific layers of the ITN to understand potential complementarities and substitutability (Barigozzi et al. 2010; Dalin et al. 2012; D'Odorico et al. 2014). We will also focus on approaches that do not discriminate between trade for goods and trade for services. Indeed, as discussed in De Benedictis et al. (2014), service trade is still poorly analyzed from a complex-network perspective, mostly because of a lack of reliable data that cover a large number of countries for a sufficiently long period of time.[10] Finally, due to space constraints, we will not be able to properly account for the vast literature on trade networks and international relations (Wilkinson 2002; Hafner-Burton et al. 2009), including conflicts and military alliances (Polachek 1980; Dorussen and Ward 2010; Kinne 2012; Jackson and Nei 2015).

2.2 *The International-Financial Network*

Despite its undeniable economic importance, merchandise trade represents only one of the many possible economic linkages existing between world countries. Another substantial role is played by financial relationships, which typically channel a much higher value than what merchandize trade does. For example, in 2012, the dollar value of world merchandise exports was close to US$18.5 trillion,[11] whereas total cross-border holdings of securities reached US$43.6 trillion.[12] It is therefore extremely important quantify and explore the network of such bilateral financial relationships, and possibly compare them with those of the ITN.[13]

---

from individual researchers, e.g. Andrew Rose (http://faculty.haas.berkeley.edu/arose/), Kristian Gleditsch (privatewww.essex.ac.uk/~ksg/), Robert Feenstra (cid.econ.ucdavis.edu), and Arvind Subramanian and Shang-Jin Wei (users.nber.org/~wei/data.html), among others. See also De Benedictis et al. (2014) for a tutorial-like presentation of the main properties of the ITN using the BACI dataset. See also the WIOD dataset (wiod.org), which provides time-series data of world input-output tables for 40 countries worldwide.

[10] See, however, Egger et al. (2016) for a recent attempt bridging trade in goods and trade in services.

[11] See wto.org/english/news_e/pres14_e/pr721_e.htm.

[12] To this figure, one should also add the value of total foreign-direct investments (FDIs) flows, which in 2012 reached US$ 1.4 trillion, cf. oecd.org.

[13] Due to space constraints and the focus on macroeconomic relations, we cannot survey here the extremely interesting and influential literature on micro and meso financial networks, especially the contributions addressing systemic risk (Cf. e.g. the work of Stefano Battiston and co-authors: See for instance Battiston et al. 2012a,b). The interested reader is referred to the reviews by Hasman (2013) and Chinazzi and Fagiolo (2013).

Financial data that can be used to build a network representing cross-border financial relationships among world countries is provided by the IMF in its Coordinated Portfolio Investment Survey (CPIS). Data include cross-border portfolio investment holdings of equity securities, long-term debt securities and short-term debt securities listed by country of residence of issuer. Overall, one has complete bilateral data for roughly 70 countries for the period 2001–2010 (Schiavo et al. 2010; Chinazzi et al. 2013), which can be employed to define a multi-graph representation of the International Financial Network (IFN), with three disaggregated layers and an aggregate one.

More precisely, existing data allow to build the IFN in five different cases: (1) all financial investments (Total Portfolio Investments, TPI); (2) equity securities (ES); (3) debt securities (TDS); (4) long-term debt securities (LTDS) and (5) short-term debt securities (STDS). More formally, one can build a 5-layer weighted-directed multigraph, where each directed link is weighted by the value of security—in millions of current dollars—issued by the origin node and held by the target. This involves aggregating first the debt layers (iv) and (v) to generate an aggregate debt layer; and then merging equity and debt to get the TPI layer. At any level of aggregation, the generic entry $w^t_{ij}(k)$ of the corresponding weight matrix $W^t(k)$ for layer k at time t represents the actual stock of assets k issued by country j, and held by country i at time t.

Notice that data used to build the IFN record year's end holdings of securities reported at the economy level, from the asset side (more reliable than liability side), like equity, long-term and short-term debt instruments, securities held as reserve assets and securities held by international organizations. Data do not record instead FDIs,[14] loans, holdings of domestic securities (issued and held by residents of the same country) and securities acquired under reverse repurchase agreements.

An additional source of network data that can be employed to study financial relations among countries comes from the Bank for International Settlements (BIS) locational statistics on exchange-rate adjusted changes in cross-border bank claims. Data record flows of financial capital channeled through the banking system in every country, and are well-suited for an analysis of geographical patterns in financial linkages across countries (Reyes and Minoiu 2011). Using these data, one may build a network representation of the global banking network (GBN), where weighted links describe estimates of flows, obtained as changes in cross-border banking stocks (aggregated at the country level) including loans, deposits, debt securities, and other bank assets.

[14]FDIs are another important channel of interaction between world countries. Typically a large part of cross-border FDIs are done in terms of direct M&A between firms of two countries. Using alternative data sources (e.g., Thomson Reuters Mergers and Acquisitions database, see http://thomsonreuters.com/en/products-services/financial/hedge-funds/mergers-and-acquisitions.html), one may build, for each given time window, an international M&A network where nodes are countries and (direct) links represent acquirer-target M&A operations (in number of in total value). Interesting issues that can be addressed with this network concern, e.g., the geographical localization of clusters of countries engaged in M&A activities and the persistency over time of investment flows.

2.3 The International Networks of Permanent Migration and Temporary Mobility

Beside commercial and financial transactions, world countries interact also through cross-border movement of people. If one considers legal permanent migration alone, existing statistics show an unprecedented level of cross-border flows in the last years, leading to an overall migrant world population of about 190 million in 2010.[15] Current estimates predict that in 2050 the population of migrants will achieve 405 million, more than twice the figure for 2010. Quantifying international migration in a globalized world becomes therefore crucial in order to provide policy makers with the right tools.

A network approach to international migration must however face the fact that finding detailed bilateral data is extremely difficult. Data problems are especially acute when compared to higher-frequency data on international trade and finance flows. Nevertheless, thanks to the combined efforts by the United Nations Population Division, the Statistics Division of the United Nations, the World Bank and the University of Sussex, a reliable source about bilateral international *permanent* migration compiled using the United Nations Global Migration Database has been made available to the community of researchers (Ozden et al. 2011).

Starting from about 3500 individual census and population register records from more than 230 destination countries and territories from across the globe, the final database comprises five origin-destination 226×226 matrices for each decade in the period 1960–2000. For each year $t = 1960, \ldots, 2000$, the generic element (i, j) of each matrix records the stock of migrants (corresponding to the last completed census round) originating in country i and present in destination j. One can therefore employ these five origin-destination 226×226 matrices W^t to build a time-sequence of weighted-directed networks describing bilateral migration stocks among $N = 226$ countries. Therefore, the International-Migration Network (IMN) at time $t = 1960, \ldots, 2000$ is defined by a weighted matrix whose generic element (i, j) represents the stock of migrants originated in country i and present at time t in country j.

Permanent (legal) migration does not of course account for all existing cross-border people movements. In addition to illegal migration, which is almost by definition not measurable, people move also temporarily across borders for business or leisure purposes. Data about temporary international human mobility do actually exist and allow for a complex-network analysis of the phenomenon. In particular, the World Tourism Organization (UNWTO, www2.unwto.org) collects data about arrivals and departures of people traveling to a different country with respect to their usual place of residence, and stay there for less than one consecutive year. Outbound data are based on incoming visitors registered by the destination country and encompass both leisure and professional travelers, excluding border and seasonal

[15] See http://publications.iom.int/bookstore/free/WMR_2010_ENGLISH.pdf.

workers as well as long term students. Data are available for $N = 213$ countries from 1995 onwards, and allow for the construction of the international human temporary mobility network (IHTMN). This is defined as the network characterized in each year t by the $N \times N$ weight matrix W^t, whose generic entry w^t_{ij} records the number of travellers who left from country i and arrived in country j during year t.

3 Empirical Evidence

This section surveys some of the empirical evidence on macroeconomic networks. In particular, we will discuss differences and similarities in the observed topology of the ITN, IFN and IMN. We shall begin describing research on the ITN, as historically most of the efforts have been initially addressed towards the exploration of the web of international trade.

3.1 Topological Properties of the ITN

Despite the ITN is a very dense graph as compared to other real-world networks (its density is close to $\frac{1}{2}$), from an international-trade perspective one is left with the puzzle that half of all possible bilateral relations are not exploited. In other words, most countries do not trade with all the others, but they rather select their partners. In the period 1950–2000, the ITN has shown a marked increase in the number of directed linkages and a (weak) positive trend in density (De Benedictis and Tajoli 2011; Garlaschelli and Loffredo 2005). This occurs irrespective of whether one factors in or not any increase in the number of countries in the sample, due e.g. to improvements in data collection or new-born countries. Therefore, trade globalization has not only increased the connections among countries that were already trading back in 1950, but it did so by embedding in the trade web the newcomers over the years, inducing a stronger trade integration.

The ITN is also a very heterogenous network. For example, the distribution of the number of export and import partners of each country (i.e., in-degree and out-degree in network jargon) has become more and more bimodal over the years, with a group of very tightly connected countries co-existing with another group holding a smaller number of inward and outward links, thus preventing one to talk of a representative country in terms of trade patterns. Furthermore, the distribution of country imports, exports and total trade all follow log-normal densities (Fagiolo et al. 2008), implying that a few countries exporting and importing a lot exist side-by-side many countries characterized by very low trade levels.

Another relevant feature of the ITN is its disassortative nature: countries that hold many trade partners typically trade with countries holding a few links (Fagiolo et al. 2010). This is relatively less true from a weighted perspective: countries that import or export a lot tend to do so from and to countries characterized by low export and import levels, but there is a small number of very intensively connected countries trading with very similar partners.

Despite trade globalization, the ITN is still a strongly modular network. Due to geographic, economic and political reasons, countries have been forming over time relatively stable modular patterns of multilateral trade relations, possibly interacting among them, which can be easily identified through network analysis. A first interesting property is that countries that trade more tend to form intense trade triangles in their neighborhoods (i.e., clustering patterns, cf. Fagiolo 2007). This hints to the presence of a core of tightly connected countries in the ITN (Fan et al. 2014). Indeed, at least in year 2000, it turns out that the ten richest countries in terms of total trade are responsible of about 40 % of the total trade flows, a quite strong indication in favor of the existence of a rich club in the weighted ITN. More generally, community-detection techniques (Fortunato 2010) allow to identify several clusters of countries forming tightly-connected trade groups, each one relatively disconnected with respect to others (Barigozzi et al. 2011; Piccardi and Tajoli 2015). These groups tend to mimic geographical partitions of the world in macro areas but are less overlapping with existing preferential trade agreements (PTAs), confirming previous findings hinting to an ambiguous role of PTAs in explaining trade (Rose 2004). Despite communities of countries in the ITN are easy to identify, their statistical significance is still an open issue (Piccardi and Tajoli 2012). Indeed, inter-community linkages are far from being irrelevant, providing support for the ITN as a globalized trading system.

As mentioned above, the ITN has undergone some structural changes over the second decade of the last century. Trade globalization has occurred through intensive and extensive processes leading to denser but more bimodal network, with a stronger core. This does not mean, however, that the periphery of the network has become more and more marginal (De Benedictis and Tajoli 2011). Indeed, both the overall betweenness centralization of the network (Vega Redondo 2007) and the average path length between the countries (Albert and Barabási 2002) have been decreasing over time, meaning that hubs have become less important and countries formerly located in the periphery moved closer to the core, not necessarily through exclusive trade connections made with the hubs.

Notwithstanding trade globalization has induced structural changes in the ITN in the period 1950–2000, one can still learn from the past evolution of the network to project its future evolution. Indeed, as shown in Fagiolo et al. (2009), the Markovian nature of the ITN dynamics allows to predict its long-run state. Their analysis suggests that the architecture of the ITN will probably evolve towards a more polarized (Pareto) distribution for link weights (i.e., export flows), implying an increasingly large majority of links carrying moderate trade flows and a small bulk of very intense trade linkages.

3.2 *Finance, Migration and Trade*

We now discuss differences and similarities between the ITN and other two macroeconomic networks that we have introduced so far, namely the IFN and the IMN.

3.2.1 IFN vs. ITN

Comparing the IFN with the ITN can give interesting insights as to the degrees of integration of real vs. financial world markets. The existing contributions (see e.g. Schiavo et al. 2010) stress the fact that real markets are typically more integrated than financial ones, and that the international movement of financial assets tends to be mediated by a small number of financial centers. Indeed, the IFN is much less dense than the ITN. Furthermore, the vast majority of countries have a very large number of partners in the ITN, whereas the IFN has a more core-periphery structure, where an elite of countries connected with everybody else coexists with a second group of nodes characterized by average connectivity, and a peripheral group featuring poorly-connected countries. Another interesting difference between the ITN and the IFN concerns the heterogeneity of country portfolios of link weights. Results suggest that the intensity of financial links is less homogeneous than in the ITN. Once again, this is consistent with the fact that trade in financial assets is channelled through a few large financial centers, whereas trade for goods occurs more directly.

The IFN and the ITN share a strongly disassortative nature (as measured e.g. by the correlation between node degree or strength and node average nearest-neighbor degree or strength). In the case of the IFN this hints to the presence of financial centers intermediating a large fraction of trades in financial assets, or with the existence of benchmark securities entering almost every portfolio. The fact that disassortativity is much lower in the weighted case suggests that the bulk of capital flows occurs between a small subgroup of financial centers: since the connections between hubs and spokes are not very strong, the resulting correlation between node strength and average nearest-neighbor strength is likely to decrease.

Both networks exhibit a strong rich-club effect, especially when one explicitly considers link weights. Indeed, as mentioned, the top ten countries in terms of node strength account for more than 40 % of world trade in goods. This share grows to above 60 % in the case of the IFN. In general, countries belonging to the core appear to be those with higher per-capita GDP in both networks. Interestingly, these countries are also the most central in the network, e.g. according to measures of random-walk betweenness centrality (Fisher and Vega-Redondo 2006).

3.2.2 IMN vs. ITN

When comparing the IMN with the ITN over the period 1960–2000, several differences stand out (Fagiolo and Mastrorillo 2014). First, despite both networks are extremely dense, the ITN has gone through a steady density increase over the years, and became more dense than the IMN in 2000. As expected, the ITN is also more symmetric than the IMN, as testified, for instance, by the percentage of reciprocated directed links. This is because a trade channel is easier to reciprocate than a migration corridor. Second, as already noticed in Fagiolo and Mastrorillo (2013), the IMN features a much more marked small-world and modular structure, with average-path lengths smaller than in the ITN.

As far as weighted topology is concerned, a very strong and positive correlation is typically observed between ITN and IMN link weights: if any country *i* exports a higher trade value to country *j*, in *j* there is also a larger stock of migrants originated in *i*. This positive association, however, is far from being perfect, as the cloud of points describing ITN-IMN link weights displays a lot of noise. Nevertheless, such a variation can be explained by larger country economic/demographic sizes and smaller distances in a gravity-like fashion. This suggests that traditional country-level explanatory variables such as real GDP and population, as well as geographical distance, may drive much of the observed correlation in the two networks.

A positive correlation also emerges when one compares node-specific network statistics (e.g., node degree and strength, average nearest-neighbor degree and strength, node clustering coefficients, etc.) between the two networks. For example, if a country has more trade channels (respectively, trades more), it also carries more migration channels (respectively, holds larger immigrant/emigrant stocks). Again, it is easy to see that this positive relation is mostly explained by country demographic and economic size. Furthermore, countries trading with countries that either trade with many other partners or trade a lot are also connected to countries that hold a lot of migration channels or stocks, i.e. both average nearest-neighbor degree and strength are positively correlated in the two networks.

However, unlike what happens for degrees and strength, smaller levels of average nearest-neighbor degree and strength are associated to larger demographic and economic country sizes. This is because both networks display a marked (binary and weighted) disassortative behavior: the partners of more strongly connected nodes are weakly connected. However, larger countries (i.e. with higher levels of real GDP and population) also hold larger degrees and strengths. Therefore, countries with larger levels of average nearest-neighbor degree and strength are smaller, in both economic and demographic terms.

4 Impact on Macro-Economic Dynamics

In the previous section, we have discussed some empirical evidence related to the topological properties of macroeconomic networks describing country linkages concerning international trade, finance and migration. We now ask whether the structure of these networks can affect macroeconomic dynamics. More precisely, we are interested in investigating if the overall position and embeddedness of world countries in these networks, as well as their direct and indirect connectivity, can constrain and influence the processes going on *over* the networks. Such processes may include, for example, economic growth and development of countries and macro-regions, as well as diffusion of shocks that originate locally and possibly percolate globally.

As we shall see, the answers to these questions are in generally encouraging, indicating that networks matter in explaining macroeconomic dynamics. However, the identification of causal linkages going from network structure to dynamic processes over the network can be strongly limited by endogeneity issues. Indeed, network structure can affect macroeconomic dynamics, but the latter is likely to impact, in turn, the structure of the network over time. This conceptual issue poses several methodological hurdles to both theoretical and empirical research trying to single out the net effect of network structure on node behaviors.[16]

4.1 Diffusion of Shocks in the International Trade Network

Since international trade is one of the most important channels of interaction among world countries, and data are easily available at a sufficient level of commodity disaggregation for a long time span, the ITN has been often used as a testbed to understand how locally-originated shocks diffuse throughout the system.[17] The idea is very simple. Suppose that countries are connected via weighted trade links, as proxied by a time-snapshot of the ITN, and that a negative shock hits a given country. Assume a set of rules that govern the way in which this initial shock is possibly transmitted to the neighbors of the shocked country, to the neighbors of neighbors, and so on. By shocking one after the other all world countries, and observing each time how shock diffusion evolves, impact other countries, and possibly dies away, one may understand the relative importance of each country as a crisis propagator.

Following this intuition, Lee et al. (2011) study a simple dynamic model of shock diffusion over the ITN. In the model, countries are characterized by their capacity (proxied by their GDP). Every time a negative shock hits a country, all its incoming and outgoing link weights are decreased by a certain percentage. If the decrease in total country trade exceeds some fraction of its capacity, the shock is transmitted

[16]Another subtle and potentially important issue arising in dealing with econometric models involving networks is the existing interdependency between dyadic observations. This might bias results in e.g. gravity-like estimations due to the omissions of higher-level correlation between triads and, more generally, cliques; see, e.g., Ward et al. (2013).

[17]See also Foti et al. (2013). They study a simple model of diffusion where, after the system is shocked, a local rebalancing of supply and demand is assumed to occur in order to mitigate the effects of the shock.

to all its trade neighbors. This may initiate an avalanche of shocks, as also some of the neighbors can then transmit it to their neighbors. The process terminates when all countries hit by the shock do not transmit it to any other additional country. An interesting statistic describing the diffusion process is the number of countries that are eventually hit by an initial shock originated from a given country (call it "avalanche size"). Interestingly, the Authors show that there exist a certain range of model parameters that allow the avalanche-size distribution to become a power law (i.e., a Pareto distribution). This implies that countries play very heterogeneous roles in their ability to propagate local crises to the system, and there exists a small but not irrelevant number of countries that, once hit by a shock, are able to diffuse it worldwide. Big countries (in terms of GDP) tend to be the most disruptive, but this is not the end of the story. Indeed, the position of the country in the ITN and its local embeddedness in the web of indirect connections plays a very crucial role in explaining avalanche size. This is because the way in which countries may be hit by a shock and transmit it to their neighbors may be either direct or indirect. It is direct if the link with the neighbor that has transmitted the shock is so strong, as compared to its GDP, that the capacity threshold is exceeded right away. Conversely, the shock transmission may be indirect if, for example, country A withstands a first shock transmitted by neighbor B, but then it is hit by a second shock transmitted by neighbor C, who is also neighbor of B, which was hit by the shock transmitted by B, and did not withstand it, thus transmitting it to its neighbors, among which there is A. All countries belonging to any single avalanche can then be associated to a direct vs indirect chain of diffusion. By repeating this exercise for all major avalanches generated in the simulations, Lee et al. (2011) show that indirect patterns account for a very large percentage of chains of reaction. This confirms that second and third order effects in the ITN are crucial to understand how shocks propagate in the system (Abeysinghe and Forbes 2005).

4.2 Embeddedness in the IFN and Post-crisis Country Performance

The recent financial crisis has clearly stressed the potential problems arising from increasing financial market interconnectedness. However, the impact of higher degrees of connectivity on the players in a financial network is far from being straightforward. On the one hand, indeed, a more connected network may favor diffusion of small shocks and therefore be conducive to systemic crises. On the other hand, players that are more connected and central in the network may more easily dissipate the shocks that hit them thanks to a sort of portfolio-diversification effect. Furthermore, despite the probability of contagion is small when connectivity is high, the system-level consequences of defaults may be widespread and difficult to isolate (Gai and Kapadia 2010).

In order to understand the interplay between player connectivity and network embeddedness in the macroeconomic financial network, Chinazzi et al. (2013) have performed an econometric study to examine the ability of network-based measures to explain cross-country differences in the way countries in the IFN have been hit by the recent financial crisis. More specifically, two indicators of country "crisis intensity" are considered, one real (i.e., the 2009–2008 Change in real GDP) and the other financial (i.e., volatility-adjusted stock-market returns between Sep 15, 2008 and Mar 31, 2009). These measures, following the literature on early-warning systems (Lane and Milesi-Ferretti 2011), are regressed against a number of country controls (e.g., credit market regulation, real GDP per capita, bank credit to private sector over GDP, current account over GDP) and a set of network-based measures controlling for country position in the IFN, including node degree and strength, clustering coefficients and centrality indicators. The Authors perform two sets of regression exercises. In the first one, a cross-section specification is fitted to the data, where crisis measures (referring to the post-crisis period) are regressed against controls and network measures in year 2006. Despite the timing chosen for the cross-section regression, this exercise may still suffer form omitted variable biases and endogeneity issues. Therefore, a second set of regressions is performed, this time in a dynamic panel framework, using a Generalized Method of Moments (GMM) estimator to reduce endogeneity biases.

Overall, the results of these two sets of econometric exercises are consistent. To begin with, country network indicators exert a significant, nonlinear, and stable role in explaining both real and financial impact of the crisis on a country. Higher local connectivity seem to shield countries from severe impact via a risk diversification effect. However, a higher global embeddedness in the IFN (e.g., a higher binary clustering or centrality) exposes a country to a higher vulnerability, especially if the country is not within the rich-club of the IFN. This result also indicates that first (e.g., node degree) and higher (e.g., clustering or centrality) order network indicators are both important to fully characterize the position of a country in the network, and can offer interesting insights about the way local and global network properties interact in influencing node behavior.

4.3 Temporary Human Mobility and Country Income

Distinguishing between local and global network properties is very important to understand the effect that network topology can have on macroeconomic dynamics. In graphs characterized by a sufficient heterogeneity, e.g. in the way link weights are distributed across pairs of nodes, local node connectivity (e.g., measured by node degree or strength) and global node importance (e.g., measured by centrality indicators) can indeed strongly differ. For instance, a node that is not strongly connected locally, may be indeed linked with very globally important nodes in the network, thus becoming itself very important despite holding a few connections.

Conversely, very locally connected nodes may end up being not that central from a global point of view in the network.

From an econometric perspective, this means that global centrality indicators may increase the explanatory power of regressions where country characteristics like income, growth or productivity are described in terms of country-specific characteristics and local country connectivity in the network. This intuition is exploited in Fagiolo and Santoni (2015), who explore the network determinants of country per-capita income and labor productivity. Traditional explanations have stressed the importance of physical and human capital, the efficiency with which capital is used, and international technological diffusion. In particular, the latter is known to be enhanced by cross-border flows of trade, people and ideas. Therefore, net of trade openness and other factors, the level of integration of world countries in the international network of human mobility is a good candidate to explain country income and productivity. How can such an integration level be measured? Starting from temporary human mobility data (see Sect. 2.3), one can consider countries in the IHTMN and define two related set of integration indicators. The first one is simply country *mobility openness* in the network, i.e. the sum of arrivals and departures from and to a given country, divided by its population.[18] Mobility openness is a local network proxy for foreign technology exposure, as it considers only first-order links with direct partners, and has been shown to significantly explain the variation in country income and productivity, net of trade openness and other factors by Andersen and Dalgaard (2011). The second integration measure is a set of country *global centrality indicators* (i.e., eigenvector and Katz centrality) that assign to each country a score that is increasing in its overall relative connectivity with respect to the whole network. These are global integration measures insofar the importance of a country is defined in terms of how much it is connected with other countries that are themselves important, and so on. Therefore, country openness takes into account only a limited subset of all the information contained in the network, which is instead fully accounted for by global centrality indicators.

Including global centrality measures in regressions explaining country income and productivity—together with standard country controls, and trade/mobility openness—gives interesting insights. Indeed, once all potential endogeneity problems are dealt with, either with an instrumental-variable approach or via a GMM estimation, one finds that, net of country mobility openness, being more globally central in the IHTMN consistently induces higher income and productivity. This implies that the impact of human mobility in the international technological-diffusion process depends not only on how many direct partners a country has (and how strongly it is connected with them), but mostly on whether such a country is embedded in a web of relationships that connect her with other influential partners in the network.

[18]This parallels trade openness, defined as the sum of import and exports divided by country GDP.

4.4 International Migration and Trade

In the previous examples, we have discussed econometric frameworks wherein one can identify, net of possible endogeneity issues, the impact of first and higher order network properties on country-specific performance indicators. More generally, similar methodological techniques can be employed to single out the causal effect that the position of a country in a certain macroeconomic network may have on the behavior of countries in other macroeconomic networks.[19]

An interesting application of such an approach concerns the relationship between international migration and trade. Several studies, indeed, find quite a robust evidence suggesting that bilateral migration affects international-trade flows (Gaston and Nelson 2011; Egger et al. 2012). As argued in Gould (1994), for example, trade between any two countries (i, j) may be enhanced by the stock of immigrants present in either country and coming from the other one (m_{ji} and m_{ij}). This is because migrants originating in j and present in i (and vice versa) may foster imports of goods produced in their mother country (bilateral consumption-preference effect) or reduce import transaction costs thanks to their better knowledge of both home- and host-country laws, habits, and regulations. Again, such a *bilateral information effect* only takes into account the direct impact of migrants from either countries present in the other one to explain bilateral trade, i.e. a first-order effect. However, in line with the discussion in the previous section, one may posit that trade between any two countries can be fostered not only by bilateral-migration effects, but also thanks to migrants coming from other "third parties" and, more generally, by the overall connectivity and centrality of both countries in the IMN (Rauch 1999; Felbermayr et al. 2010; Felbermayr and Toubal 2012). This is because the better a pair of countries is connected in the IMN, the larger the average number of third countries that they share as origin of immigration flows and the more likely the presence of strong third-party migrant communities in both countries. This may further enhance trade via both preference and information effects. Moreover, it may happen that two countries are relatively well connected in the IMN (in both binary and weighted terms) even if they share a very limited number of non-overlapping third parties. In such a case, one may ask whether a cosmopolitan environment engendered by the presence of many ethnic groups in both countries can be trade enhancing—and if so why.

To test this idea, Fagiolo and Mastrorillo (2014) fit a battery of gravity models of trade where country centrality in the IMN is added as a further explanatory factor.[20] They find that pairs of countries that are more central in the IMN also trade more. This mainly occurs through a third-country effect: the more a pair of countries is

[19]Of course here causality is exogenously assumed by means of theoretical arguments, and not tested econometrically.

[20]See Sgrignoli et al. (2015) for a complementary analysis that explores similar issues using a product-specific trade perspective.

central in the IMN, the more they share immigrants coming from the same third-country, and the stronger the impact of forces related to consumption preferences and transaction-cost reduction. Furthermore, results suggest that also inward third-party migrants coming from corridors that are not shared by the two countries can be trade enhancing, in addition to common inward ones. This can be due to either learning processes of new consumption preferences by migrants whose origins are not shared by the two countries (e.g. facilitated by an open and cosmopolitan environment) or by the presence in both countries of second-generation migrants belonging to the same ethnic group.

5 Concluding Remarks

This chapter has surveyed some of the recent literature on macroeconomic networks, with particular emphasis on the networks of international trade, finance, permanent migration and temporary mobility. We have argued that describing interactions among world countries using a complex-network approach offers several empirical and theoretical insights. Overall, considering world countries as embedded in a complex web of relationships allows one to identify a wealth of additional and non-trivial empirical facts concerning the patterns of interactions at the macroeconomic level. Furthermore, econometric exercises show that these higher-order structures, and more generally the relative positions of countries in the networks, have substantial implications as to the dynamics of country performance and shock diffusion. In other words, macroeconomic networks do matter: direct and indirect connections among countries are indeed relevant to better understand macroeconomic dynamics.

Despite these very promising results, research on macroeconomic networks is still in its infancy and much remains to be done. A first important area that requires more efforts concerns the theory behind empirically-observed properties and econometric evidence. Indeed, theoretical models, possibly micro-founded, delivering as their (equilibrium) outcomes predictions about the topology of the networks should be developed and taken to the data, in order to validate the internal mechanisms proposed as explanations for the observed network regularities. Some effort in this direction has been made in the case of the ITN. Examples are the work on null statistical network models (Squartini et al. 2011a,b; Fronczak and Fronczak 2012) and stochastic models of trade network evolution (Riccaboni and Schiavo 2010), as well as the contributions by Fernando Vega-Redondo and co-authors on the dynamics of globalization (Dürnecker and Vega-Redondo 2012).

Another interesting avenue for further research is the integration of multi-layer network techniques (Kivelä et al. 2014) in the study of macroeconomic networks. Indeed, existing contributions have so far investigated the properties of different macroeconomic networks as they were independent from each other. In reality, world countries are connected at the same time through different types of linkages, including international trade, finance, investment, migration and mobility, infrastructures. From a complex-network perspective, considering all these interaction dimensions together means building a time-sequence of multi-layer networks where every time snapshot of the multi-layer is composed of a fixed number of nodes (i.e., countries) that may be connected by several different types of links, each representing a different interaction channel. Studying how multi-layer macroeconomic networks evolve over time would allow to better understand how different interaction channels correlate among them and cause each other, and eventually to dig deeper into the relationship between the role of a country in the global macroeconomic network and its economic performance.

Finally, a very promising line of research attempts to go beyond the spatial disaggregation of nodes in terms of countries by providing a finer level for the geographical breakdown of spatial units. For example, instead of building networks where nodes are countries, one may think, data permitting, to study macro networks where nodes are sub-national entities such as regions or other administrative units (see, for example, other chapters in this volume dealing with complex networks and geographical economics). If data about both intra-national and across-country links are available, such a perspective could greatly enhance our understanding of community structures and shock diffusion mechanisms.

References

Abeysinghe T, Forbes K (2005) Trade linkages and output-multiplier effects: a structural VAR approach with a focus on Asia. Rev Int Econ 13:356–375

Acemoglu D, Carvalho VM, Ozdaglar A, Tahbaz-Salehi A (2012) The network origins of aggregate fluctuations. Econometrica 80:1977–2016

Albert R, Barabási A-L (2002) Statistical mechanics of complex networks. Rev Mod Phys 74: 47–97

Andersen T, Dalgaard C-J (2011) Flows of people, flows of ideas, and the inequality of nations. J Econ Growth 16:1–32

Baldwin RE, Martin P (1999) Two waves of globalisation: superficial similarities, fundamental differences. NBER working papers 6904, National Bureau of Economic Research, Inc.

Barigozzi M, Fagiolo G, Garlaschelli D (2010) The multi-network of international trade: a commodity-specific analysis. Phys Rev E 81:046104

Barigozzi M, Fagiolo G, Mangioni G (2011) Identifying the community structure of the international-trade multi network. Physica A 390:2051–2066

Barrat A, Barthélemy M, Pastor-Satorras R, Vespignani A (2004) The architecture of complex weighted networks. Proc Natl Acad Sci 101:3747–3752

Battiston S, Gatti DD, Gallegati M, Greenwald B, Stiglitz JE (2012a) Liaisons dangereuses: increasing connectivity, risk sharing, and systemic risk. J Econ Dyn Control 36:1121–1141; quantifying and Understanding Dysfunctions in Financial Markets

Battiston S, Puliga M, Kaushik R, Tasca P, Caldarelli G (2012b) DebtRank: too central to fail? financial networks, the FED and systemic risk. Sci Rep 2:541

Breiger R (1981) Structure of economic interdependence among nations. In: Blau PM, Merton RK (eds) Continuities in structural inquiry. Sage, Newbury Park, CA, pp 353–380

Cerina F, Zhu Z, Chessa A, Riccaboni M (2015) World input-output network. PLoS ONE 10:e0134025

Chinazzi M, Fagiolo G (2013) Systemic risk, contagion, and financial networks: a survey. LEM working papers 2013/08, Scuola Superiore Sant'Anna, Pisa, Italy

Chinazzi M, Fagiolo G, Reyes J, Schiavo S (2013). Post-mortem examination of the international financial network. J Econ Dyn Control 37:1692–1713

Dalin C, Konar M, Hanasaki N, Rinaldo A, Rodriguez-Iturbe I (2012) Evolution of the global virtual water trade network. Proc Natl Acad Sci 109:5989–5994

De Benedictis L, Tajoli L (2011) The world trade network. World Econ 34:1417–1454

De Benedictis L, Nenci S, Santoni G, Tajoli L, Vicarelli C (2014) Network analysis of world trade using the BACI-CEPII dataset. Glob Econ J 14:57

De Martí J, Zenou Y (2009) Social networks. CEPR discussion papers 7599

D'Odorico P, Carr JA, Laio F, Ridolfi L, Vandoni S (2014) Feeding humanity through global food trade. Earth's Future 2:458–469, 2014EF000250

Dorussen H, Ward H (2010) Trade networks and the Kantian peace. J Peace Res 47:29–42

Dürnecker G, Vega-Redondo F (2012). Social networks and the process of globalization. Working papers, University of Mannheim, Department of Economics.

Easley D, Kleinberg J (2010) Networks, crowds, and markets: reasoning about a highly connected world. Cambridge University Press, New York, NY

Egger PH, von Ehrlich M, Nelson DR (2012) Migration and trade. World Econ 35:216–241

Egger PH, Francois J, Nelson DR (2016). The role of goods-trade networks for services-trade volume. World Econ; article first published online: 8 Nov 2015, doi:10.1111/twec.12331

Fagiolo G (2007) Clustering in complex directed networks. Phys Rev E 76:026107

Fagiolo G, Mastrorillo M (2013) International migration network: topology and modeling. Phys Rev E Stat Nonlinear Soft Matter Phys 88:012812

Fagiolo G, Mastrorillo M (2014) Does human migration affect international trade? A complex-network perspective. PLoS ONE 9:1–11

Fagiolo G, Santoni G (2015) Human-mobility networks, country income, and labor productivity. Netw Sci 3:377–407

Fagiolo G, Schiavo S, Reyes J (2008) On the topological properties of the world trade web: a weighted network analysis. Physica A 387:3868–3873

Fagiolo G, Schiavo S, Reyes J (2009) World-trade web: topological properties, dynamics, and evolution. Phys Rev E 79:036115

Fagiolo G, Schiavo S, Reyes J (2010) The evolution of the world trade web: a weighted-network approach. J Evol Econ 20:479–514

Fan Y, Ren S, Cai H, Cui X (2014) The state's role and position in international trade: a complex network perspective. Econ Model 39:71–81

Felbermayr GJ, Kohler W (2006) Exploring the intensive and extensive margins of world trade. Rev World Econ (Weltwirtschaftliches Archiv) 142:642–674

Felbermayr GJ, Toubal F (2012) Revisiting the trade-migration nexus: evidence from new OECD data. World Dev 40:928–937

Felbermayr GJ, Jung B, Toubal F (2010). Ethnic networks, information, and international trade: revisiting the evidence. Ann Econ Stat 10:41–70

Fisher E, Vega-Redondo F (2006). The linchpins of a modern economy. Working paper, Cal Poly

Fortunato S (2010) Community detection in graphs. Phys Rep 486:75–174

Foti NJ, Pauls S, Rockmore DN (2013) Stability of the world trade web over time - an extinction analysis. J Econ Dyn Control 37:1889–1910

Fronczak A, Fronczak P (2012) Statistical mechanics of the international trade network. Phys Rev E 85:056113

Gai P, Kapadia S (2010) Contagion in financial networks. Proc R Soc A Math Phys Eng Sci 466:2401–2423

Galvão AB, Artis M, Marcellino M (2007) The transmission mechanism in a changing world. J Appl Econ 22:39–61

Garlaschelli D, Loffredo M (2005) Structure and evolution of the world trade network. Physica A 355:138–144

Gaston N, Nelson D (2011) International migration. In: Bernhofen D, Falvey R, Greenaway D, Kreickemeier U (eds) Handbook of international trade. Palgrave, London

Gould DM (1994) Immigrant links to the home country: empirical implications for U.S. bilateral trade flows. Rev Econ Stat 76:302–316
Guerrero OA, Axtell RL (2013) Employment growth through labor flow networks. PLoS ONE 8:e60808
Hafner-Burton EM, Kahler M, Montgomery AH (2009) Network analysis for international relations. Int Organ 63:559–592
Hasman A (2013) A critical review of contagion risk in banking. J Econ Surv 27:978–995
Hidalgo CA, Klinger B, Barabási AL, Hausmann R (2007) The product space conditions the development of nations. Science 317:482–487
Hua Y, Zhu D (2009). Empirical analysis of the worldwide maritime transportation network. Phisica A 388:2061–2071
Jackson MO (2010) Social and economic networks. Princeton University Press, Princeton
Jackson MO, Nei S (2015) Networks of military alliances, wars, and international trade. Proc Natl Acad Sci 112:15277–15284
Jackson MO, Zenou Y (2015) Games on networks. In: Handbook of game theory with economic applications, vol 4, chapter 3. Elsevier, Amsterdam, pp 95–163
Kaluza P, Kölzsch A, Gastner M, Blasius B (2011). The complex network of global cargo ship movements. J R Soc Interface 388:1093–1103
Keller W (2004). International technology diffusion. J Econ Lit 42:752–782
Kim S, Shin E-H (2002) A longitudinal analysis of globalization and regionalization in international trade: a social network approach. Soc Forces 81:445–471
Kinne BJ (2012) Multilateral trade and militarized conflict: centrality, openness, and asymmetry in the global trade network. J Polit 74:308–322
Kivelä M, Arenas A, Barthelemy M, Gleeson JP, Moreno Y, Porter MA (2014) Multilayer networks. J Complex Networks 2:203–271
Lane PR, Milesi-Ferretti GM (2011) The cross-country incidence of the global crisis. IMF Econ Rev 59:77–110
Lee K-M, Yang J-S, Kim G, Lee J, Goh K-I, Kim I-M (2011) Impact of the topology of global macroeconomic network on the spreading of economic crises. PLoS ONE 6:e18443
Mahutga MC (2006) The persistence of structural inequality?: a network analysis of international trade, 1965–2000. Soc Forces 84:1863–1889
May RM, Levin SA, Sugihara G (2008) Complex systems: ecology for bankers. Nature 451:893–895
Nemeth R, Smith D (1985) International trade and world-system structure: a multiple network analysis. Review (J Fernand Braudel Center) 8:517–560
Ozden C, Parsons CR, Schiff M, Walmsley TL (2011) Where on earth is everybody? The evolution of global bilateral migration 1960–2000. World Bank Econ Rev 25:12–56
Piccardi C, Tajoli L (2012) Existence and significance of communities in the World Trade Web. Phys Rev E 85:066119
Piccardi C, Tajoli L (2015) Are preferential agreements significant for the world trade structure? A network community analysis. Kyklos 68:220–239
Polachek SW (1980) Conflict and trade. J Confl Resolut 24:55–78
Rauch JE (1999) Networks versus markets in international trade. J Int Econ 48:7–35
Reyes JA, Minoiu C (2011) A network analysis of global banking: 1978–2009. IMF working papers 11/74, International Monetary Fund
Riccaboni M, Schiavo S (2010) Structure and growth of weighted networks. New J Phys 12:023003
Rose AK (2004) Do we really know that the WTO increases trade? Am Econ Rev 94:98–114
Sacks M, Ventresca M, Uzzi B (2001) Global institutions and networks: contingent change in the structure of world trade advantage, 1965–1980. Am Behav Sci 44:1579–1601
Saito YU, Watanabe T, Iwamura M (2007) Do larger firms have more interfirm relationships? Physica A 383:158–163
Schiavo S, Reyes J, Fagiolo G (2010) International trade and financial integration: a weighted network analysis. Quant Finan 10:389–399

Schweitzer F, Fagiolo G, Sornette D, Vega-Redondo F, Vespignani A, White DR (2009) Economic networks: the new challenges. Science 325:422–425
Sgrignoli P, Metulini R, Schiavo S, Riccaboni M (2015) The relation between global migration and trade networks. Physica A 417:245–260
Smith D, White D (1992) Structure and dynamics of the global economy: network analysis of international trade, 1965–1980. Soc Forces 70:857–893
Snyder D, Kick E (1979) Structural position in the world system and economic growth 1955–1970: a multiple network analysis of transnational interactions. Am J Sociol 84:1096–1126
Squartini T, Fagiolo G, Garlaschelli D (2011a) Randomizing world trade. I: a binary network analysis. Phys Rev E 84:046117
Squartini T, Fagiolo G, Garlaschelli D (2011b) Randomizing world trade. II: a weighted network analysis. Phys Rev E 84:046118
Vega Redondo F (2007) Complex social networks. Cambridge University Press, New York
Ward MD, Ahlquist JS, Rozenas A (2013). Gravity's rainbow: a dynamic latent space model for the world trade network. Netw Sci 1:95–118
Wilkinson D (2002) Civilizations as networks: trade, war, diplomacy, and command-control. Complexity 8:82–86
Woolley-Meza O, Thiemann C, Grady D, Lee J, Seebens H, Blasius B, Brockmann D (2011). Complexity in human transportation networks: a comparative analysis of worldwide air transportation and global cargo-ship movements. Eur Phys J B 84:589–600

Bank Insolvencies, Priority Claims and Systemic Risk

Spiros Bougheas and Alan Kirman

Abstract We review an extensive literature debating the merits of alternative priority structures for banking liabilities put forward by financial economists, legal scholars and policymakers. Up to now, this work has focused exclusively on the relative advantages of each group of creditors to monitor the activities of bankers. We argue that systemic risk is another dimension that this discussion must include. The main message of our work is that when bank failures are contagious then when regulators assign priority rights need also to take into account how the bankruptcy resolution of one institution might affect the survival of other institutions that have acted as its creditors. When the network structure is fixed the solution is straightforward. Other banks should have priority to minimize the risk of their downfall. However, if the choice of policy can affect the structure of the network, policy design becomes more complex.This is a fruitful avenue for future research.

Keywords Banks • Priority rules • Systemic risk

JEL: G21, G28

1 Introduction

There is a hierarchy among a firm's creditors that is relevant when the firm becomes insolvent. The hierarchy reflects the allocation of priority rights among the creditors such that those higher in the hierarchy are paid in full before any other parties below

S. Bougheas (✉)
School of Economics, University of Nottingham, Nottingham, UK
e-mail: spiros.bougheas@nottingham.ac.uk

A. Kirman
GREQAM, EHESS, Aix-Marseille University, Marseille, France
e-mail: alan.kirman@univ-amu.fr

P. Commendatore et al. (eds.), *Complex Networks and Dynamics*,
Lecture Notes in Economics and Mathematical Systems 683,
DOI 10.1007/978-3-319-40803-3_8

receive any compensation.[1] This differential protection offered by the allocation of property rights has been designed to optimize the ability of the firm to raise funds from financial markets. Since the 2008 global financial crisis there has been a lot of interest in the design of bankruptcy resolution procedures and priority rules for banks.[2] What is striking is the large variety of both bankruptcy procedures (Berkovitch and Israel 1999) and priority rules applied across the globe (Lenihan 2012; Wood 2011). Some countries have had for some time some form of depositor preference rule (e.g. Australia, Switzerland and United States). Other countries have either only recently introduced or are in the process of introducing such rules. These include Greece, Portugal, Hungary, Latvia and Romania that have to implement such rules as part of the conditions that they need to meet in order to participate in EU/IMF programmes. In the UK the Vickers report recommends the introduction of a depositor preference rule (ICB Report 2011).

Most of the arguments offered for the support of proposals concerning priority rules are based on the incentives that these rules provide to depositors and other creditors to monitor the activities of bank managers. However, as Dewatripont and Freixas (2012) point out bankruptcy rules that might be optimal responses to individual bank failures might not be efficient when the crisis is systemic. In particular, they observe that adequate liquidity provision to solvent institutions might be sufficient to avert contagion throughout the system in the case of a single bank failure but not so during a systemic crisis. In the latter case, liquidity shortages and the depression of asset prices used as collateral (fire sales) might demand support for both solvent and insolvent institutions.

In this paper we review various arguments put forward by both economists and legal scholars supporting either existing or new proposals for priority rules in banking. Our main focus is on the relative positions on the ladder of depositors and other financial institutions linked through the interbank market. In our review we include both theoretical arguments and related empirical evidence. Reading this literature we were surprised by the absence of any arguments related to systemic risk issues. In the penultimate section of the paper we argue that the choice of priority rules can have considerable implications for the propagation of failures across the financial system.

2 Priority Rules in Practice

As we indicated above there are variations in bankruptcy procedures and rules applied around the globe. To focus the discussion we begin by taking a close look at one such priority structure, namely, that of US bank balance sheets as presented in Wood (2011).

[1]There is an extensive literature in financial economics that studies the optimal design of bankruptcy procedures; see von Thadden et al. (2010) for a recent review of the relevant literature.

[2]See Walter (2004) for a description of the actual process of bankruptcy resolution followed in US during the financial crisis.

1. Super-priority creditors (secured creditors)
   (a) creditors with security interests over collateral
   (b) sale and repurchase agreements (repos)
2. Priority creditors
   (a) retail depositors
   (b) life/pension insurance claimants
   (c) employee remuneration and benefits
   (d) unpaid taxes
3. Pari passu creditors
   (a) banks
   (b) bondholders
4. Subordinated creditors (tier structure)
   (a) senior subordinated
   (b) junior subordinated
   (c) preferred shares
5. Equity shareholders
6. Expropriated creditors
   (a) foreign currency creditors

Right at the top of the list (most senior instruments) we find contracts secured by collateral. During systemic events it is the collapse of the prices of the underlying assets pledged as collateral that dries up the liquidity of the financial system. Before the 2008 financial crisis many banks had pledged as collateral very similar assets created though the securitization of mortgages. One of the causes of the crisis has been the enhanced uncertainty that surrounded the valuation of these assets. As some institutions attempted to obtain liquidity by selling these assets, they drove their prices down, directly affecting the value of collateral pledged by other institutions. This led to further drops in prices (fire sales). This phenomenon has been extensively researched in recent years and also lies behind the Dewatripont and Freixas (2012) argument for a differential treatment of failing banks during a systemic crisis.[3] In case of insolvency, assets not pledged as collateral will be distributed to other creditor following the above seniority structure.

What is most relevant for our purposes is the relative positions of retail depositors and banks. The 'banks' entry in the above table mainly captures transactions in the interbank market (loans of durations from 1 day to 3 months). The interbank

[3]For a general analysis, see Shleifer and Vishny (1992). More recently, this work has been applied to banking to explain fire sales, market freezes, market spirals and related phenomena (see, for example, Acharya et al. 2011; Bebchuk and Goldstein 2011; Brunnermeier and Pedersen 2009; Caballero and Simsek 2013; Diamond and Rajan 2011) For a more thorough review of this literature, see Shleifer and Vishny (2010).

market provides the links that connect the banking network. The severity, in terms of aggregate losses, of a financial crisis depends on the exact structure of the network and the magnitude of initial losses. There is an extensive literature studying the structure of such a network and its implications for systemic risk.[4] While the relationship between connectedness and systemic risk is complex some general patterns have been identified: for example, for low values of initial losses a higher degree of connectedness is good news as the losses are spread out and thus the impact on any particular institution is minimized; in contrast, when initial losses are large a high degree of connectedness can be harmful as it increases the likelihood of multiple failures (see Acemoglu et al. 2015a).

The particular structure shown above reflects the enactment by the US Congress of the 1991 Federal Deposit Insurance Corporation Improvement Act that was followed by the 1993 Depositor Preference Act. Both acts were part of the policy response to the 1980s Savings and Loans crisis. The purpose for introducing the 1991 Act was to shift some of the risk of bank failures away from taxpayers and uninsured depositors and more to other creditors thus reducing the cost of federally provided insurance. Similar concerns led to the introduction of explicit rules in the Single Resolution Mechanism specifying protective measures for the depositor guarantee scheme.[5] Thus, deposits are senior to bonds and interbank market loans which, in turn, are senior to subordinated debt.[6] As we observed earlier, the above structure is not universal and the relative positions of uninsured depositors and other creditors varies from country to country.

There is a variety of both theoretical and informal arguments that have been advanced in support of various priority rules.

3 Theoretical Arguments

There is a long debate about whether uninsured depositors have the incentives to monitor the activities of banks.[7] Calomiris and Kahn (1991) have argued that by its very nature demandable debt (demand deposits), that allows depositors to withdraw their funds at will, offers the required market discipline device. As Diamond and Dybvig (1983) have shown the role of demand deposits is to provide insurance to depositors against idiosyncratic liquidity risk. More specifically, the contract offers

[4]For reviews of the literature see Allen and Babus (2009) and Bougheas and Kirman (2015a).

[5]See Regulation (EU) No 806/2014 of the European Parliament Council of July 2014. http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32014R0806.

[6]By tier structure we imply that the entries under subordinated debt are also ordered according to seniority.

[7]Beyond their effects on the incentives to monitor, changes in priority rules can have other consequences. Such changes would affect the prices of those claims whose priority has been affected, potentially changing their ownership and thus the entities affected in the case of bankruptcy (see Danisewicz et al. 2015).

risk-averse depositors flexibility with the timing of their withdrawals while at the same time allows banks to invest in long-term illiquid projects. However, inherent in the design is the possibility of a bank run where all depositors withdraw their funds at the same time. These runs are not only rational, given the beliefs that each depositor holds about the actions of other depositors, but can also be ex ante optimal (Allen and Gale 2007). That is the decision of depositors to trust their funds to banks can be ex ante efficient as long as the probability of runs is relatively small. Runs in the Diamond and Dybvig (1983) framework are pure sunspot phenomena. Put differently, they arise as because of coordination failures and it is not clear why in such environments depositors would be appropriate monitors. However, Jacklin and Bhattacharya (1988) allow the investment of banks to be risky and show that widespread runs can be generated by a small number of informed depositors who receive early signals about the bank's performance. It seems in that model informed depositors are performing the monitoring role.

Rochet and Tirole (1996) offer support for the argument that the most suitable monitors for banks are other banks and therefore interbank loans should be junior to deposits. They argue that interbank exposures generated through transactions in the interbank market provide strong incentives for banks to monitor other banks.[8] Clearly, the effectiveness of such incentives would depend on whether or not banks believe that the government will intervene in their favour during a crisis. If they believe that the government is likely to come to the rescue, of at least large institutions, then they might consider that some transactions in that market do not bear any risk. Since the 1998 global financial crisis a growing literature is attempting to address the vulnerability of financial systems to institutions that are 'Too-Big-To-Fail' (Kaufamn 2014).

Along similar lines, Birchler (2000) has argued in favor of depositor preference on the grounds that other creditors, like banks, have an informational advantage relative to a large number of small depositors.[9] Moreover, he argues that offering a standardized product to depositors with priority rights is a more efficient way of raising funds than having each depositor sign a bilateral contract with a bank. Therefore, his framework explains why the balance sheets of borrowers include a whole variety of debt instruments that differ according to their seniority status. The introduction of a priority list reduces the amount of resources devoted to socially inefficient information gathering. Such an arrangement it seems is ideal for banks that raise funds from a large number of uninformed investors.

While each of the above studies clearly supports either depositor or bank preference, Freixas et al. (2004) offer a more mixed view. In their model banks provide two types of services. They screen potential applicants thus improving the pool of loans that they offer and monitor firms that receive loans to ensure they

[8]Their argument bears some similarity to the one used for supporting the seniority of bank claims on the balance sheets of other firms (see Longhofer and Santos 2000).

[9]His work is an application to banking of earlier theoretical work on the role of seniority on corporate balance sheets (see, for example, Diamond 1993; Hart and Moore 1995).

perform well. Banks are subject to both liquidity and solvency shocks. The role of the interbank market is to redistribute funds from liquid to illiquid institutions, however, insolvent institutions cannot be prevented from using the market to gamble for resurrection. The optimal seniority status of interbank market loans depends on which of the two moral hazard problems associated with two services provided by banks is the most severe. When market discipline is weak then monitoring services become important. In this case, the only banks that seek funds from the interbank market are those that are illiquid and solvent banks and should not be penalized by excessive risk premia. Thus, it is optimal that interbank market loans are either secured or senior to other claims. In contrast, when the screening constraint binds then the interbank market loans cannot be secured and the premia must reflect the cost of insolvency.

The majority of studies that analyze the seniority structure of bank loans focus on the interbank market where loans are not secured. However, on the liability side of the balance sheets of banks we find other claims by financial institutions that are secured and therefore occupy the top step in the hierarchy ladder. Bolton and Oehmke (2015) analyze the seniority status of derivatives. They conclude that while these claims enhance value by providing risk management solutions, their seniority status can lead to inefficiencies as it transfers risk to other liabilityholders, such as depositors.

Lastly, there are also studies arguing that the most suitable monitors of bank activities are subordinated debtholders. The idea is that the market will provide the discipline required for reducing risk taking activities.[10] Theoretical work by Blum (2002) sheds some doubt about the efficacy of this policy. Requiring banks to hold some prespecified amount of subordinated debt may not prevent banks from pursuing high-risk activities and even worse might induce them to undertake even higher-risk activities. The reason is that protection by limited liability offers incentives to banks to decrease the cost of debt by increasing the amount of their borrowing as soon as the interest rate is fixed by the market. Thus there is a trade-off between the benefits derived from obtaining information about what banks do and the costs associated with the increase in balance sheet risk.

At this point we notice that, with the one exception the work by Rochet and Tirole (1996), research in this area does not directly address the issue of contagion.

4 Informal Arguments

Overall, the types of arguments that have been offered in favour of one priority rule over another follow the theoretical literature discussed above by advocating that the party most suited to monitor the activities of banks should be relatively low in the priority hierarchy. This particular debate has focused on five types of bank

[10] See Evanoff (1993) and Herring (2004) for support of this view.

creditors: insured depositors, uninsured depositors, international depositors, other banks and subordinated debtholders. However, there are many researchers and legal scholars who put more emphasis on the implementation of rules arguing that often preference rules have unintended consequences.

Among domestic depositors only those with uninsured claims have an incentive to monitor their banks.[11] Do they do it? The evidence is mixed. Jordan (2000) studying a sample of banks that failed in New England during the 1990s finds that uninsured depositors respond to bad news. At times these depositors not only react in a severe fashion but also start as early as 2 years before the bank is closed. The author concludes that the ability of banks to raise funds in the insured deposit market delays the closure of banks by dampening the effects due to the actions of uninsured depositors. A similar conclusion is reached by Billett et al. (1998) who analyzed announcements of credit rating changes for bank holding companies (BHCs) for the period January 1990 through December 1995. They find that banks increase their use of insured deposits after they have been downgraded by Moody's. Thus they conclude that an increase in the interest rate that they have to pay to attract uninsured deposits, or even the withdrawal of uninsured deposits, may not have a significant effect on banks' risk taking decisions.

Other scholars have warned about unintended consequences of depositor preference rules. For example, Kaufman (1997) criticizing the 1993 Depositor Preference Act observed that depositor priority rules can be circumvented by nonpreferred claimants who effectively become preferred claimants when the borrower secures their funding by offering them collateral.[12] Thomson (1994) and Marino and Bennett (1999) have argued that while the regulation seems to have worked with small bank failures it had unintended consequences with troubled larger institutions. Because the latter have a higher proportion of unsecured and international deposits, they faced a greater risk from the actions of those parties' national governments to protect them.[13]

Concerns have also been expressed more recently by Partnoy and Skeel (2007) and Perotti (2010) in response to bankruptcy privileges granted in 2005 in both the US and Europe to overnight secured credit and derivatives that have effectively allowed these lenders to claim priority over all other creditors in case of default. They assert that while such regulations reduce considerably the cost of borrowing at the same time they eliminate all the incentives the privileged creditors have to monitor the borrowers. In times of financial trouble these are the creditors who keep providing funds to stressed institutions exactly because the last minute loans that they offer are secured. Along the same lines Hirschhorn and Zervos (1990) argue that if a large enough proportion of nondepositor claims becomes secured, depositor preference could increase the cost of bank failures to the deposit insurance

[11] See Mantripragada (1992) for support of this view from a legal perspective.

[12] There is a similarity between this argument and the theoretical argument put forward by Bolton and Oehmke (2015) related to the role of derivatives.

[13] The 1993 Act placed international deposits very low on the priority ladder.

agency. Their empirical analysis indicates that depositor preference will lead to a considerable increase in collateralization thus taking away funds during a resolution that would have been available for distribution to depositors..

As we have already observed unintended consequences are also associated with proposals aiming to delegate the monitoring role to subordinated debtholders. The evidence comes primarily from comparing yields of subordinated bonds and the performance of the issuing banks and, once more, is mixed (see Evanoff and Wall 2002; Flannery and Sorescu 1996; Goyal 2005; Hancock and Kwast 2001; Sironi 2003).

Lastly, we turn our attention to the interbank market that is the main focus of our work. Evidence about the monitoring role played by creditors in this market comes from Furfine (2001) who collected every Fedwire funds transfer made during the first quarter of 1998. The main empirical findings of this study are: (a) banks with higher profitability, higher capital ratios, and fewer problem loans are charged lower interest rates on federal funds loans, and (b) larger institutions have an advantage as they pay lower interest rates on borrowed funds and charge higher interest rates on their loans. The evidence seems to suggest that banks can efficiently monitor other banks, however, there are also some potential problems. Firstly, the advantage of larger banks is consistent with the belief that these banks are 'Too-Big-To-Fail'. Secondly, the rates reflect only counterparty risk. Thirdly, the study was conducted during a calm period in financial markets. Taken together these three arguments raise concerns about the ability of banks to monitor themselves during systemic events.

5 Priority Claims and Systemic Risk

The literature on the optimal design of the priority structure of banking liabilities has exclusively focused on the incentives that alternative structures provide for risk taking and monitoring at the institutional level. As a consequence, the main arguments put forward are based at the relative abilities of various creditors to monitor the activities of bank managers. However, we argue that given the interconnectedness of the banking system, restricting the scope of the design at the institutional level might be potentially socially harmful. Cross-banking exposures through the interbank market imply that a failure of one institution can harm other directly linked institutions potentially leading to a cascade of failures throughout the system. In general, the level of systemic risk (potential aggregate losses) is not independent of the priority structure of bank liabilities.[14] In order to keep the argument as simple as possible, in what follows, we are going to ignore all other

[14]Our analysis might also be relevant for other sectors of the economy as interconnectedness is not an exclusive feature of the financial system. For example, Acemoglu et al. (2015b) study how interindustry input-output linkages can magnify small idiosyncratic shocks to produce macroeconomic tail risk. But this is attributed to the fact that they observe the emergence of a strongly skewed distribution of firm sizes. They argue as did Gabaix (2011) that a small shock to a

Table 1 Bank balance sheet

Assets	Liabilities
L^F: Loans to firms	D^H: Deposits by households
L^B: Loans to other banks	D^B: Deposits by other banks
R: Reserves	E: Equity

reasons for generating a priority structure mentioned above and concentrate on systemic risk. Therefore, we will concentrate on total losses ignoring their division between depositors and bank equityholders. In fact, from a welfare point of view we need to compare total losses. There might be strong arguments to protect depositors (this can be the case, for example, if the goal is to protect the intermediation process by ensuring that depositors trust their savings with the financial system) but in such cases there are other instruments (e.g. deposit insurance) that can be employed to address such objectives.

For our analysis we consider a network of banks linked through the interbank market. Table 1 show a typical bank's balance sheet.

The entries L^B and D^B correspond to the links of the network. As it turns out our main arguments do not depend on the exact structure of the interbank network. However, it is important to keep in mind that the sum of all the interbank loans across the banking system is equal to the sum of all deposits by other banks across the banking system. Equity is defined as $E \equiv L^F + L^B + R - D^H - D^B$. As long as $E \geqslant 0$ the bank is solvent. However, when $E < 0$ the value of the assets falls below the value of liabilities and the bank becomes insolvent. In the latter case, the bankruptcy procedure will decide the division of assets among the bank's liabilityholders. In particular, a bankruptcy procedure specifies rules to allocate the remaining assets to the failing bank's liabilityholders, in our case, other banks and depositors.[15] There are two broad rules that every bankruptcy procedure must satisfy:

Definition 1 (Priority Rules) They specify a hierarchy among creditors such that in liquidation a group of creditors must be satisfied in full before any other group of creditors lower in the ladder receive any payments.

Definition 2 (Pro-Rata Rule) All creditors belonging, according to priority rules, to the same level are compensated proportionately to the amount of their individual claims.

Given that our main interest is in understanding the relationship between priority rules and systemic losses, in the following discussion, we treat each group as a single agent.

large firm can produce major events. To do the same here would require considering also the size distribution of banks and their place in the network.

[15] Actual capital requirement regulations imply that there will be regulatory intervention as soon as equity falls below a prespecified threshold.

We consider a bank that has to write-off some of its loans to firms and we assume that these losses are higher than the value of its equity so that the bank becomes insolvent. We would like to figure out how alternative priority rules affect not only losses born by the liabilityholders of the failing bank but also by the liabilityholders of other affected banks. The structure of the interbank network will determine which banks will be affected but this will not have any effect on our results. As we explained above, in terms of social welfare, ultimately what matters are the total losses to depositors and equityholders throughout the banking system.

Dividing the failing bank's loans to other banks, L^B, among its liabilityholders is, in principle, straightforward. These loans represent deposits of the failed bank at other banks and they can be reallocated at full value. However, the allocation of loans to firms, L^F, and reserves, R, where the latter might include a variety of assets differing to their degree of market liquidity, might be problematic. As we explained above these two groups of assets might have to be liquidated at depressed market prices below corresponding book values (fire sales) further magnifying initial losses. Let l denote the fraction of the book value of assets recovered by liquidation (for simplicity we assume that is the same for all assets) and $V(> E)$ the value of loans written-off.

We are going to consider two cases. Firstly, we are going to analyze the model for the case when book and market values are the same. Put differently, we will ignore fire sales. For this case will show that priority rules do not matter. Then, we will introduce fire sales and show that the choice of priority rules can affect the magnitude of welfare losses due to systemic events.

5.1 No Fire Sales

We first consider the case when $l = 1$ (no fire sales). Then the total losses suffered by depositors and other banks is equal to $V - E$. The exact division of these losses between the two groups of liabilityholders will depend on the priority rule. Let $x \in [0, 1]$ denote the fraction of these losses born by depositors. Therefore, the total losses for the failing bank are equal to $E + x(V - E)$. The analysis of what happens with other affected banks who were creditors of the failing bank and those who were affected because of subsequent failures is greatly simplified by the existence of a unique clearing vector of payments that settles the obligations of all members of the banking system (see Acemoglu et al. 2015a; Eisenberg and Noe 2001). This important result implies that our main conclusions follow directly from what we know about the bank that failed originally. The total losses of the banks that were direct creditors to the failing bank due to this first round of liquidation are at most equal to $(1-x)(V-E)$. If the losses are equal to the last expression implies that this second round of liquidations was sufficient to absorb the losses. Clearly, the losses were born by their equityholders and maybe, depending on the priority rule, also by their depositors. If, in contrast, some of the losses were absorbed by other banks then the process is repeated. Notice that some banks that survived earlier rounds of liquidations might not do so in subsequent rounds. What it is clear is that at the end of the clearing process the total losses will be equal to V that is equal to the initial losses. Clearly, total losses are independent of the structure of the network and the priority rules of bankruptcy. However, the priority rules matter for the division of these losses between depositors and equityholders.

Proposition 1 *In the absence of fire sales neither the structure of the interbank network nor the priority rules matter for total losses. Priority rules matter for the division of losses between depositors and equityholders.*

5.2 Fire Sales

Next, we consider what happens when $l < 1$ (fire sales). In the following analysis we assume that the network structure is independent of the priority rule. Our only objective in this paper is to show that the design of priority rules has potentially serious implications for the magnitude of systemic losses. Nevertheless, a complete analysis needs to consider that the choice of priority rules might affect the formation of the interbank network.

For the moment we focus on the bankruptcy procedure of the initial failing bank ignoring any subsequent rounds.[16] The post-liquidation value of the failing bank's assets is equal to $L^B + l(R + L^F - V)$. Notice that the losses are equal to $(1-l)(R + L^F - V)$ and are decreasing in l. Clearly, the losses borne by other banks are greater under depositor priority. This matters for the value of total losses of the financial system because of fire sales. The higher the losses borne by banks the higher the probability that other banks will become insolvent and the higher the value of total losses given that liquidations are costly. Now, the structure of the financial network matters for the value of total losses, however, for a fixed network structure the value of total losses is higher under depositor priority. The following Proposition summarizes.

Proposition 2 *Suppose that the formation of the banking network is independent of the structure of priority rules. Then, when liquidation is costly (fire sales) the value of total losses under depositor priority is at least as high as the total losses under bank priority.*

The intuition is straightforward. When bank claims are senior to depositor claims the likelihood of further liquidations declines. When liquidation is costly (fire sales) the total losses of the banking system increase with the number of failing banks.

While the above analysis is too simplistic, as it ignores the incentives that priority structure offer for creating links in the network as well as the incentives that priority rules offer to different parties to monitor the bank's activities, we hope that it makes clear that ignoring systemic risk considerations when designing policy rules might be unwise.

---

[16]The existence and uniqueness of the clearing vector is not violated when $l < 1$, see Acemoglu et al. (2015a).

6 Conclusion

The recent global financial crisis has made it painfully clear how important the design of the regulatory framework, that encompasses both rules and institutions, is for reducing the economy-wide losses associated with systemic events.[17] We have argued that we need to consider carefully those rules that allocate priority rights among the various groups of bank creditors. There is an ongoing literature on this subject, however, it has mainly been concentrated on single bank resolutions rather than systemic events. The choice of priority rules can have a considerable effect on the total losses in the economy due to a systemic event. We have demonstrated how important this choice is for the simple case where the network structure is unaffected by the choice of priority rules. Future research should aim to explore this issue for the case when the interbank network is endogenous. It might be very well the case that when we allow for the choice of priority rues to affect the formation of links in the interbank market, our simple results above do not hold anymore. If different priority rules encourage or discourage certain entities from investing in the assets of certain others then rather than monitoring when faced with an enhanced risk, banks may prefer to invest elsewhere. While to analyse this might be a formidable task, given its significance for systemic risk policy design, cannot be ignored.[18]

References

Acemoglu D, Ozdaglar A, Tahbaz-Salehi A (2015a) Systemic risk and stability in financial networks. Am Econ Rev 105:564–608

Acemoglu D, Ozdaglar A, Tahbaz-Salehi A (2015b) Microeconomic origins of macroeconomic tail risks. NBER working paper No. 20865

Acharya V, Gale D, Yorulmazer T (2011) Rollover risk and market freezes. J Financ 66:1177–1209

Allen F, Babus A (2009) Networks in finance. In: Kleindorfer P, Wind Y, Gunther R (eds) The network challenge: strategy, profit, and risk in an interconnected world. Pearson Education, New Jersey, pp 367–382

Allen F, Gale D (2007) Understanding financial crises. Oxford University Press, New York

Bebchuk L, Goldstein I (2011) Self-fulfilling credit market freezes. Rev Financ Stud 24:3519–3555

Berkovitch E, Israel R (1999) Optimal bankruptcy laws across different economic systems. Rev Financ Stud 12:347–377

Billett M, Garfinkel J, O'Neal E (1998) The cost of market versus regulatory discipline in banking. J Financ Econ 48:333–358

Birchler U (2000) Bankruptcy priority for bank deposits: a contract theoretic explanation. Rev Financ Stud 13:813–840

Blum J (2002) Subordinated debt, market discipline, and banks' risk taking. J Bank Financ 26:1427–1441

[17]This has also implications for the design of monetary policy. The Federal reserve in its attempt to provide liquidity to troubled institutions had to implement a number of non-traditional policies with unknown long-term consequences (Cecchetti 2008).

[18]See Bougheas and Kirman (2015b) for some preliminary work in this direction.

Bolton P, Oehmke M (2015) Should derivatives be privileged in bankruptcy? J Financ 70:2353–2394
Bougheas S, Kirman A (2015a) Complex financial networks and systemic risk: a review. In: Commendatore P, Kayam S, Kubin I (eds) Complexity and geographical economics: topics and tools. Springer, Heidelberg
Bougheas S, Kirman A (2015b) Systemic risk and the optimal seniority structure of banking liabilities. CFCM working paper 15/15, University of Nottingham
Brunnermeier M, Pedersen L (2009) Market liquidity and funding liquidity. Rev Financ Stud 22:2201–2238
Caballero R, Simsek A (2013) Fire sales in a model of complexity. J Financ 68:2549–2587
Calomiris C, Kahn C (1991) The role of demandable debt in structuring optimal banking arrangements. Am Econ Rev 81:497–513
Cecchetti S (2008) Crisis and responses: the federal reserve and the financial crisis of 2007–2008. NBER working paper no. 14134
Danisewicz P, McGowan D, Onali E, Schaeck K (2015) Monitoring matters: debt seniority, market discipline and bank conduct. Available at SSRN: http://dx.doi.org/10.2139/ssrn.2550866
Dewatripont M, Freixas X (2012) Bank resolution: lessons from the crisis. In: Dewatripont M, Freixas X (eds) The crisis aftermath: new regulatory paradigms. Centre for Economic Policy Research, London, pp 105–143
Diamond D (1993) Seniority and maturity of debt contracts. J Financ Econ 33:341–368
Diamond D, Dybvig P (1983) Bank runs, deposit insurance, and liquidity. J Polit Econ 91:401–419
Diamond D, Rajan R (2011) Fear of fire sales, illiquidity seeking and credit freezes. Q J Econ 126:557–591
Eisenberg L, Noe T (2001) Systemic risk in financial systems. Manage Sci 47:236–249
Evanoff D (1993) Preferred sources of market discipline. Yale J Regul 10:447–467
Evanoff D, Wall L (2002) Measures of the riskiness of banking organizations: subordinated debt yields, risk-based capital, and examination ratings. J Bank Financ 26:989–1009
Flannery M, Sorescu S (1996) Evidence of bank market discipline in subordinated debenture yields: 1983–1991. J Financ 51:1347–1377
Freixas X, Rochet J-C, Parigi B (2004) The lender of last resort: a twenty-first century approach. J Eur Econ Assoc 2:1085–1115
Furfine C (2001) Banks as monitors of other banks: evidence from the overnight federal funds market. J Bus 74:33–57
Gabaix X (2011) The granular origins of aggregate fluctuations. Econometrica 79:733–772
Goyal V (2005) Market discipline of bank risk: evidence from subordinated debt contracts. J Financ Intermed 14:318–350
Hancock D, Kwast M (2001) Using subordinate debt to monitor bank holding companies: is it feasible? J Financ Serv Res 20:147–187
Hart O, Moore J (1995) Debt and seniority: an analysis of the role of hard claims in constraining management. Am Econ Rev 85:567–585
Herring R (2004) The subordinated debt alternative to Basel II. J Financ Stab 1:137–155
Hirschhorn E, Zervos D (1990) Policies to change the priority of claimants: the case of depositor preference laws. J Financ Serv Res 4:111–125
Jacklin C, Bhattacharya S (1988) Distinguishing panics and information-based bank runs: welfare and policy implications. J Polit Econ 96:568–592
Jordan J (2000) Depositor discipline at failing banks. N Engl Econ Rev (March/April):15–28
Kaufman G (1997) The new depositor preference act: time inconsistency in action. Manag Financ 23:56–61
Kaufamn G (2014) Too big to fail in banking: what does it mean? J Financ Stab 13:214–223
Lenihan N (2012) Claims of depositors, subordinated creditors, senior creditors and central banks in bank resolutions. Speech delivered at the Association Européenne pour le Droit Bancaire et Financière Conference, Athens, 5–6 October
Longhofer S, Santos J (2000) The importance of bank seniority for relationship lending. J Financ Intermed 9:57–89

Mantripragada K (1992) Depositor as a source of market discipline. Yale J Regul 9:543–574

Marino J, Bennett R (1999) The consequences of national depositor preference. FDIC Bank Rev 12:19–38

Partnoy F, Skeel D (2007) The promise and perils of credit derivatives. Univ Cincinnati Law Rev 75:1019–1051

Perotti E (2010) Systemic liquidity risk and bankruptcy exceptions. Duisenberg School of Finance Policy Paper No. 8

Rochet J-C, Tirole J (1996) Interbank lending and systemic risk. J Money Credit Bank 28:733–762

Shleifer A, Vishny R (1992) Liquidation values and debt capacity: a market equilibrium approach. J Financ 47:1343–1366

Shleifer A, Vishny R (2010) Fire sales in finance and macroeconomics. NBER working paper No. 16642

Sironi A (2003) Testing for market discipline in the European banking industry: evidence from subordinated debt issues. J Money Credit Bank 35:443–472

Thomson J (1994) The national depositor preference law. Federal Reserve Bank of Cleveland. Econ Commentary (15 February)

von Thadden E-L, Berglöf E, Roland G (2010) The design of corporate debt structure and bankruptcy. Rev Financ Stud 23:2648–2679

Walter J (2004) Closing troubled banks: how does the process work? Federal Reserve Bank of Richmond. Econ Q 90:51–68

Wood P (2011) The bankruptcy ladder of priorities and the inequalities of life. Hofstra Law Rev 40(1), Article 9

Complex Networks in Finance

Anna Maria D'Arcangelis and Giulia Rotundo

Abstract The present paper can be considered as divided in two parts: in the first one, we provide a review of the methods of complex networks that have been mainly used in the applications to the analysis of financial data. We focus on the following topics: the usage of the correlation matrix, systemic risk, integrated ownership and control, board of directors, interbank networks, and mutual funds holdings structure. The second part shows this last subject and provides new analyses.

The main findings outline that there are substantial differences in geographical allocation among the different European fund managers. Five larger European countries dominate the market of mutual funds. The belonging of UK and Swiss opt-outs of the eurozone could be a probable explanation for our results on community detection, that give a snapshot of a sort of "geographical organization" of the core of mutual funds portfolios.

Keywords Complex networks • Correlation matrix • Financial markets • Integrated ownership • Mutual funds • Systemic risk

1 Introduction

The nouns *graphs* and *networks* refer to the same abstract structure, although they are used in different scientific areas for different purposes. Indeed, networks became popular after the exploitation of social networks in the ′30s and ′50s (Borgatti et al. 2009), while the foundation of the formal building of a graph dates back to Euler, who first presented his results on the Koenigsberg bridges problem in 1735. The

A.M. D'Arcangelis (✉)
Department of Economics and Business, University of Tuscia, DEIM, via del Paradiso 47, 01100 Viterbo, Italy
e-mail: adarcangelis@unitus.it

G. Rotundo
Department of Methods and Models for Economics, Environment and Finance, La Sapienza University of Rome, via del Castro Laurenziano 9, 00161 Rome, Italy
e-mail: giulia.rotundo@uniroma1.it

P. Commendatore et al. (eds.), *Complex Networks and Dynamics*,
Lecture Notes in Economics and Mathematical Systems 683,
DOI 10.1007/978-3-319-40803-3_9

river crossing the old Prussian city of Koenigsberg had two islands connected by seven bridges, and the problem was to find a path to make a complete tour of the city crossing each bridge only once.

The formalization proposed by Euler differed from classic geometrical problems because it did not involve distances of meters: the land was represented through circles (nodes, units, elements), and the bridges through lines (edges, arches, links). This abstraction highlights the main characteristics of the problems. The proof that was given of the impossibility of the existence of such a path is based on the count of the number of edges connecting each node (node degree). Still now, the detection of paths with specific features constitutes a relevant task in graph theory, and it has a wide range of practical applications, from the optimal design of databases and print of electronic circuits to the travelling salesman problem.

The approach outlined above differs from the methods of combinatorics, which date back to a few centuries before, and leads to probability theory. In the middle of the twentieth century, the insertion of the probability theory into graphs boosted a new field of studies, *random graphs*.

Paul Erdos and Alfréd Rényi's famous cooperation generated a series of papers, the most well known of which introduced the Erdos-Rényi model of random graphs (Newman et al. 2006). Targets in random graph theory are the detection of the probability of the presence of a specific property (defined through a variable) in graphs drawn from a particular distribution. This constitutes a meeting point with problems rising from Physics, like the percolation theory that characterizes the connectedness of random graphs. Physicists were already familiar with regular graphs/networks (lattices), mostly in the framework of ferromagnetism and statistical mechanics. The main input for passing from "simple" networks to *complex networks* raised from the studies in social sciences (Albert and Barabasi 2002). Empirical data evidenced that, besides the randomness, the network was not showing a trivial structure, but revealed features that do not occur in simple networks. Hence, the term "complex". Nowadays, complex networks constitute an active and promising area of scientific research, widely inspired by the empirical analysis of real-world networks. It is part of *network science*, coded by the United States National Research Council as "the study of network representations of physical, biological, and social phenomena leading to predictive models of these phenomena".

Excellent reviews are available on the theory and applications of complex networks in several different areas (Albert and Barabasi 2002; Barrat et al. 2004; Boccaletti et al. 2006; Borgatti et al. 2009; Bougheas and Kirman 2014; Pastor-Satorras et al. 2003; Varela Cabo et al. 2015; Newman et al. 2006; Nature, focus issue 2013).

In financial markets research, literature records a fast growth in scientific production, mostly grounded in Physics, and an increase in the level of cross-disciplinary perspectives. First, concepts for the analysis of social networks found a proper representation, and, later, the literature recorded a burst of studies tackling research issues that can be conveniently managed through complex networks. Most of the studies have an empirical approach and provide a good base for the development of new mathematical models, econometric analysis, and open new perspectives for understanding large-scale phenomena such as the contagion channels in the financial system, the interconnections among financial institutions and markets and the analysis of systemic risk and financial stability. In other words, the relevance and role of single elements in the network can be evidenced, and critical areas identified.

In this chapter, the discussion of the papers analysing financial data with methods of complex networks will be conducted into four sections: analysis of the correlation matrix, systemic risk, integrated ownership and control, and the most recent studies on mutual funds holdings structure. The last topic will be discussed in detail with the presentation of new results on mutual funds holdings connections.

Since we are interested in financial markets, we do not delve on theoretical results that are not applied to our specific focus, nor on many applications of complex networks on Economics: such as GDP, considering clustering (Ausloos and Gligor 2008; Gligor and Ausloos 2007, 2008), focusing on the dynamic evolution of the system (Miskiewicz and Ausloos 2006, 2010), and Granger causality (Caraiani 2013), just to cite a few.

For the same reason, we are not extending our review to the rapidly expanding literature on International Trade Network (Bhattacharya et al. 2008; De Benedictis and Tajoli 2011; Garas et al. 2010; Garlaschelli and Loffredo 2004; Schweitzer et al. 2009).

2 Correlation Matrix

How strongly correlated are the stock markets? What is the level of market randomness and dependence? How does the structure change during expansions and recessions? These are the main questions addressed by the papers that use complex networks for the study of the correlation matrix. Correlation matrices play a relevant role in the paramount financial problem of optimal portfolio selection (Elton et al. 2014; Markowitz 1952). In Econometrics, several tools have been developed for their analysis: from ARCH/GARCH models to vector autoregression, principal component analysis and copulas. The perspective of complex networks, besides offering a different approach for a proper correction of the correlation matrix (Aste and Di Matteo 2010; Pantaleo et al. 2011), mainly uses the network approach to build a network structure among financial quantities, and introduces a distance inversely dependent on correlation.

In 1999, Mantegna proposed the distance is $d_{i,j} = \frac{1}{2}\sqrt{1-\rho_{i,j}}$, where $\rho_{i,j}$ $\forall i, j$ are the correlation coefficient computed between all pairs of stocks of the portfolio by considering the synchronous time evolution of the difference of the logarithm of daily stock price (Mantegna 1999).

It can be proved that $d_{i,j}$ is a mathematical distance. Another benefit of using $d_{i,j}$ instead of $\rho_{i,j}$ is that the most correlated stocks are the closest, which means that the distance is the shortest. Since $\rho_{i,j}$ are gathered into a matrix, also $d_{i,j}$ constitute the distance matrix D (Caldarelli 2008).

Soon, it was clear that correlation matrices, as well as their deterministic transform into distance matrices, are far from being random networks (Bonanno et al. 2003; Caldarelli 2008). Therefore, it was straightforward to look for the origins of the dependence. Of course, D constitutes a complete network, gathering too much information. A proper analysis must evidence and filter main features and characteristics. One of the most used quantities for such a filtering is the Minimum Spanning Tree (MST). The MST is a sub-network that keeps all the n nodes of the network, but only the $n-1$ links with the minimum weight, provided that the connected components remain connected. Originally used to detect the lightest routes on a graph, it is calculated through a recursive algorithm. One advantage of using the MST is the possibility of building proper visualization of the structure of the closest stocks in distance matrix D.

The minimal spanning tree (MST) is attractive because it provides an arrangement of stocks, which selects the most relevant connections of each element of the set, and hierarchies can be settled. In Mantegna (1999), the technique is applied to a portfolio of stocks of the S&P 500 index, and it provides a taxonomy that shows the clustering of many groups of stocks, which are homogeneous from an economic point of view.

An interesting aspect on the network analysis of correlation is given by the progressive change of the graph structure, as the time horizon decreases, from a complex organisation to a simple form (where clusters are sparser), so adding further insights and empirical evidence for discussion of the hypotheses of dependence and independence that are most used in financial market models (Bonanno et al. 2004). The MST also proves that during crises the distance among markets decreases (Sandoval and De Paula Franca 2012), and investment signals may be detected (Brookfield et al. 2013). Further techniques for filtering information from the correlation network, like the Planar maximally filtered graph, have been explored to overcome the strong dependence of the presence of links in the MST on the time lag selected for the analysis (Pozzi et al. 2008).

The MST has been applied also to time series of global currencies. In this case, the geographical proximity plays a key role in showing the differences of European and Asian clusters, and the interdependence of the currencies of countries at E.U. borders. As expected, the key currencies belong to major economic countries and the U.S. dollar plays the role of primary currency for its remarkable influence. Therefore, each currency depends on the U.S. dollar and on the key currency of the region where this belongs. The predominance of the U.S. dollar is also proved in other studies (Naylor et al. 2007), which analyse a different sample of worldwide currencies and use another metric distance function, the Gower one (Gower 1986). This result is in line with the U.S. hub role detected by the MST in the international trade networks (Maeng et al. 2012).

The MST has been applied to the distance matrix calculated on correlations among GDP. The structural topology of the MST, sampled at different times, allows the identification of different clusters of countries based on their indebtedness and economic ties. The main results show that with the debt crisis, the less and most affected Eurozone's economies are shaped as a cluster in the MST. In recent papers

(Ausloos and Miskiewicz 2010; Miskiewicz and Ausloos 2010), MST, entropy and other indices are used to prove that the mean distance between the most developed countries, decreased from 1960 to 2000, which can be considered a proof of economic globalization of these countries.

It can be concluded that the correlation matrix has inspired several studies, mostly conducted by physicists, and has contributed to the development of the study of distances, clusters and induced hierarchies on networks.

3 Systemic Risk

Banking has attracted a number of dedicated studies, mainly fostered by the 2008 subprime crisis. Terms like “Too big to fail” soon became part of common talks, and had a relevant role in the public debate on bank saving policies for the recent crises.

In finance, systemic risk is the default risk of an entire financial system. The concept expressed by the phrase “Too big to fail” (TBTF) is that a single financial institute may hold so much credit, that saving it -instead of letting it fail- becomes economically convenient to prevent the failure of the entire financial system. The phrasing was already in use when the crisis in 2008 made the concept prominent and gave a big impulse to the reform of financial legislation (White 2014). TBTF financial institutions are not necessarily banks: in principle, any company that primarily holds financial instruments (such as stocks, bonds, loans, derivatives, etc.) as assets on its balance sheet is exposed to the risk of debtors’ insolvency. The term *cascades* outlines that the insolvency of one institution has negative consequences on others, becoming a *contagion* if the outcome is as bad as causing their insolvency. The larger the financial institution, the worse the effect on the economy, the more likely is the decision of policymakers to intervene by providing support to the financial institutions. The drawback is that these actions also create moral hazard and expectations for the institution’s owners and managers, opening the door to possibly even bigger and deeper crises.

The development of a regulatory system needs a clear understanding of risk exposure and its monitoring. In order to achieve these goals, scientific literature has mainly exploited the concept of centrality on the network, and simulation models of contagions.

Studies on the topology of networks have led to the development of techniques for ranking the *centrality* of nodes in a network. The more a node is *central*, the more it is relevant for the property under observation. Such rankings are often referred to as *centrality measures*, although most of them are not measures in accord to mathematical terms.

Financial networks mirror regional and sectorial organization (Allen and Babus 2009; Bellenzier and Grassi 2013; Bellenzier et al. 2015). Therefore, the empirical estimate of quantities that are standard in complex networks—clustering coefficient and shortest path length besides centrality—has highlighted regional disparities (Boss et al. 2004).

Specific features of banks can be captured by ad hoc models that assess the systemic relevance of a given institution to the contribution of heterogeneity in network structures and concentration of counterparty exposures (Cont et al. 2013; Bougheas and Kirman 2014).

The problem of the optimal network design is quite relevant for the propagation of crises (Leitner 2005; Bougheas and Kirman 2014; López-Pintado 2006), which cause financial earthquakes when triggering responses from a large portion of the financial system (Vitting Andersen et al. 2011). Usually, empirical estimates require large data set that are not so easy to retrieve and manage. Specific network-based measures for ranking the relevance of nodes have been developed (Battiston et al. 2010, 2012; Bellenzier et al. 2015), suggesting that the debate should include issues eventually even more serious than TBTF, such as *Too central to fail* (impacting those who are important via network effects) and *Too correlated to fail* (similar portfolios and/or strategies).

Likewise, in *Too interconnected to fail* (TITF) (Markose et al. 2012) the 2007 credit crisis was empirically reconstructed through data. Dense clustering and mutual exposure identify the TITF institutions, where super spreaders dominate in terms of network centrality and connectivity. Studies focusing on the role of shocks on the overall stability of the financial system (Allen and Gale 2000; D'Errico et al. 2009; Gabbi et al. 2012; Steinbacher et al. 2013) and policies for market regulation are well represented in literature (Gai et al. 2011; Gai and Kapadia 2010; Halaj and Kok 2015).

We can conclude this section with a remark about the above-mentioned models of financial contagion: they do not constitute the only contribution to understanding the extent and consequences of the failure of a financial institution as pinpointed in (Bougheas and Kirman 2014).

4 Integrated Ownership and Control: Complex Networks of the Shareholding Matrix, and Directorate Interlocks

Indirect ownership is quite a relevant issue, mostly when dealing with antitrust measures, and it is very relevant for detecting Chinese boxes and tunnelling. In order to outline the phenomenon, let us consider a company A that does not buy directly shares of a company B (direct ownership); but A holds s_{AC} shares of an intermediary company C that, in turn, owns s_{CB} shares of B. In this way, no direct ownership of A in B is recorded, but, actually, A owns $s_{AB} = s_{AC} \times s_{CB}$ share (indirect ownership). Given the cross-shareholding matrix $A = \left(s_{ij}\right) \in R^{nxn}$, the ownership through one intermediary is given by the matrix product $A^2 = A \times A$. Analogously, the ownership through two intermediaries is given by $A^2 = A \times A$, and, in general, the ownership through n intermediaries is given by A^n.

Therefore, the total ownership, through direct ownership, and through any number of intermediaries, is given by $Y = A + A^2 + \cdots + A^n + \cdots = (I - A)^{-1}A$, where I is the identity matrix. Corrections to prevent double counting due to loops (Chapelle and Szafarz 2005, 2007) give rise to the integrated ownership matrix $V = diag\left(I - \overline{A}\right) Y$, where $\overline{A} = \left(\overline{a}_{ij}\right)$, and $\overline{a}_{ij} = \sum_{k=1}^{n} s_{kj}$, so the value depends only on the column j. The elements of $V = \left(v_{ij}\right)$ represent the number of shares that the company in column j holds in the company in the row i, counting both direct and integrated ownership through any possible path in the networks. Issues on convergence do not raise because $s_{ij} < 1, \forall i, j = 1, \ldots, n$. Results are quite different, depending on the country. Italian companies listed in the MIB30 index do not show long chains of control, and the existing ones can be easily explained following the raise and settlement of the single companies. On the contrary, the Japanese market shows clear signs of tunnelling, and some examples have been detected for the German market (Flath 1992).

The availability of large databases has moved the investigation from small national data set to the international ownership network and techniques from complex networks have produced further results. For instance, the estimate of the assortativity coefficient shows the strong tendency to form high-connected groups (Rotundo and D'Arcangelis 2010a).

Transnational corporations are confirmed to form a giant bow-tie structure, where the central nodes belong to a strongly connected component. Nodes in the strongly connected component are connected by cross-shareholdings, since each of them owns some shares of the others. Such strong component constitutes a small tightly-knit core of financial institutions, eventually TITF; therefore, raising important issues on market contagions, resilience and concentration both for researchers and policy makers (Rotundo and D'Arcangelis 2010a, b, 2014; D'Arcangelis and Rotundo 2014; Rotundo 2011; Bougheas and Kirman 2014; Vitali et al. 2011). For instance, the node out-degree is a quantitative measure of portfolio diversification of the company corresponding to the node. The detection of the eventual power law decay in the histogram of the quantities of interest becomes a standard estimate in Econophysics, since it opens the way to models. Quite interestingly, the histogram of the out-degree and the change of the value of the exponent of the power law clearly show the disappearance of the middle-sized investments through ownership (in favour to the return to the core business), in line with practical managerial issues more than to instances of market expansion.

The concept of integrated ownership is quite different from control. Let us consider the following example: a chain of ownership A-B-C where $s_{AB} = 51\ \%$ and $s_{BC} = 51\ \%$. Thus, $s_{AC} = (51)^2 \cong 26\ \%$. This means that A controls B, B controls C, but A does not have a sufficient number of shares to control C. Therefore, different ways to achieve control must be considered (Chapelle and Szafarz 2005, 2007; Rotundo and D'Arcangelis 2010b).

Interlocked directorates refer to the practice of members of corporate boards of serving on the boards of multiple companies. As early as 1969, the debate on the flaw of fair competition through the interlocking directorate was quite active, especially in reference to laws enacted in 1914. In fact, the inter-organizational élite co-optation can be seen as a cooperative strategy between economic organizations for reducing sources of uncertainty.

Empirical evidences show that links -although dynamically evolving- are persistent, with a stable core, and they involve companies managed by families strong at the local level (Bellenzier and Grassi 2013). A recent analysis (Rotundo and D'Arcangelis 2010b) shows that companies listed in the stock exchange for a long time share the board of directors, whilst newcomers enter the market buying shares. Interlocks among companies can be described as links, and the analysis of the subsequent network is straightforward.

The application of centrality measures shows positive correlation among the rank of interlock and firm value, but positive correlation with betweenness and flow-betweenness, representing the intensity of the relationship between companies, capturing the volume of information flowing from one company to another through the interlocks (Croci and Grassi 2014; Grassi 2010; Grassi et al. 2008).

Understanding a complex system is quite different from controlling it. We may conclude that the role of social interaction is quite relevant, since the board of directors have their weight in achieving control (Chapelle and Szafarz 2005, 2007; Rotundo and D'Arcangelis 2010b) and further studies are available on the personal connections among important managers.

5 Investment Decisions and Institutional Investors

Complex network methodologies have been applied to the field of financial markets for many purposes. The methodology is suitable to highlight the impact of networks of investors and managers on investment decisions or directly on stock prices. Besides the applications to the interdependence of stock markets, to systemic risks and to integrated ownership and control, complex network methodologies have recently begun to be used in the area of investments and managed portfolios.

Aiming to construct an index of attractiveness of various capital markets, a recent research (Cetorelli and Peristiani 2013) carries out an in-depth analysis of the "patterns of relationships" among financial centres. Using network analysis, the Authors show that although the London Stock Exchange, the Deutsche Börse, and the Hong Kong Stock Exchange became more competitive, the U.S. exchanges remained the favourite destination for foreign issuers which wish to cross–list on multiple exchanges. Along the same line, other Authors (Lucarelli et al. 2012) apply network centrality measures to the indirect network between trading venues (regulated Stock Exchanges and Alternative Trading Venues, ATVs) with the aim to observe the dynamics of simultaneously traded European stocks from 2005 to 2009. Their results show that the advancement of Alternative Trading Venues eroded the isolated-centrality of major Stock Exchanges (above all London); in contrast, degree-centrality significantly increased for the majority of the Stock Exchanges analysed in their sample. The introduction of multi-trading venues does not deteriorate connectivity of cross-listed relationships, unveiling that multi-trading co-exists with secondary market cross-listing.

The investment management industry has been playing an increasingly important role in the financial system, especially in the most advanced economies. In recent decades, credit intermediation has been progressively shifting from the banking to the non-banking sector, particularly to the asset management industry (through different investment vehicles such as mutual funds, hedge funds, exchange-traded funds, private equity funds, pension funds). Focusing on mutual funds, the number of mutual funds in the U.S. reached 79,669 in 2014 with a value of 31.38 trillion U.S. dollars for the total assets under management.[1] U.S. mutual funds account for roughly half the global asset under management in the world. Europe has an equally important role with 9.576 trillion of asset under management. Such a crucial role for the asset management industry has obvious paybacks for financial intermediation: investors gain a better diversification of their portfolios and grant a more stable financing of the real economy even during periods of distressed market conditions. Other benefits of the asset management vehicles over banks concern the stability of the financial system: the banks are exposed to solvency and liquidity risks (due to their typical short term funding) whereas the investment risk of the shares issued by mutual funds relapses on end investors.

Focusing on this field, complex network methodologies have been firstly used to analyse how information is disseminated among mutual fund managers in financial markets and how the diffusion of such news influences stock prices. Focusing on connections between mutual fund managers and corporate board members via shared education (the connection is their attendance of the same school), a research (Cohen et al. 2008) identifies the transmission of insider information and finds that managers favour the investment on companies they are connected through their network. Placing larger bets on "connected firms", which have performed significantly better than non-connected ones, mutual fund managers have significantly improved their performance.

The availability of data on holdings of a sample of US actively managed equity mutual funds (Augustiani et al. 2015) allows other Authors to examine the effect of mutual fund connections, through managerial sharing, on performance and stock holding commonalities. Their analysis of return correlations and portfolio holdings shows that more interconnected funds managers tend to buy and sell similar stocks, hence increasing the similarity of portfolio holdings and undermining the distinctiveness of their investment strategy. Assessing performance effects, the Authors find that highly connected funds significantly underperform weakly connected funds by about 1.4 % on a yearly risk-adjusted basis. Conversely, fund family performance remains almost unaffected by the intensity of fund connections, and greater fund connections can significantly enhance family-level profit margins.

A subsequent research focused on tie between analysts and companies (Cohen et al. 2010) investigate the dissemination of information in security markets through the recommendations of sell-side equity analysts used to study the impact of

[1]The largest fund management companies worldwide as of December 2014 are Blackrock, Vanguard Asset Management, State Street Global Advisors and Fidelity Investments.

social networks on agents' ability to gather superior information about firms. The hypothesis is here that school ties between analysts and senior corporate officers provide comparative information advantages in the production of analyst research. The main result is that equity analysts outperform on their stock recommendations when they have an educational link to that company. Results are strong, so much that a strategy of going long (for the recommendations to purchase given by analysts with school ties) and going short (for the buy recommendations of analysts without ties) returns a performance of 6.60 % per year. The Authors' conclusion is that analysts' social networks facilitate the direct transfer of information, or alternatively that these networks simply allow analysts to better assess managerial quality. Even if the Authors do not always use complex network measures, the merit of these papers is to highlight the importance of network structures and interactions among agents in the analysis of information flow and price evolution in security markets.

The metrics from network analysis allows to analyse the impact of Sovereign Wealth Fund (SWF) equity investments on target firm operating performance (Del Giudice et al. 2014). The aim of the paper is to investigate whether target firms, which are better connected to each other by means of the SWF investments, gain benefits in terms of higher performance. The results indicate that more central firms in the SWF-target firm network have better operating performance and that the effect is related to the size of the stake acquired, is larger if the investment is direct and in the domestic country and if the SWF is run by a politician.

The analysis of the relationship between the location of the fund in its network and the investment performance, risk taking, and flows is the object of a paper focused on pension funds (Rossi et al. 2015). Using data on a large set of UK pension fund accounts over the period 1984–2004, the Authors investigate whether network centrality explains managers' investment performance, risk-taking behaviour, and flows. The centrality of the management company providing the fund is derived from the number of connections it has with other management companies through their commonality in managing for the same fund sponsors or through the same fund consultants. In detail, individual pension fund accounts can be connected by their sharing of the same consultant and/or the same manager. The results show that a fund-manager's (relative) degree of centrality in a network positively affects risk-adjusted returns and growth in assets under management and that this effect is particularly strong for large fund managers, even after controlling for size. Once a central position has been established, the manager tends to reduce risk-taking behaviour and reduces the chances of getting fired by institutional clients.

The fact that many mutual funds around the world have suffered from negative returns during the global financial crisis emphasizes how volatility can rapidly spread among previously unrelated assets in times of high uncertainty and turbulence. This observation provides the starting point for an investigation (Azmi and Smith 2010) of the spread of the current crisis in the correlation networks amongst a sample of mutual funds across seven regions globally. Using the data of equity funds in ten countries representing seven regions, the Authors select two funds from each country based on the highest net asset value and built a correlation network of the weekly mutual fund log-returns over the period from April 7, 2006 to April 27,

2009. They show that the losses in financial assets within certain countries could spread and follow a cascade or epidemic flow like model along their correlations. A first conclusion that that can be drawn from the analysis is the rapid spread of the credit crisis amongst previously uncorrelated markets and countries. The correlation networks under examination do not cause the transmission chain of collapse, but they are tied to it. Such architecture encourages the excesses of the global financial crisis, motivates aggressive risk taking and pushes some asset prices to unsustainable levels, increasing financial fragility.

A second stream of research uses complex networks techniques in order to detect connections among mutual fund holdings and relates complex networks measures to the dynamic of risk and return. The collection of the ten largest positions of 18 Vanguard and Fidelity family funds (Solis 2009) provides a new approach to visualizing the way stocks are affiliated to mutual funds as a bipartite graph, and computes network summary statistics. The stock network has a high clustering coefficient (indicating "prominent" stock hubs), which suggests that the managers' selection of stocks is not made independently as if the network were that of a purely random graph with similar expected number of links. The higher diameter and average degree distance between two vertices (6 and 2.91 vs. 3 and 2.01 for the random graph) suggests a small-world behaviour, due to the highly connected network of stocks, mostly blue-chips, which populate the mutual funds sample.

In a more complete study, focusing their attention on the indirect connections among holdings resulting from common ownership of a sample of mutual funds from 1980 to 2008, Anton and Polk (2013) find that pairs of stocks held in many mutual funds' portfolios show future excess correlation between stock returns. The Authors demonstrate that this "common ownership effect" is stronger for common owners who are experiencing extreme positive or negative flows in low-float stocks. Based on these results, the paper supports a cross-stock-reversal trading strategy that exploits the information in ownership connections and generates significant abnormal returns of more than 9 % per year, controlling for market, size, value, momentum, and other characteristics. Following a similar approach in identifying pairs of mutual funds linked by common portfolio holdings, Blocher shows that spillover effects associated with the fund flows of an investor's network neighbours account for roughly 2 % per quarter and are the result of crowded trades, since they are completely reversed in the subsequent year (Blocher 2013).

In the same stream, Braverman and Minca demonstrate that the network of common asset holdings is useful to identify systemic funds (Braverman and Minca 2014). Using quarterly equity mutual fund holdings data ranging from January 2003 to December 2012, they analyse the interrelations due to common asset holdings and construct a measure of fund vulnerability for the shocks of their neighbours in the network. The Authors demonstrate that this "vulnerability index" is useful in predicting returns in periods of mass liquidations, since it helps in identifying vulnerable funds based on asset holdings and the liquidity characteristics of the stocks.

Guo, Minca and Wang analyze the topology of the network of common asset holdings, a network in which nodes represent managed portfolios and edge weights capture the impact of liquidations (Guo et al. 2015). Focusing only on the subgraph of weak links (those that lead to significant liquidations), the Authors analyze the degree centrality and find that this measure follows power law distribution and is correlated with returns. For individual portfolios, higher degree is associated with future higher return in the long run, but it may negatively affect the performance during financial crises. At the aggregate level, stronger connectivity among portfolios is associated with higher systemic risk. Exploring network clustering, they identify a small number of communities, densely linked, that concentrate a significant proportion of the portfolios.

D'Arcangelis and Rotundo explore the commonalities in the holdings of Italian funds investing in domestic stocks (D'Arcangelis and Rotundo 2014). Following the empirical evidence that shows that fund managers take common decisions on stock holdings, both for benchmark constraints and for style management decisions, the Authors use the methodology of complex network analysis, to describe the way in which stocks are related to mutual funds and to detect the implications of the interactions. The results highlight a large core group of portfolios that have many stocks in common; while other funds invest in a wider variety of stocks. These results confirm empirical findings on US market (Solis 2009) and show that Italian mutual funds holdings are highly interconnected, suggesting the existence of a small-world behaviour and the tendency of mutual funds managers to steadily invest in a restricted number of well-established high capitalization stocks (blue chips). They also test the impact of overlap on performance and risk (raw performance, Sharpe Ratio, standard deviation, beta coefficient and fund tracking error), and find that the overlap is predictive of performance similarity, at least for the samples involving blue chip stocks, even though comparison with and influence of many other factors are not to be completely disregarded. The intersection of small cap holdings, usually exploited for tactical asset allocation purposes, is not feasible to conveniently differentiate final performance of funds. The results support the thesis of substantial passive management of institutional portfolios, realized through the investment in a portfolio of a limited number of stocks, which are selected by mathematical algorithms in order to optimize the solution of the index-tracking problem.

5.1 Mutual Funds Holdings: A Network Analysis

The data used for the analysis are the holdings of 215 European mutual funds, as listed in the database made available by Morningstar Italy in December 31st, 2015. The funds legal address is in Europe, and their investment focus is the European equity market. In detail, the parameters used in the data query from Morningstar request that the funds belong to one of the following categories: EURO Large Cap, Europe Large Cap Blend, Europe Large Cap Growth, Europe Large Cap Value. The Bloomberg database has been used to check the equity exposure for each of the 272 mutual funds coming out from the query. This check caused the reduction from 272

Table 1 Funds primary benchmark of the funds of the sample (Source: Bloomberg)

Fund primary benchmark	No of funds	Sample (%)
MSCI Europe NR USD	96	44.65
MSCI EMU NR USD	72	33.49
MSCI Europe Value NR USD	29	13.49
MSCI Europe Growth NR USD	17	7.91
MSCI Europe NR EUR	1	0.47

to the final 215 list of funds, because some funds do not authorize the disclosure of the composition of their portfolios. The selected funds belong to investment houses of 14 European countries.[2]

For each fund in the different categories, we have registered all the stock holdings ranked in terms of weight: the equity sample contains 1603 stocks belonging to 51 different countries.[3] The sampled funds adopt similar benchmarks, confirming their belonging to comparable categories: 92.5 % of funds are related to MSCI Europe Indices (44.2 % to MSCI Europe NR and 21.4 % to MSCI Euro NR EUR), STOXX Europe Indices and FTSE Indices, 4.2 % are not benchmarked and only 3.26 % are related to minor benchmarks. Following the benchmark attribution by the site Morningstar Italy, 44.65 % and 33.49 % of the funds are represented by the MSCI Europe and by the MSCI EMU indices (Table 1).

We have started our analysis by building a network on the mutual funds stock holdings. A link is drawn from mutual fund i to stock j if mutual fund i owns stocks of company j. Links are only drawn from a mutual fund to a stock. Therefore, starting and ending nodes of each link belong to different sets. Networks showing this property are named bipartite.

We represent the funds-stocks network through a matrix A. Rows of $A = (a_{ij})$ correspond to funds, and columns relate to stocks. The weights of the network links a_{ij} are the percentages of stock j held by fund i. In our sample, matrix A has 215 rows and 1603 columns. The binary matrix $B = sign(A)$ is well suitable for projecting the network into the space of relationships among mutual funds as follows: two funds are connected if they have at least one stock in common. The funds-funds matrix C is then created through matrix multiplication: $C = BB^T$. For any two funds i and j, the elements c_{ij} report the number of stocks common to both funds i and j.

Lastly, the stocks-stocks matrix $D = (d_{ij})$ is defined as $D = B^T B$. For any two stocks i and j, the elements d_{ij} report the number of funds that own shares of both stocks i and j.

[2] Five European countries mostly contribute to the group of 215 mutual funds: Germany (18 funds), France and Italy (45 funds each), United Kingdom (59 funds), Switzerland (26 funds). Other 22 funds belong to investment houses of 9 other European countries. Due to the exiguity of the samples, they have been gathered into the category "Others".

[3] In the sample of stocks, the ones belonging to the five most represented countries (DE, FR, IT, SW, UK) account for the 57.08 % of the sample.

5.1.1 Analysis of the Overall Funds-Funds Network

The funds-funds network is a non directed matrix, as witnessed by the symmetry of matrix C, which shows the overlap of investments among fund i and fund j. However, this information is biased by the dimension of both funds i and j, since funds with more holdings could be more overlapped only due to the number of different stocks in which they are invested. Considering matrix A, the size of funds can be calculated at once as $(sign(A^T)\ e)$, where $e = (1,1,1,\ \ldots\ 1)^T$ is a vector with 1063 components. In matrix B, the size of each fund is reported in the elements on the diagonal. In order to overcome the issue, we have divided each element in matrix C by the maximum between the dimension of funds i and j, so obtaining the symmetric matrix $F = c_{ij}/max(b_{ii}, b_{jj})$.

The elements on the diagonal were set equal to zero to avoid counting the size of the fund. Each of the 215 rows of F represents the average overlap of the fund on row i with the remaining 214 funds of the sample. The result is a vector of 215 cells (each representing the mean overlap of fund in a row with the other funds in the columns), whose statistics (mean, median and mode, skewness and kurtosis) are summarized in Table 2 (column full sample).

The analysis has been repeated calculating the mean of each of the rows of F within the columns of funds belonging to the single country (DE, FR, IT, SW, UK and "Other"); this average represents the mean overlap of the fund of row i with the other funds of that country. The statistics on these country vectors are shown in the respective columns of Table 2.

The Jaque-Bera test[4] is used to check the null hypothesis that data come from a normally distributed population. The results show that normality holds on the full sample and does not hold on subsamples.

Table 2 Mean overlap of the F normalized funds-funds matrix

Funds-funds matrix (normalized sample, average values)							
	Full sample	DE	SW	FR	IT	UK	Other
Mean	0.0526	0.0643	0.0577	0.0537	0.0498	0.0499	0.0475
Median	0.0545	0.0656	0.0608	0.0553	0.0497	0.051	0.0605
Mode	0.0606	0.0397	0.0514	0.0514	0.0514	0.027	0.0606
Skewness	−0.0282	−0.007	−0.3082	−0.188	0.3507	0.1348	−0.4827
Kurtosis	2.0354	1.9451	1.9045	2.0763	2.5952	1.7755	1.8971
Jarque Bera							
h	1	0	0	0	0	1	1

[4] If h = 1, the hypothesis of normality is rejected, h = 0 means that it is accepted.

5.1.2 The Analysis of Centrality

On the basis of this sample, we have calculated nowadays standard measures of centrality on networks. At a first analysis, the node degree or "degree centrality" (that indicates the number of neighbours of a node) reveals that our network is nearly a complete one. The betweenness centrality is the number of shortest paths going through the node under examination in relation to the total number of shortest paths of the network. As this measure of centrality indicates if a node can be "intermediary" within the network, it assumes a specific meaning for portfolios like mutual funds. A high value of the betweenness centrality unveils funds that "bridge" different groups: a fund with the higher betweenness owns stocks that are owned by different groups, which would not overlap otherwise.

Whether the network is not showing well distinct groups, the fund with the maximal betweenness is not investing in the core investment target of most mutual funds, so its eventual crash is not as relevant in other applied problems, for instance in the spread of epidemics or in the true collapse of a bridge. This is exactly our case. The entire group is so highly connected that the values of the betweenness centrality do not provide an informative discriminant analysis. Similarly, those funds high on eigenvector centrality are linked to well-connected portfolios and may influence many others in the network either directly or indirectly through their connections.

Another centrality measure that can be applied to the funds-funds matrix is the k-shell decomposition (Garas et al. 2012), which tries to detect the most influential nodes in the network of funds. Therefore, our analyses on K-shell does not lead to meaningful partitions. The application of the Louvain method of community detection (Blondel et al. 2008) reaches a similar conclusion.

In conclusion, results state that the funds in our network are highly connected and their portfolios exhibit close overlaps. These results can be the consequence of the adoption of similar investment policies in presence of different but highly correlated benchmarks.

5.1.3 Analyses Considering the Benchmark

Such a conclusion suggested we should look into it further. Therefore, we have repeated the analysis on two sub-portfolios obtained by insulating the stocks belonging to the benchmark and those out of the benchmark. The aim is to detect the possibility of different behaviour of fund managers in the management of the market based and the tactical components of the portfolios under management.

We have extracted the constituents of the MSCI Europe index from the MSCI site[5]: of the 442 constituents of the MSCI Europe, as many as 440 of them are present in the sample of 1603 stocks of our funds-stock matrix. Then we have built two separate networks: the funds-stock in the benchmark (215 rows × 440 columns)

[5] Available at https://www.msci.com/constituents.

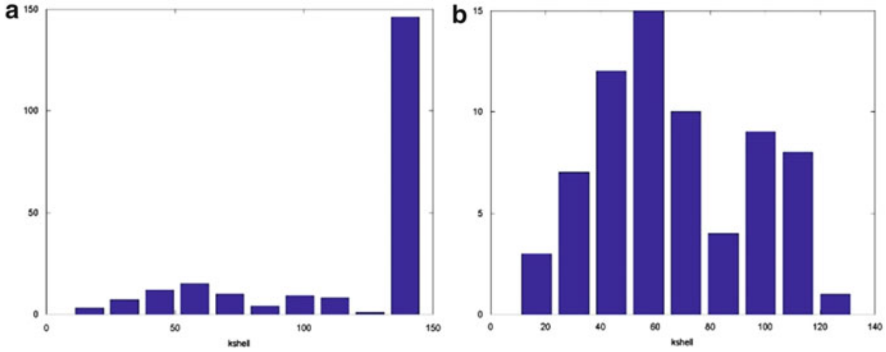


Fig. 1 Bar diagram of the k-shell of the funds-funds network calculated on the out-of-benchmark stock sample. (**a**) Entire distribution (**b**) having removed the peak at 146

and the funds stocks out of benchmark (215 rows × 1163). Looking at the results on the first subsample, the conclusions on the degree centrality and on the k-shell decomposition do not change much. The fluctuations of the values do not lead to significant differences. A quite different picture emerges when considering the second network built on out-of-benchmark stocks. The k-shell degree shows a large group of 146 funds showing high peak at 146, a second group of 18 funds with values between 122 and 100 and a residual group of funds with k-shell under the value of 100 (see Fig. 1a, b).

This is quite different from the nearly complete connection of the previous networks, and also quite far from the ubiquitous power law behaviour. In fact, there are many nodes with a high node-degree. This means that the connections among funds are very dense also in the out-of-benchmark sample, which means that even if the investment is diversified with a tactical and unsystematic component of the portfolio, such diversification is not strong enough: a situation completely different from what expected in presence of a power law distribution (for an example, refer to Fig. 2b of D'Arcangelis and Rotundo 2014).

Starting from the percentage of weights of the five most significant countries of the sample,[6] a geographical analysis of the three k-shell groups reveals that

- apart from Great Britain funds, whose frequency falls from 27.4 % in the full sample to 21.9 %, the subsample of the most connected stocks (k-shell value equal to 146) does not signal any outlier;
- the percentage incidence of UK holdings increases with the decreasing value of k-shell, supporting the existence of much *bigger diversification* benefits in UK funds;

[6]Germany 8.4 %, Switzerland 12.1 %, Italy and France 20.9 % each, UK 27.4 % and the residual Countries 10.1 %.

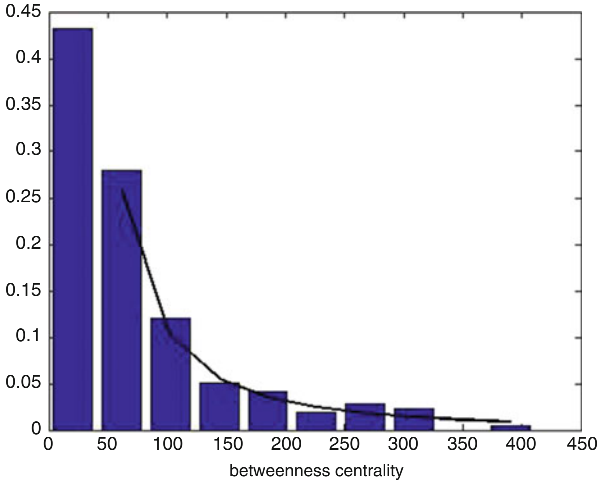


Fig. 2 Histogram of the betweenness (fund-funds matrix on out-of-benchmark sample). The tail of the distribution is well fit by a power law $f(x) = cx^{-\alpha}$ with coefficients (with 95 % confidence bounds): $c = 488.7$ (27.2, 950.1), $\alpha = 1.81$ (1.589, 2.031). Goodness of fit: SSE: 0.0005668, R-square: 0.991, Adjusted R-square: 0.9897, RMSE: 0.008999

– Italy and Germany show an antithetic behaviour relative to UK, with a strong underweight in the group of funds with minimum k-shell (under 100), supporting the evidence of more concentrated portfolios.

In contrast, the betweenness centrality shows a power law behaviour (see Fig. 2). This means that there is a continuum of values in the relative overlap of funds, with the usual implications when it comes to resilience to spread of volatility, contagions and financial fragility.

5.1.4 A Country-Based Analysis of Fund-Fund Network

Moreover, the analysis of the funds-funds matrix was performed also on separate subsamples of funds belonging to each of the five single countries under examination (Germany, Switzerland, France, Italy and UK, and the residual class "Others"). In this case, indeed, the mean has been calculated for each row of the funds-funds matrix, whose dimension is related to the number of funds of the country in the network. The result of this operation is a vector with 215 components, where each value shows the mean of the stocks that overlap the fund in the rest of the sample (the elements on the diagonal were set equal to 0). Table 3 shows the mean of this vector.

Germany funds qualify as the most connected network. In fact, the German funds in the subsample funds-funds matrix (18 rows and columns) have the highest overlap (0.0781 is the highest value on the main diagonal, and also the maximum

Table 3 Mean overlap on the single country normalized funds-funds matrices

Country	DE	SW	FR	IT	UK	Others
DE	0.078	0.068	0.069	0.057	0.061	0.066
SW	0.068	0.063	0.062	0.055	0.055	0.051
FR	0.069	0.062	0.057	0.049	0.049	0.049
IT	0.057	0.055	0.049	0.047	0.050	0.046
UK	0.061	0.055	0.049	0.050	0.047	0.046
Others	0.063	0.048	0.047	0.044	0.044	0.051

Table 4 Mean overlap of the single country stocks-stocks matrices

	DE	SW	FR	IT	UK
DE	3.052	1.978	2.954	1.442	1.497
SW	1.978	2.291	1.930	0.961	1.644
FR	2.954	1.930	3.223	1.423	1.441
IT	1.442	0.961	1.423	0.944	0.717
UK	1.497	1.644	1.441	0.717	1.293

element in the matrix). Moreover, the overall connection of German funds to the funds of the other countries is the higher (the sum of the connections to FR, IT, SW, UK and "Others" is equal to 0.3211, and is highest than all the other sums). Similar conclusions hold for the Swiss funds, with the second highest value on the main diagonal, that shows the total connection among the Swiss funds. The overall connection to the funds of other countries is the higher and equals 0.29. Italian funds reveal an opposite behaviour, showing the lower value of internal connection (0.443) and also the minimum overlap with the funds of the other countries (the sum of the connections with German, French, Swiss, UK and "Other" funds equals 0.253).

5.1.5 Analysis of the Overall Stocks: Stocks Network

The stocks-stocks matrix D is defined as $D = B^T B$. For any two stocks i and j, the elements d_{ij} report the number of funds that own shares of both stocks i and j. The stocks-stocks network is a non directed matrix, as witnessed by the symmetry of matrix D.

The mean of each of the rows of D represents the mean overlap of the stock on row i with the remaining of the stocks of the sample, divided by countries. It is the mean of all the values in the submatrix corresponding to the rows and columns listed for each block in Table 4, that sums up the findings of the calculation of the means on the country blocks of the vector of the means of the overlap values in each row.

The results show that German stocks reveal the maximum overlap both with foreign and domestic stocks. The conclusion is that in the sample under examination German stocks are diffused to the utmost degree. A similar behaviour is detectable also for French stocks, which reveal the maximum overlap with other French stocks (France/France = 3.22); the mean value of the overlap with stocks of other countries is 7.748 and is the second value (France has the minimum overlap 0.944 in the

Table 5 Mean overlap on the single country stocks-stocks matrices

	DE	SW	FR	IT	UK
DE	0.915	0.670	0.884	0.670	0.625
SW	1.491	0.801	1.040	1.102	0.722
FR	1.544	1.166	1.736	0.694	1.079
IT	1.082	0.779	1.199	0.942	0.765
UK	0.592	0.500	0.598	0.418	0.432

sample with Italian stocks). Going through the rows of Table 4, we can notice that also Swiss stocks are more connected with the other Swiss stocks (Switzerland-Switzerland = 2.29) than with the stocks of other countries, but the overlap with German stocks is also very high (1.97).

Of the five countries examined, only Italy and the United Kingdom have a value on the main diagonal that is not the maximum in their row; therefore, for these countries, the overlap among domestic stocks is lower than with those of other countries. Italy and United Kingdom stocks are mainly bought jointly with stocks of other geographical areas: the maximum overlap of United Kingdom stocks is with Swiss stocks; the maximum overlap of Italian stocks is with German and French ones.

Lastly, the overlap between Italy and the UK is always very weak. Further, we may add the comment that Italy has a marginal weight in international benchmarks and that the result of the United Kingdom could be partially due to its disposition to minimize exchange rate risks and avoid costly hedging strategies.

Following the methodology used for the funds-funds network, we have built the stocks-stocks matrix on separate subsamples based on the domicile of the funds in the sample (Germany, France, Italy, Switzerland, and UK, and the residual class "Others").

The difference between the country-based analyses in Tables 4 and 5 lies in the different data of each cell in the stocks-stock matrix: in Table 4, the cell contains the mean overlap of domestic and international funds that share the couple of stocks in the corresponding row and column; indeed, in Table 5, such funds are only domestic (ref. to the row). For instance, in the cell corresponding to DE (row) and UK (column) the value 0.625 is the overlap between DE stocks and UK stocks, considered as a block, owned exclusively by DE funds. The results converge towards the conclusions detected from previous tests: French mutual fund managers confirm to be inclined to overweight the percentage of French stocks in their portfolios and are at the same time exposed to German stocks; Italian funds combine mainly French and German stocks, the same conclusion is valid for Swiss managers. The exposition of UK funds to other countries holdings is systematically low (the row corresponding to United Kingdom has the minimum values of the entire matrix), the consequence is that we detect an intrinsic tendence of these actors towards a strong diversification.

The node-degree in the stocks-stocks network is the number of links connected to the nodes. It can be calculated through the $K_{out} = sign(D)\, e$, where $e = (1,1,1,\ldots 1)^T$

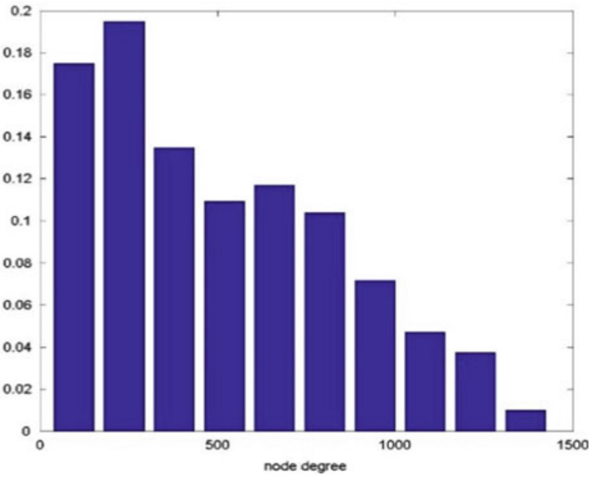


Fig. 3 Node degree of the single country stocks-stocks matrices

is a vector with 215 rows. A node i showing a high degree signals a stock that is associated to many other stocks in the portfolios of the sample mutual funds. A high mean node-degree means that stock i is connected to many others because they appear together in the portfolios of many funds. Figure 3 shows a behaviour different from the power law, although showing a small number of couples of stocks belonging to many funds, and a higher number of couples of stocks belonging to a few funds.

Due to the high connectivity of the matrix, the results of betweenness and k-shells are not performing a clear discriminant analysis; therefore, we have proceeded with the analysis of two well distinct sub-groups: the stocks belonging to the benchmark and the ones out-of-the-benchmark.

5.1.6 Analyses of the Stocks-Stocks Matrix for Benchmark Constituents

On the benchmark, the node-degree is far from the classic power law. Counting the nodes whose degree equals zero, we can notice that there are 1161 stocks that do not belong to the ownership of the same mutual fund with other stocks in the benchmark. The node degree has a peak concentrated on the mode at 428 (343 nodes). Since the remaining nodes have values ranging nearly uniformly from 50 to 437, we may conclude that the stocks in the benchmark either do not belong to the same mutual funds of other stocks or are the most frequently bought with other 343 stocks of the benchmark. Non zero elements in the betweenness show a power law behaviour, which implies that there are a few elements (with high betweenness) that lay in paths connecting communities (Fig. 4).

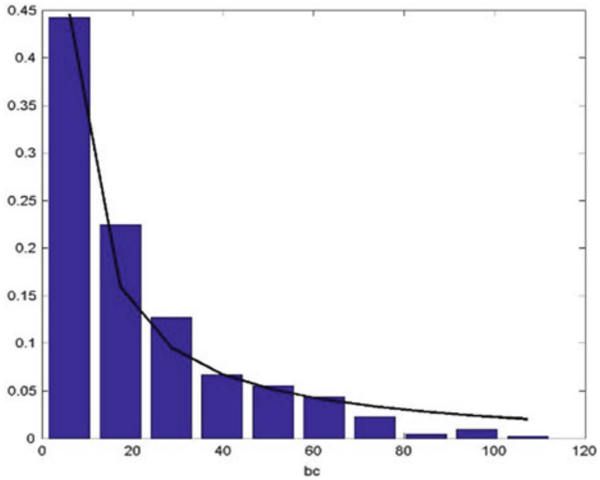


Fig. 4 Histogram of the betwenness for stocks belonging to the benchmark. General model Power1: $f(x) = cx^{-\alpha}$ Coefficients (with 95 % confidence bounds): $c = 2.463$ (1.402, 3.525), $\alpha = 0.9394$ (0.7462, 1.133). Goodness of fit: SSE: 0.0068, R-square: 0.9605, Adjusted R-square: 0.9556, RMSE: 0.02916

Table 6 Geographical allocation of the community structure—benchmark sample

Group	n.	DE (%)	SW (%)	FR (%)	IT (%)	UK (%)	Other (%)
1	147	17.69	0.68	27.89	11.56	1.36	40.82
2	120	0.00	21.67	0.00	0.00	46.67	31.67
3	88	1.14	17.05	1.14	0.00	50.00	30.68
4	85	28.24	0.00	36.47	8.24	2.35	24.71

The analysis of k-shells shows that there are two very different groups: one at 400 and one at 50. The one around 400 contains 419 elements, confirming the existence of a large group of highly connected nodes.

The presence of groups is well outlined by the Louvain method for communities' detection (Blondel et al. 2008). Table 6 shows four communities with a specific geographic concentration. Starting from the geographical percentage coverage of the sample of stocks (DE 11.59 %, SW 9.55 %, FR 16.59 %, IT 5.45 and UK 23.64), the analysis shows that Group 2 and Group 3 have a strong concentration on Swiss and UK stocks, whereas the stocks of the other countries (Germany, France and Italy) are concentrated in Group 1 and Group 4. The fact that the UK and Switzerland are non-Eurozone countries may be a tentative explanation for the results that give a snapshot of a sort of "geographical organization" of the communities. The output of the Louvain method for the out-of-benchmark matrix is strongly different, as it does not show a clear geographical allocation within the groups.

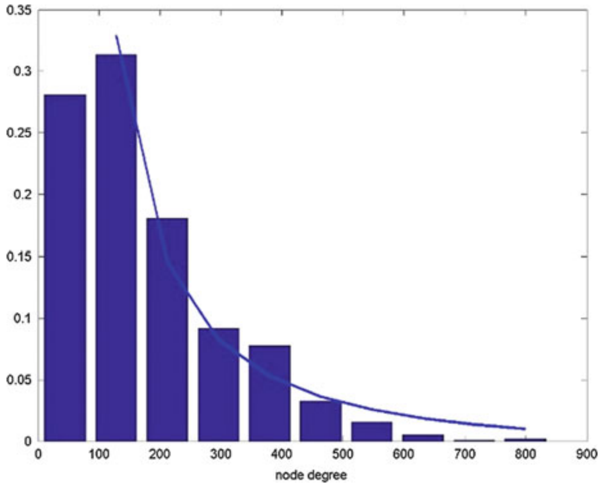


Fig. 5 Histogram of the node degree of the stocks-stocks matrix of the out-of-benchmark stocks. General model Power1: $f(x) = cx^{-\alpha}$. Coefficients (with 95 % confidence bounds): a = 624.9 (509.7, 1760), $\alpha = 1.559$ (1.2, 1.918). Goodness of fit: SSE: 0.003182, R-square: 0.9643, Adjusted R-square: 0.9592, RMSE: 0.02132

5.1.7 Analysis of the Stocks-Stocks Matrix: Out-of-Benchmark Stocks

The statistics of the stocks out-of-benchmark are quite different from the ones in the benchmark. The node degree has a power law tail distribution. This is in accord with the strategic asset allocation of non-benchmark stocks. In fact, there are a few that are bought by the many mutual funds, and many that are bought by a few mutual funds (Fig. 5). The comparison with the node degree of the entire stocks-stocks network shows lower values in the middle part of the distribution.

The values of betweenness are all concentrated around the mode, with a very few elements with high betweenness. This is in favour of the hypothesis of the presence of groups, with some elements bridging them. The analysis of k-shells shows a quite interesting behaviour, where the highest value at 200 reveals a k-core of overlapping stocks, while the minimum of the k-shell is 3 and the highest number of stocks are in the first bin of the bar diagram (Fig. 6).

The presence of scaling and the absence of well-insulated groups of specific dimension is confirmed also by the Louvain method for communities detection (Blondel et al. 2008). Twenty groups are detected, and the presence of many small groups and a few large ones is clear, although the peaks around 140 and 160 deviate from the power law (Fig. 7). Such groups do not reflect any geographical grouping homogeneity. Therefore, the result is quite different from the result on the benchmark. The absence of country-based grouping confirms the fact that out-of-benchmark stocks are bought by mutual funds mostly for tactical asset allocation purposes without a particular interest in geographical distribution. We leave the investigation of other causes for such grouping to future work.

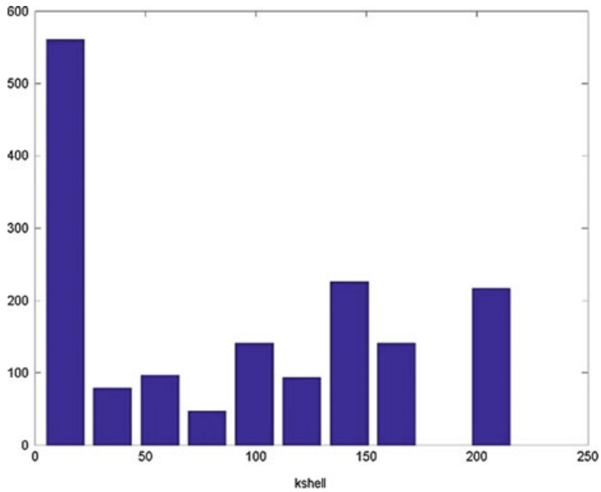


Fig. 6 Bar diagram of the k-shell of the stocks-stocks matrix calculated on the not-benchmark

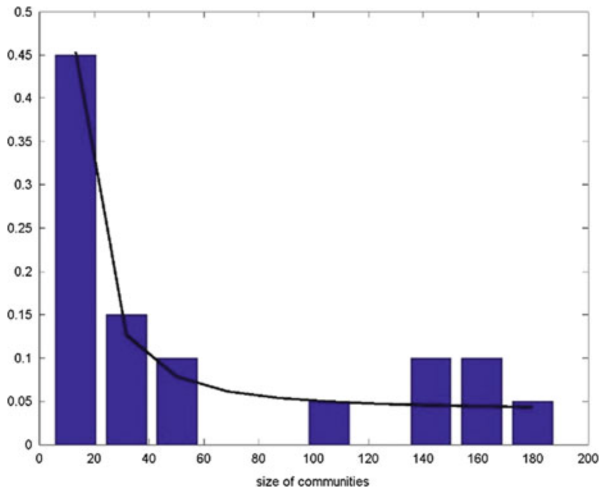


Fig. 7 Size of communities (20 communities with sizes 189, 157, 155, 147, 134, 101, 58, 50, 40, 33, 32, 14, 10, 9, 8, 8, 6, 4, 4, 4). The peaks around 140 and 160 deviate from the power law, although the presence of many small groups and a few large communities is clear. The regression line is $f(x) = b*x^{-\alpha}c$, Coefficients (with 95 % confidence bounds): $b = 41.27$ (−138.9, 221.4), $\alpha = 1.781$ (0.06896, 3.494), $c = 0.03919$ (−0.02474, 0.1031). Goodness of fit: SSE: 0.01607, R-square: 0.8996, Adjusted R-square: 0.8709, RMSE: 0.04791

6 Conclusions

The present work has examined a sample of equity mutual funds investing in European stocks and presents various analyses mainly based on the complex network approach applied to stock holdings. The main results show that stocks are connected through the mutual fund owners they have in common and that there are substantial differences in geographical allocation among the different European fund managers. Five larger European countries dominate the market of mutual funds. United Kingdom and Switzerland are the less overlapped, while the highest overlap is among Germany and France. Italy is quite close to Germany and France for the overlap and selection of stocks, but it shows a lower diversification, like the United Kingdom, by looking at funds as a whole. The belonging of UK and Swiss opt-outs of the Eurozone, whose consequence is the exchange rate risk for non domestic investors, could be a probable explanation for our results on community detection that give a snapshot of a sort of "geographical organization" of the core of mutual fund portfolios, the part associated to the benchmark. The results of the different analyses provide valuable input for further research. In particular, we leave to future work the analysis of managers' behaviour non compliant with geographical allocation of holdings.

Acknowledgements The authors thank COST Action IS1104 for fruitful networking and financial support and Dr. Anna Romagnuolo for editing and proofreading.

References

Albert R, Barabasi AL (2002) Statistical mechanics of complex networks. Rev Mod Phys 74:47–97
Allen F, Babus A (2009) Networks in finance. In: Kleindorfer P, Wind J (eds) The network challenge. Wharton School Publishing, pp 367–382
Allen F, Gale D (2000) Financial contagion. J Polit Econ 108:1–33
Anton M, Polk C (2013) Connected stocks. J Finance LXIX(3):1099–1127
Aste T, Di Matteo T (2010) Introduction to complex and econophysics systems: a navigation map. In: Complex physical, biophysical and econophysical systems, pp 1–35
Augustiani C, Casavecchia L, Gray J (2015) Managerial sharing, mutual fund connections, and performance. Int Rev Finance 15:427–455
Ausloos M, Gligor M (2008) Cluster expansion method for evolving weighted networks having vector-like nodes. Acta Phys Pol A 114:491–499
Ausloos M, Miskiewicz J (2010) Entropy correlation distance method applied to study correlations between the gross domestic product of rich countries. Int J Bifurcat Chaos 20:381–389
Azmi RA, Smith R (2010) Contagion within financial markets and networks across the globe: evidence from equity mutual funds during the current crisis. Available at SSRN http://ssrn.com/abstract=1691157 or http://dx.doi.org/10.2139/ssrn.1691157
Barrat A, Barthelemy M, Pastor-Satorras R, Vespignani A (2004) The architecture of complex weighted networks. Proc Natl Acad Sci U S A 101:3747–3752
Battiston S, Glattfelder JB, Garlaschelli D, Lillo F, Caldarelli G (2010) The structure of financial networks. In: Estrada E, Fox M, Higham D, Oppo G-L (eds) Network science: complexity in nature and technology. Springer, London, pp 131–163

Battiston S, Puliga M, Kaushik R, Tasca V, Caldarelli V (2012) DebtRank: too central to fail? Financial networks, the FED and systemic risk, scientific reports 2, article number: 541 doi:10.1038/srep00541
Bellenzier L, Grassi R (2013) Interlocking directorates in Italy: persistent links in network dynamics. J Econ Interact Coord 9:183–202
Bellenzier L, Vitting Andersen J, Rotundo G (2015) Contagion in the world's stock exchanges seen as set of coupled oscillators. Economic Modelling (in press)
Bhattacharya K, Mukherjee G, Saramaki J, KaskiV, Manna SS (2008) The International Trade Network: weighted network analysis and modelling. J Stat Mech, P02002
Blocher J (2013) The externalities of crowded trades. Technical report, Vanderbilt Owen Graduate School of Management Research Paper No. 1968488. Available at SSRN: http://ssrn.com/abstract=1968488 or http://dx.doi.org/10.2139/ssrn.1968488
Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech Theory Exp (10), 10008
Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang DU (2006) Complex networks: Structure and dynamics. Phys Rep 424(4–5):175–308
Bonanno G, Caldarelli G, Lillo F, Mantegna RN (2003) Topology of correlation-based minimal spanning trees in real and model markets. Phys Rev E 68(4):046130
Bonanno G, Caldarelli G, Lillo F, Micciché S, Vandewalle N, Mantegna RN (2004) Networks of equities in financial markets. Eur Phys J B 38:363–371
Borgatti S, Mehra A, Brass D, Labianca G (2009) Network analysis in the social sciences. Science 323(5916):892–895
Boss M, Elsinger H, Summer M, Thurner S (2004) The network topology of the interbank market. Quant Finance 4:677–684
Bougheas S, Kirman AP (2014) Complex financial networks and systemic risk: a review. CESifo Working Paper Series 4756, CESifo Group, Munich
Braverman A, Minca A (2014) Networks of common asset holdings: aggregation and measures of vulnerability. SSRN: http://ssrn.com/abstract=2379669
Brookfield D, Boussabaine H, Su C (2013) Identifying reference companies using the book-to-market ratio: a minimum spanning tree approach. Eur J Finance 19(6):466–490
Caldarelli G (2008) Scale free networks. Oxford University Press, ISBN: 9780199211517
Caraiani P (2013) Using complex networks to characterize international business cycles. PLoS One 8(3), e58109. doi:10.1371/journal.pone.005810924
Cetorelli N, Peristiani S (2013) Prestigious stock exchanges: a network analysis of international financial centers. J Bank Finance 37(5):1543–1551
Chapelle A, Szafarz A (2005) Controlling firms through the majority voting rule. Phys A 355(2):509–529
Chapelle A, Szafarz A (2007) Control consolidation with a threshold: an algorithm IMA J Manag Math 18(3):235–243
Cohen L, Frazzini A, Malloy C (2008) The small world of investing: board connections and mutual fund returns. J Polit Econ 116(5):951–979
Cohen L, Malloy C, Frazzini A (2010) Sell side school ties. J Finance 65:1409–1437
Cont R, Moussa A, Santos EB (2013) Network structure and systemic risk in banking systems. No. hal-00912018
Croci E, Grassi R (2014) The economic effect of interlocking directorates in Italy: new evidence using centrality measures. Comput Math Organ Theory 20(1):89–112
D'Arcangelis AM, Rotundo G (2014) Mutual funds relationships and performance analysis. Qual Quant 49(4):1573–1584
D'Errico M, Grassi R, Stefani S, Torriero A (2009) Shareholding networks and centrality: an application to the Italian financial market. In: Naimzada AK, Stefani S, Torriero A (eds) Networks, topology and dynamics: theory and applications to economics and social systems. Lecture notes in economics and mathematical systems. Springer, Heidelberg, pp 215–228
De Benedictis L, Tajoli L (2011) The world trade network. World Econ 34:1417–1454

Del Giudice A, Marinelli N, Vitali S (2014) Sovereign wealth funds and target firms: does 'networking' matter? J Finance Manag Mark Inst 2:185–206
Elton EJ, Gruber MJ, Brown SJ, Goetzmann WN (2014) Modern portfolio theory and investment analysis, 9th edn
Flath D (1992) Indirect shareholding within Japan's business groups. Econ Lett 38:223–227
Gabbi G, Germano G, Hatzopoulos V, Iori G, Politi M (2012) Market microstructure, bank's behaviour and interbank spreads (report no. 12/06). Department of Economics, City University London
Gai P, Kapadia S (2010) Contagion in financial networks. Proc R Soc A. doi:10:1098=rspa:2009:0410
Gai P, Haldane A, Kapadia S (2011) Complexity, concentration and contagion. J Monet Econ 58(5):453–470
Garas A, Argyrakis P, Rozenblat C, Tomassini M, Havlin S (2010) Worldwide spreading of economic crisis. New J Phys 12:113043
Garas A, Schweitzer F, Havlin S (2012) A k-shell decomposition method for weighted networks. New J Phys 14, 083030
Garlaschelli D, Loffredo MI (2004) Fitness-dependent topological properties of the World Trade Web. Phys Rev Lett 93
Gligor M, Ausloos M (2007) Cluster structure of EU-15 countries derived from the correlation matrix analysis of macroeconomic index fluctuations. Eur Phys J B 57:139–146
Gligor M, Ausloos M (2008) Convergence and cluster structures in EU area according to fluctuations in macroeconomic indices. J Econ Integr 23:297–330
Gower JC (1986) Metric and Euclidean properties of dissimilarity coefficients. J Classif 3:5–48
Grassi R (2010) Vertex centrality as a measure of information flow in Italian Corporate Board Networks. Phys A 389(12):2455–2464
Grassi R, Patarnello A, Szpilska V (2008) Corporate board network and information flows in the Italian Stock Exchange (short paper). In: Proceedings of methods, models and information technologies for decision support systems (MTISD), Università del Salento, pp 110–112
Guo W, Minca A, Wang L (2015) The topology of overlapping portfolio networks. Available at SSRN: http://ssrn.com/abstract=2619514
Halaj G, Kok C (2015) Modeling emergence of the interbank networks. Quant Finance 15:653–671
Leitner Y (2005) Financial networks: contagion, commitment, and private sector bailouts. J Finance 60(6):2925–2953
López-Pintado D (2006) Contagion and coordination in random networks. Int J Game Theory 34(3):371–381
Lucarelli C, Molyneux P, Vitali S (2012) Network features of European trading venues. Available at SSRN: http://ssrn.com/abstract=1981843 or http://dx.doi.org/10.2139/ssrn.1981843
Maeng SE, Choi HW, Lee JW (2012) Complex networks and minimal spanning trees in international trade networks. Int J Mod Phys Conf Ser 16:51–60
Mantegna RN (1999) Hierarchical structure in financial markets. Eur Phys J B 11:193–197
Markose S, Giansante S, Shaghaghi AR (2012) Too interconnected to fail financial network of US CDS market: topological fragility and systemic risk. J Econ Behav Organ 83(3):627–646
Markowitz H (1952) Portfolio selection. J Finance 7(1):77–91
Miskiewicz J, Ausloos M (2006) G7 country gross domestic product (GDP) time correlations. A graph network analysis. In: Takayasu H (ed) Practical fruits of econophysics. Springer, Tokyo, pp 312–316
Miskiewicz J, Ausloos M (2010) Has the world economy reached its globalization limit? Phys A 389:797–806
Nature, focus issue on Complex networks in finance (2013) http://www.nature.com/nphys/focus/finance/index.html
Naylor M, Rose L, Moyle B (2007) Topology of foreign exchange markets using hierarchical structure methods. Phys A 382:199–208
Newman MEJ, Barabasi AL, Watts DJ (2006) The structure and dynamics of networks. Princeton University Press, Princeton

Pantaleo E, Tumminello M, Lillo F, Mantegna RN (2011) When do improved covariance matrix estimators enhance portfolio optimization? An empirical comparative study of nine estimators. Quant Finance 11(7):1067–1080

Pastor-Satorras R, Rub M, Diaz-Guilera A (eds) (2003) Statistical mechanics of complex networks. Springer, Berlin

Pozzi F, Aste T, Rotundo G, Di Matteo T (2008) Dynamical correlations in financial systems. In: Abbott D, Aste T, Bachelor M, Dewar R, Di Matteo T, Guttmann T (eds) Complex systems II. Proceedings SPIE 6802, 68021E. ISBN 978-1-60560-322-3

Rossi AG, Blake D, Timmermann A, Tonks I, Wermers R (2015) Network centrality and pension fund performance. CFR working papers no 15-16, University of Cologne, Centre for Financial Research

Rotundo G (2011) Centrality measures in shareholding networks. In: Duman E, Atiya A (eds) Use of risk analysis in computer-aided persuasion. NATO Science for Peace and Security Series E: Human and Societal Dynamics, 88, pp. 12–28. ISBN 978-1-60750-827-4 (print) ISBN 978-1-60750-828-1 (online), ISSN 1874-6276

Rotundo G, D'Arcangelis AM (2010a) Ownership and control in shareholding networks. J Econ Interact Coord 5(2):191–219

Rotundo G, D'Arcangelis AM (2010) Network analysis of ownership and control structure in the Italian Stock market. Advances and applications in statistical sciences, ISSN 0974-68119, Special Issue 2, pp 255–273

Rotundo G, D'Arcangelis AM (2014) Network of companies: an analysis of market concentration in the Italian stock market. Qual Quant 48(4):1893–1910

Sandoval L Jr, De Paula Franca I (2012) Correlation of financial markets in times of crisis. Phys A 391(1–2):187–208

Schweitzer F, Fagiolo G, Sornette D, Vega-Redondo F, Vespignani A (2009) Economic networks: the new challenges. Science 325(5939):422

Solis R (2009) Visualizing stock-mutual fund relationships through social network analysis. Global J Finance Bank Issues 3(3):8–22

Steinbacher M, Steinbacher M, Steinbacher M (2013) Credit contagion in financial markets: a network-based approach. Available via SSRN. http://papers.ssrn.com/sol3/papers.cfm? Abstract id = 2068716

Varela Cabo LM, Rotundo G, Ausloos M, Carrete J (2015) Complex networks analysis in socioeconomic models. In: Commendatore P, Kayam SS, Kubin I (eds) Complexity and geographical economics – topics and tools. Springer series; Dynamic modeling and econometrics in economics and finance, 19. ISBN 978-3-319-12805-4, pp 209–245

Vitali S, Glattfelder JB, Battiston S (2011) The network of global corporate control. PLoS One. doi:10.1371/journal.pone.0025995

Vitting Andersen J, Nowak A, Rotundo G, Parrott L, Martinez S (2011) Price-quakes shaking the world's stock exchanges. PLoS One 6(11):e26472. doi:10.1371/journal.pone.002647232

White L (2014) The basics of "Too Big to Fail". In: Schultz PH (ed) Perspective in Dodd-Frank and Finance. MIT Press

Part III
Dynamical Systems

A Formal Setting for Network Dynamics

Ian Stewart

Abstract This chapter is an introduction to coupled cell networks, a formal setting in which to analyse general features of dynamical systems that are coupled together in a network. Such networks are common in many areas of application. The nodes ('cells') of the network represent system variables, and directed edges ('arrows') represent how variables influence each other. Cells and arrows are assigned types, which determine the form of admissible differential equations-those compatible with the network structure. By analogy with the modern theory of dynamical systems, emphasis is placed on phenomena that are typical of entire classes of model equations with a given network structure, rather than on specific models. Such phenomena include symmetry and synchrony relations among cells, leading to a clustering effect embodied in a quotient network described by a balanced colouring. Rigid patterns of synchrony (those preserved by admissible perturbations) for equilibria and periodic states are classified by the balanced colourings. Bifurcations in which network structure can cause anomalous power-law growth rates are briefly mentioned. The formal concepts are motivated and explained in terms of typical examples.

Keywords Bifurcation • Dynamics • Network • Symmetry • Synchrony

1 Introduction

In recent years it has become increasingly apparent that networks play a highly significant role in many areas of science and technology. Examples include the spread of epidemics, food webs in ecosystems, gene regulation, intercellular signalling, neuroscience, market trading, control, and communications.

I. Stewart (✉)
Mathematics Institute, University of Warwick, Coventry CV4 7AL, UK
e-mail: I.N.Stewart@warwick.ac.uk

P. Commendatore et al. (eds.), *Complex Networks and Dynamics*,
Lecture Notes in Economics and Mathematical Systems 683,
DOI 10.1007/978-3-319-40803-3_10

The defining features of a network are a set of nodes, which interact through a system of connections. In mathematics, such a structure has traditionally been called a graph, but ‘network’ is more evocative. Nodes are also known as vertices or dots, and connections as edges or lines; these may or may not be directed. We will shortly rename nodes as ‘cells’ and directed edges as ‘arrows’ to emphasise the extra structure that will be brought into play.

Many different aspects of network structure and behaviour have been studied, ranging from statistical features to dynamics. Applications include the rate of spread of an epidemic, stability of the population distribution in an ecosystem, the development of organisms, broken connections in communications networks, ‘small world’ phenomena, stock market crashes, and internet search engines. The literature is vast, with many different viewpoints and philosophies, and we make no attempt to summarise it here. Instead, we focus on one specific area: the nonlinear dynamics of networks of coupled dynamical systems. By a dynamical system we mean a system of ordinary differential equations in one or more variables, which we abbreviate to ‘ODE’.

Just over a decade ago an analogy between symmetric dynamical systems (Golubitsky et al. 1988) and network dynamics began to be explored (Golubitsky and Stewart 2006; Golubitsky et al. 2005; Stewart et al. 2003). The aim was to apply, in a network context, the modern philosophy of nonlinear dynamics. This approach was pioneered by Poincaré (1881, 1882, 1885, 1886) in his work on the qualitative theory of differential equations. Among other things, this viewpoint led him to discover chaotic dynamics in the three-body problem for Newtonian gravitation (Poincaré 1892, 1893, 1899). His qualitative approach to differential equations was developed into a systematic theory by several mathematicians, especially in the Soviet Union, and became firmly established as a new branch of mathematics with the work of Arnold (1963), Smale (1967), and others. The central idea in this approach to dynamics is that significant structural phenomena are invariant under appropriate coordinate changes, and are thus determined purely by the *topology* of the trajectories in phase space—the phase portrait. For example, the presence of a time-periodic state (limit cycle) is a topological feature, but the detailed waveform, the period, and the shape of the cycle in phase space are not.

This approach deliberately ignores many details of the system, which have to be supplied by other means—typically numerical solutions, because few interesting nonlinear ODEs can be solved explicitly. So why do we need topological dynamics when any specific problem can be understood by numerical computation? Often we do not: numerical simulations provide all the answers required. However, numerical solutions sometimes make little sense on their own—they reveal some form of behaviour, but do not explain why it is occurring, or whether it is typical or unusual in the appropriate context. Moreover, most real-world models include numerical parameters that can take on many values, and it is often important to understand how the solutions change as parameters vary. Some numerical schemes exist that can explore such issues, but in general such questions may require infeasibly lengthy calculations. Moreover, it can be difficult to organise the results into a sensible description of the system’s behaviour. Topological dynamics can help here, because it provides a systematic framework for organising, classifying, and recognising the basic types of behaviour. It relates them to each other, and allows insights to be transferred from one area of application to others. It is in some ways a coarse instrument, but that is a virtue as well as a vice, because it removes inessential information.

The pioneers of the subject realised that the topological approach can be a highly effective approach to a basic, *general* question: ‘what can dynamical systems do?’. The effect of this change of viewpoint was a bit like the zoological move from butterfly-collecting to Linnaean taxonomy. Post-Linnaeus, you still had to collect butterflies to find out what existed in nature, but you began to appreciate how they related to other butterflies—and, more crucially, to other species.

Many special classes of dynamical system have extra structure. For example Hamiltonian systems are defined by a Hamiltonian function, which is conserved along trajectories and induces a symplectic structure, Smale (1967). Symmetric dynamical systems are defined by ‘equivariant’ vector fields with specific symmetry properties. In networks of coupled dynamical systems, the variables that appear in the differential equation, and the form of that equation, respect the network architecture. When the system has special structure, it is sensible to require the permissible coordinate changes to preserve this structure. This restriction can lead to new phenomena, invariant under this more limited type of coordinate change. Examples, in these three contexts, are the topology of energy levels, the symmetry group of a solution, and synchrony of specific nodes of a network.

Our focus here is on the network case. Network dynamics has been widely studied in many specific settings. Often the network structure is treated informally. However, it makes sense to develop a general overview by defining an appropriate formal structure, analogous to that for general dynamical systems. Here we survey some of the basic ideas in one systematic approach to this issue (Golubitsky and Stewart 2006; Golubitsky et al. 2005; Stewart et al. 2003). The main motivation in those papers was to seek analogies with symmetric dynamics (Golubitsky et al. 1988) and to devise alternatives when new issues arose. As they did.

1.1 Outline of Chapter

We begin by describing some examples of networks and their dynamics, to act as motivation. This leads to a formal definition of a coupled cell network and the corresponding class of ‘admissible’ differential equations. A key concept here is the input set of a node, which determines how the rest of the network is coupled to (drives) that cell.

Analogies with other special classes of dynamical systems help to motivate some basic questions and concepts. In particular, we take inspiration from symmetric dynamics, where the ODE respects a group of symmetry transformations.

An immediate obstacle arises, which causes technical difficulties, but cannot easily be avoided. In symmetric dynamics, the composition of two symmetric (that is, equivariant, see Golubitsky et al. 1988) maps is always symmetric. The analogue for network dynamics is false in general: the composition of two admissible maps need not be admissible. However, there is a partial substitute: strongly admissible maps. The composition of a strongly admissible map and an admissible map, in either order, is always admissible.

We include a brief discussion of symmetries of networks, an important area that combines (often in an uneasy alliance) features of symmetric dynamics and network dynamics.

One important issue in network dynamics is the possibility of synchrony, in which two (or more) cells have identical time series. One way to approach synchrony is through the concept of a balanced colouring of the cells. Suppose that the state of the network exhibits some pattern of synchrony; that is, certain cells are synchronous with others. Assign the same colour to all cells that are synchronous with each other. Intuitively, synchronous cells should receive the same input from the network: if not, the synchrony would be destroyed. The most natural way to ensure this is if the corresponding input sets match up in a manner that preserves colours. That is, cells with the same colour have inputs that are related by a colour-preserving permutation. This is the balance condition.

(An alternative is that some kind of cancellation of inputs takes place, but this would be 'accidental' and would disappear after a small admissible perturbation—unless the network equations have extra special features. In such cases, a generalised form of the balance condition must still apply.)

The above statement can be made precise. Balanced colourings define a distinguished class of subspaces of phase space that are invariant under any admissible map. The dynamics on this subspace leads to the pattern of synchrony determined by the colours. In contrast, an unbalanced colouring does not have this invariant subspace property.

If cells with the same colour are identified, the result is a 'quotient network' on a smaller number of cells, whose dynamics corresponds to synchronous dynamics in the original network with the corresponding pattern of synchrony.

There are some stronger results which apply to suitable equilibrium and periodic states. Say that a pattern of synchrony is rigid if it persists after any sufficiently small admissible perturbation of the ODE. Then rigid synchrony of equilibria defines a balanced colouring. So does rigid synchrony of periodic states; the current proof uses a mild technical hypothesis but it seems likely that this can be removed. There is also a version of this theorem for patterns of phase-related cells, rather than synchronous ones. It leads to a characterisation of conditions under which clusters of synchronous cells have a 'rotating wave' spatio-temporal symmetry.

The final topic is bifurcation theory, where states of the system undergo qualitative changes as some parameter varies. In particular we remark that network architecture can create anomalous bifurcation behaviour—that is, different form the typical bifurcations in general dynamical systems. As an illustration we exhibit an example of anomalous power-law growth of the amplitude of bifurcating branches of periodic states in a three-cell feed-forward network.

2 Network Diagrams and Admissible Maps

2.1 Motivation

The theory of networks goes back to the work of Euler (1741) on the puzzle of the Königsberg bridges. Contrary to common belief, he did not introduce the concept of a graph in its familiar geometric form; instead, he employed a symbolic representation of paths and argued combinatorially, Wilson (1985). However, the standard graphical representation soon followed. The main ingredients for a *graph* are a set of nodes, represented by dots, connected by a set of edges, represented by lines. The edges may be undirected (line segments) or directed (arrows). The main objects of study initially were the combinatorics and topology of graphs. As the subject developed, extra structure was imposed: directed edges were assigned numerical probabilities, connection strengths, flow rates, or durations (for example in critical path analysis).

In applications, especially to neuroscience, ODEs are associated with a given network, and the form of these equations reflects the network architecture. For example in a neuroscience model, nodes might represent neurons and edges axons, coupled via electrical signals passing along the axons. The state of each node i is represented by a variable x_i, which might be a scalar or a vector. Each node typically has an internal dynamic, an ODE that determines how it would behave if it were not coupled to other nodes. Connections from one node to another lead to coupling terms in the equations: if there is an input from node j to node i, then $\mathrm{d}x_i/\mathrm{d}t$ is a function of both x_i and x_j.

Example 1 The βIG model of diabetes, Topp et al. (2000), takes the form

$$\begin{aligned}
\dot{G} &= a - (b + cI)G \\
\dot{I} &= \beta\left(\frac{dG^2}{e + G^2}\right) - fI \\
\dot{\beta} &= (-g + hG - iG^2)/\beta
\end{aligned}$$

Here dots are time derivatives. The terms G = glucose level, I = insulin level, and β = beta-cell mass depend on time t. The other terms $a, b, c, d, e, f, g, h, i$ are parameters, whose value is constant during any particular run of the model or in any particular real system.

The network structure arises when we consider which variables depend on which. Here:

- The change in G depends on G, I but not on β.
- The change in I depends on G, I, and β.
- The change in β depends on G, β but not on I.

It is natural to encode these relationships as the network (called a block diagram in some areas of application) shown in Fig. 1. Here each variable is represented by a cell symbol (circle, square, hexagon) and arrows show which variable affects any given cell variable. The different cell symbols indicate different 'cell types', meaning that the form of the equation is different for those cells. The different arrow symbols (solid, dotted, and so on) indicate different 'arrow types', meaning that the form of the coupling is different for those cells.

In such a representation individual cell or arrow symbols have no further meaning on their own. Their interpretation depends on the entire diagram. For example, in this case the coupling from G and β to I is not a sum of terms in G and β separately, but a combination of both variables. Coupling terms need not be additive; for example in the equation for G the variable I appears as a product cIG.

Example 2 Consider an ODE representing three coupled FitzHugh-Nagumo neurons:

$$\begin{aligned}\dot{v}_1 &= v_1(a - v_1)(v_1 - 1) - w_1 - cv_2 \quad \dot{w}_1 = bv_1 - \gamma w_1\\ \dot{v}_2 &= v_2(a - v_2)(v_2 - 1) - w_2 - cv_3 \quad \dot{w}_2 = bv_2 - \gamma w_2\\ \dot{v}_3 &= v_3(a - v_3)(v_3 - 1) - w_3 - cv_1 \quad \dot{w}_3 = bv_3 - \gamma w_3\end{aligned} \tag{1}$$

Here v_i is the membrane potential of cell i, w_i is a surrogate for an ionic current, and a, b, γ are parameters with $0 < a < 1, b > 0, \gamma > 0$.

In (1) the dynamic equations are the same for each neuron, subject to appropriate permutations of the variables. In other words, the individual neurons are identical, and the couplings are also identical. So in this case the natural diagram is a ring of three identical cells (same cell symbol) with identical unidirectional coupling (same arrow symbol). See Fig. 2.

The state space of cell i is now 2-dimensional, with variables (v_i, w_i). Because the variables enter the equations in the same manner for each i, subject to the cyclic

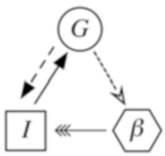


Fig. 1 Network representation of the βIG model

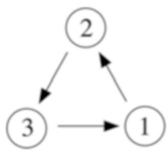


Fig. 2 Network representation of a ring of three identical FitzHugh-Nagumo neurons with identical unidirectional coupling

permutation, the cells have the same type and so do the arrows. In the diagram, we represent this by using circles for all three cells and the same kind of arrow for all three couplings.

2.2 Modelling

A network diagram does not specify an ODE as such. In particular it tells us nothing about the functional form of the equations. Instead, the diagram acts as a schematic representation of which variables affect which, and specifies when the same equation arises for corresponding variables. So each diagram determines a *class* of ODEs that 'respect the network structure'. Moreover, certain dynamical features may be common to all ODEs in this class, and are thus typical features for that network. These include possible patterns of synchrony, phase relations in periodic states, and a singularity-theoretic interpretation of homeostasis (Golubitsky and Stewart 2016a,b). Other features depend on the precise equations. So the formal theory separates the features that are typical for all networks with a given diagram from those that are special, and depend on the precise terms in the equations.

In a conventional approach to modelling, the equations are set up from the beginning using specific terms that reflect known (or presumed) aspects of the biology or physics of the system being modelled. For example, the term $dG^2/(e + G^2)$ in the βIG model tends to a constant d for large G, modelling a feature of the insulin response to large glucose levels. Other ODEs consistent with the network architecture need not behave in that manner, but would probably not be appropriate to model diabetes. Having set up specific equations that incorporate various assumptions of this kind, they can then be studied analytically or numerically to see how solutions behave. In circumstances when there is strong justification for choosing a particular formula, this type of model is an accurate representation of the real system.

However, especially in biology, there is often a lot of flexibility in the choice of formula, and the literature typically considers many variants. This is where the 'model-independent' philosophy presented here differs from this conventional 'model-dependent' approach. It offers some advantages by distinguishing between aspects of the solution that are sensitive to the precise formula employed, and those that are relatively robust and depend mainly on the network architecture. Specific models are still important; for example, to work out which parameter values lead to particular types of behaviour. But they can be used in the context of knowledge of what kind of behaviour should be expected on the basis of the network structure. This avoids the danger of attributing predicted behaviour to a specific formula, when it is mainly a result of the network structure and would occur for other formulas.

This viewpoint shifts the emphasis to a two-stage approach. First, understand model-independent features. Second, consider model-dependent features in the context of the model-independent ones to find out what extra information or insight the specific choice of model adds. The first step motivates defining a formal setting for network dynamics and working out the general principles that apply. The initial aim is to use the network structure to define a natural *class* of differential equations whose structure is compatible with a given network. We say that these ODEs are 'admissible' for that network.

There are several general formulations in the literature. For example, Kuramoto (1984) considers nonlinear internal dynamics plus linear coupling:

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = f_i(x_i) + \sum_j a_{ij} x_j \tag{2}$$

with nonlinear f_i and constants a_{ij} for some set of input nodes j. The idea is that each cell has a nonlinear internal dynamic f_i, and the couplings are linear, given by the matrix (a_{ij}). The form (2) can be motivated as a pragmatic low-order approximation to more complicated equations, where linearity corresponds to weak coupling, but this form of coupling is very special. In particular it is not preserved by any obvious type of coordinate change beyond linear maps, contrary to the spirit of topological dynamics. However, it also has some advantages: a specific internal dynamic $f_i(x_i)$, and removal of couplings by setting the relevant a_{ij} to zero.

Another common choice is to assume that the nodes represent 'phase oscillators', whose state space is a circle $\mathbb{S}^1$, and a state $\theta \in \mathbb{S}^1$ describes the phase of the oscillator. In this model the amplitudes of the oscillations are ignored.

Which formalism is appropriate depends on the questions being asked. The choice described in this chapter avoids restrictive assumptions on the form of the ODEs. It therefore provides a suitable context to study 'generic' or 'typical' phenomena in network dynamics, offering a useful perspective on more specific models, and it helps to explain some of their features.

3 Coupled Cell Networks and Systems

We now begin to set up a formal structure for network dynamics.

For reasons loosely related to the motivating examples, and to distinguish the topic from standard graph theory, the terms 'node' and 'directed edge' were replaced by 'cell' and 'arrow' in early work. For consistency with the literature, we do the same here.

Definition 1 A *coupled cell network* satisfies the following conditions:

(1) There is a finite set $\mathscr{C}$ of *cells*, usually identified with the standard set $\mathscr{C} = \{1, 2, \ldots, n\}$.
(2) There is a finite set $\mathscr{E}$ of *arrows*.
(3) Each arrow e has a *head cell* $\mathscr{H}(e) \in \mathscr{C}$ and a *tail cell* $\mathscr{T}(e) \in \mathscr{C}$.
(4) Cells are classified into *types*. Formally, this is done by defining an equivalence relation $\sim_C$ on $\mathscr{C}$, called *cell equivalence*. Cells are equivalent if they have the same type.

(5) Arrows are also classified into *types* by defining an equivalence relation $\sim_E$ on $\mathscr{E}$, called *arrow equivalence*. Arrows are equivalent if they have the same type.
(6) Types satisfy two compatibility conditions. If $e_1, e_2 \in \mathscr{E}$ are arrow-equivalent, then $\mathscr{H}(e_1)$ and $\mathscr{H}(e_2)$ are cell-equivalent, and $\mathscr{T}(e_1)$ and $\mathscr{T}(e_2)$ are cell-equivalent.

From now on we often shorten 'coupled cell network' to 'network'. A network can be represented graphically by its *diagram*. Here cells are drawn as dots, circles, squares, hexagons, and so on, with a different symbol for each type. Arrows are drawn as arrows, similarly decorated to distinguish types by using dotted or wavy lines, different shapes of arrowhead, and so on. Each arrow e runs from $\mathscr{T}(e)$ to $\mathscr{H}(e)$. The diagram is a directed labelled graph, where the 'labels' are graphical representations of the cell and arrow types.

Warning: An arrow can have the same head and tail, forming a *self-connection* from a cell to itself. Two distinct arrows (of the same or different types) can have the same head and the same tail, giving *multiple connections* between the two cells. Arrows of this kind arise naturally in connection with a basic construction, the 'quotient network', which is related to synchrony; the entire formalism works much better if they are permitted from the start. See Sect. 7.

Figure 3 shows a few examples, and we take the opportunity to illustrate some basic types of network *architecture* (that is, topology) at the same time.

Tacit conventions are often used to simplify such diagrams. For example in the 'all-to-all' network, pairs of equivalent arrows in opposite directions are shown as

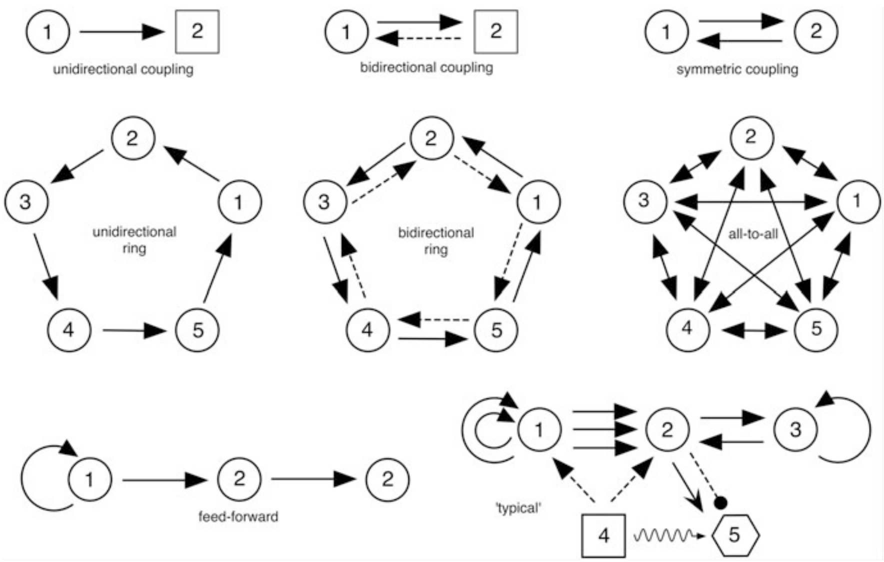


Fig. 3 A sample of coupled cell networks

a single line with two heads. Some examples have a single cell type and a single arrow type; others do not. The final 'typical' example illustrates a few possibilities consistent with the formalism, and has no special significance.

A network is *connected* if the underlying graph (ignoring cell and arrow types and arrow directions) is connected; that is, any two nodes are joined by a path of mutually adjacent edges. It is *path-connected* (another term widely used is *transitive*) if any two nodes of the underlying graph are joined by a directed path of mutually adjacent edges. It is *disconnected* if it is not connected, in which case it breaks up into connected components. The examples include some self-connections, multiple arrows (of the same type or different types), and a multiple self-connection.

3.1 Global Symmetries

Symmetries of ODEs have a strong effect on their solutions (Golubitsky et al. 1988; Golubitsky and Stewart 2002a). We therefore make a few remarks about symmetries here and expand on them later.

A *(global) symmetry* of a network is a permutation of its cells that preserves the network architecture: how many arrows of each type input to each cell, and how they are connected in the network. Among the examples in Fig. 3, the two-cell network labelled 'symmetric coupling' has symmetry group $\mathbb{Z}_2$, generated by the transposition (12). The unidirectional ring has cyclic group symmetry $\mathbb{Z}_5$ generated by the 5-cycle (12345). So does the bidirectional ring, as drawn, because it has two types of arrow. If the dotted arrows were of the same type as the solid ones, the symmetry group would be the dihedral group $\mathbb{D}_5$. The all-to-all connected network has symmetry group $\mathbb{S}_5$, consisting of all permutations of the cells.

The other networks illustrated have trivial symmetry group.

Networks can also have 'local' symmetries, known formally as input isomorphisms, see Definition 2. These have a significant influence, and are central to network dynamics, but their role is less transparent.

4 Admissible Maps

We repeat that the central role of a coupled cell network is to encode a space of ODEs whose couplings model the architecture of the network. We then seek features that are 'typical' for all ODEs in this space, as explained below. The formalism does not, and is not intended to, pin down a specific ODE. Instead, it determines a class of ODEs compatible with the network, allowing us to distinguish features that are typical of this class from those that are not.

4.1 Cell Phase Spaces

In order to define an ODE, or a class of them, we need to specify the variables, or *phase space*, and the functions that appear as components of the vector field. For networks, we choose variables that respect the network structure. For each cell $c \in \mathscr{C}$ define a *cell phase space* P_c. In general, this should be a smooth manifold. To avoid too much manifold formalism (tangent bundles on the like) we will assume for most of this chapter that $P_c = \mathbb{R}^{n_c}$ is a real vector space. For local bifurcation theory, this case is all we need. However, for some purposes other choices are necessary; in particular, systems of *phase oscillators* correspond to choosing $P_c = \mathbb{S}^1$, the circle.

The role of cell-equivalence is to identify the phase spaces of equivalent cells. That is, if cells c, d are cell-equivalent then P_c and P_d are required to be equal. The overall phase space of the network is the direct sum

$$P = \bigoplus_{c \in \mathscr{C}} P_c$$

4.2 Input Sets

Networks have a new feature, compared to symmetric systems. Not only can they have global symmetries: they can have 'partial symmetries' in which some subnetwork has the same structure as some other subnetwork. This concept is most useful when the subnetworks concerned encode the inputs to cells, because a partial symmetry of this type in effect states that the cells concerned 'have the same kinds of couplings'.

Definition 2 Let $c, d \in \mathscr{C}$. The *input set* of c is the set $I(c)$ of all arrows e such that $\mathscr{H}(e) = c$.

An *input isomorphism* $\beta : I(c) \to I(d)$ is a bijection between their input sets that preserves arrow type. That is, e is arrow-equivalent to $\beta(e)$ for all β and all $e \in I(c)$. (It follows that $\beta^{-1}(f)$ is arrow-equivalent to f for all $f \in I(d)$.)

If there exists an input-isomorphism $\beta : I(c) \to I(d)$ we say that c, d are *input-isomorphic* or *input-equivalent*.

The input set is important because it encodes which cells are connected to which, and by which type of arrow. Input-equivalent cells receive the same couplings from the relevant cells of the network.

Example 3 Let $\mathscr{G}$ be the network of Fig. 4. The input sets of the five cells are shown in Fig. 5.

It is clear from the figure that cells 1 and 2 are input-isomorphic, and so are cells 3 and 5. However, cells 1 and 3 are not input-isomorphic. Although they both receive two inputs, the arrow-types are different.

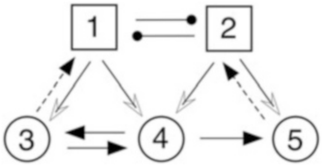


Fig. 4 A 5-cell network

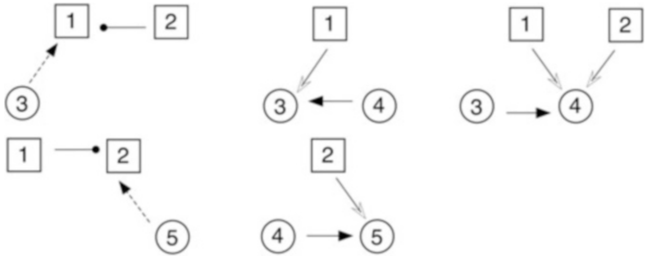


Fig. 5 Input sets of the 5-cell network. *From left to right*: $I(1), I(3), I(4), I(2), I(5)$. Strictly, the *arrows* constitute the input set, but is convenient to show the head and tail cells as well

The set of all input-isomorphisms from cell c to cell d is denoted by

$$B(c, d)$$

These maps are closed under composition in the following sense. If a, b, c are input-equivalent cells, and $\alpha \in B(a, b), \beta \in B(b, c)$, then $\beta\alpha \in B(a, c)$. Composition is not always defined, but when it is, it is associative.

It follows that for any $c \in \mathscr{C}$ the set $B(c, c)$ is a finite group, the *vertex group* of c. 'Cell group' might be a more more consistent choice of terminology, but this choice avoids overusing the word 'cell'.

The union $\mathscr{B} = \bigcup_{c,d} B(c, d)$ is in general not a group, because its elements may not compose. Technically, it is a *groupoid*, Brandt (1927), Higgins (1971) and Brown (1987). The groupoid structure can be viewed as a side-effect of the formalism rather than a vital ingredient. It does have a few useful implications, but its main influence to date has been through the vertex groups. See Golubitsky and Stewart (2006) for further discussion.

4.3 *Admissible Maps*

To each network $\mathscr{G}$, and each specific choice of cell coordinates x_c that preserves cell type,we associate the space of all ODEs that are compatible with the network architecture. Such ODEs are called *coupled cell systems* or *network ODEs*.

To define these ODEs, we associate to $\mathscr{G}$ a space of *admissible vector fields*. When all P_i are real vector spaces, we refer to these as *admissible maps*. (The tangent space of $\mathbb{R}^n$ at any point is $\mathbb{R}^n$.)

Example 4 Consider once more the network of Fig. 4. Here there are two cell types and four arrow types.

Choose coordinates $(x_1, x_2, x_3, x_4, x_5)$ for cells $1, 2, 3, 4, 5$. By cell-equivalence, $P_2 = P_1$ and $P_3 = P_5$. Admissible ODEs take the following form:

$$\begin{aligned} \dot{x}_1 &= f(x_1, x_2, x_3) \\ \dot{x}_2 &= f(x_2, x_1, x_5) \\ \dot{x}_3 &= g(x_3, x_1, x_4) \\ \dot{x}_4 &= h(x_4, \overline{x_1, x_2}, x_3) \\ \dot{x}_5 &= g(x_5, x_2, x_4) \end{aligned} \tag{3}$$

for arbitrary smooth functions

$$\begin{aligned} f &: P_1 \times P_2 \times P_3 \to P_1 \\ g &: P_3 \times P_1 \times P_4 \to P_3 \\ h &: P_4 \times P_1 \times P_2 \times P_3 \to P_4 \end{aligned}$$

(The overline in the fourth equation indicates symmetry, see below.)

First, we explain how this form is obtained from the network. Consider the first equation, for cell 1. The vector field component is $f(x_1, x_2, x_3)$. The first entry x_1 is the cell coordinate, and it represents the internal state of that cell. The other two entries x_2, x_3 are the input coordinates—those of the tail cells of the two input arrows to cell 1, as in Fig. 5. Similarly the equations for cells 2–5 comprise the cell coordinate and the input coordinates, with the cell coordinated being distinguished. We do this because the cell coordinate is not represented by an arrow. (It would be possible to add an explicit self-connection to represent this variable; however, this arrow would naturally be distinguished from any other self-connections in any case.)

A glance at Fig. 5 shows that cells 1 and 2 are input-equivalent; that is, they have have the same input sets aside from the numbering of cells. Each cell receives one dashed arrow and one arrow with a dot. Admissibility means that the same function f occurs for cells 1 and 2. The variables are written in an order that respects this equivalence: corresponding variables come from tail cells of arrows of the same type.

The equation for cell 3 has a different function g, because cell 3 is not input-equivalent to cells 1 or 2. Because cell 5 is input-equivalent to cell 3, we use the same g in that equation, with variables again corresponding via the input isomorphism.

In cell 4 we encounter a new feature. Two input arrows are equivalent, those from cells 1 and 2. Therefore there exists an input-isomorphism from $I(4)$ to itself, which swaps these two arrows. Admissibility requires h to be symmetric in those two variables; that is, $h(x_4, x_1, x_2, x_3) \equiv h(x_4, x_2, x_1, x_3)$. Conventionally the overline on the variables x_1, x_2 in (3) indicates this symmetry.

We mention one feature of the formalism that is sometimes misunderstood. When symmetries of this kind are not appropriate in a model, they should be removed by drawing the network using distinct arrow types. Symmetry is an *option*, not a general requirement.

We now describe, informally, a procedure for writing down admissible maps. Formal definitions are given in Golubitsky et al. (2005), Sect. 3.

For each cell $c \in \mathscr{C}$, choose *cell coordinates* x_c on P_c. (In general, x_c may be multidimensional.) Phase space P then comprises all n-tuples

$$x = (x_c)_{c \in \mathscr{C}}$$

A vector field on P, adapted to cell coordinates, comprises components $f_c, c \in \mathscr{C}$ such that

$$f_c : P \to P_c$$

For admissibility we impose extra conditions on the f_c that reflect network architecture, as follows:

Definition 3 Let $\mathscr{G}$ be a network. A vector field $f : P \to P$ is $\mathscr{G}$*-admissible* if:

(1) *Domain Condition*: For every cell c, the component f_c depends only on the cell variable x_c and the input variables $x_{\mathscr{T}(e)}$ where $e \in I(c)$.
(2) *Symmetry Condition*: If c is a cell, f_c is invariant under all permutations of tail cell coordinates for equivalent input arrows.
(3) *Pullback Condition*: If cells $c \neq d$ are input-equivalent, the components f_c, f_d are identical as functions. The variables to which they are applied correspond under some (hence any, by condition (2)) input-isomorphism.

Formally, conditions (2) and (3) are combined into a single pullback condition applying to any pair c, d of cells, equal or different.

Example 4 exhibits consequences of all three conditions.

Associated with any admissible map f is an *admissible ODE* or *coupled cell system*

$$\frac{dx}{dt} = f(x) \tag{4}$$

If f also depends on a (possibly multidimensional) parameter λ, and is admissible as a function of x for any fixed λ, we have an *admissible family* of maps and ODEs. Such families arise in bifurcation theory.

4.4 Strongly Admissible Maps

A special class of admissible maps plays a key role in the theory, mainly as a technical tool in proofs. Equivariant dynamics has a very useful feature: composing two equivariant maps yields an equivariant map. However, simple examples show that admissible maps often lack this property. It can be regained by considering a more restrictive class of maps:

Definition 4 A *strongly admissible* map is a map g such that:

(1) g_c depends only on x_c for each cell c.
(2) If c, d are cell-equivalent then $g_c = g_d$.

It follows that $g(x) = (g_1(x_1), \ldots, g_n(x_n))$, where $g_c = g_d$ whenever c, d are cell-equivalent.

Proposition 1 *Let $f : P \to P$ be admissible and let $g : P \to P$ be strongly admissible. Then*

(1) *If g is invertible (that is, a diffeomorphism) then g^{-1} is also strongly admissible.*
(2) *Both fg and gf are admissible.*

For some networks, other types of map can compose with admissible maps to give admissible maps. See Golubitsky and Stewart (2016b).

5 Global Symmetries

The formalism for networks introduced in Golubitsky et al. (2005) and Stewart et al. (2003) originally emerged from symmetric dynamics, specifically symmetric networks of coupled oscillators, for example Golubitsky and Stewart (1986). We enlarge on our earlier remarks about global symmetry and make them more precise.

Definition 5 Let $\mathcal{G}$ be a network with cells $\mathcal{C}$ and arrows $\mathcal{E}$. A *(global) symmetry* of $\mathcal{G}$ is a permutation π of $\mathcal{C}$ such that the set of arrows from cell c to cell d is isomorphic to the set of arrows from cell $\pi(c)$ to cell $\pi(d)$. That is, the number of arrows of given type is the same in both cases. (It therefore extends naturally to a permutation acting on $\mathcal{E}$ that preserves arrow-type, but it is more convenient to consider the action on cells. The two formulations are equivalent.)

The*(global) symmetry group* of $\mathcal{G}$ is the group formed by all such permutations π, and is denoted by $\mathrm{Sym}(\mathcal{G})$.

The action of π on arrows induces one on cells, by requiring $\pi(\mathcal{T}(e)) = \mathcal{T}(\pi(e))$, or $\pi(\mathcal{H}(e)) = \mathcal{H}(\pi(e))$, or both. (These conditions are consistent because equivalent arrows have equivalent heads and tails.)

There is a connection between admissible maps and symmetric (that is, *equivariant*) maps. These satisfy

$$f(\pi(x)) = \pi f(x)$$

where π acts by permuting indices on x_c and f_c.

Theorem 1 *Any $\mathcal{G}$-admissible map is* $\mathrm{Sym}(\mathcal{G})$*-equivariant.*

Example 5 In general the converse is not true: equivariant maps need not be admissible. The 'easy' way for this to occur is when the functions have the wrong domains. But satisfying the domain condition and being equivariant need not imply admissibility. To see why, consider Fig. 6.

This network has dihedral group $\mathbb{D}_5$ symmetry, determined by all rotations and reflections of the pentagon. There are two types of arrow: short-range (solid) and long-range (dashed).

Consider a global symmetry that fixes cell1. It is either the identity, or it acts on cells by the reflectional permutation (25)(34).

The vertex group $B(1, 1)$ is larger. Because there are no multiple arrows, we can define its action on arrows by considering the effect on their tail cells. It contains the identity, (25)(34), but also (25) and (34) on their own. Here (25) interchanges the short arrows inputting to cell 1, and (34) interchanges the long arrows inputting to cell 1.

The map

$$f = \begin{bmatrix} x_2x_4 + x_3x_5 \\ x_1x_4 + x_3x_5 \\ x_1x_4 + x_2x_5 \\ x_1x_3 + x_2x_5 \\ x_1x_3 + x_2x_4 \end{bmatrix}$$

is $\mathbb{D}_5$-equivariant but not admissible. It is obtained by making f_1 invariant under (25)(34) but not under (25) or (34). Then we use pullback to define the other components.

An analogous admissible map would have $f_1(x) = x_2x_4 + x_2x_3 + x_3x_5 + x_4x_5$, invariant under the whole of $B(1, 1)$.

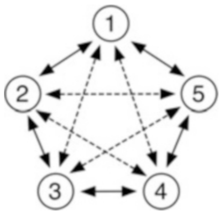


Fig. 6 Network with dihedral group $\mathbb{D}_5$ symmetry

Examples like this need to be borne in mind when applying equivariant dynamics and bifurcation theory to symmetric networks. In principle the extra constraints on admissible maps could change the generic behaviour. This effect occurs, for example, in steady-state bifurcation for some regular networks (Stewart 2014; Stewart and Golubitsky 2011), causing higher singularities to be generic. Such networks, however, are very unusual.

5.1 Fixed-Point Subspaces

In equivariant bifurcation theory, it is proved that any symmetric ODE possesses a class of subspaces that are invariant under any equivariant map. These are the *fixed-point subspaces* of subgroups Σ of the overall symmetry group Γ, defined by

$$\mathrm{Fix}(\Sigma) = \{x : \sigma x = x \ \ \forall \sigma \in \Sigma\}$$

Suppose that the system concerned is an admissible ODE for a symmetric network. Since all admissible maps are equivariant, $\mathrm{Fix}(\Sigma)$ is invariant under all admissible maps. Antoneli and Stewart (2006, 2007, 2008) explore links between symmetry and synchrony in networks, showing in particular that there can be subspaces other than fixed-point subspaces with this invariance property—even when arrows are deemed equivalent if and only if they are related by a symmetry. This again shows that it is necessary to be careful when applying equivariant dynamics to symmetric networks; however, examples of this type are also rare.

6 Quotient Networks and Synchrony

A basic question in network dynamics is: when are two cells synchronous? We define synchrony by identical time-series: if $x = x(t)$ is a solution of an admissible ODE, we say that cells c, d are *synchronous* on x if $x_c(t) = x_d(t)$ for all times t. This definition is a strong one, and many applications employ a weaker version in which the time series are close together, or are equal most of the time. However, it lets us prove precise theorems that yield useful insights.

A very strong kind of synchrony occurs for *any* admissible ODE, and is associated with a subspace of phase space that is invariant under all admissible maps f. Here cells synchronise in clusters, so that all cells in a given cluster have identical time-series. To introduce this idea we return to Example 4.

Example 6 In Fig. 7 (right) we have assigned 'colours' to the cells, shown as grey shading and diagonal hatching. In this example, cells 1 and 2 have the same colour, and cells 3 and 5 have the same colour. So the set of cells $\mathscr{C}$ is partitioned into three subsets, determined by 'same colour'; namely $\{1, 2\}, \{3, 5\}, \{4\}$. (Technically

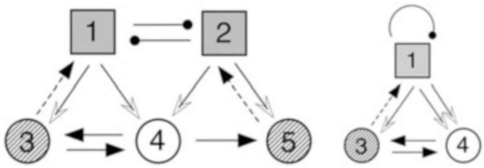


Fig. 7 *Left*: Balanced colouring of the 5-cell network. *Right*: Corresponding quotient network

these can be considered as the equivalence classes for the equivalence relation 'same colour', but intuitively it seems simpler to think about colours.)

A given network can be coloured in many ways, but this choice has a special feature, which becomes apparent if we look for solutions in which cells of the same colour are synchronous. That is, we set $x_1 = x_2 = u, x_3 = x_5 = v, x_4 = w$, so

$$(x_1, x_2, x_3, x_4, x_5) = (u, u, v, w, v)$$

The admissible ODE (3) now becomes

$$\begin{aligned} \dot{u} &= f(u, u, v) \\ \dot{u} &= f(u, u, v) \\ \dot{v} &= g(v, u, w) \\ \dot{w} &= h(w, u, u, v) \\ \dot{v} &= g(v, u, w) \end{aligned} \tag{5}$$

Although we have five equations in only three unknowns, the system is not overdetermined because the second equation is the same as the first, and the fifth is the same as the third.

If we project (u, u, v, w, v) to (u, v, w) we get a *restricted ODE*

$$\begin{aligned} \dot{u} &= f(u, u, v) \\ \dot{v} &= g(v, u, w) \\ \dot{w} &= h(w, u, u, v) \end{aligned}$$

We recognise this as an admissible ODE for a smaller network, in which cells of the same colour are identified with a single cell, and input sets of arrows remain unchanged (but tail cells with the same colour are identified). This *quotient network* is shown in Fig. 7 (left).

This construction works because the space

$$\Delta = \{(u, u, v, w, v) : u \in P_1, v \in P_3, w \in P_4\}$$

is invariant under all admissible maps, hence under the flow of the corresponding ODEs. It has the pleasant feature that the space of restricted ODEs is *precisely* the space of admissible ODEs for the quotient network, provided the same cell coordinates are used.

Here the quotient network has a double arrow from cell 1 to cell 4, and a self-connection from cell 1 to itself. However, the original network does not have multiple arrows (pointing in the same direction). Multiple arrows and self-connections are natural consequences of the restricted ODE. The equation for $\dot{w}$ involves two entries u, corresponding to the two arrows from cell 1 to cell 4; the equation for $\dot{u}$ has two entries u: one for the cell coordinate and another for the input coordinate from cell 2. The 'single-arrow' network formalism in Stewart et al. (2003) failed to take proper account of this effect, leading to complications when characterising restricted ODEs (Dias and Stewart 2004). The modified 'multi-arrow' formalism of Golubitsky et al. (2005) relates the space of restricted ODEs to a network in a satisfactory manner by permitting multiple arrows and self-connections.

7 Balanced Colourings

It so happens that in Fig. 7 cells are coloured according to input-equivalence. However, this type of colouring does not always produce a consistent synchrony relation. The next step is to characterise those that do.

Definition 6 A *colouring* of a network $\mathscr{G}$ is a map

$$k : \mathscr{C} \to K$$

where K is a finite set, whose members are called *colours*.

We say that c, d *have the same colour* if $k(c) = k(d)$, and write $c \sim_k d$.

A colouring k of a network is *balanced* if whenever cells c, d have the same colour, there exists an input isomorphism $\beta : I(c) \to I(d)$ such that i and $\beta(i)$ have the same colour for all $i \in \mathscr{T}(I(i))$.

Informally, a colouring is balanced if there exists a *colour-preserving* input isomorphism for any two cells of the same colour. In particular, cells of the same colour must be input-equivalent, so a balanced colouring is a refinement of input equivalence. That is, if $c \sim_k d$ then $c \sim_I d$.

Definition 7 The *polydiagonal* defined by a colouring k of $\mathscr{G}$ is the space

$$\Delta_k = \{x \in P : k(c) = k(d) \implies x_c = x_d\}$$

That is, cells of the same colour are synchronous for $x \in \Delta$.

Theorem 2 *A polydiagonal Δ_k is invariant under every admissible map if and only if k is balanced.*

One consequence is that when k is balanced, initial conditions that have the pattern of synchrony defined by k (that is, lie in Δ_k) give rise to solutions that remain inside Δ_k. However, this result does not guarantee that the pattern of synchrony is stable: perturbations that break the synchrony could cause the solution to deviate from Δ_k instead of returning close to it. This kind of stability depends on the admissible vector field; more precisely, on its component transverse to Δ_k.

Definition 8 Let k be a balanced colouring on $\mathscr{G}$, with colour set K. The associated *quotient network* $\mathscr{G}_k$ has K as its set of cells (that is, there is one cell per colour).

The cell type of cell $i \in K$ is that of any cell $c \in \mathscr{C}$ with colour i (that is, $k(c) = i$).

The arrows in $I(i)$ in $\mathscr{G}_k$ are obtained from the input set $I(c)$ of any cell c with colour i by copying each arrow e to create an arrow with head $k(\mathscr{H}(e)$ and tail $k(\mathscr{T}(e))$, of the same type as e.

The set of arrows of $\mathscr{G}_k$ is the union of the $I(i)$ as i runs through K.

Deville and Lerman (2015) have reformulated the notion of quotient in a more general manner, in terms of network fibrations. Nijholt et al. (2016) have developed this idea in a very interesting manner to set up a form of semigroup equivariance for some classes for networks, which explains many hitherto puzzling phenomena.

Example 7 We now return to Example 6 in the light of the above definition of a balanced colouring.

First, we check that the colouring in Fig. 7 (left) is balanced.

Cells 1 and 2 have the same colour. So we must check that their input sets are coloured in the same manner.

Cell 1 has two input arrows: one from cell 2 (with a dot for its head) and one from cell 3 (dashed line).

Cell 2 has two input arrows: one from cell 2 (with a dot for its head) and one from cell 3 (dashed line).

The tail cells are (2,3) and (1,5) respectively. Corresponding cells 1 and 2 have the same colour, and corresponding cells 3 and 5 have the same colour.

Similarly, cells 3 and 5 have the same colour and their input sets match up in a way that preserves colours.

Finally, cell 4 has a different colour from all other cells so there is nothing more to check.

Figure 7(right) shows the corresponding quotient network. This has one cell for each colour. For convenience we label these by representatives 1, 3, 4 of those colours. Arrows are drawn to mimic the input sets in the original network, Fig. 7(left).

We emphasise that although in this particular case colours correspond to input-equivalence classes of cells, colouring by input-equivalence need not be balanced. On the other hand, many other balanced colourings may exist, depending on the network.

Theorem 2 is the first and weakest in a series of results that demonstrate the central role played by balanced colourings. Intuitively, the result is straightforward: if two cells remain synchronised as time passes, the inputs to those cells must also be synchronised. however, this does not necessarily imply that the *states* of those input cells are synchronised. Nonetheless, this ought to be the case for most admissible vector fields, and the proof of Theorem 2 is relatively straightforward: it just requires a sensible choice of admissible vector field.

Theorem 3 *Let k be a balanced colouring of $\mathscr{G}$. Then*

(1) *The restriction of any $\mathscr{G}$-admissible map to Δ_k is $\mathscr{G}_k$-admissible.*
(2) *Every $\mathscr{G}_k$-admissible map is a restriction to Δ_k of a $\mathscr{G}$-admissible map.*

Another way to say (2) is that every $\mathscr{G}_k$-admissible map on Δ_k *lifts* to a $\mathscr{G}$-admissible map on P.

If f is $\mathscr{G}$-admissible, the restricted map $f|_{\Delta_k}$ determines the dynamics under f of the synchronous clusters determined by the colouring k.

8 Rigid Synchrony for Equilibria

In dynamical systems theory an equilibrium x^0 of an ODE $\dot{x} = f(x)$ is said to be *hyperbolic* if no eigenvalues of the derivative (or Jacobian) $D_x f|_{x^0}$ lie on the imaginary axis. It can then be proved that if g is a small perturbation of f there exists a unique equilibrium y^0 of the ODE $\dot{x} = g(x)$ with y^0 near x^0. See Hirsch and Smale (1974) and Guckenheimer and Holmes (1983).

Definition 9 A hyperbolic equilibrium x^0 of a network ODE $\dot{x} = f(x)$ is *rigid* if its pattern of synchrony is preserved by any sufficiently small admissible perturbation. That is, suppose that $g = f + \varepsilon p$ is any admissible perturbation of f and ε is sufficiently small. Let y^0 be the unique perturbed equilibrium near x^0. Then whenever $x^0_c = x^0_d$, we have $y^0_c = y^0_d$.

Golubitsky et al. (2005) prove the Rigid Equilibrium Theorem:

Theorem 4 *Let x^0 be a hyperbolic rigid equilibrium of a network ODE. Define the relation $\sim$ by $c \sim d \iff y^0_c = y^0_d$ for the perturbed equilibrium y^0 of any sufficiently small admissible perturbation of f. Then $\sim$ is balanced.*

Briefly: rigid synchrony patterns of equilibria are balanced. Another very different proof can be found in Aldis (2010).

9 Rigid Synchrony and Phase Relations for Periodic States

The Rigid Equilibrium Theorem 4 has an analogue for periodic states. We introduce this idea with an example, the coupled FitzHugh-Nagumo equations (1) represented by a ring of three identical cells with unidirectional coupling as in Fig. 2. This network has $\mathbb{Z}_3$ symmetry, which has implications for periodic states.

When $a = b = \gamma = 0.5$ and $c = 0.8$, the origin is a stable equilibrium for the full six-dimensional system, and the cells undergo a *synchronous* oscillation. That is, their time-series are identical. However, when $a = b = \gamma = 0.5$ and $c = 2$, the system has a stable periodic state in which successive cells are one third of a period out of phase. Figure 8, which shows the pattern for the v_j; the same pattern occurs for the w_j. This state is a *discrete rotating wave*. It has *spatio-temporal* symmetry:

$$x_2(t) = x_1(t - T/3) \qquad x_3(t) = x_1(t - 2T/3)$$

That is, $x(t)$ is invariant if we permute the labels using the 3-cycle $\rho = (123)$ *and* shift phase by $T/3$. So

$$\rho x(t + T/3) = x(t)$$

where $x_i = (v_i, w_i)$. Thus $x(t)$ is fixed by $(\rho, T/3) \in \Gamma \times \mathbb{S}^1$, where $\mathbb{S}^1$ is the circle group of phase shifts modulo the period.

The Equivariant Hopf Theorem (Golubitsky and Schaeffer 1985; Golubitsky and Stewart 2002b; Golubitsky et al. 1988) provides conditions under which phase-related states of this type occur; the H/K Theorem (Buono and Golubitsky 2001) classifies the possible spatio-temporal symmetries. This theorem has been applied to analyse central pattern generators for quadruped locomotion. Different gait patterns exhibit different phase relations between various legs, and these can be read off from the network structure of the central pattern generator by considering symmetries. See Buono (2001), Buono and Golubitsky (2001), Collins and Stewart (1993a,b), Golubitsky and Stewart (2002a).

Example 8 Figure 9 shows a chain of 7 identical cells with identical couplings, driven by a ring of three cells 1, 2, 3. (There is nothing special about the numbers here, and both 3 and 7 can be replaced by arbitrary positive integers for appropriate chains.)

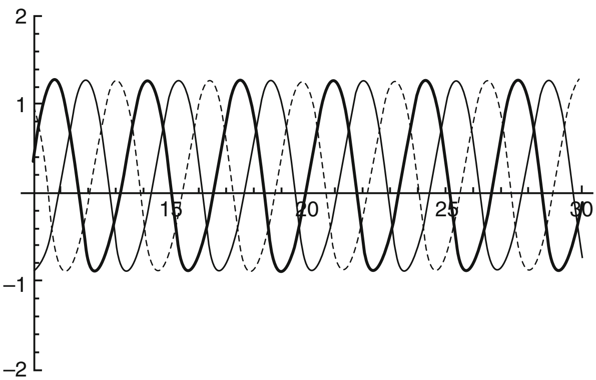


Fig. 8 Periodic oscillations of the 3-cell ring exhibiting a $\frac{1}{3}$-period out of phase periodic solution. Time series of v_1 (*thick solid*), v_2 (*thin solid*), v_3 (*dashed*)

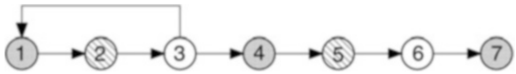


Fig. 9 Balanced colouring of a feed-forward chain leading to travelling wave

The colouring shown is balanced, and the corresponding quotient network is the $\mathbb{Z}_3$-symmetric ring of Fig. 2. With suitable admissible equations, this ring supports a rotating wave with 1/3 period phase shifts as above. Therefore, lifting, the original chain supports a state with three synchronous clusters, formed by cells $\{3k+1\}, \{3k+2\}, \{3k\}$, with x_i, x_{i+1} being synchronous except for a phase shift of one third of a period. The effect is similar to a travelling wave in which cells 1, 2, 3, 4, ... 'fire' in turn, and cells $i, i-3, i-6, \ldots$ are synchronous.

The 7-cell chain has no global symmetry, but its symmetric 3-cell quotient implies that certain synchronised states in the chain can behave in a manner that is typical of symmetric rings of cells.

This example motivates (and illustrates the answer to) an interesting converse question: if certain cells have identical time-series apart from a phase shift, does this imply some kind of global symmetry of the network? Remarkably, the answer, subject to reasonable conditions, is 'yes'. But, as Example 8 shows, we mist first pass to a quotient.

The main condition required is rigidity: the phase relation must stay unchanged (as a proportion of the period) after any sufficiently small admissible perturbation of the underlying ODE. To state this precisely, we need the following concept form dynamical systems theory. A periodic state $x(t)$ is *hyperbolic* if it has no Floquet exponent on the imaginary axis. Hyperbolicity implies that after a small perturbation of the vector field there exists a unique periodic orbit near $x(t)$ in the C^1 topology, Katok and Hasselblatt (1995), and its period is near that of $x(t)$. Thus we may talk of 'the' perturbed periodic state.

Definition 10 Suppose that $x(t)$ is a hyperbolic periodic state of period T of a $\mathscr{G}$-admissible ODE. A phase relation

$$x_c(t) = x_d(t-\theta) \qquad c,d \in \mathscr{C}, \quad \theta \in \mathbb{R}/T\mathbb{Z} \tag{6}$$

is *rigid* if for all sufficiently small admissible perturbations the perturbed periodic state $\tilde{x}(t)$ satisfies

$$\tilde{x}_c(t) = \tilde{x}_d(t-\theta) \qquad c,d \in \mathscr{C}, \quad \theta \in \mathbb{R}/\tilde{T}\mathbb{Z}$$

where $\tilde{T}$ is the period of $\tilde{x}(t)$.

The $T/3$ and $2T/3$ phase shifts in the rotating wave state for the coupled FitzHugh-Nagumo system is rigid. This can be proved for any rotating wave state arising by Hopf bifurcation in a symmetric system, indeed for any such state consistent with the H/K Theorem.

When $\theta = 0$ in (6) we say that cells c and d are *rigidly synchronous*.

Intuitively, whenever (6) holds, we expect the states $x_{I(c)}(t)$ and $x_{I(d)}(t)$ of the input sets of cells c and d to be phase-related by the same θ, up to some input isomorphism. Taken literally, this statement is false: the inputs states could differ in a way that does not affect the coupling to cells c, d. But we expect such a relationship to be destroyed by most small perturbations. For several years this belief was conjectural (Stewart and Parker 2007); the main difficulty in proving it was to keep track of how the periodic state perturbed.

We now introduce a mild technical condition, which some authors include in the definition of a coupled cell network:

Definition 11 A network is *cell-homogeneous* if all cell-equivalent cells are input-equivalent.

Assuming this condition, Golubitsky et al. (2010) proved the Rigid Synchrony Theorem:

Theorem 5 *Suppose that $\mathcal{G}$ is a cell-homogeneous path-connected network and two cells c, d are rigidly synchronous. Then there exists an input isomorphism $\beta : I(c) \to I(d)$ such that for all $j \in \mathcal{T}(d)$ cells j and $\beta^*(j)$ are rigidly synchronous.*

Corollary 1 *Suppose that $\mathcal{G}$ is a cell-homogeneous path-connected network. Then the colouring K in which cells have the same colour if and only if they are rigidly synchronous is balanced.*

Their method is inspired by singularity theory, and requires studying a space of perturbations large enough to destroy any spurious synchrony but small enough to control. Shortly afterwards, Golubitsky et al. (2012) extended their methods to handle nonzero phase shifts, obtaining the Rigid Phase Theorem:

Theorem 6 *Suppose that $\mathcal{G}$ is a cell-homogeneous path-connected network and two cells c, d are rigidly phase related by a phase shift that is a proportion θ of the period of the perturbed periodic state. Then there exists an input isomorphism $\beta : I(c) \to I(d)$ such that for all $j \in \mathcal{T}(d)$ cells j and $\beta^*(j)$ are phase related by a phase shift that is the same proportion of the period of the perturbed periodic state.*

It is conjectured that the condition of cell-homogeneity can be removed, and it seems likely that the methods of Golubitsky et al. (2010, 2012) can be modified to prove this, but this issue is currently unresolved.

A key consequence had already been observed in Stewart and Parker (2008):

Theorem 7 *Suppose that $\mathcal{G}$ is a cell-homogeneous path-connected network and two cells c, d are rigidly phase related by a phase shift that is a proportion θ of the period of the perturbed periodic state. Let $\mathcal{G}_1$ be the quotient of $\mathcal{G}$ by the balanced coloring corresponding to rigid synchrony of cells. Then there exist integers m, k such that $\theta = m/k$, $\mathcal{G}_1$ has a global group of symmetries that is the cyclic group $\mathbb{Z}_k$, and all rigid phase relations between cells are determined by a discrete rotating wave consistent with these symmetries.*

Informally: whenever a rigid phase shift is observed in a periodic state for a path-connected network, it is a consequence of a *global* cyclic-group symmetry of the quotient network in which rigidly synchronous cells are identified.

10 Bifurcations

Informally, a bifurcation occurs in a family of ODEs

$$\frac{\mathrm{d}x}{\mathrm{d}t} = f(x, \lambda)$$

with a parameter λ when the qualitative description of states changes near some parameter value λ_0. For example the number of steady states may change as λ passes through λ_0, or a stable steady state may become unstable and throw off a periodic cycle.

Local bifurcation, where the states branch along different curves in (x, λ)-space, is governed by the eigenvalues of the Jacobian matrix $J = \mathrm{D}_x f|_{(x,\lambda)}$. If an eigenvalue of J is zero at some point (x_0, λ_0) then typically a new branch of steady states appears. If a complex conjugate pair of eigenvalues are purely imaginary, equal to $\pm i\omega$, then typically there is a Hopf bifurcation to a branch of time-periodic states with frequency close to $2\pi/\omega$, Hassard et al. (1981). Such eigenvalues are said to be critical.

For standard dynamical systems, 'typically' here requires the critical eigenvalues to be simple. Moreover, they should pass through the imaginary axis with nonzero speed as λ passes through λ_0. In equivariant dynamics, symmetry constraints can force eigenvalues to be multiple, and new phenomena occur. A notable one is spontaneous symmetry-breaking, where solutions have less symmetry than the equations (Golubitsky and Stewart 2002a; Golubitsky et al. 1988).

In networks, local bifurcation is more complicated. The network architecture can have a strong effect not only on the eigenvalues, but also on the nonlinearities along the bifurcating branch. For example, there exist networks for which 'typical' steady-state bifurcation is more degenerate, in a singularity-theoretic sense, than the usual transcritical or pitchfork bifurcations. This affects the typical growth rate of the bifurcating branch (Stewart 2014; Stewart and Golubitsky 2011).

Instead of symmetry-breaking bifurcations, networks can exhibit synchrony-breaking bifurcations. Here a state with some pattern of synchrony loses stability and the pattern of synchrony changes: some cells that were synchronous cease to be synchronous. The interplay between network architecture and eigenvalues (and eigenvectors) of the Jacobian plays a central role in the theory of synchrony-breaking bifurcations. Rink and Sanders (2012, 2013a, 2014) explain this relationship in terms of a modified type of equivariance, using semigroups rather than groups.

A very surprising synchrony-breaking bifurcation occurs at Hopf bifurcation in a 3-cell feed-forward network, Fig. 10. Generic Hopf bifurcation in a general

Fig. 10 A 3-cell feed-forward network

dynamical system creates a bifurcating branch of equilibria whose amplitude grows like $\lambda^{1/2}$. However, Elmhirst and Golubitsky (2006) proved that typically there is a bifurcating branch of periodic states in which cell 1 is steady, the amplitude of cell 2 grows like $\lambda^{1/2}$, and the amplitude of cell 3 has the anomalous growth rate $\lambda^{1/6}$.

There is an analogous result for a feed-forward chain of m nodes. Hopf bifurcation can then lead to states that grow like $\lambda^{1/18}$ in the fourth node, $\lambda^{1/54}$ in the fifth node, and so on. This has been proved by Rink and Sanders (2013b) using a far-reaching generalisation of the notion of symmetry for networks. See also Rink and Sanders (2012, 2013a, 2014).

There is also a potential application to a nonlinear filter that selects and amplifies periodic oscillations close to a specific frequency (Golubitsky et al. 2009; McCullen et al. 2007).

11 Conclusions

The main message of this chapter is very simple. Networks are becoming increasingly important as models of many real systems, across the whole range of sciences. Moreover, the dynamics of networks has its own special flavour and differs considerably from the standard theory of dynamical systems. There is now a growing understanding of network dynamics, which in particular makes it possible to distinguish typical phenomena common to many networks with a given architecture from special phenomena that depend on the modelling equations. Among them are patterns of synchrony and phase relations, but the approach is not limited to these types of behaviour.

Placing network dynamics in a formal, abstract setting makes the above distinction clear, and offers several benefits, which are already sufficiently interesting to justify setting up such a formalism. Many types of behaviour become comprehensible and natural within this setting. On the other hand, it is important to recognise that the general abstract results must be augmented by special considerations, either for classes of networks with extra structure, or for specific models. The area of network dynamics is developing rapidly with many new results and open questions.

References

Aldis JW (2010) On balance. Ph.D. thesis, University of Warwick

Antoneli F, Stewart I (2006) Symmetry and synchrony in coupled cell networks 1: fixed-point spaces. Int J Bifurcation Chaos 16:559–577

Antoneli F, Stewart I (2007) Symmetry and synchrony in coupled cell networks 2: group networks. Int J Bifurcation Chaos 17:935–951

Antoneli F, Stewart I (2008) Symmetry and synchrony in coupled cell networks 3: exotic patterns. Int J Bifurcation Chaos 18:363–373

Arnold VI (1963) Proof of a theorem of A. N. Kolmogorov on the preservation of conditionally periodic motions under a small perturbation of the Hamiltonian. Uspehi Mat Nauk 18:13–40

Brandt H (1927) Über eine Verallgemeinerung des Gruppenbegriffes. Math Ann 96:360–366

Brown R (1987) From groups to groupoids: a brief survey. Bull Lond Math Soc 19:113–134

Buono P-L (2001) Models of central pattern generators for quadruped locomotion II: secondary gaits. J Math Biol 42:327–346

Buono P-L, Golubitsky M (2001) Models of central pattern generators for quadruped locomotion: I. Primary gaits. J Math Biol 42:291–326

Collins JJ, Stewart I (1993a) Hexapodal gaits and coupled nonlinear oscillator models. Biol Cybern 68:287–298

Collins JJ, Stewart I (1993b) Coupled nonlinear oscillators and the symmetries of animal gaits. J Nonlinear Sci 3:349–392

Deville L, Lerman E (2015) Modular dynamical systems on networks. J Eur Math Soc 17:2977–3013. http://arxiv.org/abs/1303.3907

Dias APS, Stewart I (2004) Symmetry groupoids and admissible vector fields for coupled cell networks. J Lond Math Soc 69:707–736

Elmhirst T, Golubitsky M (2006) Nilpotent Hopf bifurcations in coupled cell systems. SIAM J Appl Dyn Syst 5: 205–251

Euler L (1741) Solutio problematis ad geometriam situs pretinentis. Commentarii Academiae Scientiarum Petropolitanae 8:128–140. Reprinted in Opera Omnia: Series 1, 7 1–10. Reproduced at http://math.dartmouth.edu/euler/docs/originals/E053.pdf

Golubitsky M, Schaeffer DG (1985) Singularities and groups in bifurcation theory I. Applied mathematics series, vol 51. Springer, New York

Golubitsky M, Stewart I (1986) Hopf bifurcation with dihedral group symmetry: coupled nonlinear oscillators. In: Golubitsky M, Guckenheimer J (eds) Multiparameter bifurcation theory. Proceedings of the AMS-IMS-SIAM joint summer research conference, July 1985, Arcata. Contemporary mathematics, vol 56. American Mathematical Society, Providence RI, pp 131–173

Golubitsky M, Stewart I (2002a) The symmetry perspective: from equilibria to chaos in phase space and physical space. Progress in mathematics, vol 200. Birkhäuser, Basel

Golubitsky M, Stewart I (2002b) Patterns of oscillation in coupled cell systems. In: Holmes P, Newton P, Weinstein A (eds) Geometry, dynamics and mechanics: 60th birthday volume for J.E. Marsden. Springer, New York, pp 243–286

Golubitsky M, Stewart I (2006) Nonlinear dynamics of networks: the groupoid formalism. Bull Am Math Soc 43:305–364

Golubitsky M, Stewart I (2016a) Homeostasis, singularities, and networks. J Math Biol. doi:10.1007/s00285-016-1024-2

Golubitsky M, Stewart I (2016b) Coordinate changes for network dynamics (to appear)

Golubitsky M, Stewart I, Schaeffer DG (1988) Singularities and groups in bifurcation theory II. Applied mathematics series, vol 69. Springer, New York

Golubitsky M, Stewart I, Török A (2005) Patterns of synchrony in coupled cell networks with multiple arrows. SIAM J Appl Dyn Syst 4:78–100

Golubitsky M, Postlethwaite C, Shiau L-J, Zhang Y (2009) The feed-forward chain as a filter amplifier motif. In: Josíc K, Matias M, Romo R, Rubin J (eds) Coherent behavior in neuronal networks. Springer, New York, pp 95–120

Golubitsky M, Romano D, Wang Y (2010) Network periodic solutions: full oscillation and rigid synchrony. Nonlinearity 23:3227–3243

Golubitsky M, Romano D, Wang Y (2012) Network periodic solutions: patterns of phase-shift synchrony. Nonlinearity 25:1045–1074

Guckenheimer J, Holmes P (1983) Nonlinear oscillations, dynamical systems, and bifurcations of vector fields. Springer, New York

Hassard BD, Kazarinoff ND, Wan Y-H (1981) Theory and applications of Hopf bifurcation. Cambridge University Press, Cambridge

Higgins PJ (1971) Notes on categories and groupoids. Van Nostrand Reinhold mathematical studies, vol 32. Van Nostrand Reinhold, London

Hirsch MW, Smale S (1974) Differential equations, dynamical systems, and linear algebra. Academic, New York

Katok A, Hasselblatt B (1995) Introduction to the modern theory of dynamical systems. Cambridge University Press, Cambridge

Kuramoto Y (1984) Chemical oscillations, waves, and turbulence. Springer, Berlin

McCullen NJ, Mullin T, Golubitsky M (2007) Sensitive signal detection using a feed-forward oscillator network. Phys Rev Lett 98:254101

Nijholt E, Rink B, Sanders J (2016) Graph fibrations and symmetries of network dynamics. J Differ Equ (to appear)

Poincaré H (1881) Mémoire sur les courbes définiés par une équation différentielle. J Math 7:375–422

Poincaré H (1882) Mémoire sur les courbes définiés par une équation différentielle. J Math 8:251–296

Poincaré H (1885) Mémoire sur les courbes définiés par une équation différentielle. J Math 1:167–244

Poincaré H (1886) Mémoire sur les courbes définiés par une équation différentielle. J Math 2:151–217

Poincaré H (1892) Les Méthodes Nouvelles de la Mécanique Céleste, vol 1. Gauthier-Villars, Paris

Poincaré H (1893) Les Méthodes Nouvelles de la Mécanique Céleste, vol 2. Gauthier-Villars, Paris

Poincaré H (1899) Les Méthodes Nouvelles de la Mécanique Céleste, vol 3. Gauthier-Villars, Paris

Rink B, Sanders J (2012) Coupled cell networks: semigroups, Lie algebras, and normal forms. arXiv:1209.3209 [math.DS]

Rink B, Sanders J (2013a) Coupled cell networks and their hidden symmetries. arXiv:1304.1460 [math.DS]

Rink B, Sanders J (2013b) Amplified Hopf bifurcations in feed-forward networks. SIAM J Appl Dyn Syst 12:1135–1157

Rink B, Sanders J (2014) Coupled cell networks: semigroups, Lie algebras and normal forms. Trans Am Math Soc. doi:http://dx.doi.org/10.1090/S0002-9947-2014-06221-1

Smale S (1967) Differentiable dynamical systems. Bull Am Math Soc 73:747–817

Stewart I (2014) Synchrony-breaking bifurcations at a simple real eigenvalue for regular networks 2: higher-dimensional cells. SIAM J Appl Dyn Syst 13:129–156. doi:10.1137/130917636

Stewart I, Golubitsky M (2011) Synchrony-breaking bifurcations at a simple real eigenvalue for regular networks 1: 1-dimensional cells. SIAM J Appl Dyn Syst 10:1404–1442. doi:10.1137/110825418

Stewart I, Parker M (2007) Periodic dynamics of coupled cell networks I: rigid patterns of synchrony and phase relations. Dyn Syst 22:389–450

Stewart I, Parker M (2008) Periodic dynamics of coupled cell networks II: cyclic symmetry. Dyn Syst 23:17–41

Stewart I, Golubitsky M, Pivato M (2003) Symmetry groupoids and patterns of synchrony in coupled cell networks. SIAM J Appl Dyn Syst 2:609–646

Topp B, Promislow K, De Vries G, Miura RM, Finegood DT (2000) A model of β-cell mass, insulin, and glucose kinetics: pathways to diabetes. J Theor Biol 206:605–619

Wilson RJ (1985) Introduction to graph theory, 3rd edn. Longman, Harlow

Dynamics on Large Sets and Its Applications to Oligopoly Dynamics

Jose S. Cánovas and María Muñoz Guillermo

Abstract In this chapter we analyze discrete dynamical systems with a phase space on a high dimensional space. We explain some techniques which allow us to make some approaches to the system analysis with special emphasis to oligopoly dynamics.

Keywords Chaos • Oligopoly dynamics • Synchronization • Topological entropy.

1 Introduction

Oligopolies are models in which several firms compete in one market in such a way the interaction between them plays a crucial role in the market evolution. The models can be stated either on continuous or discrete time, but with the same common problem: when the number of firms increases, the dynamics, that is, the market evolution is hard to analyze because the number of equations increases, which implies that analytical results are unknown and numerical tools, including simulations valid for one and two dimensional systems are not valid.

In this chapter we study oligopoly models given by systems of difference equations, that is, in discrete time. Our aim is to give some ideas and techniques which can be useful for obtaining partial results on oligopoly models for an arbitrary number of firms. Of course, our approach is partial and there are many Open questions and problems involving this topic, some of them presented along the chapter.

Our results follow two different lines: On one hand, we look for symmetries in the models in order to reduce the dimension of the phase space. Then we may study a simplified model, embedded in the original one. The idea is that such simplified

J.S. Cánovas (✉) • M.M. Guillermo
Departamento de Matemática Aplicada y Estadística, Universidad Politécnica de Cartagena, C/ Doctor Fleming sn, 30202 Cartagena, Spain
e-mail: Jose.Canovas@upct.es; Maria.mg@upct.es

P. Commendatore et al. (eds.), *Complex Networks and Dynamics*, Lecture Notes in Economics and Mathematical Systems 683,
DOI 10.1007/978-3-319-40803-3_11

model can be analyzed using well-known techniques and later on, we study whether the whole system, at least locally, follows the dynamics of the simplified one.

On the other hand, we look for dynamical results which may have some meaning in economics as, for instance, what conditions must fulfill a firm to disappear from the market, that is, the firm will no longer produce in future. In these results we work with particular cases of the model.

The chapter is organized as follows. First, we introduce some basic mathematical tools on discrete dynamical systems and difference equations. The dynamical notions will be applied to analyze well-known oligopoly models when the number of firms increases. First, we introduce a general framework, which is applied to each concrete model, which is piecewise linear in Sect. 3 and non piecewise linear in Sect. 4.

2 Mathematical Tools

The models considered in this chapter are given by *difference equations*, which are expressions with the form

$$\begin{cases} x(t+1) = f_t(x(t)), \\ x(0) = x_0, \end{cases}$$

where $f_t : X \to X$, $t \in \mathbb{N}$, is a sequence of maps on a metric space X into itself and $x_0 \in X$. The solution of the above difference equation is called *orbit* or *trajectory* of x_0 under f_t. When the sequence of maps is constant, that is, $f_t = f$, $t \in \mathbb{N}$, we have an autonomous difference equation, which is usually seen as a *discrete dynamical system*, usually denoted by the pair (X,f). Then, the orbit of x_0 under f, denoted $\mathrm{Orb}(x_0,f)$ is given by the sequence $f^t(x_0)$, $t \geq 0$, where $f^t = f \circ f^{t-1}$, $t > 1$, $f^1 = f$, and f^0 is the identity on X.

Although one can study topological properties of dynamical systems, in this chapter we are interested in the case $X = \mathbb{R}^n_{\geq}$, where $\mathbb{R}_{\geq}$ represents the set of non negative real numbers. There is a huge literature on discrete dynamical systems either for the one dimensional case, when $n = 1$ (see e.g. Block and Coppel 1992, Alsedá et al. 1993 or de Melo and van Strien 1993) or for higher dimensions and even general topological (metric) spaces (see e.g. Aoki and Hiraide 1994 and Devaney 1989). Here, we introduce some basic results and notation on dynamical systems on general metric spaces which can be easily translated for real maps.

2.1 Periodic Orbits and Topological Dynamics

We consider a metric space (X,d), which is usually compact, and a continuous map $f : X \to X$, and recall that (X,f) denotes a discrete dynamical system. Note that all the definitions below can be expressed either in terms of the map f or the system (X,f).

To understand the dynamics of f, we have to introduce some definitions, which have topological roots, to obtain some knowledge of the system (see e. g. Block and Coppel 1992 or Sharkovsky et al. 1997). A point $x \in X$ is *periodic* when $f^t(x) = x$ for some $t \geq 1$. The smallest positive integer satisfying this condition is called the *period* of x. Periodic points of period 1 are called *fixed points*. Denote by $\mathrm{F}(f)$, $\mathrm{P}(f)$ and $\mathrm{Per}(f)$ the sets of fixed and periodic points and periods of f, respectively.

Periodic orbits are the simplest orbits that a discrete dynamical system can generate, but there are many other classes of orbits that make the dynamics richer. For $x \in X$, define its *ω-limit set*, $\omega(x,f)$, as the set of limit points of its orbit $\mathrm{Orb}(x,f)$. If $\omega(x,f)$ is finite, then it is a periodic orbit, but often, the dynamical behavior of a single orbit can be very complicated or unpredictable, and usually the word chaos is used to refer to dynamical systems which are able to produce such complicated orbits as we discuss below.

Previously, note that to understand the dynamics it is enough to do it on small subsets of X called *attractors*, which are non empty compact sets A that attracts all trajectories starting in some neighborhood $\mathscr{U}$ of A, that is, for all $x \in \mathscr{U}$ we have that

$$\lim_{t\to\infty} dist(f^t(x), A) = 0,$$

where $dist(x, A) = \min\{d(x, y) : y \in A\}$. When $\mathscr{U}$ is the whole space X we have a *global attractor*. The existence of attractors makes easier the understanding of the dynamics, which in principle may be very complex. The existence and approximate location of attractors are usually given by the *absorbing sets*, namely, a subset $B \subset X$ is an *absorbing set* if for any bounded set D of X there is $t_0 = t_0(D)$ such that $f^t(D) \subset B$ for all $n \geq t_0$.

There are many definitions of chaos, but we will focus our interest in the following well-known ones. The map f is *chaotic in the sense of Li and Yorke* (*LY-chaotic*) (Li and Yorke 1975) if there is an uncountable set $S \subset X$ (called *scrambled set* of f) such that for any $x, y \in S$, $x \neq y$, we have that

$$\liminf_{t\to\infty} d(f^t(x), f^t(y)) = 0,$$

$$\limsup_{t\to\infty} d(f^t(x), f^t(y)) > 0.$$

Li and Yorke's definition of chaos became famous because of the famous result *period three implies chaos* which linked periodic orbits and unpredictable dynamical behavior for continuous interval maps. Note, that the definition implies the comparison between two orbits or limit points of orbits. Another well-known chaos definition, inspired by the notion of sensitivity respect to the initial conditions (Guckhenheimer 1979), was given by Devaney (1989) as follows. The map f is said to be *chaotic in the sense of Devaney* (*D-chaotic*) if it fulfills the following properties:

- The map f it is *transitive*, which in absence of isolated points means that there is $x \in X$ such that $\omega(x,f) = X$.
- The set of periodic points $\mathrm{P}(f)$ is dense on X.
- It has *sensitive dependence on initial conditions*, that is, there is $\varepsilon > 0$ such that for any $x \in X$ there is an arbitrarily close $y \in X$ and $t \in \mathbb{N}$ such that $d(f^t(x), f^t(y)) > \varepsilon$.[1]

Both Li–Yorke chaos and sensitivity to initial conditions are in the dynamical systems folklore.

There are a lot of different dynamics between periodic orbits and chaotic behavior, so it is interesting to explain what simple dynamics are. In fact, sometimes, the chaotic behavior can be also taken as the opposite of simple (or ordered) behavior. We say that f is *strongly simple* (*ST-simple*) if any ω-limit set is a periodic orbit of f. We say that an orbit $\mathrm{Orb}(x,f)$, $x \in X$, is approximated by periodic orbits if for any $\varepsilon > 0$ there is $y \in \mathrm{P}(f)$ and $t_0 \in \mathbb{N}$ such that $d(f^t(x), f^t(y)) < \varepsilon$ for all $t \geq t_0$. The map f is *LY-simple* (Smítal 1986) if any orbit is approximated by periodic orbits. Finally f is *Lyapunov stable* (*L-simple*) (Fedorenko et al. 1990) if it has equicontinuous powers.

The above definitions are quite difficult to verify and, specially when we are working with models which in principle may depend on several parameters, we need some practical methods to try to measure the dynamical complexity of the system. One of them is given by *topological entropy*, which was introduced in the setting of continuous maps on compact topological spaces by Adler et al. (1965) and Bowen (1971).[2] It is remarkable that both definitions agree when the set X is metric and compact. It is a conjugacy invariant[3] which is usually taken as a criterion to decide whether the dynamic is complicated or not according to the topological entropy $h(f)$, which will be defined below, is greater than zero or not. Here we introduce the equivalent definitions by Bowen (1971) when (X, d) is a compact metric space. Given $\varepsilon > 0$, we say that a set $E \subset X$ is (t, ε, f)-*separated* if for any $x, y \in E$, $x \neq y$, there exists $k \in \{0, 1, \ldots, t-1\}$ such that $d(f^k(x), f^k(y)) > \varepsilon$. Denote by $s(t, \varepsilon, f)$ the biggest cardinality of any maximal (t, ε, f)-separated set in X. Then the topological entropy of f is

$$h(f) = \lim_{\varepsilon \to 0} \limsup_{t \to \infty} \frac{1}{t} \log s(t, \varepsilon, f).$$

---

[1] It is proved in Banks et al. (1992) that the first two conditions in Devaney's definition implies the third one. The definitions is presented in the original form because of the dynamical meaning of sensitive dependence on initial conditions.

[2] Dinaburg (1970) gave simultaneously a Bowen like definition for continuous maps on a compact metric space.

[3] Two continuous maps $f : X \to X$ and $g : Y \to Y$ are said to be topologically conjugate if there is an homeomorphism $\varphi : X \to Y$ such that $g \circ \varphi = \varphi \circ f$. In general, conjugate maps share many dynamical properties.

There is an equivalent definition using spanning sets as follows. We say that a set $F \subset X$ (t, ε, f)-*spans* X if for any $x \in X$ there exists $y \in F$ such that $d(f^i(x), f^i(y)) < \varepsilon$ for any $i \in \{0, 1, \ldots, t-1\}$. Denote by $r(t, \varepsilon, f)$ the smallest cardinality of any minimal (t, ε, f)-spanning set in X. Then, topological entropy can be computed as

$$h(f) = \lim_{\varepsilon \to 0} \limsup_{t \to \infty} \frac{1}{t} \log r(t, \varepsilon, f).$$

The above definitions do not depend on the metric d, and give us a nice interpretation of topological entropy (see Alsedá et al. 1993, p. 188) as follows. Imagine that we have a magnifying glass through which we can distinguish two point if and only if they are more than ε-apart. If we know t points of two orbits given by x and y, that is, $(x, f(x), \ldots, f^{t-1}(x))$ and $(y, f(y), \ldots, f^{t-1}(y))$, then we can distinguish between x and y if and only if $\max_{1 \leq i \leq t} d(f^i(x), f^i(y)) > \varepsilon$. Hence, $s(t, \varepsilon, f)$ gives us how many points of the space X we can see if we know the pieces of orbits of length t. Then we take the exponential growth rate with t of this quantity, and finally the limit of this as we take better and better magnifying glasses. Then we obtain the topological entropy.

In general, the above chaos definitions are not equivalent and their relations with topological entropy are not homogeneous. For instance, it has been proved that D-chaotic maps are LY-chaotic (Huang and Ye 2002), but the converse is false (Smítal 1986). On the other hand, positive topological entropy implies LY-chaos (Blanchard et al. 2002)[4] and the converse is also false (Smítal 1986). In Balibrea and Snoha (2003) and Kwietniak and Misiurewicz (2005) it is studied the relationship between topological entropy and D-chaos. ST-simple maps are LY-simple maps but the converse is false (Smítal 1986).

More popular than topological entropy are the so-called Lyapunov exponents (see Oseledets 1968), which are defined when differentiable structures are considered. Namely, assume that X is a smooth finite dimensional manifold and $f : X \to X$ is a $C^{1+\alpha}$ map. Denote, as usual, by T_xX the tangent space at x and the derivative $d_xf : T_xX \to T_{f(x)}X$. The Lyapunov exponent at $x \in X$ in the direction of $\mathbf{v} \in T_xX \setminus \{\mathbf{0}\}$ by

$$\text{lyex}(x, \mathbf{v}) = \lim_{t \to \infty} \frac{1}{t} \log ||d_x f^t(\mathbf{v})||$$

if this limit exists. An invariant measure μ is a probability measure on the Borel sets of X such that $\mu(f^{-1}(A)) = \mu(A)$ for any Borel set $A \subseteq X$. This invariant measure μ is ergodic if the equality $f^{-1}(A) = A$ implies that $\mu(A)$ is either 0 or 1. The multiplicative ergodic Theorem states that the above limit exists for μ-almost all point in X. We use Lyapunov exponents in particular cases where chaos is associated to have positive Lyapunov exponents.

[4]See also Sumi (2003) which almost simultaneously states the same result for C^2 diffeomorphisms on compact manifolds of dimension greater than one.

Next, we study the particular case of real maps, starting by the one dimensional case. We will see how the above results are sharpened for continuous interval maps. In addition, we will give some notions on the dynamics of several dimensions real maps.

2.2 Dynamics of Continuous Interval Maps

In general, for one dimensional maps, the relevant results are given when $X = [a, b] \subset \mathbb{R}$ is a compact interval, usually $[0, 1]$ due to linear conjugacy. In this setting, Sharkovsky's Theorem is a remarkable result which helps to distinguish between simple and complicated dynamics. Recall Sharkovsky's order of natural numbers

$$3 >_s 5 >_s 7 >_s \ldots >_s 2 \cdot 3 >_s 2 \cdot 5 >_s \ldots >_s 2^2 \cdot 3 >_s 2^2 \cdot 5 >_s \ldots$$

$$\ldots >_s 2^k \cdot 3 >_s 2^k \cdot 5 >_s \ldots >_s 2^3 >_s 2^2 >_s 2 >_s 1.$$

Applying Sharkovsky's Theorem (see Sharkovsky et al. (1997) or Alsedá et al. (1993). Also Du (2004) for an "easy" proof) one can see that for any continuous map $f : \mathbb{R} \to \mathbb{R}$ with one periodic point holds that either $\mathrm{Per}(f) = \mathrm{S}(m) = \{k : m >_s k\} \cup \{m\}$, with $m \in \mathbb{N}$, or $\mathrm{Per}(f) = \mathrm{S}(2^\infty) = \{2^n : n \in \mathbb{N} \cup \{0\}\}$. A map is of *type* $m \in \mathbb{N} \cup \{2^\infty\}$ if $\mathrm{Per}(f) = \mathrm{S}(m)$. A map f is called *S-chaotic* if $\mathrm{Per}(f) = \mathrm{S}(m)$, $m = 2^r q$, $r \geq 0$ and $q > 1$ odd.

On the other hand, for one dimensional dynamics the topological entropy is an useful tool to check the dynamical complexity of a map because it is strongly connected with the notion of *horseshoe* (see Alsedá et al. 1993, p. 205). We say that the map $f : [0, 1] \to [0, 1]$ has a k-horseshoe, $k \in \mathbb{N}$, $k \geq 2$, if there are k disjoint subintervals J_i, $i = 1, \ldots, k$, such that $J_1 \cup \ldots \cup J_k \subseteq f(J_i)$, $i = 1, \ldots, k$.[5]

The following result shows some equivalences among the above definitions of chaos and order (see Sharkovsky et al. 1997, Smítal 1986 and Block and Coppel 1992). Note that the situation is simpler than in the general case.

Theorem 1 *Let $f : [0, 1] \to [0, 1]$ be a continuous map. Then*

(a) The map f has positive topological entropy if and only if the map f is S-chaotic.
(b) If f is D-chaotic, then $h(f) > 0$.
(c) If f is either ST-simple or L-simple, then $h(f) = 0$.
(d) If $h(f) > 0$, then f is LY-chaotic, but the converse is false in general. If f is LY-simple, then $h(f) = 0$. The union of LY-chaotic and LY-simple continuous maps is the set of continuous interval maps.

[5] Since Smale's work (see Smale 1967), horseshoes have been in the core of chaotic dynamics, describing what we could call random deterministic systems.

The nature of the above result is topological. If we consider another points of view, we can obtain more information giving rise to apparently strange paradoxes. For instance, there exist maps with positive entropy, and therefore chaotic in some sense, such that the orbit of almost all points in $[0,1]$ (with respect to the Lebesgue measure) converges to a periodic orbit.

Although we will come back to this point later, let us show how to get such example. Consider f a C^3 *unimodal* map such that $f(0) = f(1) = 0$. Recall that a map f is said to be unimodal if there is $c \in [0,1]$, called *turning point* such that $f|_{[0,c)}$ is strictly increasing and $f|_{[c,1]}$ is strictly decreasing. The *Schwarzian derivative* (see Singer 1978 or Thunberg 2001) is then given by

$$S(f)(x) = \frac{f'''(x)}{f'(x)} - \frac{3}{2}\left(\frac{f''(x)}{f'(x)}\right)^2,$$

at those points whose first derivative does not vanish. Assume that $S(f)(x) < 0$ and that there is a locally attracting periodic orbit, that is, a periodic orbit $P = \{x_1, \ldots, x_p\}$ for which there exists a neighborhood V of P such that for any $x \in V$ the distance $d(f^t(x), P) = \min_{1 \le i \le p} d(f^t(x), x_i)$ tends to zero as t tends to infinity. The logistic map $f(x) = 3.83\,x(1-x)$ is a good example of such behavior; almost all trajectory converges to a periodic orbit of period 3, while the topological entropy is positive (see e.g. Block et al. 1989). This example, and many others in the literature, shows that it is important to study the dynamics from several points of view.

2.3 Piecewise Monotone Maps: Entropy and Attractors

Usually, one dimensional difference equations models in science are given by piecewise monotone maps. A continuous interval map is *piecewise monotone* if there is a finite partition of $[0,1]$, $0 = x_0 < x_1 < \ldots < x_k = 1$, such that $f|_{[x_i,x_{i+1}]}$ is monotone for $0 \le i < k$. Note that a piecewise monotone map may have constant pieces. The extreme points, which can be isolated or contained in a subinterval of extreme points, of f will be called *turning points* (turning intervals if the extreme points form a subinterval). For a piecewise monotone map f, let $c(f)$ denote the number of pieces of monotonicity of f. If g is another piecewise monotone map, it is easy to see that $c(f \circ g) \le c(f)c(g)$. Hence, the sequence $c(f^t)$ gives the number of monotonicity pieces of f^t and the following result due to Misiurewicz and Szlenk (see Misiurewicz and Szlenk 1980), shows that for piecewise monotone maps topological entropy can be easily understood.

Theorem 2 *Let* $f : [0,1] \to [0,1]$ *be a continuous and piecewise monotone map. Then*

$$h(f) = \lim_{t\to\infty} \frac{1}{t} \log c(f^t).$$

Note that $c(f^t) \leq c(f)^t$, and so $h(f) \leq \log c(f)$. Hence, a consequence of Misiurewicz–Szlenk Theorem is that homeomorphisms on the interval have zero topological entropy. On the other hand, following Theorem 2, we can easily see that the logistic map $f(x) = 4x(1-x)$ and the tent map $g(x) = 1 - |2x-1|$ have topological entropy $\log 2$, since $c(f^t) = c(g^t) = 2^t$ for all $t \in \mathbb{N}$. However, computing topological entropy can be a very complicated task, but we will see in what follows how to make these computations for a suitable class of maps.

The dynamics of smooth enough piecewise monotone maps are well-known in the following sense. Following (Milnor 1985), a metric attractor is a subset $A \subset [0,1]$ such that $f(A) \subseteq A$, $O(A) = \{x : \omega(x,f) \subset A\}$ has positive Lebesgue measure, and there is no proper subset $A' \subsetneq A$ with the same properties. The set $O(A)$ is called the *basin* of the attractor.

By van Strien and Vargas (2004), the regularity properties of f imply that there are three possibilities for its metric attractors for a class of piecewise monotone maps, called *multimodal maps*, fulfilling the following assumptions. There are $c_1 < c_2 < \ldots < c_k$, creating a partition on [0, 1], such that f is strictly monotone on each element of the partition. f is C^3 and f is non flat on the turning points $c_1, \ldots, c_k$,, that is, for x close to c_i, $i = 1, 2, \ldots, k$,

$$f(x) = \pm|\phi_i(x)|^{\beta_i} + f(c_i),$$

where ϕ_i is C^3, $\phi_i(c_i) = 0$ and $\beta_i > 0$. Then, the metric attractors of such multimodal maps can be of one of the following types:

(A1) A periodic orbit.

(A2) A solenoidal attractor, which is basically a Cantor set in which the dynamic is quasi periodic. More precisely, the dynamic on the attractor is conjugated to a minimal translation, in which each orbit is dense on the attractor. The dynamic of f restricted to the attractor is simple, neither positive topological entropy nor Li–Yorke chaos can be obtained. Its dynamic is often known as quasi-periodic.

(A3) A union of periodic intervals $J_1, \ldots, J_k$, such that $f^k(J_i) = J_i$ and $f^k(J_i) = J_j$, $1 \leq i < j \leq k$, and such that f^k is topologically mixing. Topologically mixing property implies the existence of dense orbits on each periodic interval (under the iteration of f^k).

Moreover, if f has an attractor of type (A2) and (A3), then they must contain the orbit of a turning point, and therefore its number is bounded by the turning points. In addition, if $Sf(x) < 0$, then the total number of attractors is bounded by k. From a practical point of view, in a computer simulation we are able to show the existence of attractors of type (A1) and (A3), and only attractors of type (A3) are able to exhibit unpredictable dynamics. As a conclusion of this, if all the turning points of f are attracted by periodic orbits, then the map f will not exhibit physically observable chaos, although it can be topologically chaotic.

The Lyapunov exponents on the turning points can be computed by

$$\mathrm{lyex}(c_i) = \lim_{t\to\infty} \frac{1}{t} \log |(f^t)'(c_i)| = \lim_{t\to\infty} \frac{1}{t} \log |f'((f^{t-1})(c_i))|,$$

for $i = 1, 2, \ldots, k$, and all of them are negative when the map f is free of attractors of type (A3). So, positive Lyapunov exponents imply the existence of observable chaos.

2.4 Computing Topological Entropy

The above definition of topological entropy is not useful in practice, and counting monotone pieces of an iterated map f^t is not easy. In addition, an exact computation of topological entropy for continuous interval maps cannot be done in general, but there are several papers devoted to compute it approximately for unimodal maps (see Block et al. 1989) bimodal maps, that is, with three monotone pieces (see Block and Keesling 1992) and four monotone pieces (see Cánovas and Muñoz-Guillermo 2014c). In general, it is possible to make computations for arbitrarily large monotone pieces whenever the number of so-called kneading sequences will not be big enough (see Cánovas and Muñoz-Guillermo 2014b).

Now, we introduce the unimodal case where the topological entropy can be computed by using kneading sequences as follows. Let f be an unimodal map with maximum (turning point) at c. Let $k(f) = (k_1, k_2, k_3, \ldots)$ be its kneading sequence given by the rule

$$k_i = \begin{cases} R & \text{if } f^i(c) > c, \\ C & \text{if } f^i(c) = c, \\ L & \text{if } f^i(c) < c. \end{cases}$$

We fix that $L < C < R$. For two different unimodal maps f_1 and f_2, we fix their kneading sequences $k(f_1) = (k_n^1)$ and $k(f_2) = (k_n^2)$. We say that $k(f_1) \le k(f_2)$ provided there is $m \in \mathbb{N}$ such that $k_i^1 = k_i^2$ for $i < m$ and either an even number of $k_i^1{}'s$ are equal to R and $k_m^1 < k_m^2$ or an odd number of $k_i^1{}'s$ are equal to R and $k_m^2 < k_m^1$. Then it is proved in Block et al. (1989) that if $k(f_1) \le k(f_2)$, then $h(f_1) \le h(f_2)$. In addition, if $k_m(f)$ denotes the first m symbols of $k(f)$, then if $k_m(f_1) < k_m(f_2)$, then $h(f_1) \le h(f_2)$.

The algorithm for computing the topological entropy is based in the fact that the tent family

$$g_k(x) = \begin{cases} kx & \mathit{if}\ x \in [0, 1/2], \\ -kx + k & \mathit{if}\ x \in [1/2, 1], \end{cases}$$

with $k \in [1,2]$, holds that $h(g_k) = \log k$. The idea of the algorithm is to bound the topological entropy of an unimodal maps between the topological entropies of two tent maps. The algorithm is divided in four steps:

Step 1. Fix $\varepsilon > 0$ (fixed accuracy) and an integer n such that $\delta = 1/n < \varepsilon$.
Step 2. Find the least positive integer m such that $k_m(g_{1+i\delta})$, $0 \leq i \leq n$, are distinct kneading sequences.
Step 3. Compute $k_m(f)$ for a fixed unimodal map f.
Step 4. Find r the largest integer such that $k_m(g_{1+r\delta}) < k_m(f)$. Hence $\log(1 + r\delta) \leq h(f) \leq \log(1 + (r+2)\delta)$.

The algorithm is easily programmed. We usually use Mathematica, which has the advantage of computing the kneading invariants of tent maps without round off errors, improving in practice the accuracy of the method. We will show in Sect. 4 some examples of computing the topological entropy for our models.

2.5 Dynamics in Higher Dimension

Things are more complicated when $n > 1$, and discrete dynamical systems as far to be understood in an analytic way. The common agreement among researchers is that in general one dimensional results cannot be extended to general higher dimension dynamical systems. As a keynote example, one can easily check that Sharkovsky's Theorem does not hold for two dimensional maps: rational rotations on the plane are a good example of that. Although there are some results on limit sets (see Agronsky and Ceder 1991/1992b or Agronsky and Ceder 1991/1992a) and good result for some types of two dimensional maps like triangular or skew product ones (see Kloeden 1979, Kolyada 1992 or Kolyada and Snoha 1992/1993) and antitriangular ones (see Cánovas and Linero 2001 or Balibrea et al. 2004), the dynamics on two dimensional maps is still quite unexplored and usually papers dealing with models constructed on higher dimensional spaces have to show numerical experiments and simulations.

In this paper we are going to analyze the models trying to reduce the dimension. This can be done if the system has a global attractor with dimension smaller than the ambient space. Another way is to work with models that have some symmetry properties. The problem can be stated as follows. Assume that $f : X \to X$ is a $C^{1+\alpha}$ map on a manifold X with dimension n and there is a submanifold $Y \subset X$ with dimension m such that $f(Y) \subseteq Y$, due to symmetric properties. Hence, any orbit starting with an initial condition $x \in Y$ will remain in Y along the trajectory. Since $m < n$, it is possible that the dynamics of f on Y can be understood, and from this knowledge, we can derive some properties on the dynamics of f on the whole space X.

For instance, assume that $f|_Y$ is chaotic in the sense of Li and Yorke, that is, there exists an uncountable scrambled subset $S \subset Y \subset X$. It is clear that f itself is also chaotic in the sense of Li and Yorke. The same happens if $f|_Y$ has positive topological entropy, but it is not true in general if $f|_Y$ is Devaney chaotic. Moreover, it may happen that $f|_Y$ is Li–Yorke chaotic, but for any neighborhood N of Y one has that the trajectory of any $x \in N \setminus Y$ converges to a periodic orbit, which will make unobserved the existence of Li–Yorke chaos. So, we are interested in analyzing not only the dynamics of $f|_Y$ but also when the trajectories outside Y may converge or synchronize with trajectories inside Y.

This can be easily done if the attractor inside Y is a periodic orbit, because at least locally, Jacobian matrices along the periodic orbit give you the key: the spectral radius of the product of such Jacobian matrices has modulus smaller than one.[6] The problem arises when the attractor is chaotic. This paper mainly considers one-dimensional chaotic attractors of piecewise monotone maps. We refer the reader to Ashwin et al. (1996) for a precise description on transverse and normal Lyapunov exponents that we will use in the next section. Anyway, we should mention that all the results are local and, as far as we know, no global results are known.

2.6 Spectral Theory for a Type of Symmetric Matrix

Before finishing this mathematical section, we will analyze a matrix type that appears in several problems. In particular, we are interested in the eigenvalues of a class of square matrices with n rows, and the form

$$\mathbf{A}(a,b) = \begin{pmatrix} a & b & b & \dots & b & b \\ b & a & b & \dots & b & b \\ b & b & a & \dots & b & b \\ \dots & \dots & \dots & \dots & \dots & \dots \\ b & b & b & \dots & a & b \\ b & b & b & \dots & b & a \end{pmatrix},$$

where $a, b \in \mathbb{R}$, $b \neq 0$. The characteristic polynomial is

$$p(t) = \begin{vmatrix} a-t & b & b & \dots & b & b \\ b & a-t & b & \dots & b & b \\ b & b & a-t & \dots & b & b \\ \dots & \dots & \dots & \dots & \dots & \dots \\ b & b & b & \dots & a-t & b \\ b & b & b & \dots & b & a-t \end{vmatrix}$$

[6] A periodic orbit can be an attractor when spectral radius has modulus one, but in general the converse is not true.

$$= \begin{vmatrix} a-t+(n-1)b & b & b & \dots & b & b \\ a-t+(n-1)b & a-t & b & \dots & b & b \\ a-t+(n-1)b & b & a-t & \dots & b & b \\ \dots & \dots & \dots & \dots & \dots & \dots \\ a-t+(n-1)b & b & b & \dots & a-t & b \\ a-t+(n-1)b & b & b & \dots & b & a-t \end{vmatrix}$$

$$= (a-t+(n-1)b) \begin{vmatrix} 1 & b & b & \dots & b & b \\ 1 & a-t & b & \dots & b & b \\ 1 & b & a-t & \dots & b & b \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & b & b & \dots & a-t & b \\ 1 & b & b & \dots & b & a-t \end{vmatrix},$$

and hence it is clear that solving the equation

$$a-t+(n-1)b = 0,$$

we obtain the eigenvalue

$$t = (n-1)b + a.$$

On the other hand, solving the equation

$$a-t = b,$$

which implies $t = a-b$, we have that

$$\mathbf{A}(a,b) - (a-b)\mathbf{I}_n = \begin{pmatrix} b & b & b & \dots & b & b \\ b & b & b & \dots & b & b \\ b & b & b & \dots & b & b \\ \dots & \dots & \dots & \dots & \dots & \dots \\ b & b & b & \dots & b & b \\ b & b & b & \dots & b & b \end{pmatrix},$$

where $\mathbf{I}_n$ is the identity matrix of n rows. Then, $t = a-b$ is an eigenvalue of $\mathbf{A}(a, b)$. The equation of its eigenspace is obtained by equating

$$\begin{pmatrix} b & b & b & \dots & b & b \\ b & b & b & \dots & b & b \\ b & b & b & \dots & b & b \\ \dots & \dots & \dots & \dots & \dots & \dots \\ b & b & b & \dots & b & b \\ b & b & b & \dots & b & b \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \dots \\ x_{n-1} \\ x_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \dots \\ 0 \\ 0 \end{pmatrix},$$

and then $x_1 + x_2 + \ldots + x_n = 0$. Then, the eigenspace of eigenvalue $a - b$ has dimension $n - 1$. As a consequence, the eigenspace of the eigenvalue $(n - 1)b + a$ has dimension 1. Therefore, $a - b$ and $(n - 1)b + a$ are the eigenvalues of $\mathbf{A}(a, b)$ with multiplicity $n - 1$ and 1, respectively.

3 Oligopoly Dynamics: Theocharis' Model

In oligopoly models, a number of firms compete in a market in such a way the interaction between them plays a crucial role in the market evolution. The rules are usually given by assumptions that we make on e.g. demand functions, cost functions or decisions on future productions. For a wide range of different scenarios the reader can see Bischi et al. (2010). The basic idea is that, if we have n firms and Π_i is the profit for each firm, which will be assumed to be smooth enough, the optimization of profits can be the key for describing ways of how firms organize their future productions. At the end of the process, we have a system of difference equations

$$x_i(t + 1) = f_i(x_1(t), \ldots, x_n(t)), \ i = 1, \ldots, n,$$

where x_i is the variable that firm i wants to control, basically quantities or price, and f_i are called reaction functions, which give you the evolution of variables x_i with time.

Of course, the reaction functions need not be linear maps and then, the analysis of the dynamics of the systems, that is, the evolution with time of all the possible initial states is quite hard to analyze in general. However, there are several ideas that can be used to give some partial results on the dynamics, as it is shown below.

To fix ideas, we will always assume that q_i are the quantities of goods produced by each firm.[7] The reaction functions can have different forms according to the way that firms will organize their future productions. Perhaps the simplest case is naive expectations on future productions, in such a way firms choose the maximum of Π_i

[7]From now on, we denote the production with the letter "q" instead of "x" because this is the usual notation for that.

at time t to be the quantity produced in time $t + 1$. Then, maximizing the profit, that is, solving the equation

$$\frac{\partial \Pi_i}{\partial q_i}(q_1, \ldots, q_n) = 0,$$

we obtain the reaction function for firm i, which will have the general form

$$f_i(q_1, \ldots, q_n) = g_{\mathbf{C}_i}(Q - q_i),$$

where $Q = \sum_{i=1}^{n} q_i$ is the total market supply, and $g_{\mathbf{C}_i} : \mathbb{R}^+ \to \mathbb{R}^+$, $\mathbb{R}^+ = [0, \infty)$, is a one-dimensional map depending on parameter $\mathbf{C}_i \in \mathbb{R}^m$. In addition, the closure of the support of the map $g_{\mathbf{C}_i}$ is given by

$$\text{Cl}\{q \in \mathbb{R}^+ : g_{\mathbf{C}_i}(q) > 0\} = [0, q_{\mathbf{C}_i}],$$

and therefore the map $g_{\mathbf{C}_i}$ has an absolute maximum value x_{M_i}.

Of course, there are more sophisticated ways of generating the function f_i, $i = 1, 2, \ldots, n$. For instance, under adaptive expectations we may assume that

$$f_i(q_1, \ldots, q_n) = (1 - \lambda_i)q_i + \lambda_i g_{\mathbf{C}_i}(Q - q_i),$$

where $\lambda_i \in [0, 1]$. Notice that $\lambda_i = 1$ gives us the case of naive expectations. Another alternatives which do not imply a optimization process can be see e.g. Bischi and Baiardi (2015) as for instance

$$f_i(q_1, \ldots, q_n) = q_i + \lambda_i \Pi_i(q_1, \ldots, q_n),$$

or

$$f_i(q_1, \ldots, q_n) = q_i + \frac{\partial \Pi_i}{\partial q_i}(q_1, \ldots, q_n).$$

Below, we will show some examples on different oligopoly models that we can construct by assuming different economic conditions. We restrict ourselves to the case of naive and adaptive expectations.

3.1 Theocharis' Model

In Theocharis' model (see Theocharis 1959), it is assumed that inverse demand function gives the price in the form $p = a - bQ$ and cost functions $C_i(q_i) = c_i q_i$, $i = 1, 2, \ldots, n$. Let $Q_i = Q - q_i$, for $i = 1, 2, \ldots, n$ denote the residual supply.

Hence, the profit functions are given by

$$\Pi_i(q_1, \ldots, q_n) = (a - bQ)q_i - c_i q_i$$

and then one easily obtains from the first order condition that

$$g_{a,b,c_i}(q_1, \ldots, q_n) = \frac{a - c_i}{2b} - \frac{1}{2}Q_i,$$

and hence, for naive expectations we have that

$$\begin{aligned} f_i(q_1, \ldots, q_n) &= \max\{0, g_{a,b,c_i}(q_1, \ldots, q_n)\} \\ &= \max\left\{0, \frac{a - c_i}{2b} - \frac{1}{2}Q_i\right\}, \end{aligned}$$

because obviously a firm cannot produce a negative quantity. For adaptive expectations we have

$$\begin{aligned} f_i(q_1, \ldots, q_n) &= \max\{0, (1 - \lambda_i)q_i + \lambda_i g_{a,b,c_i}(q_1, \ldots, q_n)\} \\ &= \max\left\{0, (1 - \lambda_i)q_i + \lambda_i\left(\frac{a - c_i}{2b} - \frac{1}{2}Q_i\right)\right\}. \end{aligned}$$

For other kinds of above mentioned adjustments we will have

$$f_i(q_1, \ldots, q_n) = \max\{0, q_i + \lambda_i[(a - bQ)q_i - c_i q_i]\},$$

and

$$f_i(q_1, \ldots, q_n) = \max\{0, q_i + \lambda_i(a - c_i - bQ_i - 2bq_i)\}.$$

In Theocharis' model, all the firms react following naive expectations, but in general, firms can react in different ways, not all of them following the same strategy and, even more, they could change their strategies with time. Therefore, even in this simple case we are forced to make assumptions to simplify the different models that we can generate.

Basically, these simplifications consist of making the market more homogeneous, that is, assuming that all the constants characterizing each firms are the same for the firms which share a common strategy. In this case, this means that some of the firms following e.g. naive expectations have the same marginal costs c_i while if they follow adaptive expectations, then the adaptive constant λ is the same for all the firms. Then, the model has some symmetries which lead us to a simplified model defined on a subset (submanifold) on the phase space. The idea is to try to analyze and obtain some information of the whole model from the simplified one. Let us show it with an example based on Theocharis' assumptions.

First, we consider the naive expectations case, where the system of difference equations reads as

$$q_i(t+1) = \max\left\{0, \frac{a-c_i}{2b} - \frac{1}{2}Q_i\right\},$$

for $i = 1, 2, \ldots, n$. Note that at those points which make the reaction functions smooth enough, the Jacobian matrix is equal to

$$\mathbf{J} = \begin{pmatrix} 0 & -1/2 & -1/2 & \ldots & -1/2 \\ -1/2 & 0 & -1/2 & \ldots & -1/2 \\ -1/2 & -1/2 & 0 & \ldots & -1/2 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ -1/2 & -1/2 & -1/2 & \ldots & 0 \end{pmatrix},$$

which is symmetric and therefore with non complex eigenvalues. By Sect. 2.6 we see that $\mathbf{J}$ has two eigenvalues $\frac{1}{2}$ and $\frac{1-n}{2}$ with multiplicity $n-1$ and 1, respectively. The second eigenvalue can be obtained as well by noticing that the linear system of difference equations

$$\mathbf{y}(t+1) = \mathbf{J} \cdot \mathbf{y}(t)$$

leaves invariant the subset $\{\mathbf{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n : y_1 = y_2 = \ldots = y_n\}$ which is a linear subspace of dimension 1. On this subspace, the above system reads as

$$y(t+1) = \frac{1-n}{2} y(t).$$

When $n > 1$, that is, the model is not a monopoly, the determinant $|\mathbf{J} - \mathbf{I}_n| \neq 0$ and hence the system of equations

$$q_i = \frac{a-c_i}{2b} - \frac{1}{2}Q_i, \; i = 1, 2, \ldots, n,$$

has a unique solution, called Cournot equilibrium, which will be stable provided $n = 3$ and asymptotically stable when $n < 3$. If $n > 3$, then $\frac{1-n}{2} < -1$ and Cournot equilibrium is unstable. The existence of Cournot points allows us to define another reaction for the firms, that is,

$$f_i(q_1, \ldots, q_n) = \overline{q}_i,$$

where $\overline{q}_i$ is the ith coordinate of Cournot equilibrium (see Hommes et al. 2011).

Note that Cournot equilibrium is also a fixed point when we consider adaptive expectations and solve the system of equations

$$q_i = (1-\lambda_i)q_i + \lambda_i \left(\frac{a-c_i}{2b} - \frac{1}{2}Q_i \right),\ i = 1, 2, \ldots, n.$$

In this case, the Jacobian matrix is

$$\mathbf{J} = \begin{pmatrix} 1-\lambda_1 & -\lambda_1/2 & -\lambda_1/2 & \ldots & -\lambda_1/2 \\ -\lambda_2/2 & 1-\lambda_2 & -\lambda_2/2 & \ldots & -\lambda_2/2 \\ -\lambda_3/2 & -\lambda_3/2 & 1-\lambda_3 & \ldots & -\lambda_3/2 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ -\lambda_n/2 & -\lambda_n/2 & -\lambda_n/2 & \ldots & 1-\lambda_n \end{pmatrix},$$

which is no longer symmetric. To obtain significative results, we must to make extra assumptions to simplify the model, which can be easily done by assuming that all firms make their adjustment with the same constant $\lambda_i = \lambda$ for $i = 1, \ldots., n$. Hence, the Jacobian reads as

$$\mathbf{J} = \begin{pmatrix} 1-\lambda & -\lambda/2 & -\lambda/2 & \ldots & -\lambda/2 \\ -\lambda/2 & 1-\lambda & -\lambda/2 & \ldots & -\lambda/2 \\ -\lambda/2 & -\lambda/2 & 1-\lambda & \ldots & -\lambda/2 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ -\lambda/2 & -\lambda/2 & -\lambda/2 & \ldots & 1-\lambda \end{pmatrix},$$

and eigenvalues can be easily obtained as shown in Sect. 2.6. Namely, solving the equation $1-\lambda-t = -\lambda/2$ we obtain the eigenvalue $t = 1-\lambda/2$, which has an eigenspace of dimension $n-1$, while $1-\lambda\frac{n+1}{2}$ is the other eigenvalue. This eigenvalue can be obtained as well by noticing that the linear system of difference equations

$$\mathbf{y}(t+1) = \mathbf{J} \cdot \mathbf{y}(t)$$

leaves invariant the subset $\{\mathbf{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n : y_1 = y_2 = \ldots = y_n\}$ which is a linear subspace of dimension 1. On this subspace, the above system reads as

$$y(t+1) = \left(1 - \lambda\frac{n+1}{2}\right) y(t).$$

Recall that $\lambda \in [0, 1]$, and then the absolute value $|\lambda/2 - 1| < 1$. On the other hand, for the other eigenvalue we have that $|1-\lambda\frac{n+1}{2}| \leq 1$ provided $\lambda \in [0, \frac{4}{n+1}]$. Note that for $n = 2, 3$ the eigenvalues are bounded by 1 in modulus for any value of λ. For $n \geq 4$, then we have to reduce the value of λ to obtain that Cournot equilibrium is asymptotically stable.

Note however that our models are not linear because of the maximum operator. Because of this, the quantity that a single firm may produce is bounded by $\lambda\frac{a-c_i}{2b}$, and so the production is contained in the interval $[0, \lambda\frac{a-c_i}{2b}]$. If we take the equations

$$q_i(t+1) = (1-\lambda)q_i(t) + \lambda\left(\frac{a-c_i}{2b} - \frac{1}{2}Q_i(t)\right),\ i = 1, 2, \ldots, n$$

and sum up them, we obtain

$$Q(t+1) = (1-\lambda)Q(t) + \lambda\left(\frac{na-nc}{2b} - \frac{n-1}{2}Q(t)\right),$$

where $c = (c_1 + \ldots + c_n)/n$ is the average marginal cost. The fixed point of above equation is $\overline{Q} = \frac{n(a-c)}{(n+1)b}$. Hence, the set

$$\mathscr{S} = \{(q_1, \ldots, q_n) : q_1 + \ldots + q_n = \overline{Q},\ q_i \geq 0,\ i = 1, \ldots, n\}$$

is invariant and any orbit with initial conditions on $\mathscr{S}$ converges to Cournot equilibrium. To obtain clear results for the evolution of initial conditions outside $\mathscr{S}$ we must make another additional assumption, which is that firms are homogeneous, that is, $c_i = c$ for all $i = 1, 2, \ldots, n$. Then, it is proved in Cánovas et al. (2008) that, under naive expectations, any orbit starting outside $\mathscr{S}$ converges to a 2-periodic point $(0, 0, \ldots, 0)$ and $(\frac{a-c}{2b}, \frac{a-c}{2b}, \ldots, \frac{a-c}{2b})$, that is, in one time all the firms will produce nothing, while in the next period of time they produce the maximum quantity allowed. It is interesting to point out that the average profit for the 2-periodic point is given by

$$\begin{aligned} prof1 &= \frac{\Pi_i(0, 0, \ldots, 0) + \Pi_i(\frac{a-c}{2b}, \frac{a-c}{2b}, \ldots, \frac{a-c}{2b})}{2} \\ &= (a - n\frac{a-c}{2})\frac{a-c}{4b} - c\frac{a-c}{4b} \\ &= \frac{(a-c)^2}{4b} - n\frac{(a-c)^2}{8b} = \frac{(a-c)^2}{4b}\left(1 - \frac{n}{2}\right), \end{aligned}$$

while the profit at Cournot equilibrium $\left(\frac{a-c}{(n+1)b}, \frac{a-c}{(n+1)b}, \ldots, \frac{a-c}{(n+1)b}\right)$ is

$$\begin{aligned} prof2 &= \Pi_i\left(\frac{a-c}{(n+1)b}, \frac{a-c}{(n+1)b}, \ldots, \frac{a-c}{(n+1)b}\right) \\ &= \left(a - n\frac{a-c}{(n+1)}\right)\frac{a-c}{(n+1)b} - c\frac{a-c}{(n+1)b} \\ &= \frac{(a-c)^2}{(n+1)b} - n\frac{(a-c)^2}{(n+1)^2 b} = \frac{(a-c)^2}{(n+1)b}\left(1 - \frac{n}{n+1}\right). \end{aligned}$$

In the case of Theocharis' model, when $\lambda = 1$, we have that *prof* 1 is always negative, while *prof* 2 is positive, although decreasing with the number of firms n. It is unclear whether such scheme is repeated under adaptative expectations. When all the firms are homogeneous, then the set

$$\Delta = \{(q, q, \dots, q) \in \mathbb{R}^n : q \geq 0\}$$

is invariant, and then the system of difference equations on Δ reduces to the single difference equation

$$q(t+1) = \max\left\{0, \lambda \frac{a-c}{2b} + (1 - \frac{\lambda}{2}(n+1))q(t)\right\}.$$

Clearly, then hyperplane $q_1 + \ldots + q_n = 0$ is orthogonal to Δ. Since this hyperplane is in fact the eigenspace $\ker(\mathbf{J} - \frac{1}{2}\mathbf{I}_n)$, then the dynamics outside Δ synchronizes to Δ, that is, the dynamics of Δ characterizes the dynamics of the n-dimensional model. Let us remark that the model is not linear, but it is locally linear in the sense that for any point in the interior of the feasible set $[0, \lambda \frac{a-c_i}{2b}]^n$ the model is linear.

3.2 Non Homogeneous Models

Now, we consider two different ways to make firms non homogeneous. First, they have the same reaction function but their marginal cost may be different, namely $c_1 \geq c_2 \geq \ldots \geq c_n$. Then, the symmetric properties of the system do not hold, Δ is no longer invariant and Cournot equilibrium is not placed on it. In addition, under some assumptions, Cournot coordinates may become negative, which is not possible since we assumed each firm produces at leat zero, and then firms may disappear from the market when naive expectations are assumed (see Cánovas 2009). In particular, if $a + c_n \leq 2c_i$, $i = 1, \ldots, n-1$, then, under some assumptions, the model evolves to a monopoly in which the only active firm produces the constant output $(a - c_n)/2b$. The above mentioned assumptions are the following:

- $n = 2, 3, 4$.
- $n = 5$ and $a + c_5 < 2c_1$.
- $n \geq 6$ and $n(a - c) < 3(a - c_n)$.

One can see that, when the number of firms increases, the conditions to evolve to a monopoly are more difficult to be fulfilled by the system. If $a + c_n > 2c_{n-1}$, the system still can reduce the number of firms and evolve to a duopoly when $a + c_n + c_{n-1} \leq 3c_i$, $i = 1, \ldots, n-2$, and one of the following conditions are satisfied:

- $n = 3$.
- $n = 4$ and $a + c_4 < c_1 + c_2$.
- $n = 5$ and $a + c_5 < 2c_1$.
- $n \geq 6$ and $n(a - c) < 3(a - c_n)$.

In both cases, monopoly and duopoly, the results are given by the fact that for $n = 1$ or 2 the system converges to the Cournot equilibrium. If $a + c_n + c_{n-1} > 3c_{n-2}$, then we could investigate whether the system may reduce the number of firms, but we have the problem that the model with three firms does not converge to the Cournot equilibrium unless their production were contained in the set $\mathscr{S}$ (for three firms). Therefore, reducing firms from the system is only possible when firms collaborate for that. Let us consider the following example when only two marginal costs c_1 and c_n, $c_1 > c_n$, are possible. Suppose that m firms produce at marginal cost c_1. Then, we have two different Cournot coordinates

$$\overline{q}_1 = \frac{a + nc - (n+1)c_1}{(n+1)b} = \frac{a - (n-m+1)c_1 + (n-m)c_n}{(n+1)b},$$

and

$$\overline{q}_n = \frac{a + nc - (n+1)c_n}{(n+1)b} = \frac{a + mc_1 - (m+1)c_n}{(n+1)b}.$$

The condition $\overline{q}_1 \leq 0$ is necessary to have that the first m firms disappear from the market. Such condition is fulfilled when

$$a + (n-m)c_n \leq (n-m+1)c_1. \tag{1}$$

However, if we have that $q_1(t) = \ldots = q_m(t) = 0$ for some time t, then the remaining firms have to produce at this time $q_{m+1}(t) + \ldots + q_n(t) = \frac{n-m}{n-m+1}\frac{a-c_n}{b}$. Otherwise, the piecewise linear model given by firms $m+1, \ldots, n$ will converge to the two periodic point given by $(0, 0, \ldots, 0)$ and $(\frac{a-c}{2b}, \frac{a-c}{2b}, \ldots, \frac{a-c}{2b})$, and zero production implies that the first m firms will start to produce again. Note that the system has new symmetries and now the set

$$\Delta_2 = \{(\overbrace{q_1 \ldots, q_1}^{m}, \overbrace{q_n, \ldots, q_n}^{n-m}) : q_1, q_n \in \mathbb{R}^+\}$$

is invariant, and on it, the system reads as

$$\begin{cases} q_1(t+1) = \max\left\{0, \frac{a-c_1}{2b} - \frac{(m+1)q_1(t)+(n-m)q_n(t)}{2}\right\}, \\ q_n(t+1) = \max\left\{0, \frac{a-c_n}{2b} - \frac{mq_1(t)+(n-m+1)q_n(t)}{2}\right\}. \end{cases}$$

If condition (1) is fulfilled, then $(0, \frac{a-c_n}{b(n-m+1)})$ is a fixed point and hence a Cournot point of the system. Now, the Jacobian matrix

$$\mathbf{J} = \begin{pmatrix} -\frac{m+1}{2} & -\frac{n-m}{2} \\ -\frac{m}{2} & -\frac{n-m+1}{2} \end{pmatrix}$$

has eigenvalues $1/2$ and $-(n+1)/2$, as it is expected.

Question 1 It is unclear whether the above results can be translated for adaptative expectations. More precisely, what are the conditions on the marginal costs to guarantee that the production $q_i(t)$ of the less efficient firms satisfy that $\lim_{t\to\infty} q_i(t) = 0$.

The second possibility of making the model non homogenous is by assuming that firms make their adjustments following different strategies, that is, different reaction functions. To fix ideas, denote by f_N the reaction function under naive expectations, f_A under adaptive expectations and f_C when the firm plays Cournot. Assume that n_1 first firms react by using f_N, n_2 firms do it using f_N and $n - n_1 - n_2$ firms play Cournot. Then the system can be written by

$$\begin{cases} q_i(t+1) = f_N(q_1(t),,,.q_n(t)), & i = 1,\dots,n_1, \\ q_i(t+1) = f_A(q_1(t),,,.q_n(t)), & i = n_1+1,\dots,n_1+n_2, \\ q_i(t+1) = f_C(q_1(t),,,.q_n(t)), & i = n_1+n_2+1,\dots,n. \end{cases}$$

The Jacobian $\mathbf{J}$ at Cournot point is

$$\begin{pmatrix} 0 & -1/2 & \dots & -1/2 & -1/2 & -1/2 & \dots & -1/2 & -1/2 & \dots & -1/2 \\ -1/2 & 0 & \dots & -1/2 & -1/2 & -1/2 & \dots & -1/2 & -1/2 & \dots & -1/2 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ -1/2 & -1/2 & \dots & 0 & -1/2 & -1/2 & \dots & -1/2 & -1/2 & \dots & -1/2 \\ -\lambda/2 & -\lambda/2 & \dots & -\lambda/2 & 1-\lambda & -\lambda/2 & \dots & -\lambda/2 & -\lambda/2 & \dots & -\lambda/2 \\ -\lambda/2 & -\lambda/2 & \dots & -\lambda/2 & -\lambda/2 & 1-\lambda & \dots & -\lambda/2 & -\lambda/2 & \dots & -\lambda/2 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ -\lambda/2 & -\lambda/2 & \dots & -\lambda/2 & -\lambda/2 & -\lambda/2 & \dots & 1-\lambda & -\lambda/2 & \dots & -\lambda/2 \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & \dots & 0 \end{pmatrix}$$

which has eigenvalues 0, with an eigenspace of dimension $n - n_1 - n_2$, $1/2$ with an eigenspace of dimension $n_1 - 1$, $1 - \lambda/2$, which has an eigenspace of dimension $n_2 - 1$, and two unknown eigenvalues v_1 and v_2. The values of v_1 and v_2 will give us the stability conditions of Cournot equilibrium. However, they are hard to compute, so we consider the system defined in the space

$$\Delta_3 = \{(\overbrace{q_1\dots,q_1}^{n_1},\overbrace{q_2,\dots,q_2}^{n_2},\overbrace{\overline{q},\dots,\overline{q}}^{n-n_1-n_2}) : q_1,q_2 \in \mathbb{R}^+\},$$

where $\overline{q}$ is the Cournot equilibrium. Then Δ_3 is invariant and the system on it reads as

$$\begin{cases} q_1(t+1) = \max\left\{0, \frac{a-c}{2b} - \frac{n-n_1-n_2}{2}\overline{q} - \frac{(n_1-1)q_1(t)+n_2q_2(t)}{2}\right\}, \\ q_2(t+1) = \max\left\{0, q_2(t) + \lambda\left(\frac{a-c}{2b} - \frac{n-n_1-n_2}{2}\overline{q} - \frac{n_1q_1(t)+(n_2+1)q_2(t)}{2}\right)\right\}, \\ q_3(t+1) = \overline{q}. \end{cases}$$

As the third function is constant which implies zero rows in the Jacobian matrix, we can consider the system as 2 dimensional with Jacobian matrix

$$\mathbf{J} = \begin{pmatrix} \frac{1-n_1}{2} & -\frac{n_2}{2} \\ -\lambda\frac{n_1}{2} & 1-\lambda\frac{n_2+1}{2} \end{pmatrix}.$$

Clearly, the stability of the reduced system will imply the stability of the general system. The eigenvalues of $\mathbf{J}$ are

$$\frac{3-n_1-\lambda(1+n_2) \pm \sqrt{n_1^2+2n_1(1+\lambda(n_2-1))+(\lambda(n_2+1)-1)^2}}{4},$$

which are always real numbers depending on n_1 and n_2, that is, the firms playing Cournot do not have influence in the system stability. They are equal to 1 when $\lambda = 0$, which is not an interesting case, and equal to -1 when

$$\lambda = \frac{4(n_1-3)}{n_1-3(n_2+1)}.$$

It can be easily seen numerically that

$$\left|\frac{3-n_1-\lambda(1+n_2) + \sqrt{n_1^2+2n_1(1+\lambda(n_2-1))+(\lambda(n_2+1)-1)^2}}{4}\right|$$

is bounded by 1, and then, the stability of the system depends on the number

$$\left|\frac{3-n_1-\lambda(1+n_2) - \sqrt{n_1^2+2n_1(1+\lambda(n_2-1))+(\lambda(n_2+1)-1)^2}}{4}\right| \tag{2}$$

which is hard to study analytically in detail. When $n_1 = 0$, that is, all the firms play naive expectations, the above value is

$$\left|\frac{3-\lambda(1+n_2) - \sqrt{(\lambda(n_2+1)-1)^2}}{4}\right| = \begin{cases} \left|\frac{2-\lambda(1+n_2)}{2}\right| \ \textit{if}\ \lambda(n_2+1) \geq 1, \\ \frac{1}{2}\ \textit{if}\ \lambda(n_2+1) \leq 1. \end{cases}$$

The value $\left|\frac{2-\lambda(1+n_2)}{2}\right|$ can be greater than 1 when $\lambda(1+n_2) \geq 2$, that is, $\frac{\lambda(1+n_2)-2}{2} = 1$, which implies

$$\lambda = \frac{4}{1+n_2},$$

and again the condition $\lambda \in [0, \frac{4}{1+n_2}]$ must be fulfilled to have a stable system. Similarly, when $n_2 = 0$, we get that $n_1 \leq 3$. For both values greater than one, the following table shows the stability pairs (n_1, n_2), that is the pairs (n_1, n_2) such that (2) is smaller than 1, for some values of λ.

- For $\lambda = 0.9$, the pairs (n_1, n_2) are $(1,1)$, $(1,2)$, $(2,1)$.
- For $\lambda = 0.8$, the pairs (n_1, n_2) are $(1,1)$, $(1,2)$, $(2,1)$.
- For $\lambda = 0.7$, the pairs (n_1, n_2) are $(1,1)$, $(1,2)$, $(1,3)$, $(2,1)$.
- For $\lambda = 0.6$, the pairs (n_1, n_2) are $(1,1)$, $(1,2)$, $(1,3)$, $(2,1)$.
- For $\lambda = 0.5$, the pairs (n_1, n_2) are $(1,1)$, $(1,2)$, $(1,3)$, $(1,4)$, $(2,1)$, $(2,2)$.
- For $\lambda = 0.25$, the pairs (n_1, n_2) are $(1,1)$, $(1,2)$, $(1,3)$, $(1,4)$, $(1,5)$, $(1,6)$, $(1,7)$, $(1,8)$, $(2,1)$, $(2,2)$, $(2,3)$, $(2,4)$.

In the above case, the reaction functions remain unchanged along the process, but firms may decide to change their reaction functions with time (see Hommes et al. 2011, Droste et al. 2002, Bischi et al. 2015 or Baiardi et al. 2015). This can be defined using a skew product map as follows. Consider $\Sigma = \{(x_n)_{n=0}^{\infty} : x_n \in \{N, A, C\}, n \geq 0\}$ be the infinite sequences of letters in the alphabet $\{N, A, C\}$.[8] The shift map $\sigma : \Sigma \to \Sigma$ is then defined by $\sigma((x_n)_{n=0}^{\infty}) = (x_{n+1})_{n=0}^{\infty} = (x_n)_{n=1}^{\infty}$. Let $T : \Sigma^n \times (\mathbb{R}^+)^n \to \Sigma^n \times (\mathbb{R}^+)^n$ be given by

$$T((x_i^1), \ldots (x_i^n), q_1, \ldots, q_n) = (\sigma(x_i^1), \ldots, \sigma(x_i^n), f_{x_0^1}(q_1, \ldots, q_n), \ldots, f_{x_0^n}(q_1, \ldots, q_n)).$$

The map T contains all the possible trajectories for initial productions when the firms have the freedom of choosing the reaction function. Of course, we should restrict the map T by some rules, endogenous or not, which reduces the size of Σ^n.

Question 2 It is an open question to provide tools to analyze the above map T, as well as finding rules to reduce the size of Σ^n.

4 Non (Piecewise) Linear Oligopolies

The above section was devoted to analyze a model which is piecewise linear. There are many oligopoly models which are not piecewise linear. Here, we consider two of them given by two different reaction functions which are unimodal. In particular,

[8]From now on we denote $(x_n)_{n=0}^{\infty}$ by (x_n) if there is no ambiguity on that.

we assume that we are working with maps $f : \mathbb{R}^+ \to \mathbb{R}^+$ smooth enough and such that the following hypothesis are fulfilled:

h0. There is a map $g : \mathbb{R}^+ \to \mathbb{R}^+$ such that there is $x_M \in (0, x_0)$ such that $g|_{(0,x_M)}$ is strictly increasing and $g|_{(x_M,x_0)}$ is strictly decreasing. The maximum x_M is called the turning point of g.
h1. $g^{-1}(0) = \{0, x_0\}, x_0 \in \mathbb{R}^+$.
h2. $g(x) > 0$ for $x \in (0, x_0)$.
h3. Then $f(x) = \max\{0, g(x)\}$.

The two examples we are going to analyze are Puu's model (see Puu 1991), where $f(x) = \max\{0, \sqrt{x/c} - x\}$ and Kopel's model (see Kopel 1996) where $f(x) = \max\{0, ax(1-x)\}$ is given by the well-known logistic map. The oligopoly is then defined under naive expectations by

$$q_i(t+1) = f(Q_i(t)),\ i = 1, \dots, n,$$

and under adaptive expectations by

$$q_i(t+1) = \max\{0, (1-\lambda)q_i(t) + \lambda g(Q_i(t))\},\ i = 1, \dots, n.$$

In the next section, we are going to analyze both oligopolies showing some differences between them. Previously, we gave a general framework for analyzing these models following the ideas used in the piecewise linear Theocharis' model. If all the firms are homogeneous, the space

$$\Delta = \{(q, q, \dots, q) \in \mathbb{R}^n : q \geq 0\}$$

is invariant by the model. On Δ the model reads as

$$q(t+1) = f((n-1)q(t)) = \max\{0, g((n-1)q(t))\}$$

under naive expectations and

$$q(t+1) = \max\{0, (1-\lambda)q(t) + \lambda g((n-1)q(t))\}$$

for adaptive expectations. Let $q_0 > 0$ be such that $g(q_0) = 0$. Then, under naive expectations we get positive productions when $q \in (0, \frac{q_0}{n-1})$. Since g is unimodal, let q_M be the turning point of g, and note that $g(q_M)$ is the maximum output given by the system. When $g(q_M) \geq \frac{q_0}{n-1}$, we have that g has a 2-horseshoe, and therefore we prove that the dynamics on Δ is topologically chaotic because its topological entropy is equal to $\log 2$. In addition, if $g(q_M) > \frac{q_0}{n-1}$, then there is an interval J containing the turning point such that $f(J) = \{0\}$ and numerical simulations will show that all the orbits go eventually to zero (moreover, when we take initial conditions outside Δ, numerical simulations show that all the outputs go to zero). In other words, the set $\cup_{n\geq 0} f^{-n}(J)$ seems to have full Lebesgue measure on $(0, \frac{q_0}{n-1})$ and therefore the chaotic dynamics lies in a residual set of apparently zero Lebesgue measure.

When adaptive expectations are assumed, the linear part $(1-\lambda)q$ goes to infinite as q tends to infinite, while the nonlinear part $\lambda g((n-1)q)$ tends to minus infinite, and so it is not guaranteed the existence of q_0 as above. If such number q_0 exists, that is, there is $q_0 = q_0(n)$ such that $(1-\lambda)q + \lambda g((n-1)q) = 0$, then there is $q_M \in (0, q_0)$ such that $(1-\lambda)q + \lambda g((n-1)q)$ attains its maximum value at q_M, and the above reasoning made for naive expectations makes sense in the adaptive expectations case.

On Δ, the Jacobian matrix is

$$\mathbf{J} = \begin{pmatrix} 1-\lambda & \lambda g'((n-1)q) & \lambda g'((n-1)q) & \dots & \lambda g'((n-1)q) \\ \lambda g'((n-1)q) & 1-\lambda & \lambda g'((n-1)q) & \dots & \lambda g'((n-1)q) \\ \lambda g'((n-1)q) & \lambda g'((n-1)q) & 1-\lambda & \dots & \lambda g'((n-1)q) \\ \dots & \dots & \dots & \dots & \dots \\ \lambda g'((n-1)q) & \lambda g'((n-1)q) & \lambda g'((n-1)q) & \dots & 1-\lambda \end{pmatrix}$$

under adaptive expectations. Naive expectations are obtained by letting $\lambda = 1$. Consider the vector $(1, 1, \dots, 1)$ and note that

$$\mathbf{J} \cdot \begin{pmatrix} 1 \\ 1 \\ \dots \\ 1 \end{pmatrix} = (1-\lambda+\lambda(n-1)g'((n-1)q)) \begin{pmatrix} 1 \\ 1 \\ \dots \\ 1 \end{pmatrix}.$$

On the other hand, the subspace generated by the vectors

$$\{(1,-1,0,0,\dots,0),(0,1,-1,0,\dots,0),\dots,(0,0,\dots 0,1,-1)\}\}$$

is orthogonal to Δ and

$$\mathbf{J} \cdot \begin{pmatrix} 0 \\ \dots \\ 0 \\ 1 \\ -1 \\ 0 \\ \dots \\ 0 \end{pmatrix} = (1-\lambda-\lambda g'((n-1)q)) \begin{pmatrix} 0 \\ \dots \\ 0 \\ 1 \\ -1 \\ 0 \\ \dots \\ 0 \end{pmatrix}.$$

Hence, on Δ, the Lyapunov exponent is given by

$$lyex_{||}(q) = \lim_{m\to\infty} \frac{1}{m} \sum_{i=0}^{m-1} \log |1 - \lambda + \lambda(n-1)g'((n-1)q(i))|$$

while the normal Lyapunov exponent is

$$lyex_{\perp}(q) = \lim_{m\to\infty} \frac{1}{m} \sum_{i=0}^{m-1} \log |1 - \lambda - \lambda g'((n-1)q(i))|$$

where $q(i)$ ranges the orbit of $(q, q, \ldots, q)$ for some $q \geq 0$. Fixing an invariant measure μ on Δ, note that $lyex_{||}(q)$ and $lyex_{\perp}(q)$ are well defined for almost all q related to μ. In addition, $lyex_{||}(q)$ will provide us information on the dynamics on Δ, while $lyex_{\perp}(q)$ informs us on when initial conditions outside Δ converge to attractors on Δ, that is, when firms locally synchronize (see Ashwin et al. (1996) for more information). The fact that firms do synchronize is very important because it is a commonly accepted fact, although we will show in our examples that sometimes they fail to synchronize.

In the remaining of this chapter, we are going to consider the above general description to see how Puu's and Kopel's oligopolies behave. We will show that there are two big differences between these models: the role of the parameters and the synchronization properties.

4.1 Puu's Oligopoly

Puu's oligopoly was introduced in fact as a duopoly in Puu (1991). The assumptions for the model are a demand function (in inverse form) given by $p = \frac{1}{Q}$ and linear cost functions $C_i = c_i q_i$ for each firm $i = 1, 2, \ldots, n$. Then the profit function is given by

$$\Pi_i = \frac{q_i}{Q} - c_i q_i.$$

Then, the reaction function is $g_{c_i}(Q_i) = \sqrt{Q_i/c} - Q_i$.

The first important thing regarding this model is that, when all the firms are homogenous, the parameter c does not play any role, it is just a change of scale. To see this, consider the homeomorphism $\varphi(x) = x/c$, and note that if $g_c(x) = \sqrt{x/c} - x$, then

$$(\varphi^{-1} \circ g_c \circ \varphi)(x) = g_1(x) = \sqrt{x} - x,$$

and

$$\varphi^{-1}((1-\lambda)\varphi(x)+\lambda g_c(\varphi(x))) = (1-\lambda)x+\lambda g_1(x) = (1-\lambda)x-\lambda(\sqrt{x}-x),$$

for all $x \geq 0$. Then, it is straightforward to see that the model with parameter $c \neq 1$ and the model with parameter 1 are conjugated via the linear homeomorphism $\varphi \times \varphi \times \ldots \times \varphi$. So, we may assume that $c = 1$ without any restriction, and we have to take care on λ and the number of firms n.

Once we have reduced the number of significative parameters, we are going to analyze how the map shapes. Consider the auxiliary map

$$\begin{aligned}\overline{g}(x) &= (1-\lambda)x+\lambda\left(\sqrt{(n-1)x}-(n-1)x\right)\\ &= (1-n\lambda)x+\lambda\sqrt{(n-1)x},\end{aligned}$$

which we can call the best reply to the rest of the industry. The maximum x_M of $\overline{g}$ has to satisfy the condition

$$1-\lambda n+\frac{\lambda(n-1)}{2\sqrt{(n-1)x_M}} = 0,$$

which reduces to

$$\sqrt{(n-1)x_M} = \frac{\lambda(n-1)}{2(\lambda n-1)},$$

which is positive if and only if $\lambda n-1 > 0$, which implies that $\lambda > 1/n$. In that case

$$x_M = \frac{\lambda^2(n-1)}{4(\lambda n-1)^2}.$$

Hence, we have two possible shapes for $\overline{g}$ as follows:

- If $\lambda \leq 1/n$, the map $\overline{g}$ is an homeomorphism on $[0,+\infty)$ with two fixed points, 0 which is repulsive, and $\frac{n-1}{n^2}$, which does not depend on λ and attracts all the orbits on Δ starting on $(0,+\infty)$.
- If $\lambda > 1/n$, the map $\overline{g}$ is unimodal. It gives non negative outputs at the interval $\left[0, \frac{\lambda^2(n-1)}{(\lambda n-1)^2}\right]$, which ranges the interval $[0,\overline{g}(x_M)] = \left[0, \frac{\lambda^2(n-1)}{4(\lambda n-1)}\right]$. The condition

  $$\overline{g}(x_M) \leq \frac{\lambda^2(n-1)}{(\lambda n-1)^2}$$

  is equivalent to

  $$\frac{\lambda^2(n-1)}{4(\lambda n-1)} \leq \frac{\lambda^2(n-1)}{(\lambda n-1)^2},$$

which reduces to

$$\lambda \leq \frac{5}{n}. \tag{3}$$

The Cournot equilibrium can be computed easily from the equations

$$\sqrt{Q_i} - Q_i = q_i, \ i = 1, 2, \dots, n,$$

which gives us

$$\sqrt{Q_i} = Q, \ i = 1, 2, \dots, n,$$

and hence

$$\sqrt{Q_i} = \sqrt{Q_1}, \ i = 2, 3, \dots, n.$$

Taking into account that $q_i \geq 0$ for all $i = 1, 2, \dots, n$, we conclude that $q_i = q_1$ for all $i = 2, 3, \dots, n$ and therefore the Cournot equilibria lie on Δ. The equation

$$\sqrt{(n-1)q} = nq$$

gives us the Cournot coordinates $\overline{q}_0 = 0$ and

$$\overline{q} = \frac{n-1}{n^2}.$$

The Jacobian at the non negative Cournot equilibrium is

$$\mathbf{J} = \begin{pmatrix} 1-\lambda & -\lambda\frac{n-2}{2(n-1)} & -\lambda\frac{n-2}{2(n-1)} & \dots & -\lambda\frac{n-2}{2(n-1)} \\ -\lambda\frac{n-2}{2(n-1)} & 1-\lambda & -\lambda\frac{n-2}{2(n-1)} & \dots & -\lambda\frac{n-2}{2(n-1)} \\ -\lambda\frac{n-2}{2(n-1)} & -\lambda\frac{n-2}{2(n-1)} & 1-\lambda & \dots & -\lambda\frac{n-2}{2(n-1)} \\ \dots & \dots & \dots & \dots & \dots \\ -\lambda\frac{n-2}{2(n-1)} & -\lambda\frac{n-2}{2(n-1)} & -\lambda\frac{n-2}{2(n-1)} & \dots & 1-\lambda \end{pmatrix}$$

with eigenvalues $1 - \frac{n\lambda}{2(n-1)}$, with multiplicity $n-1$, and $1 + \frac{\lambda}{2}(n-4)$. The first one has modulus smaller than one if and only if

$$\frac{n\lambda}{2(n-1)} < 2,$$

which implies that

$$\lambda < \frac{4(n-1)}{n} = 4 - \frac{4}{n},$$

which is fulfilled for $\lambda \in [0, 1]$ when $n \geq 2$. The second eigenvalue has modulus greater than one when $n \geq 5$ and exactly one for $n = 4$. For $n = 2$ and 3, the modulus is smaller than one if

$$\frac{\lambda}{2}(4 - n) < 2,$$

which gives us

$$\lambda < \frac{4}{4 - n},$$

which is always fulfilled for $n = 2$ and 3. As a consequence, the Cournot equilibrium is locally asymptotically stable provided $n = 2$ and 3. Next, we compute with accuracy 10^{-5} the topological entropy and estimate the (normal) Lyapunov exponents.

The topological entropy is zero when the number of firms is equal to 2, 3 and 4. From Fig. 1, we check that when the number of firms increases, the set of λ's producing zero topological entropy decreases. This means that when the number of firms increases, the complexity also increases. Similarly, when the number of firms increases, the set of λ's producing topological entropy $\log 2$ increases, which is due to the fact that $\lambda > 5/n$. Recall that in this case, although we have positive entropy,

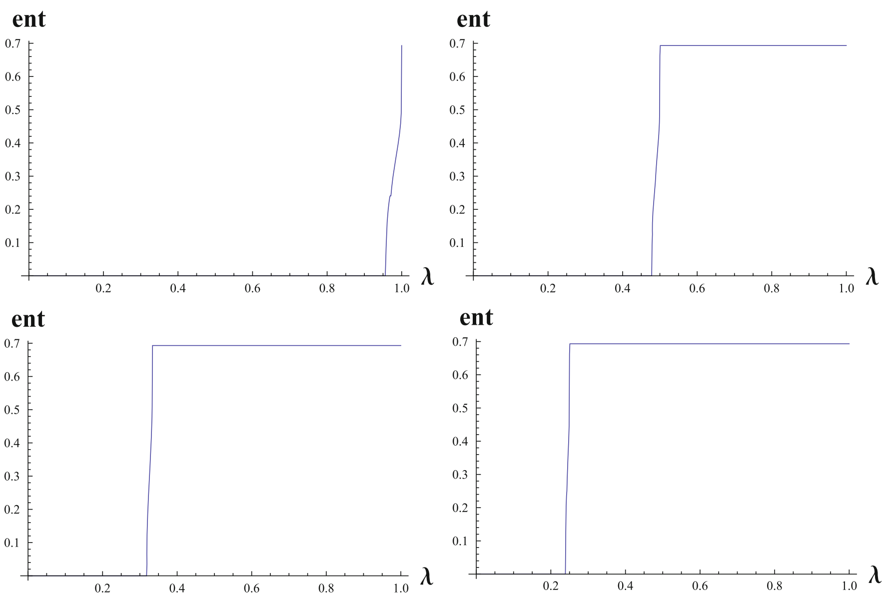


Fig. 1 The topological entropy of Puu's oligopoly on the invariant subset Δ is computed with accuracy 10^{-5} when the number of firms is 5 (*top-left*), 10 (*top-right*), 15 (*down-left*) and 20 (*down-right*), respectively

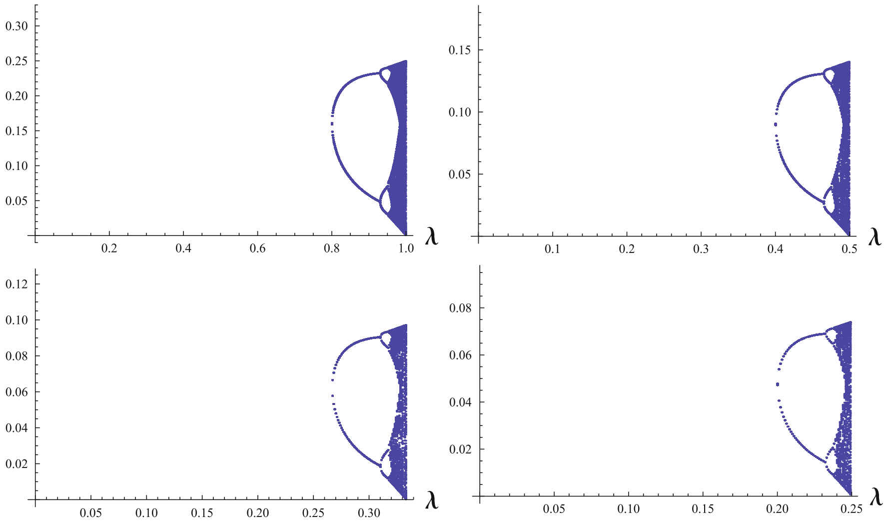


Fig. 2 Bifurcation diagrams of Puu's oligopoly on the invariant subset Δ when the number of firms is 5 (*top-left*), 10 (*top-right*), 15 (*down-left*) and 20 (*down-right*), respectively. A sample of length 25,000 of orbits starting at the turning point is computed, drawing the last 250 points

numerically one cannot observe any kind of complexity: the bifurcation diagrams of Fig. 2 show the classical double period bifurcation scheme when $\lambda \leq 5/n$, but when $\lambda > 5/n$ we obtain a fixed point at 0, where the profit function $\Pi_i : (q_1, \ldots, q_n) = \frac{q_i}{Q} - cq_i$ is not defined. In addition, the model is no longer C^1 when $\lambda > 5/n$ and then Lyapunov exponents cannot be computed (see Cánovas (2015) for a precise explanations of the above facts). So, we must concentrate to compute Lyapunov exponents when $\lambda < 5/n$, whose results are shown in Fig. 3.

Figure 3 shows that, at least for the values of λ making the model smooth enough, the normal Lyapunov exponent is negative, which will imply that at least locally, there is a synchronization of firms. This synchronization may be chaotic when the Lyapunov exponent is positive. We repeat our computations for firms ranging from 2 to 25. Since the topological entropy is zero for 2, 3 and 4 firms, there is not a chance of obtaining a chaotic synchronization. However, it appears when the number of firms is greater than 5. In particular, Fig. 4 shows the values of λ producing such chaotic synchronization.

From an economic point of view, Fig. 5 is remarkable. It shows the average profit of firms along two orbits, one of them starting at the turning point and the other one at Cournot equilibrium. We see that the average profit is higher when the orbit starts at the turning point, even when the orbit lies in a chaotic attractor. Although it is a simulation, we see a big difference with Theocharis' piecewise linear model.

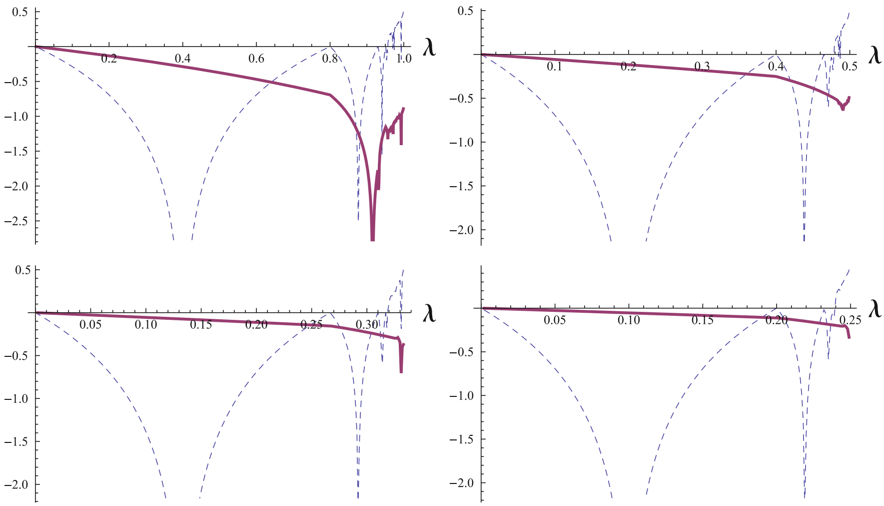


Fig. 3 Normal Lyapunov exponent estimations (*solid lines*) and the Lyapunov exponent estimations (*dashed lines*) for 5 firms (*top-left*), 10 (*top-right*), 15 (*down-left*) and 20 (*down-right*). A sample of 250,000 points starting at the turning point is computed for the estimations

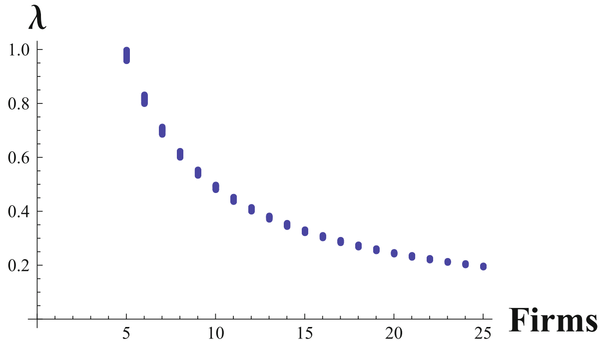


Fig. 4 For each number of firms, the range of λ's which produce local chaotic synchronization is showed. Note that obviously, λ values decrease when the number of firms increases due to the expression $\lambda < 5/n$

4.2 Kopel's Oligopoly

This model was introduced by Kopel (1996) with different economic assumptions.[9] One of them considers linear demand function (in inverse form) $p = \alpha - \beta Q$ and

[9]As in the case of Puu's oligopoly, Kopel's oligopoly was introduced first as a duopoly. In addition, the reaction function was obtained under different demand and cost functions.

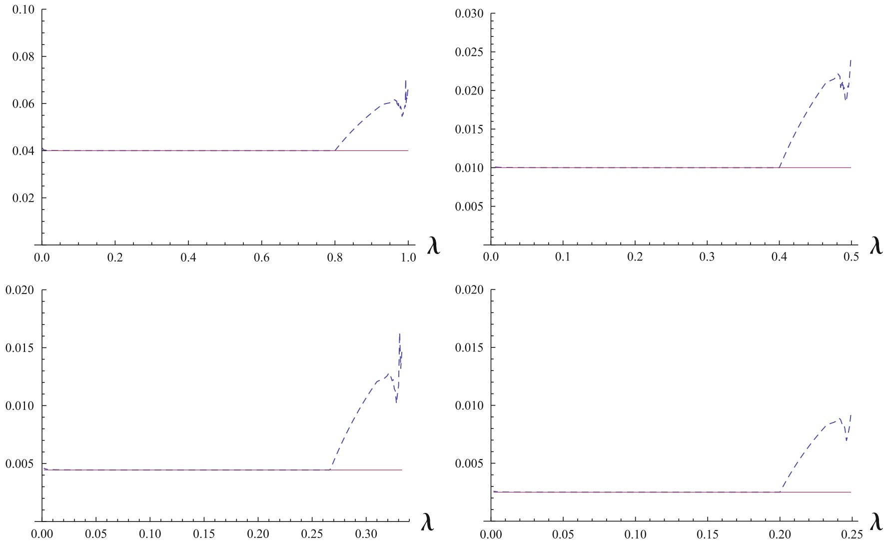


Fig. 5 Estimations of the average profit at Cournot equilibrium (*solid lines*) and at the turning point (*dashed lines*) for 5 firms (*top-left*), 10 (*top-right*), 15 (*down-left*) and 20 (*down-right*). A sample of 250,000 points of the orbit starting at the turning point is taken for making the estimations

cost functions $c_i = d + \alpha q_i - \beta q_i Q_i(1+2a) + 2\beta\alpha q_i Q_i^2$, in such a way the profit function is

$$\Pi_i = (\alpha - \beta Q)q_i - (d + \alpha q_i - \beta q_i Q_i(1+2a) + 2\beta\alpha q_i Q_i^2),$$

and maximizing the profit we obtain the reaction functions $g_a(Q_i) = aQ_i(1-Q_i)$.

When we consider an homogeneous model, it is not possible to reduce the number of parameters and then, all of them, a, λ and the number n of firms must be considered when we analyze the system. This is due to the well-known fact that the dynamics of the logistic map changes when the parameter a ranges in the parameter space. In addition, the auxiliary map

$$\overline{g}(x) = (1-\lambda)x + \lambda a(n-1)x(1-(n-1)x)$$

is always quadratic, and hence the turning point given by

$$x_M = \frac{1-\lambda+a\lambda(n-1)}{2a\lambda(n-1)^2} = \frac{1-\lambda}{2a\lambda(n-1)^2} + \frac{1}{2(n-1)}$$

always exist. The map $\overline{g}$ gives positive outputs at the interval $\left[0, \frac{1-\lambda+\lambda a(n-1)}{\lambda a(n-1)^2}\right]$, which ranges in the interval $[0, \overline{g}(x_M)] = \left[0, \frac{(1-\lambda+\lambda a(n-1))^2}{4\lambda a(n-1)^2}\right]$. The condition

$$\overline{g}(x_M) \leq \frac{1-\lambda+\lambda a(n-1)}{\lambda a(n-1)^2}$$

is equivalent to

$$\frac{(1-\lambda+\lambda a(n-1))^2}{4\lambda a(n-1)^2} \leq \frac{1-\lambda+\lambda a(n-1)}{\lambda a(n-1)^2},$$

which is fulfilled when $a \leq \frac{1}{n-1}$ and it reduces to

$$\lambda \leq \frac{3}{a(n-1)-1} \tag{4}$$

if $a > \frac{1}{n-1}$.

For Kopel's model, the number of Cournot equilibrium points can be outside the set Δ and therefore we may have more than one non null equilibrium. The Cournot points contained in Δ are computed by solving the equation

$$(1-\lambda)q + \lambda a(n-1)q(1-(n-1)q) = q,$$

which gives us $\overline{q} = 0$ and

$$\overline{q} = \frac{a(n-1)-1}{a(n-1)^2}.$$

The Jacobian matrix at the positive Cournot point is

$$\mathbf{J} = \begin{pmatrix} 1-\lambda & -\lambda\frac{2-a(n-1)}{(n-1)} & -\lambda\frac{2-a(n-1)}{(n-1)} & \dots & -\lambda\frac{2-a(n-1)}{(n-1)} \\ -\lambda\frac{2-a(n-1)}{(n-1)} & 1-\lambda & -\lambda\frac{2-a(n-1)}{(n-1)} & \dots & -\lambda\frac{2-a(n-1)}{(n-1)} \\ -\lambda\frac{2-a(n-1)}{(n-1)} & -\lambda\frac{2-a(n-1)}{(n-1)} & 1-\lambda & \dots & -\lambda\frac{2-a(n-1)}{(n-1)} \\ \dots & \dots & \dots & \dots & \dots \\ -\lambda\frac{2-a(n-1)}{(n-1)} & -\lambda\frac{2-a(n-1)}{(n-1)} & -\lambda\frac{2-a(n-1)}{(n-1)} & \dots & 1-\lambda \end{pmatrix},$$

with eigenvalues $1-\lambda(a+1)+\frac{2\lambda}{n-1}$, with multiplicity $n-1$, and $1+\lambda(1-a(n-1))$. Note that $\frac{2\lambda}{n-1} > \lambda(a+1)$ if and only if $n < 1+\frac{2}{a+1}$ which is not possible. Then, the first one has modulus smaller than one if and only if

$$\lambda(a+1) - \frac{2\lambda}{n-1} < 2,$$

which implies that

$$\lambda > \frac{2n-2}{(a+1)(n-1)-2}$$

if $(a+1)(n-1) < 2$, which is fulfilled for $\lambda \in [0,1]$, and

$$\lambda < \frac{2n-2}{(a+1)(n-1)-2},$$

if $(a+1)(n-1) > 2$. The second eigenvalue has modulus smaller than one when $a(n-1) > 1$ and

$$\lambda(a(n-1)-1) < 2,$$

which gives us

$$\lambda < \frac{2}{a(n-1)-1}.$$

When $n = 2$ and $a > 3$, there are two symmetric Cournot points at

$$\left(\frac{a+1+\sqrt{(a+1)(a-3)}}{2a}, \frac{a+1-\sqrt{(a+1)(a-3)}}{2a}\right)$$

and

$$\left(\frac{a+1-\sqrt{(a+1)(a-3)}}{2a}, \frac{a+1+\sqrt{(a+1)(a-3)}}{2a}\right).$$

The stability regions for that points can be seen in Kopel (1996). When $n > 2$, the existence of such Cournot equilibrium points can be checked by numerical computations, and it seems that finding analytical equations is complicated since we have to solve polynomial equations with degree higher than 5.

If Fig. 6 we compute the topological entropy for 2, 3 and 4 firms. We note that the topological entropy increases with the number of firms, and the region with positive topological entropy smaller than log 2 is really thin. The regions of constant topological entropy log 2 corresponds to parameter values making the model non-smooth, and therefore we cannot compute Lyapunov exponents in that parameter regions.

However, for normal Lyapunov exponents we have a completely different situation that those we found in Puu's model. Namely, the parameter region where normal Lyapunov exponents are negative is reduced when the number of firms increases and so, the dynamics observed within the set Δ are not observed outside

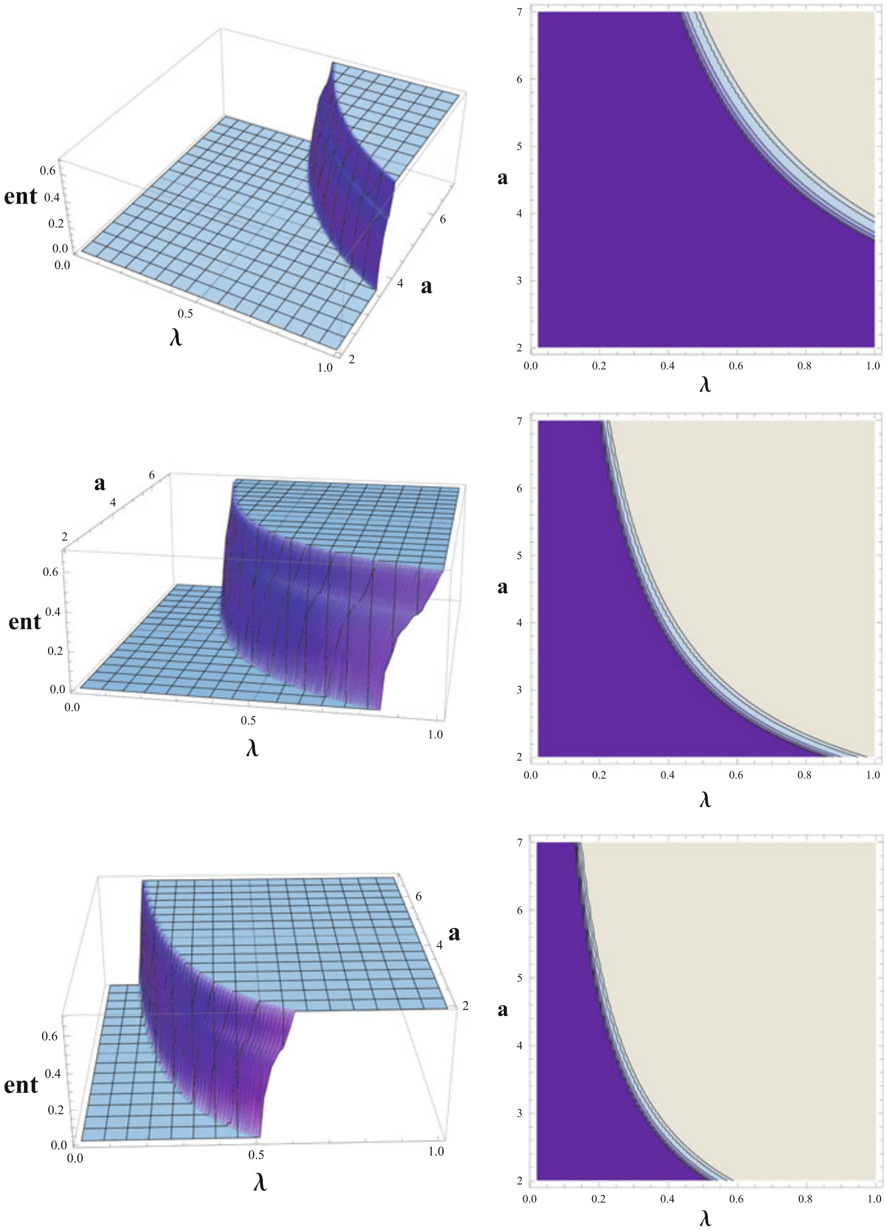


Fig. 6 The topological entropy for 2 (*up*), 3 (*middle*) and 4 firms (*down*) with accuracy 10^{-5} of Kopel's model. The topological entropy (*left-side*) and the level curves (*right-side*) are shown with parameters $a \in [2, 7]$ and $\lambda \in [0, 1]$

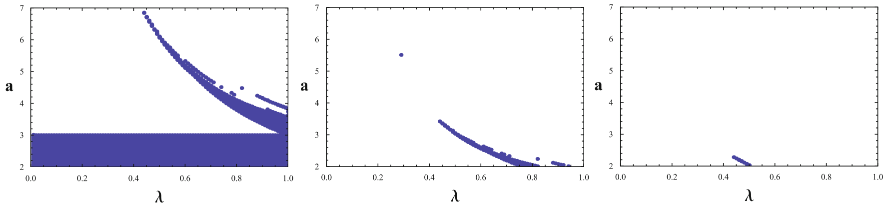


Fig. 7 Normal Lyapunov exponent estimations for 2 (*left*), 3 (*middle*) and 4 firms (*right*), with a sample of 250,000 points of the orbit starting at the turning point. The parameter regions where our estimations are negative are represented in *blue* (Color figure online)

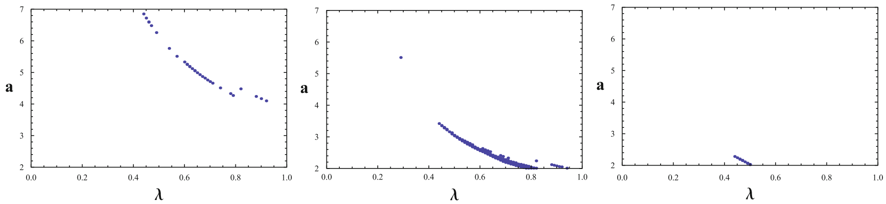


Fig. 8 In *blue* color, parameter regions where our estimations of the normal Lyapunov exponent is negative and the Lyapunov exponent is positive are shown for 2 (*left*), 3 (*middle*) and 4 firms (*right*), with a sample of 250,000 points starting from the turning point (Color figure online)

this set. In other words, usually the firms do not synchronize their productions. In Fig. 7 we represent the parameter regions where the normal Lyapunov exponents are negative and in Fig. 8 we show the subregions where, in addition, the Lyapunov exponents are positive, and therefore a chaotic synchronization can be observed. As we can see, these regions decrease when the number of firms increase. Even more, for five firms our experiments do not show the existence of negative normal Lyapunov exponents.

4.3 Non Linear Non Homogeneous Oligopolies

As in the Theocharis' case, non homogeneous oligopolies can be constructed combining two ideas. We may assume that economic constants of firms are different, for instance different costs, and all the firms follow the same adjustment process, for instance naive expectations. On the other hand, we may assume that economic constants are the same but firms use a different method to plan their future productions. Of course, we can combine both of them, which probably is more realistic but in general hard to handle when the number of firms increases. Below we analyze the first case.

4.3.1 Reduction of Firms

Assume that our model is given by

$$q_i(t+1) = \max\{0, (1-\lambda)q_i(t) + \lambda g_{\mathbf{c}_i}(Q_i(t))\},\ i = 1, 2, \ldots, n,$$

where $g_{\mathbf{c}_i}$ is the reaction function under naive expectations depending on the parameter $\mathbf{c}_i$, which is possibly a vector. It is not difficult to realize that if we want that some firms, say m of them, will not produce in a future time, then their Cournot coordinates must be zero. If we assume that the disappearing firms are for $i = 1, \ldots, m$, then all the fixed points of the system must be

$$\left(\overbrace{0, \ldots, 0}^{m}, \overline{q}_{m+1}, \ldots, \overline{q}_n\right), \tag{5}$$

where $(\overline{q}_{m+1}, \ldots, \overline{q}_n)$ are the Cournot coordinates of the oligopoly model consisting on $n-m$ firms with equations

$$q_i(t+1) = \max\{0, (1-\lambda)q_i(t) + \lambda g_{\mathbf{c}_i}(Q_i^m(t))\},\ i = m+1, \ldots, n, \tag{6}$$

where $Q_i^m = Q_i - (q_1 + \ldots + q_m)$. Hence, the dynamics of the reduced $n-m$ dimensional oligopoly is very important and its knowledge is necessary to be able to guarantee that when the m-th first firms do not produce, then they will not produce anything in future. Of course, it is easier to give negative results, that is, when the reduction of firms cannot occur. For instance, we can state the following result.

Theorem 3 *Assume that* $\mathbf{c}_i = \mathbf{c}$ *for all* $i = m+1, \ldots, n$, *and the map* $f(x) = (1-\lambda)x + \lambda g_{\mathbf{c}_i}((n-m-1)x)$ *is unimodal such that* $f(x_M) \geq x_0$, *where* x_0 *is the first positive real number such that* $f(x_0) = 0$. *Then, there exists* t_0 *such that the orbit with initial condition* $\left(\overbrace{0, \ldots, 0}^{m}, \overbrace{q, \ldots, q}^{n-m}\right)$, $q > 0$, *satisfies that* $q_i(t_0) > 0$ *for* $i = 1, 2, \ldots, m$.

Proof Clearly, the result follows if we prove that, for some $q > 0$, the solution of the difference equation

$$\begin{cases} q(t+1) = \max\{0, f(q(t))\}, \\ q(0) = q, \end{cases}$$

can take values smaller than ε, for some fixed $\varepsilon > 0$. Let $J = [a, b]$ be such that $f(x) \geq x_0$ for all $x \in J$. Since f is unimodal, we have that $[0, q_0] \subset f([0, a]) \cap f([b, q_0])$, that is, f has a 2-horseshoe. By Block and Coppel (1992), there is a compact set Λ containing 0, invariant by f, such that the dynamics of $f|_\Lambda$

is transitive. As a consequence, there are solutions of the above difference equation that goes arbitrary close to zero when time evolves, and the proof concludes.□

If we analyze the hypothesis of Theorem 3 for our two models, we realize that they are satisfied when the number of firms is big enough. In addition, the strict inequalities of expressions (3) and (4) gives us the opposite conditions for Puu's and Kopel's oligopolies, respectively. In the case of Puu's oligopoly, for a fixed λ, the number of firms $n-m$ must satisfy

$$\lambda > \frac{5}{n-m}$$

to fulfill conditions of Theorem 3. For Kopel's model the condition is

$$\lambda > \frac{3}{a(n-m-1)-1}.$$

It is more difficult to obtain positive results, which must be based on the fact that the orbits of the system are big enough for any time in such a way that the m-th firms always choose the null production as their best replies. If the map $f(x) = (1-\lambda)x + \lambda g_{\mathbf{c}}((n-m-1)x)$ is unimodal, then there is a maximum value x_M such that $[f^2(x_M), f(x_M)]$ is invariant by f and contains all its dynamic information. Hence $f^2(x_M)$ is the smallest value that the difference equation

$$\begin{cases} q(t+1) = \max\{0, f(q(t))\}, \\ q(0) = q, \end{cases}$$

can take after a sufficiently large number of iterations. If $(0, x_0^i)$ is the support of $f_i(x) = (1-\lambda)x + \lambda g_{\mathbf{c}_i}(x)$, $i = 1, 2, \ldots, m$, that is, the interval such that $f_i(x) > 0$ for all $x \in (0, x_0^i)$, then if the initial conditions are in the set

$$\mathscr{A} = \left\{ \left(\overbrace{0, \ldots, 0}^{m}, q, \ldots, q \right) : q_i \in [f^2(x_M), f(x_M)],\ i = m+1, \ldots, n \right\},$$

and

$$(n-m)f^2(x_M) \geq x_0^i,\ i = 1, 2, \ldots, m, \tag{7}$$

then all the future productions $q_i(t) = 0$ for $t \geq 1$ and $i = 1, 2, \ldots, m$. The problem is whether the set $\mathscr{A}$ attracts all the possible orbits of the oligopoly, which in general, is a very complicated question.

Hopefully, the conditions (5) and (7), which are necessary conditions for firms to disappear from the market, can be checked in some cases. For instance, consider Puu's oligopoly with n firms and assume that for $i = 1, \ldots, m$, the marginal cost

$c_i = c_1$, and $c_i = c_n$ for $i = m+1, \ldots, n$, in such a way that $c_1 \geq c_n$. Then, the non null Cournot point is

$$\overline{q}_1 = \ldots = \overline{q}_m = (n-1)\frac{c_n(n-m) - c_1(n-m-1)}{(c_1 m + c_n(n-m))^2}$$

$$\overline{q}_{m+1} = \ldots = \overline{q}_n = (n-1)\frac{c_1 m - c_n(m-1)}{(c_1 m + c_n(n-m))^2}.$$

Note that $\overline{q}_n > 0$ and $\overline{q}_1 \leq 0$ if and only if

$$c_n(n-m) \leq c_1(n-m-1),$$

which gives us condition (5). On the other hand, $x_M = \frac{\lambda^2(n-m-1)}{4c_n(1+\lambda(m-n))^2}$ and $x_0^1 = \frac{\lambda^2}{c_1(2\lambda-1)^2}$, and hence condition (7) reads as

$$(n-m)\frac{\lambda^2(n-m-1)}{4c_n}\left(1 + 2\frac{(n-m)\lambda+1}{\sqrt{((n-m)\lambda-1)^2}}\right.$$
$$\left. +2\sqrt{2\sqrt{\frac{1}{((n-m)\lambda-1)^2}} - \frac{1}{(n-m)\lambda-1}}\right) \geq \frac{\lambda^2}{c_1(2\lambda-1)^2}.$$

For naive expectations, we have $\lambda = 1$ and the above expression simplifies to

$$(n-m)\frac{(n-m-1)}{4}\left(1 + 2\frac{n-m+1}{n-m-1} + 2\sqrt{\frac{1}{n-m-1}}\right) \geq \frac{c_n}{c_1}.$$

When $n - m \geq 5$ Theorem 3 states that reducing firms is not possible. For $n = 3$ and $m = 1$ the above conditions reduces to $c_1/c_3 \geq 2$, but the a sufficient condition is $c_1/c_3 \geq \frac{1}{3}(1 + 2\sqrt{7}) \simeq 2.09716754070972$ (see Cánovas and Muñoz-Guillermo 2014a). In general, we think that it is quite difficult and technical to improve the above results.

4.3.2 Non Homogeneous Adjustments

As in Theocharis' model, we may assume that firms make their adjustments following different strategies, that is, different reaction functions. To fix ideas, recall that we denote by f_N the reaction function under naive expectations, f_A under adaptive expectations and f_C when the firm plays Cournot. Note that for Puu's oligopoly there is one possibility of choosing a Cournot coordinate, but there can be many possibilities when Kopel's model is considered. In general, we consider

$\Sigma = \{(x_n)_{n=0}^{\infty} : x_n \in \{N, A, C\}, n \geq 0\}$ be the infinite sequences of letters in the alphabet $\{N, A, C\}$. Let $T : \Sigma^n \times (\mathbb{R}^+)^n \to \Sigma^n \times (\mathbb{R}^+)^n$ be given by

$$T((x_i^1), \ldots (x_i^n), q_1, \ldots, q_n) = (\sigma(x_i^1), \ldots, \sigma(x_i^1), f_{x_0^1}(q_1, \ldots, q_n), \ldots, f_{x_0^n}(q_1, \ldots, q_n)).$$

Again, we do not know how the dynamics of the map T can be analyzed, even when firms change their strategies in a periodic way.

5 Conclusions

We present a general framework to study, at least partially, oligopoly dynamics when the number of firms increases. These ideas are applied to Theocharis' model, which is piecewise linear, and two well-known non linear models due to Puu and Kopel. Our simulations show that, when the firms are homogeneous, Puu's oligopoly can be studied from its dynamics along the diagonal set where all the firms produce the same. However, in general, the same does not happen for Kopel's oligopoly.

Our results are partial and some open problems are stated along the chapter. Probably, the most important one is to model the change of firms' strategies along the time, and develop mathematical techniques that will be useful to describe the models, finding analytical results that can go further than numerical simulations.

Acknowledgements This chapter is based upon work from COST Action ISCH COST Action IS1104 "The EU in the new complex geography of economic systems: models, tools and policy evaluation", supported by COST (European Cooperation in Science and Technology) www.cost.eu. This work has been partially supported by the grant MTM2014-52920-P from Ministerio de Economía y Competividad (Spain).

References

Adler RL, Konheim AG, McAndrew MH (1965) Topological entropy. Trans Am Math Soc 114:309–319

Agronsky S, Ceder J (1991/1992a) Each Peano subspace of E^k is anω-limit set. Real Anal Exch 17:371–378

Agronsky S, Ceder J (1991/1992b) What sets can be -limit sets in E^n? Real Anal Exch 17:97–109

Alsedá L, Llibre J, Misiurewicz M (1993) Combinatorial dynamics and entropy in dimension one. World Scientific Publishing, Singapore

Aoki N, Hiraide K (1994) Topological theory of dynamical systems: recent advances. North-Holland, Amsterdam

Ashwin P, Buescu J, Stewart I (1996) From attractor to chaotic saddle: a tale of transverse instability. Nonlinearity 9:703–737

Baiardi LC, Lamantia F, Radi D (2015) Evolutionary competition between boundedly behavioral rules in oligopoly games. Chaos Solitons Fractals 79:204–225

Balibrea F, Snoha L (2003) Topological entropy of Devaney chaotic maps. Topol Appl 133:225–239

Balibrea F, Cánovas JS, Linero A (2004) On ω-limit sets of antitriangular maps. Topol Appl 137:13–19

Banks J, Brooks J, Cairns G, Davis G, Stacey P (1992) On Devaney's definition of chaos. Am Math Mon 99:332–334

Bischi GI, Baiardi LC (2015) Fallacies of composition in nonlinear marketing models. Commun Nonlinear Sci Numer Simul 20:209–228

Bischi GI, Chiarella C, Kopel M, Szidarovszky F (2010) Nonlinear oligopolies. Springer, Heidelberg

Bischi GI, Lamantia F, Radi D (2015) An evolutionary Cournot model with limited market knowledge. J Econ Behav Organ 116:219–238

Blanchard F, Glasner E, Kolyada S, Maass A (2002) On Li-Yorke pairs. J Reine Angew Math 547:51–68

Block LS, Coppel WA (1992) Dynamics in one dimension. Lectures Notes in Mathematics, vol 1513. Springer, Berlin

Block L, Keesling J (1992) Computing the topological entropy of maps of the interval with three monotone pieces. J Stat Phys 66:755–774

Block LS, Keesling J, Li S, Peterson K (1989) An improved algorithm for computing topological entropy. J Stat Phys 55:929–939

Bowen R (1971) Entropy for group endomorphism and homogeneous spaces. Trans Am Math Soc 153:401–414

Cánovas JS (2009) Reducing competitors in a Cournot–Theocharis oligopoly model. J Differ Equ Appl 15:153–165

Cánovas JS (2015) A note on a modified Cournot-Puu duopoly. J Differ Equ 2015:9 pp. Article ID 545364

Cánovas JS, Linero A (2001) Topological dynamic classification of duopoly games. Chaos Solitons Fractals 12:1259–1266

Cánovas JS, Muñoz-Guillermo M (2014a) On the dynamics of a restricted Cournot–Puu triopoly: firms survival and complexity. Appl Math Comput 237:356–372

Cánovas JS, Muñoz-Guillermo M (2014b) Computing topological entropy for periodic sequences of unimodal maps. Commun Nonlinear Sci Numer Simul 19:3119–3127

Cánovas JS, Muñoz-Guillermo M (2014c) Computing topological entropy for periodic sequences of unimodal maps. Preprint

Cánovas JS, Puu T, Ruíz M (2008) The Cournot–Theocharis problem reconsidered. Chaos Solitons Fractals 37:1025–1039

de Melo W, van Strien S (1993) One-dimensional dynamics. Springer, New York

Devaney RL (1989) An introduction to chaotic dynamical systems. Addison-Wesley, Redwood City

Dinaburg EI (1970) The relation between topological entropy and metric entropy. Sov Math 11:13–16

Droste E, Hommes CH, Tuinstra J (2002) Endogeneous fluctuations under evolutionary pressure in Cournot competition. Games Econ Behav 40:232–269

Du BS (2004) A simple proof of Sharkovsky's theorem. Am Math Mon 111:595–599

Fedorenko VV, Sharkovsky AN, Smítal J (1990) Characterizations of weakly chaotic maps of the interval. Proc Am Math Soc 110:141–148

Guckhenheimer J (1979) Sensitive dependence to initial conditions for one-dimensional maps. Commun Math Phys 70:133–160

Hommes CH, Ochea MI, Tuinstra J (2011) On the stability of the Cournot equilibrium: an evolutionary approach (Preprints, CeNDEF Working Paper, no 11–10). Universiteit van Amsterdam, Amsterdam

Huang W, Ye X (2002) Devaney's chaos or 2-scattering implies Li-Yorke's chaos. Topol Appl 117:259–272

Kloeden PE (1979) On Sharkovsky's cycle coexistence ordering. Bull Aust Math Soc 20:171–177
Kolyada SF (1992) On dynamics of triangular maps of the square. Ergod Theory Dyn Syst 12:749–768
Kolyada SF, Snoha L' (1992/1993) On ω-limit sets of triangular maps. Real Anal Exch 18:115–130
Kopel M (1996) Simple and complex adjustment dynamics in Cournot duopoly models. Chaos Solitons Fractals 7:2031–2048
Kwietniak D, Misiurewicz M (2005) Exact Devaney chaos and entropy. Qual Theory Dyn Syst 6:169–179
Li TY, Yorke JA (1975) Period three implies chaos. Am Math Mon 82:985–992
Milnor J (1985) On the concept of attractor. Commun Math Phys 99:177–195
Misiurewicz M, Szlenk W (1980) Entropy of piecewise monotone mappings. Stud Math 67:45–63
Oseledets VI (1968) A multiplicative ergodic theorem. Lyapunov characteristic numbers for dynamical systems. Trans Mosc Math Soc 19:197–231. Moscov Mat Obsch 19:179–210
Puu T (1991) Chaos in duopoly pricing. Chaos Solitons Fractals 1:573–581
Sharkovsky AN, Kolyada SF, Sivak AG, Fedorenko VV (1997) Dynamics of one-dimensional maps. Kluwer Academic, Boston
Singer D (1978) Stable orbits and bifurcation of maps of the interval. SIAM J Appl Math 35:260–267
Smale S (1967) Differentiable dynamical systems. Bull Am Math Soc 73:747–817
Smítal J (1986) Chaotic functions with zero topological entropy. Trans Am Math Soc 297:269–282
Sumi N (2003) Diffeomorphisms with positive entropy and chaos in the sense of Li–Yorke. Ergod Theory Dyn Syst 23:621–635
Theocharis RD (1959) On the stability of the Cournot solution on the oligopoly problem. Rev Econ Stud 27:133–134
Thunberg H (2001) Periodicity versus chaos in one-dimensional dynamics. SIAM Rev 43:3–30
van Strien S, Vargas E (2004) Real bounds, ergodicity and negative Schwarzian for multimodal maps. J Am Math Soc 17:749–782

Attracting Complex Networks

G. Guerrero, J.A. Langa, and A. Suárez

Abstract Real phenomena from different areas of Life Sciences can be described by complex networks, whose structure is usually determining their intrinsic dynamics. On the other hand, Dynamical Systems Theory is a powerful tool for the study of evolution processes in real situations. The concept of global attractor is the central one in this theory. In the last decades there has been an intensive research in the geometrical characterization of global attractors. However, there still exists a weak connection between the asymptotic dynamics of a complex network and the structure of associated global attractors. In this paper we show that, in order to analyze the long-time behavior of the dynamics on a complex network, it is the topological and geometrical structure of the attractor the subject to take into account. In fact, given a complex network, a global attractor can be understood as the new attracting complex network which is really describing and determining the forwards dynamics of the phenomena. We illustrate our discussion with models of differential equations related to mutualistic complex networks in Economy and Ecology.

Keywords Complex networks • Dynamics • Structure of global attractor

1 Introduction

In any real phenomena complexity plays a crucial role, which can be characterized from the following items:

– The reality under study is composed of a set of simpler elements.
– These elements organize a network of connections, building a complex system.

G. Guerrero
Universidad Central del Ecuador, Ciudadela Universitaria, Av. América S/N, Quito, Ecuador
e-mail: guerrerogiovanny@hotmail.com

J.A. Langa (✉) • A. Suárez
Departamento de Ecuaciones Diferenciales y Análisis Numérico, Seville University, Sevilla, Spain
e-mail: langa@us.es; suarez@us.es

P. Commendatore et al. (eds.), *Complex Networks and Dynamics*, Lecture Notes in Economics and Mathematical Systems 683,
DOI 10.1007/978-3-319-40803-3_12

– The weight of the links between nodes plays an important role.
– And finally, either the elements or the relationships between them evolves in time, i.e, the network possesses an intrinsic dynamics.

The description of real phenomena as complex networks has emerged as a powerful tool to understand the behavior of models in Life and Social Sciences. In particular, it is observed that there is a strong relation between the topological structure of the network (described as nodes, links and strength in connections) and the forwards dynamics of the phenomena (Figs. 1 and 2).

On the other hand, the theory of Dynamical Systems has a long history in the Applied Mathematics. Indeed, Dynamical Systems is a very well suited methodology for modelization as we get

– From a real phenomena, a complex graph.
– From a graph, a mathematical formulation of each expression, describing the weight of links and the dynamics of connections.
– The network is then analyzed by a system of (ordinary or partial) differential equations.

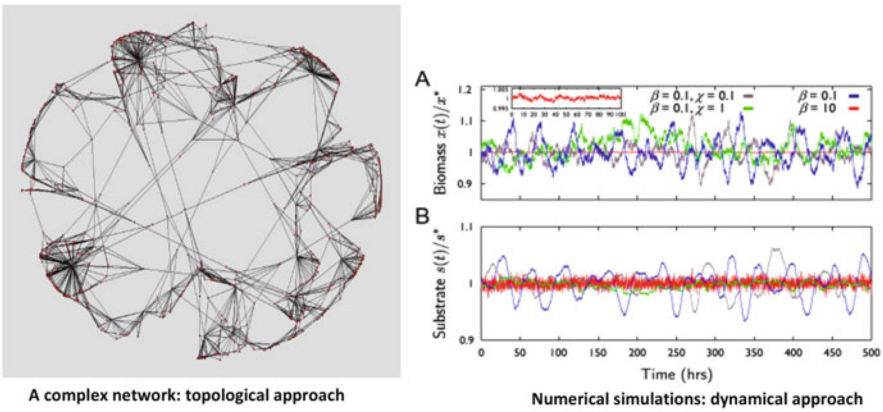


Fig. 1 From topology of the network to its dynamics

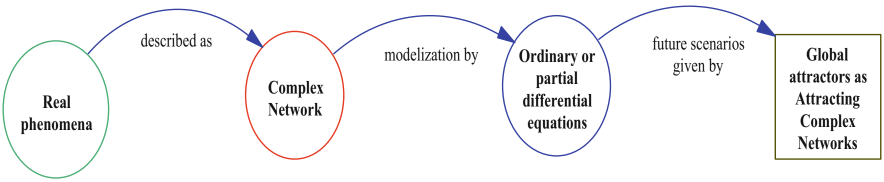


Fig. 2 The dynamics of a complex network associated to a real phenomena, modelled by a system of differential equations, can be characterized by the global attractor. When available, this attracting complex network describes all the future scenarios of the phenomena

Thus, a possible strong mathematical formalism to study a Complex Network is to describe it as a Dynamical System. Indeed, the theory of dynamical systems becomes a powerful tool for the modelization of many different and important real phenomena for multiple scientific areas. In particular, the study of compact attracting invariant sets has developed a large and deep research area, providing essential information for an increasing number of models from Physics, Biology, Economics, Engineering and others. Indeed, the analysis of qualitative properties of semigroups in general phase spaces (infinite-dimensional Banach spaces or general metric spaces) has received a lot of attention throughout the last four decades (see, for instance, Babin and Vishik (1983, 1992), Hale (1988), Ladyzhenskaya (1991), Robinson (2001), Sell and You (2002), Temam (1988) or Vishik (1992)). In this framework, the global attractor is a very consistent concept describing the long-time behavior of dynamical systems. A global attractor is an invariant compact set in the phase space determining all the asymptotic dynamics of the system under consideration. The study of the structure of the global attractor has received a lot of attention, going to a broad theory related to gradient systems, Morse Decomposition, Morse-Smale systems or chaotic dynamics in the attractor (see, for instance, Carvalho et al. 2015; Conley 1978; Hale et al. 1984; Sell and You 2002).

Our aim in this paper is to highlight the structure of a global attractor as a complex network. Indeed, there is a natural relation between a phenomenological complex network and the (even more complex) network given by the geometrical characterization of the global attractor.

For instance, consider the following two dimensional competitive Lotka-Volterra system describing the interactions between two species which compete for the same resources

$$\begin{cases} u'(t) = \lambda_1 u - \beta_1 u^2 - \gamma_1 uv, \\ v'(t) = \lambda_2 v - \beta_2 v^2 - \gamma_2 vu. \end{cases}$$

In this model the phenomenological network would be composed just by two connected nodes. However, the long-time behavior of the system is determined by the global attractor, one of its possible structures is described in Fig. 3.

In this work we give mathematical evidence on the fact that is the description of these attracting complex networks what is really determining the future scenarios of the real phenomena. Thus, the global attractor, and the analysis of its structure as a dynamical complex network, emerge as a key concept to explain how the architecture of reality is transformed into an abstract attracting network determining the future behavior of the phenomena.

In Sect. 2 we describe the main concepts and results related to dynamical systems, the existence of the global attractor and its internal characterization. In Sect. 3 we describe a general model by a system of differential equations related to mutualistic complex networks associated to Ecological and Economical phenomena in which cooperation among nodes plays an essential role. We will describe the geometrical structure of the associated global attractor. To illustrate the results, in Sect. 4 we present a simplified three dimensional (3D) model for which all the ideas

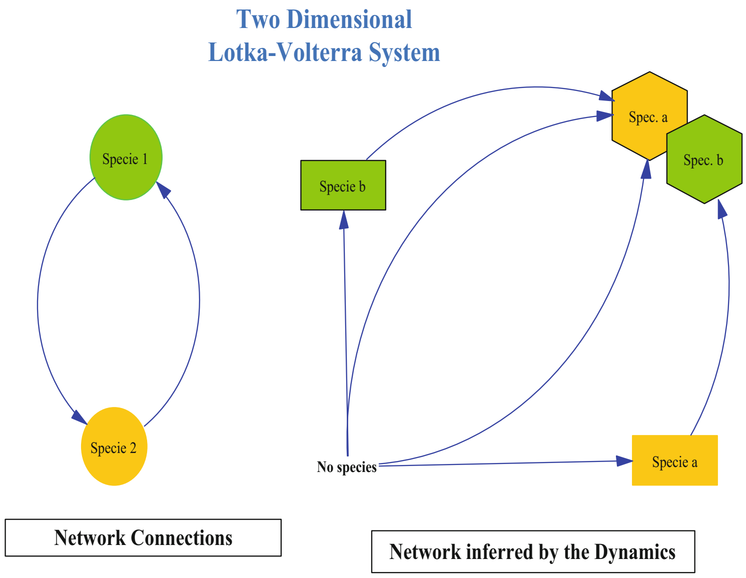


Fig. 3 From the real network to the dynamical complex network given by the structure of its global attractor

of this work can be highlighted by drawing some graphs and pictures. In a final Section we write some conclusive remarks and open questions for a further research in the near future.

2 Dynamical Systems and Attractors

Suppose we have a system of ordinary or partial differential equations defined in a (finite or infinite-dimensional) Banach space X

$$\begin{cases} \dfrac{du}{dt} = F(u(t)) \\ \\ u(0) = u_0 \in X \end{cases} \tag{1}$$

with $u(t)$ the unknown at time $t \geq 0$ and $F : X \to X$ a nonlinear operator on X. Suppose (1) has existence and uniqueness of solution $u(t; u_0)$, for all $t \geq 0$. A family $\{T(t) : t \geqslant 0\}$ is called a *continuous semigroup* if

(a) $T(0) = I_X$, with I_X being the identity in X,
(b) $T(t+s) = T(t)T(s)$, for all $t, s \in \mathbb{R}^+$ and
(c) the map $\mathbb{R}^+ \times X \ni (t, x) \mapsto T(t)x \in X$ is continuous.

$T(t)$ on X describes the dynamics of each element $u \in X$. The phase space X represents the framework in which the dynamics described by $T(t)$ is developed. In general, $T(t)u_0 = u(t; u_0)$ is the solution of (1) at time t with initial condition u_0.

2.1 Global Attractors

First we recall the definition of a global attractor for a nonlinear semigroup $\{T(t) : t \geqslant 0\}$ (Babin and Vishik 1992; Hale 1988; Ladyzhenskaya 1991; Robinson 2001; Temam 1988).

Definition 1 A set $\mathscr{A} \subseteq X$ is a global attractor for $\{T(t) : t \geq 0\}$ if it is

(i) compact,
(ii) invariant under $\{T(t) : t \geq 0\}$, i.e. $T(t)\mathscr{A} = \mathscr{A}$ for all $t \geq 0$, and
(iii) attracts bounded subsets of X under $\{T(t) : t \geq 0\}$; that is, for all $B \subset X$ bounded

$$\lim_{t\to+\infty} \operatorname{dist}(T(t)B, \mathscr{A}) \longrightarrow 0$$

where $\operatorname{dist}(D, A) := \sup_{d\in D} \inf_{a\in A} \operatorname{dist}(d, a)$ is the Hausdorff semidistance between two sets $D, A \subset X$.

A *global solution* for a semigroup $\{T(t) : t \geq 0\}$ is a continuous function $\xi : \mathbb{R} \to X$ such that $T(t)\xi(s) = \xi(t + s)$ for all $s \in \mathbb{R}$ and all $t \in \mathbb{R}^+$. We say that $\xi : \mathbb{R} \to X$ is a *global solution through* $z \in X$ if it is a global solution with $\xi(0) = z$. The global attractor can be characterized as the collection of all globally defined bounded solutions:

Lemma 1 *If a semigroup* $T(\cdot)$ *has a global attractor* $\mathscr{A}$*, then*

$$\mathscr{A} = \{y \in X : \textit{ there is a bounded global solution } \xi : \mathbf{R} \to X \textit{ with } \xi(0) = y\}.$$

It is well known that global attractors for semigroups are unique. For the existence, we have the following general result (see Carvalho et al. 2013).

Theorem 1 *There exists a* global attractor *for a semigroup* $T(\cdot)$ *if and only if there exists a compact attracting set of bounded sets, i.e., a compact set* $K \subset X$ *such that* $\operatorname{dist}(T(t)C, K) \to 0$ *as* $t \to +\infty$*, for all* $C \subset X$ *bounded.*

Definition 2 We say that $u^* \in X$ is an equilibrium point (or stationary solution) for the semigroup $T(t)$ if $T(t)u^* = u^*$, for all $t \geq 0$.

Definition 3 The unstable set of an invariant set Ξ is defined by

$$W^{\mathrm{u}}(\Xi) = \{z \in X : \text{ there is a global solution } \xi : \mathbf{R} \to X \text{ for } T(t) \text{ satisfying } \xi(0) = z \text{ and such that } \lim_{t\to-\infty} \operatorname{dist}(\xi(t), \Xi) = 0\}.$$

2.2 Attracting Complex Networks

In this section we will describe the geometrical structure of the global attractor.

Definition 4 Let $\{T(t) : t \geq 0\}$ be a semigroup on X. We say that an invariant set $E \subset X$ for the semigroup $\{T(t) : t \geq 0\}$ is an *isolated invariant set* if there is an $\epsilon > 0$ such that E is the maximal invariant subset in the neighbourhood $\mathscr{O}_\epsilon(E)$.

A disjoint family of isolated invariant sets is a family $\{E_1, \cdots, E_n\}$ of isolated invariant sets with the property that,

$$\mathscr{O}_\epsilon(E_i) \cap \mathscr{O}_\epsilon(E_j) = \varnothing, \qquad 1 \leq i < j \leq n,$$

for some $\epsilon > 0$.

2.2.1 Morse Decomposition of a Global Attractor

Next we introduce the notion of a Morse decomposition for the attractor $\mathscr{A}$ of a semigroup $\{T(t) : t \geq 0\}$ (see Conley (1978), Rybakowski (1987) or Sell and You (2002)). We start with the notion of an attractor-repeller pair.

Definition 5 Let $\{T(t) : t \geq 0\}$ be a semigroup with a global attractor $\mathscr{A}$. We say that a non-empty subset A of $\mathscr{A}$ is a *local attractor* if there is an $\epsilon > 0$ such that $\omega(\mathscr{O}_\epsilon(A)) = A$, where $\omega(B)$ is the ω-limit set of B, defined as

$$\omega(B) = \{x \in X : \ S(t_n)x_n \to x, \ \text{for some } x_n \in B, \ t_n \to \infty\}.$$

The *repeller* A^* associated with a local attractor A is the set defined by

$$A^* := \{x \in \mathscr{A} : \omega(x) \cap A = \varnothing\}.$$

The pair (A, A^*) is called an *attractor-repeller pair* for $\{T(t) : t \geq 0\}$.

Note that if A is a local attractor, then A^* is closed and invariant.

Definition 6 Given an increasing family $\varnothing = A_0 \subset A_1 \subset \cdots \subset A_n = \mathscr{A}$, of $n + 1$ local attractors, for $j = 1, \cdots, n$, define $E_j := A_j \cap A^*_{j-1}$. The ordered n- tuple $\mathbf{E} := \{E_1, E_2, \cdots, E_n\}$ is called a *Morse decomposition for* $\mathscr{A}$.

An equivalent definition of a Morse decomposition for the attractor $\mathscr{A}$ of a semigroup $\{T(t) : t \geq 0\}$ can be found at Aragão-Costa et al. (2011).

Definition 7 Let $\{T(t) : t \geq 0\}$ be a semigroup with a global attractor $\mathscr{A}$ and $\mathbf{E} = \{E_1, E_2, \cdots, E_n\}$ is a *Morse decomposition* of $\mathscr{A}$. We say that the semigroup is dynamically gradient if for a given global solution $\xi : \mathbb{R} \to \mathscr{A}$ of $\{T(t) : t \geq 0\}$

i) either $\xi(t) \in E_i$, for all $t \in \mathbb{R}$ and some $i = 1, \cdots, n$;
ii) or there exist $1 \leq i < j \leq n$ such that $E_j \overset{t\to-\infty}{\longleftarrow} \xi(t) \overset{t\to\infty}{\longrightarrow} E_i$.

2.2.2 Lyapunov Functions

Definition 8 We say that a semigroup $\{T(t) : t \geq 0\}$ with a global attractor $\mathscr{A}$ and a disjoint family of isolated invariant sets $\mathbf{E} = \{E_1, \cdots, E_n\}$ is a *gradient semigroup* with respect to $\mathbf{E}$ if there exists a continuous function $V : X \to \mathbb{R}$ such that

(i) $[0, \infty) \ni t \mapsto V(T(t)x) \in \mathbb{R}$ is non-increasing for each $x \in X$;
(ii) V is constant in E_i, for each $1 \leq i \leq n$; and
(iii) $V(T(t)x) = V(x)$ for all $t \geq 0$ if and only if $x \in \bigcup_{i=1}^{n} E_i$.

In this case we call V a Lyapunov functional related to $\mathbf{E}$.

For gradient semigroups, the structure of the global attractor can be described as follows:

Theorem 2 *Let $\{T(t) : t \geq 0\}$ a gradient semigroup with respect to the finite set $\mathbf{E} := \{E_1, E_2, \cdots, E_n\}$. If $\{T(t) : t \geq 0\}$ has a global attractor $\mathscr{A}$, then $\mathscr{A}$ can be written as the union of the unstable manifolds related to each set in* $\mathbf{E}$*, i.e,*

$$\mathscr{A} = \bigcup_{j=1}^{n} W^u\left(E_j\right). \tag{2}$$

Remark 1 When E_j are equilibria u_j^*, the attractor is described as the union of the unstable manifolds associated to them

$$\mathscr{A} = \bigcup_{j=1}^{n} W^u\left(u_j^*\right).$$

This description shows a geometrical picture of the global attractor, in which all the stationary points are ordered by connections related to its level of attraction or stability.

Observe that each node given by a partially feasible equilibrium point in the attractor represents an attracting complex network in the original one. Thus, the attractor can be understood as a new complex dynamical network describing all the possible feasible future networks. It contains all the abstract information related to future scenarios of the model.

2.2.3 Energy Levels

Any Morse decomposition $\mathbf{E} = \{E_1, \cdots, E_n\}$ of a compact invariant set $\mathscr{A}$ leads to a partial order among the isolated invariant sets E_i; that is, we can define an order between two isolated invariant sets E_i and E_j if there is a chain of global solutions

$$\{\xi_\ell,\ 1 \leq \ell \leq r\} \tag{3}$$

with

$$\lim_{t \to -\infty} \xi_\ell(t) = E_\ell$$

and

$$\lim_{t \to \infty} \xi_\ell(t) = E_{\ell+1}$$

$1 \leq \ell \leq r-1$, with $E_1 = E_i$ and $E_r = E_j$.

This implies that, given any dynamically gradient semigroup with respect to the disjoint family of isolated invariant sets $\mathbf{E} = \{E_1, \cdots, E_n\}$, there exists a partial order in $\mathbf{E}$. In Aragão-Costa et al. (2012) it is shown that there exists a Morse decomposition given by the so-called energy levels $\mathbf{N} = \{\mathscr{N}_1, \mathscr{N}_2, \cdots, \mathscr{N}_p\}, p \leq n$. Each of the levels $\mathscr{N}_i$, $1 \leq i \leq p$ is made of a finite union of the isolated invariant sets in $\mathbf{E}$ and $\mathbf{N}$ is totally ordered by the dynamics defined in (3). The associated Lyapunov function has different values in any two different level-sets of $\mathbf{N}$ and any two elements of $\mathbf{E}$ which are contained in the same element of $\mathbf{N}$ (same energy level) are not connected.

2.2.4 Structural Stability: Robustness Under Perturbation

A detailed understanding of the behaviour of isolated invariant sets and their associated unstable manifolds is one of the key facts used to prove the characterization of attractors as the union of unstable manifolds. Moreover, similar properties are used to prove that a gradient system with a finite number of hyperbolic equilibria (see Hale 1988; Henry 1981) can be completely characterized by the internal dynamics between equilibria: every global solution connects two different equilibria and there are no homoclinic structures connecting equilibria (see Carvalho and Langa 2009; Carvalho et al. 2013).

It has been proved in Aragão-Costa et al. (2011) that a semigroup $\{T(t) : t \geq 0\}$ is gradient with respect to $\mathbf{E}$ if and only if it is dynamically gradient with respect to $\mathbf{E}$. Indeed, we have the following result

Theorem 3 *Given a disjoint family of isolated invariant sets* $\mathbf{E} = \{E_1, \cdots, E_n\}$ *for a semigroup* $T(t)$*, the following three properties are equivalent:*

i) $T(\cdot)$ is dynamically gradient;
ii) there exists an associated ordered family of local attractor-repellers; and
iii) there exists a Lyapunov functional related to $\mathbf{E}$.

Moreover, in Carvalho and Langa (2009) (see also Arrieta et al. 2012 and Carvalho et al. 2007) it is shown that a dynamically gradient system (then, a gradient one or a system with a Morse decomposition of its global attractor) is robust under small perturbation of the parameters and/or the linear and nonlinear operator in the equations. This means that the structure of a gradient-like global attractor is robust under perturbation. Indeed, the results in Carvalho and Langa (2009) show dynamically gradient nonlinear semigroups are stable under perturbation, so that we conclude that gradient semigroups are stable under perturbation as well; that is, the existence of a continuous Lyapunov function is robust under perturbation of parameters (see Carvalho et al. (2015) and the references therein).

3 Mutualistic Complex Networks: Real Phenomena

In the last decade there has been an intensive interdisciplinary research on mutualistic complex networks, mainly based on Ecology models (see, for instance, Allesina and Tang 2012, Bascompte et al. 2003, Bascompte et al. 2006, Bascompte and Jordano 2007, Bastolla et al. 2009, Okuyama and Holland 2008, Rohr et al. 2014, Saavedra et al. 2011, Suweis et al. 2013 and Thébault and Fontaine 2010). The results have also been applied to complex networks in Economy, in particular to the modellization of the cooperative interactions between designers and their contractors in the New York City garment industry. Indeed, this industry is characterized by a dynamic environment where resource exchanges among firms and survival depends on mutualistic connections between firms (see Saavedra et al. 2009a,b).

In general, a mutualistic network implies dozens or even hundreds of nodes building a complex net of interdependencies. In Saavedra et al. (2009a,b), nodes correspond to an individual designer or contractor firm, and links between nodes indicate that a designer exchanged money for the contractors production services. In Bascompte et al. (2003, 2006), Bascompte and Jordano (2007), Bastolla et al. (2009) and Saavedra et al. (2011), nodes show density of population of different species in a particular environment. Mutualism means collaborative interaction between nodes (species) of mutual benefits for both of them. We found two different set of species (two modes network leading to bipartite graphs): plant and pollinators, or plant and seed dispersal, or firms of different business groups (see Fig. 4).

One of the main discovers in this framework, it that, despite of its different nature, most of these mutualistic networks show a similar structure. What is more, a common architecture (similar patter formation), which in fact explains the robustness of the network (Bascompte et al. 2003, 2006; Bascompte and Jordano 2007; Bastolla et al. 2009; Saavedra et al. 2011). Robustness is defined as the strength of the net to lose its components under perturbation. This is why these complex networks in Ecology have been defined as the architecture of Biodiversity (Bascompte and Jordano 2007).

The main characteristics of these networks are the following:

(a) Distribution degree: measures the distribution of the number of connections. It is observed heterogeneity, in the sense that

1. Most of the species are specialists: they are connected to a small number of nodes.
2. A few species are generalists, they are connected to a huge number of nodes.

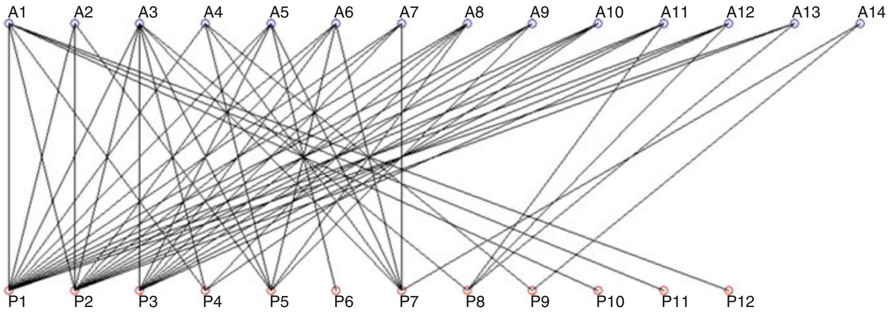


Fig. 4 Mutualistic complex network: a bipartite graph; each part (A—fourteen nodes- and P—twelve nodes) is in competition among its group. Links represent cooperative interactions between the two groups in the graph

(b) Nestedness: show a patron of interdependencies in which a specialist connects with subsets of the set of connections of a generalist.
(c) Asymmetries: The connections between two species have different weights depending the direction of connection. Specialists of one group tend to have a strong interaction with generalists of the other group.

The consequence of this network structure is robustness. In particular, for Ecological systems:

- The nestedness configuration allows alternative routes for the persistence of the system: there exists a low number of intermediate connections to join any two nodes.
- Generalists tend to be very spread and common species, very robust to changes. This is crucial, as specialists then tend to depend of abundant species, allowing their coexistence.
- New invasive species tend also to interact with generalist, being then well integrated to the whole network.
- Asymmetries allows the network to coexist even if a specialist dies.

3.1 The Abstract Model: Dynamical System Approach

Bascompte et al. (2006) (see also Rohr et al. 2014; Saavedra et al. 2011), after a description of the phenomenological properties on mutualistic complex network, introduce the following model in order to study the forwards dynamics of nodes and relations in the net.

Suppose that the P nodes of a group (and the A nodes of the other group) are in competition and P-nodes and A-nodes have cooperative links.

$$\begin{cases} \dfrac{dS_{p_i}}{dt} = \alpha_{p_i} S_{p_i} - \sum_{j=1}^{P} \beta_{p_{ij}} S_{p_i} S_{p_j} + \sum_{k=1}^{A} \dfrac{\gamma_{p_{ik}} S_{p_i} S_{a_k}}{1 + h_P \sum_{l=1}^{A} \gamma_{p_{il}} S_{a_l}} \\ \dfrac{dS_{a_i}}{dt} = \alpha_{a_i} S_{a_i} - \sum_{j=1}^{A} \beta_{a_{ij}} S_{a_i} S_{a_j} + \sum_{k=1}^{P} \dfrac{\gamma_{a_{ik}} S_{a_i} S_{p_k}}{1 + h_A \sum_{l=1}^{P} \gamma_{a_{il}} S_{p_l}} \\ S_{p_i}(0) = S_{p_{i0}}, \\ S_{a_i}(0) = S_{a_{i0}} \end{cases} \quad (4)$$

with $p_i \in \{p_1, \ldots, p_P\}$, $a_i \in \{a_1, \ldots, a_A\}$; α_{p_i} and α_{a_i} represent the intrinsic growth rates, $\beta_{p_{ij}} \geq 0$ and $\beta_{a_{ij}} \geq 0$ competitive interactions, $\gamma_{p_{ij}} \geq 0$ and $\gamma_{a_{ij}} \geq 0$ the mutualistic interactions, $h_A \geq 0$ and $h_P \geq 0$ can be interpreted as handling times.

With this mathematical model of differential equations, which has an associated dynamical system $\{T(t) : t \geq 0\}$, we are able to describe all the main structure properties of the network. Note that the original network is drawn into the abstract model, given by the positiveness or nullness of the mutualistic $\gamma_{p_{ij}}$, $\gamma_{a_{ij}}$ parameters, related to the associated adjacency matrix of the network.

From now on, we assume that $h_A = h_P = 0$. The following results have been proved in Guerrero et al. (2015):

Theorem 4

a) Assume that $\beta = \min\{\beta_{p_{ij}}, \beta_{a_{ij}}\} < 1$, $\gamma_1 = \max\{\gamma_{p_{ij}}\}$, $\gamma_2 = \max\{\gamma_{a_{ij}}\}$ *and*

$$\gamma_1 \gamma_2 < \frac{1 + \beta(P-1)}{P} \frac{1 + \beta(A-1)}{A}. \quad (5)$$

Then, there exists a unique positive solution of (4), for all $t > 0$.

b) Assume that $\beta = \beta_{p_{ij}} = \beta_{a_{ij}}$, $\gamma_1 = \gamma_{p_{ij}}$, $\gamma_2 = \gamma_{a_{ij}}$, $\alpha = \alpha_{p_i} = \alpha_{a_i} > 0$ *for all* i, j *and*

$$\gamma_1 \gamma_2 > \frac{1 + \beta(P-1)}{P} \frac{1 + \beta(A-1)}{A}. \quad (6)$$

Then, any positive solution of (4) blows up in finite time.

Furthermore, in order to simplify some of the following calculations, we consider that the competition and mutualistic matrices are of mean-field type, that is $\beta_{p_{ij}} = \beta_1$, $\beta_{a_{ij}} = \beta_2$, $\gamma_{p_{ij}} = \gamma_1$ and $\gamma_{a_{ij}} = \gamma_2$.

Let $n = P + A$. Denote by $\alpha_i = \alpha_{p_i}$ for $i = 1, \ldots, P$ and $\alpha_i = \alpha_{a_i}$ for $i = 1, \ldots, A$. Observe that in this case, (4) can also be written as

$$\frac{dS_i}{dt} = S_i(\alpha_i + M * S(t)), \quad i = 1, \ldots, n, \quad (7)$$

for $S(t) = (S_{p_1}(t), \ldots, S_{p_P}(t), S_{a_1}(t), \ldots, S_{a_A}(t)) := (S_P(t), S_A(t))$, and

$$M = \begin{bmatrix} B_1 & \Gamma_1 \\ \Gamma_2 & B_2 \end{bmatrix}_{(P+A)\times(P+A)}.$$

where, for $k = 1, 2$,

$$B_k = \begin{bmatrix} -1 & -\beta_k & \cdots & -\beta_k \\ -\beta_k & -1 & \cdots & -\beta_k \\ \vdots & \vdots & \ddots & \vdots \\ -\beta_k & -\beta_k & \cdots & -1 \end{bmatrix}$$

and

$$\Gamma_k = \begin{bmatrix} \gamma_k & \gamma_k & \cdots & \gamma_k \\ \gamma_k & \gamma_k & \cdots & \gamma_k \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_k & \gamma_k & \cdots & \gamma_k \end{bmatrix}.$$

As we are interested in positive solutions, we define the positive cone

$$\mathbf{R}^n_+ = \{u \in \mathbf{R}^n : \ u_i \geq 0, \ \text{for all } i = 1, \ldots, n\}.$$

Recall that, given an initial condition $(S_{P_0}, S_{A_0}) = \{(S_{p_{i0}}, S_{a_{i0}}) \ \{i = 1, \ldots, n\}\}$, $T(t)(S_{P_0}, S_{A_0}) = (S_P(t), S_A(t))$ denotes the solution of (4) at time t starting in (S_{P_0}, S_{A_0}) at time 0.

Theorem 5 *Suppose (5), then* (4) *has a global attractor* $\mathscr{A} \subset \mathbf{R}^n$ *associated to* $\{T(t) : \ t \geq 0\}$.

Once we know the existence of a global attractor, we can try to describe it as an attracting complex network. In this sense, the structure of the attractor becomes the crucial complex network to take into account.

Note that the set of equilibria (stationary points) for (4) is given by solving the system of equations

$$S_i(\alpha_i + M * S) = 0, \quad i = 1, \ldots, n.$$

Denote by $\mathbf{E} := \{E_1, E_2, \cdots, E_m\} \subset \mathbf{R}^{P+A}$ the set of equilibria related to (4). Observe that if M is a regular matrix, at most, there is only one equilibrium with all its components strictly positive, give by the solution of the linear system $\alpha + M * S = 0$, where $\alpha = (\alpha_1, \ldots, \alpha_n)$. Any other equilibria possesses, at least, one null component.

Definition 9 A matrix $M \in \mathbf{R}^{n\times n}$ is said to be Volterra-Lyapunov stable if there exists a positive diagonal matrix $D > 0$ such that $DM + M^tD$ is negative definite, i.e., $u^t(DM + M^tD)u < 0$ for all $u \in \mathbf{R}^n$.

Definition 10 We say that an equilibrium u^* is globally asymptotically stable in a region $D \subset \mathbb{R}^n$ if

$$\lim_{t\to+\infty} \text{dist}(T(t)u, u^*) = 0, \text{ for all } u \in D.$$

In Takeuchi (1996), the following result is proved:

Theorem 6 *Suppose M in (7) is Volterra-Lyapunov stable. Then there exists a unique globally stable equilibrium in the positive cone* $\mathbf{R}^n_+$.

The following result comes from Guerrero et al. (2015), and describes a Morse decomposition on global attractor for (4)

Theorem 7 *Assume (5). Then*

a) There exists a Morse decomposition of the global attractor on the positive cone, with $\mathbf{E} := \{E_1, E_2, \cdots, E_m\}$ *the set of positive equilibria.*
b) The associated dynamical system is gradient with respect to $\mathbf{E}$.
c) Thus, the global attractor can be described by

$$\mathscr{A} = \bigcup_{i=1}^{m} W^u(E_i). \tag{8}$$

It is important to observe that each equilibrium is a vector in $\mathbf{R}^{P+A}$ and that its $P + A$ components correspond to the $P + A$ nodes of the phenomenological complex network. In this sense, it is remarkable that each of the stationary points is highlighting a subnet of the former complex network. Indeed, the strictly positive components of each equilibrium point out a subset of nodes and connections of the original network. In particular, the first local attractor in $\mathbf{E}$, is indeed the complex network of the phenomena showing the future biodiversity of the Ecological system, or the joint success (firms which do not disappear) in an Economical framework.

Note that (a) and (8) is not only saying that all the asymptotic behavior of the system is concentrated around $\mathscr{A}$, but it is describing the way in which the attraction takes place. In particular, it is not only showing that there exists a unique globally stable equilibrium in the positive cone $\mathbf{R}^n_+$, but how this stationary point is connected to any other, building some energy levels which organize the attraction rates. In summary, all the equilibria are ordered and oriented connected, i.e., the connections are just in one direction, determined by the forwards dynamical of the system. In this sense, it is the structure of the global attractor which is showing a complex network with a natural intrinsic dynamics. In fact, as each node of this attracting network is a subnet of the former network, the global attractor in this case can be understood as a network of subnetworks of the original one, which, moreover, is dynamically organized to describe all the possible future scenarios of the phenomena.

4 A 3D Model

To illustrate our ideas, we now develop a simple model from (4) (see Guerrero et al. 2014), so that we can make some pictures in order to highlight our previous conclusions.

Thus, we consider a model consisting of three differential equations, two nodes in the first group (u_1 and u_2) and just another node in the second group (u_3) with cooperative relations with the first ones. We have

$$\begin{cases} u_1' = u_1(\alpha_1 - u_1 - \beta u_2 + \gamma_1 u_3) \\ u_2' = u_2(\alpha_2 - u_2 - \beta u_1 + \gamma_1 u_3) \\ u_3' = u_3(\alpha_3 - u_3 + \gamma_2 u_1 + \gamma_2 u_2), \\ (u_1(0), u_2(0), u_3(0)) = (u_1^0, u_2^0, u_3^0), \end{cases} \tag{9}$$

where $\alpha_1, \alpha_2, \alpha_3 \in \mathbf{R}$, $0 < \beta < 1$, $\gamma_1, \gamma_2 > 0$ and we suppose positive initial data.

Assume that

$$\gamma_1\gamma_2 < \frac{1+\beta}{2}.$$

Then, there exists a unique solution for (9), and we can define a semigroup which possesses a global attractor $\mathscr{A} \subset \mathbf{R}^3$. Observe that, in this case, the initial real network is just composed of three nodes, connected as show in the figure:

For (9), it is easy to show that the eight stationary points

$$\mathbf{E} := \left\{E_{ijk}\right\}, \quad i, j, k = 0, 1,$$

are given by

$$E_{000} = (0,0,0),\ E_{100} = (\alpha_1, 0, 0),\ E_{010} = (0, \alpha_2, 0),\ E_{001} = (0, 0, \alpha_3),$$

$$E_{011} = \left(0, \frac{\alpha_2 + \gamma_1\alpha_3}{1-\gamma_1\gamma_2}, \frac{\alpha_3 + \gamma_2\alpha_2}{1-\gamma_1\gamma_2}\right),$$

$$E_{101} = \left(\frac{\alpha_1 + \gamma_1\alpha_3}{1-\gamma_1\gamma_2}, 0, \frac{\alpha_3 + \gamma_2\alpha_1}{1-\gamma_1\gamma_2}\right),$$

$$E_{110} = \left(\frac{\alpha_1 - \beta\alpha_2}{1-\beta^2}, \frac{\alpha_2 - \beta\alpha_1}{1-\beta^2}, 0 \right),$$

$$E_{111} = \begin{pmatrix} \dfrac{\alpha_1(1-\gamma_1\gamma_2) + \alpha_2(\gamma_1\gamma_2 - \beta) + \alpha_3\gamma_1(1-\beta)}{(1-\beta)(1+\beta-2\gamma_1\gamma_2)} \\ \dfrac{\alpha_1(\gamma_1\gamma_2 - \beta) + \alpha_2(1-\gamma_1\gamma_2) + \alpha_3\gamma_1(1-\beta)}{(1-\beta)(1+\beta-2\gamma_1\gamma_2)} \\ \dfrac{(\alpha_1+\alpha_2)\gamma_2 + \alpha_3(1+\beta)}{1+\beta-2\gamma_1\gamma_2} \end{pmatrix}^t.$$

4.1 Stability of Equilibria

To analyze the stability of these points, we calculate the eigenvalues of the Jacobian matrix at a stationary point (u_1, u_2, u_3) given by

$$J(u_1, u_2, u_3) = \begin{pmatrix} \alpha_1 - 2u_1 - \beta u_2 + \gamma_1 u_3 & -\beta u_1 & \gamma_1 u_1 \\ -\beta u_2 & \alpha_2 - 2u_2 - \beta u_1 + \gamma_1 u_3 & \gamma_1 u_2 \\ \gamma_2 u_3 & \gamma_2 u_3 & \alpha_3 - 2u_3 + \gamma_2(u_1 + u_2) \end{pmatrix}.$$

We have the following result (see Guerrero et al. 2014):

Theorem 8

a) Assume that the three components of E_{111} are strictly positive, then E_{111} is locally stable.

b) When the components of E_{111} are strictly positive then the semi-trivial stationary points E_{011}, E_{101} and E_{110} are unstable.

c) Assume that E_{111} exists. Then, it is globally asymptotically stable in the interior of $\mathbf{R}^3_+$. As a consequence, system (9) is permanent, i.e., asymptotically there exists coexistence of the three nodes.

d) The global attractor $\mathscr{A} \subset \mathbf{R}^3$ is given by

$$\mathscr{A} = \bigcup_{i,j,k=0}^{1} W^u\left(E_{ijk}\right).$$

This simplified model allows us to describe in detail the dependence of the associated attracting networks on parameters, showing that, in particular, to a fix phenomenological network of three species (Fig. 5), corresponds a bigger and more complex set of possible future configurations given by different architectures of the global attractor, described from the equilibria and his oriented connections, as show in the Fig. 6.

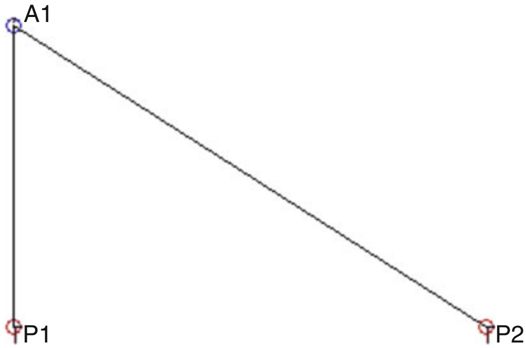


Fig. 5 Phenomenological simple network in the 3D case

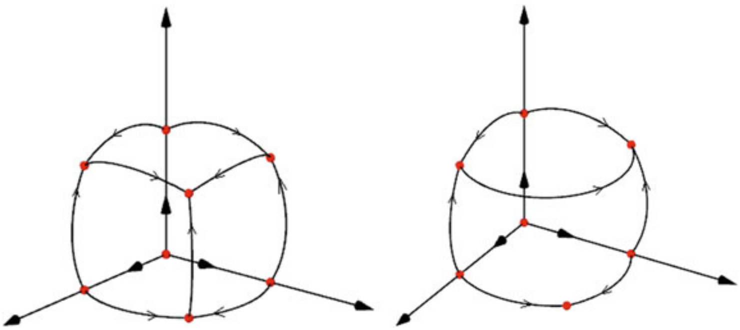


Fig. 6 Two possible attracting complex networks in the 3D case. *Left* network shows the case in which E_{111} exists and it is asymptotically globally stable; in this case E_{111} is the stationary point with the lower energy level. In the *right picture* E_{111} is not present, and the global configuration of the attracting network changes, now with E_{011} as the asymptotically globally stable equilibrium. Other networks could be also reached, depending the values of the parameters in (9). Observe that all this network of possible attracting networks correspond to the same phenomenological characterization given in Fig. 5

Item (d) of Theorem 8 shows that the global attractor is gradient-like. In fact, as the associated semigroup for (9) is dynamically gradient, the stationary points in the global attractor are the Morse sets of an associated Morse decomposition of it. In particular, we can order the equilibria by energy levels (given by the associated Lyapunov functional), which describes the hierarchy on how the long-time dynamics develops with respect to positive solutions (see Fig. 7).

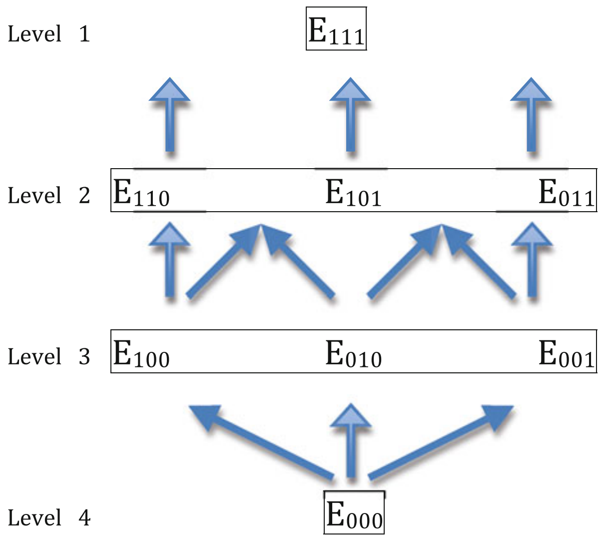


Fig. 7 Organization of stationary points in the global attractor by Energy Levels. Upper level shows the minimum energy (given by the Lyapunov function), attracting every strictly positive solution. *Blue arrows* shows the direction of the forwards dynamics. The second level is achieved if E_{111} does not exists and any of the E_{110}, E_{101} or E_{011} exists. The third level is only reached if any of the equilibria in upper levels are not present. $E_{000} = (0, 0, 0)$ is asymptotically globally stable only if any of the stationary points in an upper level exist (Color figure online)

5 Conclusions

We now enumerate some conclusions and open problems of the approach we have developed in this chapter.

(a) From a real phenomena we build a complex network of nodes and connections. To model the dynamics on this complex network we build a system of ordinary or partial differential equation with an associated dynamical system.
(b) A new complex network appears, described by the structure of the global attractor of the dynamical system, which includes dynamics and dependence on parameters.
(c) New mathematically open problems appear:

1. The dependence of the structure of an attractor on parameters of the system is usually unknown.
2. In this sense, a theory on Attractor Bifurcation on parameters would be needed.
3. It still remains the challenging problem of dynamical complex networks, i.e., a network with time dependent parameters.

Acknowledgements J.A. Langa is partially supported by FEDER and Ministerio de Economía y Competitividad grant # MTM2015-63723-P, Junta de Andalucía under Proyecto de Excelencia FQM-1492 and Brazilian-European partnership in Dynamical Systems (BREUDS) from the FP7-IRSES grant of the European Union.

A. Suárez is partially supported by FEDER and Ministerio de Economía y Competitividad grant # MTM2012-31304. This Chapter is based upon work from COST Action ISCH COST Action IS1104 "The EU in the new complex geography of economic systems: models, tools and policy evaluation", supported by COST (European Cooperation in Science and Technology) www.cost.eu.

References

Allesina S, Tang S (2012) Stability criteria for complex ecosystems. Nature 483:205–208

Aragão-Costa ER, Caraballo T, Carvalho AN, Langa JA (2011) Stability of gradient semigroups under perturbation. Nonlinearity 24:2099–2117

Aragão-Costa ER, Caraballo T, Carvalho AN, Langa JA (2012) Continuity of Lyapunov functions and of energy level for a generalized gradient system. Topol Methods Nonlinear Anal 39:57–82

Arrieta J, Carvalho AN, Langa JA, Rodríguez-Bernal A (2012) Continuity of dynamical structures for non-autonomous evolution equations under singular perturbations. J Dyn Diff Equat 24:427–481

Babin AV, Vishik M (1983) Regular attractors of semigroups and evolution equations. J Math Pures Appl 62:441–491

Babin AV, Vishik MI (1992) Attractors of evolution equations. North Holland, Amsterdam

Bascompte J, Jordano P (2007) The structure of plant-animal mutualistic networks: the architecture of biodiversity. Annu Rev Ecol Evol Syst 38:567–593

Bascompte J, Jordano P, Melián CJ, Olesen JM (2003) The nested assembly of plant-animal mutualistic networks. Proc Natl Acad Sci USA 100:9383–9387

Bascompte J, Jordano P, Olesen JM (2006) Asymmetric coevolutionary networks facilitate biodiversity maintenance. Science 312:431–433

Bastolla U, Fortuna MA, Pascual-García A, Ferrera A, Luque B, Bascompte J (2009) The architecture of mutualistic networks minimizes competition and increases biodiversity. Nature 458:1018–1020

Carvalho AN, Langa JA (2009) An extension of the concept of gradient semigroups which is stable under perturbation. J Differ Equ 246:2646–2668

Carvalho AN, Langa JA, Robinson JC, Suárez A (2007) Characterization of non-autonomous attractors of a perturbed gradient system. J Differ Equ 236:570–603

Carvalho AN, Langa JA, Robinson JC (2013) Attractors for infinite-dimensional non-autonomous dynamical systems. Applied mathematical series, vol 182. Springer, New York

Carvalho AN, Langa JA, Robinson JC (2015) Non-autonomous dynamical systems. Discret Cont Dyn Syst B 20:703–747

Conley C (1978) Isolated invariant sets and the Morse index. CBMS regional conference series in mathematics, vol 38. American Mathematical Society, Providence, RI

Guerrero G, Langa JA, Suárez A (2014) Biodiversity and vulnerability in a 3D mutualistic system. Discret Cont Dyn Syst 34:4107–4126

Guerrero G, Langa JA, Suárez A (2015) Architecture of attractor determines dynamics on mutualistic complex networks. Preprint

Hale JK (1988) Asymptotic behavior of dissipative systems. Mathematical surveys and monographs. American Mathematical Society, Providence, RI

Hale JK, Magalhães LT, Oliva WM (1984) An introduction to infinite-dimensional dynamical systems - geometric theory. Applied mathematical sciences, vol 47. Springer, Berlin

Henry D (1981) Geometric theory of semilinear parabolic equations. Lecture notes in mathematics, vol 840. Springer, Berlin
Ladyzhenskaya OA (1991) Attractors for semigroups and evolution equations. Cambridge University Press, Cambridge
Okuyama T, Holland JN (2008) Network structural properties mediate the stability of mutualistic communities. Ecol Lett 11:208–216
Robinson JC (2001) Infinite-dimensional dynamical systems. Cambridge University Press, Cambridge
Rohr RP, Saavedra S, Bascompte J (2014) On the structural stability of mutualistic systems. Science 345:1–9
Rybakowski KP (1987) The homotopy index and partial differential equations. Springer, Berlin
Saavedra S, Reed-Tsochas F, Uzzi B (2009a) A simple model of bipartite cooperation for ecological and organisational networks. Nature 457:463–466
Saavedra S, Reed-Tsochas F, Uzzi B (2009b) Asymmetric disassembly and robustness in declining networks. Proc Natl Acad Sci USA 10:16466–16471
Saavedra S, Stouffer DB, Uzzi B, Bascompte J (2011) Strong contributors to network persistence are the most vulnerable to extinction. Nature 478:233–235
Sell GR, You Y (2002) Dynamics of evolutionary equations. Applied mathematical sciences, vol 143. Springer, New York
Suweis S, Simini F, Banavar JR, Maritan A (2013) Emergence of structural and dynamical properties of ecological mutualistic networks. Nature 500:449–452
Takeuchi Y (1996) Global dynamical properties of Lotka-Volterra systems. World Scientific, River Edge, NJ
Temam R (1988) Infinite-dimensional dynamical systems in mechanics and physics. Springer, Berlin
Thébault E, Fontaine C (2010) Stability of ecological communities and the architecture of mutualistic and trophic networks. Science 329:853–856
Vishik MI (1992) Asymptotic behaviour of solutions of evolutionary equations. Cambridge University Press, Cambridge

Good Old Economic Geography

Tönu Puu

Abstract This chapter discusses classical economic modelling in continuous two-dimensional geographical space, focusing some ingenious models due to Harold Hotelling and Martin Beckmann concerning population growth and migration, and spatial market equilibrium, respectively. It also adds some modelling of business cycles, and discusses the issue the shape of market areas. Focus is on transversality and stability of structure.

Keywords Business cycle diffusion • Continuous geographical space • Interregional trade • Migration • Structural stability

1 Introduction

Upon the advent of the New Economic Geography some 15 years ago, the present author wrote that space was important, but it needed not to be reinvented so many times. Maybe he was wrong. During his student years economics textbooks, such as those by Erich Schneider in Kiel started with spatial issues such as Launhardt's "funnels" after an introduction dealing with demography. This is history, and no economics textbook mentions such issues any longer.

According to Joseph Shumpeter's monumental history (Schumpeter 1954), Johann Heinrich von Thünen's pioneering work on land use of 1826 (von Thunen 1826) provided the first true mathematical model in economics. He also rated his capacity high above David Ricardo's, and above any contemporary, maybe with the exception of Augustin Cournot.

In what follows we find reason to deplore that von Thünen was degraded to an "agricultural economist", when in truth he produced the most general theory for interregional specialization and trade ever proposed.

The founding of the Regional Science Association through Walter Isard and other pioneers marked the birth of a new interdisciplinary research field where

T. Puu (✉)
CERUM, Umeå University, SE 90187 Umeå, Sweden
e-mail: tonu.puu@umu.se

P. Commendatore et al. (eds.), *Complex Networks and Dynamics*,
Lecture Notes in Economics and Mathematical Systems 683,
DOI 10.1007/978-3-319-40803-3_13

economists and geographers collaborated. To understand economic phenomena in space needs insights from both fields. Geographers understand the nature of space through their familiarity with map projections in a way economists never do, but they tend to base models on simple ad hoc assumptions and analogies, such as the gravity hypothesis. The economist's concepts of optimizing agents can put the models on a more stable footing. At the same time the openness of this new interdisciplinary field provided for access for physicists and mathematicians. After all, Cournot had been a mathematician. These people contributed most interesting ideas to regional science. I will only exemplify with physicist Sir Alan Wilson and his most innovative "entropy model" (Wilson 1970).

But, regional science separated from economics, and space, with issues such as location and land use, no longer belonged to the core of economic theory. So, maybe economists indeed need to reinvent space.

Now, space can be treated as a discrete set of locations connected through links or as a continuous two-dimensional plane. The first alternative is good for applications, but one loses intuition for the geometry of shape and size. Every assumption of continuity, be it time or space, is always an abstraction, which itself can never be verified as an assumption. Only indirectly can continuous models be checked through the conclusions they yield, quite as in physics.

As most modelling, be it in regional science, or the new economic geography, is concerned with matrices that have no visuality, we in the present chapter intend to show some examples of the intuitive attraction of models in continuous space as von Thünen and many other pioneers saw it.

2 Spatial Trade and Pricing

One of the most innovative models ever contributed to economics is the one proposed by Martin J Beckmann in 1952 (Beckmann 1952). It is, to say the least, nonstandard, even conceptually, and therefore it never gained popularity. Probably there even was a methodological block for potential readers, as economists almost never used concepts or methods of this kind from vector analysis, calculus of variations, and partial differential equations.

The same year 1952 Paul A Samuelson treated the same problem, though space is treated as a discrete set of nodes, joined by arcs. Space is represented as abstract matrices of incidence, distance, and transportation costs. So, the intuitive notions of size and shape slip out of one's hands. In stead of arriving at two simple partial differential equations as Beckmann does, Samuelson's model ends up in a mess of conditions with no link to visualisation.

Beckmann's purpose is to model a spatially extended economy where trade moves in a continuous 2D flow. There are continuously distributed sources of excess supply and sinks of excess demand over space. Goods are entered into or withdrawn from the flow to provide equilibrium over space as a whole if the region considered is insulated, which, however, is not necessary; there may also be exports or imports all along its exterior.

2.1 Concepts

The concept of flow is central, mathematically a two-dimensional vector field over the space R considered, changing direction and volume with its basing point $(x_1, x_2) \in R \subset \mathbb{R}^2$.

$$\boldsymbol{\phi} = \begin{bmatrix} \phi_1 (x_1, x_2) \\ \phi_2 (x_1, x_2) \end{bmatrix} \tag{1}$$

The direction θ is the actual direction of trade flow on a map of the region studied

$$\frac{\boldsymbol{\phi}}{|\boldsymbol{\phi}|} = \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix} = \begin{bmatrix} \frac{dx_1}{ds} \\ \frac{dx_2}{ds} \end{bmatrix} \tag{2}$$

where s denotes a path parameter, and the volume $|\boldsymbol{\phi}|$ is the actual volume of commodities transported in the trade flow

$$|\boldsymbol{\phi}| = \sqrt{\phi_1^2 + \phi_2^2} \tag{3}$$

How excess supply/demand is related to the flow (in economics, unlike physics, incompressible, as traded merchandise is not altered in volume due to any kind of "pressure") is formalized through the divergence operator, written $\nabla \cdot \boldsymbol{\phi}$. It is defined as follows

$$\nabla \cdot \boldsymbol{\phi} = \frac{\partial \phi_1}{\partial x_1} + \frac{\partial \phi_2}{\partial x_2} \tag{4}$$

This is definitely a nonstandard differential operator to economists. However, it can be intuitively explained as follows: Consider a flow passing a small box. The flow has a horizontal component ϕ_1 and a vertical ϕ_2. While passing the box the horizontal component increases by $\frac{\partial \phi_1}{\partial x_1}$ and the vertical by $\frac{\partial \phi_2}{\partial x_2}$. If the sum is positive then the addition to the flow must be withdrawn from a local source (excess supply), if it is negative, then there is a local sink (excess demand) that sucks from the flow.

The reader must not think that the shape of a rectangular box or its size are essential. A very general theorem due to Gauss (or Green in 2D) ascertains that the divergence is related to local source/sink density, whatever the shape or size of "box". For mathematical detail see Puu (1997, 2002).

So shrinking size to zero, we equate divergence in every point to excess demand, formally $\nabla \cdot \boldsymbol{\phi} + z(x_1, x_2) = 0$, where $z(x_1, x_2)$ is local excess demand. This is an equilibrium condition for local consumption, production, and trade in the flow. It just guarantees that every not locally produced item consumed is withdrawn from the trade flow, and every not locally consumed item is entered into it. This is all well known stuff to any engineer. As one of the applications is hydrodynamics for compressible or incompressible fluids, economists who have not quite understood

Beckmann's model have called it hydrodynamic analogy. As we will see this is extremely unfair, as Beckmann's model contains other elements that have nothing whatever to do with hydrodynamics.

Stated in this way, local excess demand $z(x_1, x_2)$ is a given datum in each location. This needs not be so, we can make it dependent on local commodity price. Beckmann in his later addition actually uses this. Assuming excess demand to be a decreasing function of price, he is able to prove both uniqueness and stability for a linearized dynamic version of the model (Beckmann 1976).

But we wait with introducing price as it turns up as a Lagrange multiplier of the market equilibrium constraint as Beckmann formulates the model.

The flow $\boldsymbol{\phi}$ is a vector field with two components ϕ_1 and ϕ_2, both dependent on the space coordinates x_1, x_2. We also need a scalar field $\kappa(x_1, x_2)$ that represents local transportation cost for hauling the commodities in the flow over any given point. Written in this way transport is independent of direction θ, but it can be made direction-dependent, for instance in the case when we deal with a Manhattan distance metric instead of the Euclidean. Beckmann's model is valid also for such cases.

2.2 The Model

We are now ready to state Beckmann's model: Minimize total transportation cost on the region R:

$$C = \iint_R \kappa(x_1, x_2) \, |\boldsymbol{\phi}| \, dx_1 dx_2 \tag{5}$$

subject to the spatial equilibrium constraint

$$\nabla \cdot \boldsymbol{\phi} + z(x_1, x_2) = 0 \tag{6}$$

using a Lagrange function $p(x_1, x_2)$ associated with the constraint.

Note that such a calculus of variations problem is non-standard even in mathematical physics. Total transportation cost C is defined as the product of local transportation cost κ and the volume of flow $|\boldsymbol{\phi}| = \sqrt{\phi_1(x_1, x_2)^2 + \phi_2(x_1, x_2)^2}$ integrated over all points of the region R. It can, by the way, be shown that this cost expression equals the total of all individual transports taken from source to sink along the optimal routes.

Beckmann's optimization problem is solved by the vector equation

$$\kappa \frac{\boldsymbol{\phi}}{|\boldsymbol{\phi}|} = \nabla p = \begin{bmatrix} \frac{\partial p}{\partial x_1} \\ \frac{\partial p}{\partial x_2} \end{bmatrix} \tag{7}$$

to which we add the equilibrium constraint

$$\nabla \cdot \boldsymbol{\phi} + z(x_1, x_2) = 0$$

which we state once more for convenience.

These two equations describe the solution to the interregional trade and pricing problem. Note that though Beckmann stated optimization as a planning problem for all transports, the same result is obtained if we consider any number of independent individual transport operators.

As $p(x_1, x_2)$ is associated with a scarcity constraint, economic intuition says that it is an imputed price, which is indeed verified if we consider what the optimality condition tells us:

Goods are transported (1) in the direction of the price gradient ∇p, and (2) in this direction prices increase with transportation cost, neither more nor less. Intuitively this makes perfect sense. Prices change over space, from a given point the rate of increase/decrease depends on direction. Prices tell us where in the neighbourhood the commodity is scarce, and transporters make the largest profit by moving them in the direction where the gain is maximal, i.e. the gradient direction. Further, if the gains from trading are larger than transportation cost it would make sense to increase the flow, if the gains are less trade is unprofitable and should be decreased. The system of interregional trade can be in equilibrium if and only if the gains from trade exactly equal the cost of transportation.

Mathematically, we can now take the square of the optimality condition. Though it represents two differential equations, the unit vector squared multiplies to scalar unity, i.e., $\left(\frac{\boldsymbol{\phi}}{|\boldsymbol{\phi}|}\right)^2 = 1$, whereas $(\nabla p)^2 = \left(\frac{\partial p}{\partial x_1}\right)^2 + \left(\frac{\partial p}{\partial x_2}\right)^2$ by definition. Thus, we get one single partial differential equation for price:

$$\left(\frac{\partial p}{\partial x_1}\right)^2 + \left(\frac{\partial p}{\partial x_2}\right)^2 = \kappa(x_1, x_2)^2 \tag{8}$$

Its structure even suggests a constructive method for solution. Say that we know a curve for constant price in x_1, x_2 space. The equation suggests that if we for any point on it draw a circle with radius $1/\kappa$ it encloses the area that can be covered using up one monetary unit. We do this for many points, not forgetting to adjust the radius as κ changes from point to point in space. The envelope to such a family of circles then is the next constant price curve, and the gradient directions (= flow lines) become orthogonal to any so constructed family of constant price curves. And so it goes, from the new constant price curve we can construct the next, and so on. The idea is illustrated in Fig. 1, and obviously suggests some algorithm for solving the partial differential equation for pricing.

However, recall that a partial differential equation requires some boundary condition for a solution. In the case illustrated it was the starting curve of constant prices. The outcome, i.e. the price surface over the region depends on the boundary condition.

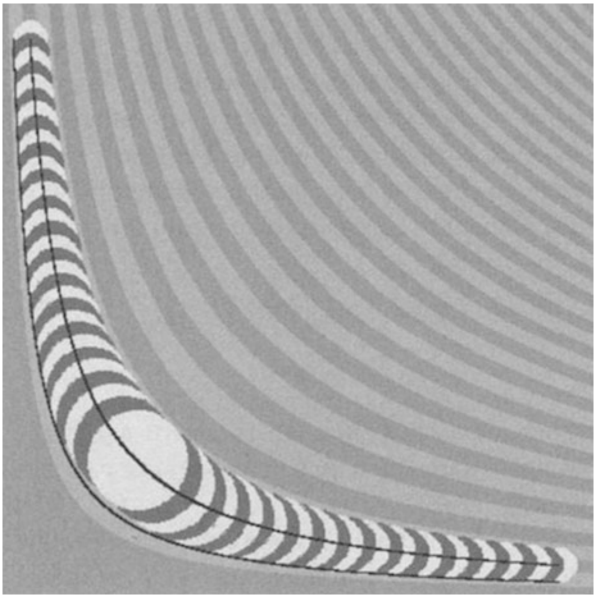


Fig. 1 Method of constructing new constant price curves starting with one initial one, in this case the hyperbola passing the centres of all the circles. Take any point on the hyperbola and draw a circle with radius $1/\kappa$. Do this for as many circles as needed, and draw an envelope touching all the circles. As transportation cost varies over space, the radius must be adjusted anew for each point. The lowest envelope is shown, and the whole shading illustrates the family of price contours. The starting curve can be regarded as a boundary condition. The accuracy of the method, of course, becomes better the smaller increments in transportation costs we take. This pedestrian procedure could be the basis for constructing an algorithm for numerical solution of price and flow structure in Beckmann's model

2.2.1 Examples

As example, suppose transportation cost is zero in the origin of x_1, x_2 space and increases with radius vector, i.e., $\kappa = \sqrt{x_1^2 + x_2^2}$, and our PDE reads $\left(\frac{\partial p}{\partial x_1}\right)^2 + \left(\frac{\partial p}{\partial x_2}\right)^2 = x_1^2 + x_2^2$. Then both $p = \frac{1}{2}\left(x_1^2 + x_2^2\right)$ and $p = \frac{1}{2}\left(x_1^2 - x_2^2\right)$ solve the equation, as illustrated in Figs. 2 and 3. In the first case we have a von Thünen type of ring shaped structure with radial flows, in the second there is a shear structure. The corresponding price surfaces are of bowl type and saddle type respectively. Which depends on the kind of boundary conditions. Also the mountain type $p = -\frac{1}{2}\left(x_1^2 + x_2^2\right)$ would fit the PDE. Actually, in Fig. 2 it is not obvious if the flow is outward (bowl) or inward (mountain). This depends on the equilibrium condition and the distribution of excess demand, which, admittedly, is a tougher issue.

As mentioned in a rejoinder Beckmann also extended the equilibrium model to a linearized dynamic version of the type current at the time. It is interesting to note that he proved that if excess demand $z(x_1, x_2, p)$ was dependent on price with

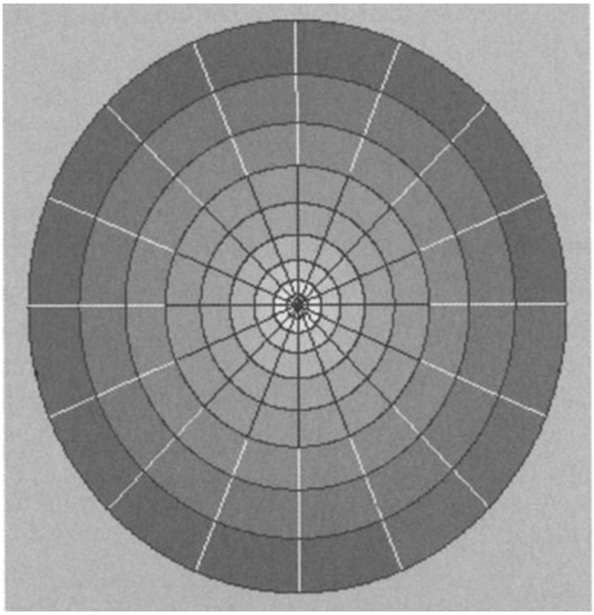


Fig. 2 Radial flow and circular price contours as one solution to the case where $\kappa = \sqrt{x_1^2 + x_2^2}$

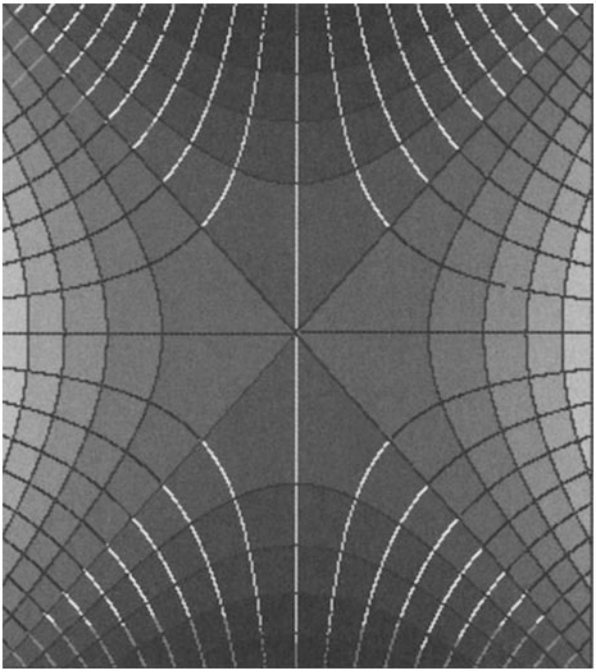


Fig. 3 Hyperbola shaped flow lines and price contours, for the same transportation cost function $\kappa = \sqrt{x_1^2 + x_2^2}$. as in the previous illustration

$\frac{\partial z}{\partial p} \leq 0$, then equilibrium was stable. By the same assumption it was also proved that equilibrium was unique. In this context it is essential that we make clear what equilibrium means. It refers to the spatial structure of commodity price $p(x_1, x_2)$, and the flow of trade $\boldsymbol{\phi} = (\phi_1(x_1, x_2), \phi_2(x_1, x_2))$ which evens out local excesses of supply and demand. No doubt, as there are goods traded and excess demand from local consumption or excess supply from local production there is something going on all the time. So, equilibrium implies that things are just repeated from one period to next, if we may say so in the abstract context of continuous time.

2.3 Multi-Commodity Production Economy

Another extension, suggested by Beckmann and the present author in 1985 (Beckmann and Puu 1985) is to extend the originally one commodity market to multiple commodities that are all produced, consumed, and transported. The most interesting outcome of this is perhaps that in each location the productive activity is chosen which is most profitable just there. This leads to a uniqueness theorem stating that everywhere just one commodity is both produced and transported. Hence production is specialized and space is broken up in production regions—just as in von Thünen's original model from 1826. von Thünen's theory has never been acknowledged as the most general model ever produced for specialization and trade, assuming neither comparative advantages (Ricardo), nor spatially trapped production resources (Heckscher/Ohlin). This still holds in our generalization from 1985—capital and labour are freely movable and hence equalize capital rents and wage rates over space, only land stays where it is and yields different rents depending on location. Note that, like von Thünen, we assume no differences in fertility or the like—the same technology applies everywhere, so land rent differences result solely from location in the whole layout of economic space. However, unlike von Thünen, we do not just refer to agricultural activities, but to any kind of economic activity.

Passing to the formal model, assume any number of commodities producable by Cobb-Douglas technologies with constant returns to scale when all inputs can be varied, $\alpha_i + \beta_i + \gamma_i = 1$. K_i, L_i denote capital and labour inputs, and M_i the input of land (or space).

$$Q_i = A_i K_i^{\alpha_i} L_i^{\beta_i} M_i^{\gamma_i}$$

As there are constant returns, we can divide through by M_i, which is the more reasonable as it converts output and inputs to *areal densities*. Thus

$$\frac{Q_i}{M_i} = A_i \left(\frac{K_i}{M_i}\right)^{\alpha_i} \left(\frac{L_i}{M_i}\right)^{\beta_i}$$

or, introducing lower case symbols for the densities:

$$q_i = A_i k_i^{\alpha_i} l_i^{\beta_i} \tag{9}$$

Now, assume costless mobility of capital and labour resulting in equalized capital rent r and wage rate w over the entire region. Profit per unit land area thus becomes:

$$g_i = p_i q_i - r k_i - w l_i$$

Optimizing g_i with respect to k_i, l_i under the constraint of the production function results in the well known Cobb-Douglas conditions

$$k_i = \alpha_i \frac{p_i}{r} q_i$$

and

$$l_i = \beta_i \frac{p_i}{w} q_i$$

Substituting in the profit expression, recalling that $\gamma_i = 1 - \alpha_i - \beta_i$,

$$g_i = \gamma_i p_i q_i$$

which as stated is land rent.

However, substituting for capital and labour inputs in the production function (9), we can solve for output (areal density)

$$q_i = \gamma_i A_i^{\frac{1}{\gamma_i}} \left(\frac{p_i}{r} \right)^{\frac{\alpha_i}{\gamma_i}} \left(\frac{p_i}{w} \right)^{\frac{\beta_i}{\gamma_i}}$$

and substitute back in the land rent expression. We get profit as dependent on (local) commodity price alone:

$$g_i = c_i p_i^{\frac{1}{\gamma_i}} \tag{10}$$

where

$$c_i = \gamma_i A_i^{\frac{1}{\gamma_i}} \left(\frac{\alpha_i}{r} \right)^{\frac{\alpha_i}{\gamma_i}} \left(\frac{\beta_i}{w} \right)^{\frac{\beta_i}{\gamma_i}} \tag{11}$$

is a compound constant depending on the parameters alone, i.e., the Cobb-Douglas parameters, and further the capital rent and wage rate.

Price, of course, like land rent, are functions of the space coordinates (as are inputs of capital and labour which have been substituted out at this stage).

Now, potential entrepreneurs will, of course, choose the production activity at each location which yields maximum profit, i.e. for which

$$g_i(x_1, x_2) = g(x_1, x_2) = \max_j g_j(x_1, x_2) \tag{12}$$

holds.

2.4 Specialization

The obvious question now arises whether there can be a mix up of activities, so that the condition (12) holds for several or even all i?

For an answer, assume that the commodities are also transported. We already have the PDE for the price of transported goods $(\nabla p)^2 = \kappa^2$. Now the goods are different. We may still assume that the same transportation cost metric applies for all, but they differ with respect to weight and bulkiness. Hence, we put transportation cost as $\kappa_i d(x_1, x_2)$, where $d(x_1, x_2)$ represents the common metric, and κ_i are commodity specific. Hence $(\nabla p_i)^2 = \kappa_i^2 d(x_1, x_2)^2$.

To connect price gradient to areal profit gradient, we take the gradient of $g_i = c_i p_i^{\frac{1}{\gamma_i}}$, substituting back for p_i in terms of g_i:

$$\nabla g_i = \frac{c_i^{\gamma_i}}{\gamma_i} g_i^{1-\gamma_i} \nabla p_i$$

Finally, substituting for ∇p_i and taking squares

$$\left(\frac{\partial g_i}{\partial x_1}\right)^2 + \left(\frac{\partial g_i}{\partial x_2}\right)^2 = \left(\frac{c_i^{\gamma_i}}{\gamma_i} g_i^{1-\gamma_i}\right)^2 \kappa_i^2 d(x_1, x_2)^2 \tag{13}$$

This is a proper PDE, even if it is not quite as simple as the one we dealt with, due to the appearance of g_i in the right hand side. It can be fitted to a curve in x_1, x_2 space—as we know it even has to in order to yield a solution. And then it results in a potential land rent surface for commodity number i. The equation for another activity can be fitted to the same curve as boundary condition, but as its PDE is different the coincidence only holds on the curve. If they do, i.e., yield the same land rent, then the curve is a "cultivation zone boundary".

That the potential land rent surfaces intersect along curves of dimension 1 only follows from transversality, emerged with catastrophe theory, and discussed in the sequel of the chapter in connection with the formation of market areas. However, transversality was tacit already in von Thünen's analysis, even if it was formalized only more than a Century later.

The choice of maximum possible land rent everywhere, breaks up space into specialization zones. Actually we have a specialization theorem: only one commodity in each location can be both locally produced and traded. There may yet be any number of goods produced for only local consumption, or any number just flowing past the location without any local production.

2.4.1 Examples

To visualize the issue of transversality and the specialization we take some illustrative examples. We need solutions to the differential equation (13), restated for convenience

$$(\nabla g_i)^2 = \left(\frac{c_i^{\gamma_i}}{\gamma_i} g_i^{1-\gamma_i}\right)^2 \kappa_i^2 d\,(x_1, x_2)^2$$

where

$$(\nabla g_i)^2 = \left(\frac{\partial g_i}{\partial x_1}\right)^2 + \left(\frac{\partial g_i}{\partial x_2}\right)^2$$

To be sure that they intersect in a specialization boundary curve, let us fix the condition $\nabla g_i = 1$ when $x_1 x_2 = 1$. Recall that we anyhow need a boundary condition to at all get a solution for a PDE—in this case it is a hyperbola at unit height.

We have lots of parameters to choose from; capital rent r and wage rate w, the parameters of the production function, $\alpha_i, \beta_i, \gamma_i = 1-\alpha_i-\beta_i$ and A_i, and, finally, the product specific transportation rate κ_i. As we do not need that many to fit examples, fix $r = w = 1$.

(i) Then, let us first take $\alpha_i = \beta_i = \gamma_i = \frac{1}{3}$, $A_i = 1$, and $\kappa_i = 1$. From (11) then $c_i = 1$, and the PDE reads: $(\nabla g_i)^2 = \left(3 g_i^{\frac{2}{3}}\right)^2 d\,(x_1, x_2)^2$. Suppose the until now unspecified transport cost metric is $d = \sqrt{x_1^2 + x_2^2}$.

$$(\nabla g_i)^2 = 9 g_i^{\frac{4}{3}} \left(x_1^2 + x_2^2\right)^2$$

and its solution, which also satisfies the boundary condition is $g_i = (x_1 x_2)^3$. This is then the profit, imputable as land rent if the activity i is chosen.

(ii) As another case take $\alpha_j = \beta_j = \frac{3}{8}$, $\gamma_j = \frac{1}{4}$, $A_j = 8$, and $\kappa_j = \frac{1}{72}$. Then

$$\left(\nabla g_j\right)^2 = 16 g_j^{\frac{3}{2}} \left(x_1^2 + x_2^2\right)^2$$

with solution $g_i = (x_1 x_2)^4$. which too satisfies the boundary condition.

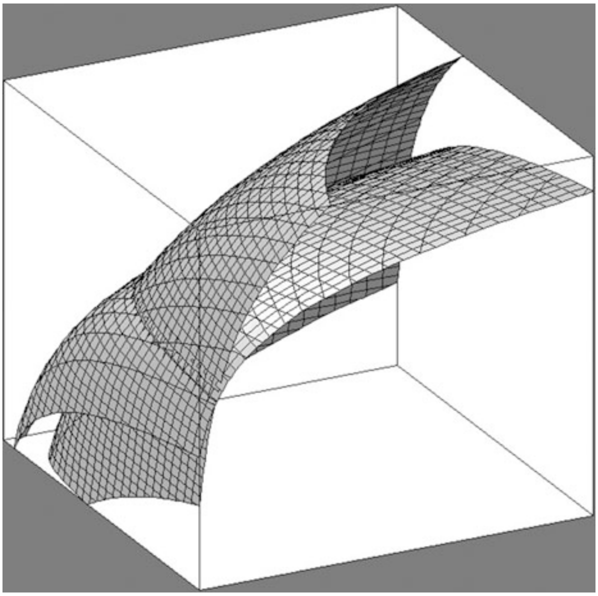


Fig. 4 The profit surfaces for the two exemplified productive activities and their intersection. Note that in order to increase visibility we display the surfaces in logarithmic scale vertically

The two surfaces are illustrated in Fig. 4. It is clear that they intersect along the common boundary curve, which hence becomes a boundary for specialization zones, as the productive activity is chosen which yields the highest rent imputed to land.

The only way specialization can be avoided would be if the surfaces not only shared the same boundary, but coincided over 2-dimensional land patches, so could this happen? The answer is no, because the surfaces are different, generated by different PDE. Note that this does not refer to our particular examples, it is a fact of intersecting surfaces in general. Two surfaces intersect along curves, they do not coincide. They can, of course, intersect again somewhere else, but this intersection again is a curve. This is regulated by the topological principle of transversality, which states that the sum of the dimensions of intersecting manifolds must be equal to the sum of the dimensions of the intersection manifold and the embedding space. Thus, $2 + 2 = 1 + 3$, surfaces on left, curve and embedding space on right. But $2 + 2 < 2 + 3$. For the same reason, a third surface, resulting from a third activity, would intersect the two previous ones along curves, but both of them only at a point $2 + 1 = 0 + 3$, new surface and intersection curve of the previous on the left, point and embedding space on the right.

We will consider transversality again below, in the context of market areas.

The choice of most profitable activity everywhere results in a unique land rent surface shown in Fig. 5. It can be considered a generalization of von Thünen's theory. In stead of an "isolated state" surrounded by cultivation zones and eventually

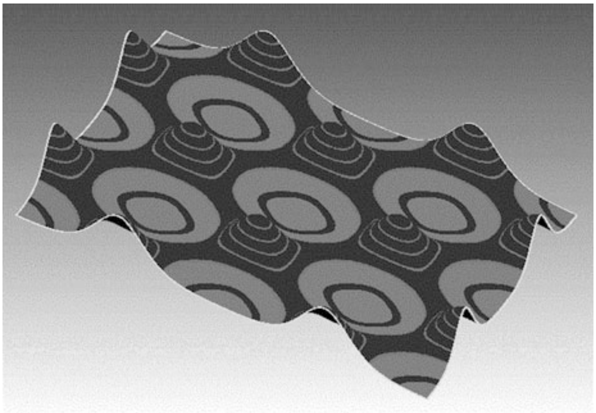


Fig. 5 Example of land rent landscape, and specialization zones for the multi-product economy, with land rent maxima and minima in a regular square pattern

"wilderness", it shows an "integrated state" with multiple "central cities" (local land rent maxima) and "wildernesses" that contract to multiple land rent minima.

We leave the interregional trade model for a while, but return to it by the end of the chapter for some topological considerations. General models such as Beckmann's may be illustrated by specific geometric examples, but only when considering topology can we enter into qualitative issues pertinent to structural stability.

3 Population Growth and Dispersal

Concepts from vector analysis were used even before, by Harold Hotelling in 1921 (Hotelling 1921). The contribution appeared as a master degree thesis. It is so little known that it was not even included in the tiny volume of Hotelling's (always ingenious) collected works in economics. Ultimately it was published by Sir Alan Wilson in his journal Environment and Planning. To the knowledge of the present author it was never commented on by any economist except Martin Beckmann and the present author (Beckmann and Puu 1990). The model was reinvented in 1951 by Skellam (1951) in mathematical ecology, 30 years later, and gave rise to thousands of scientific contributions.

The model combines a Malthusian (logistic) type of population growth and spatial dispersal modelled through the Laplacian differential operator. The arguments for growth are the same for Hotelling and Skellam, whereas the arguments for dispersal from densely to sparsely populated areas differ.

Both argue that close to zero population it tends to grow exponentially, whereas close to the limiting sustainable size the rate becomes drastically smaller. Should it ever overshoot the sustainable size, it is bound to decrease, quite as Thomas Malthus once suggested.

As for dispersal, Hotelling claims that whenever production occurs under decreasing returns, it is better to move to another place where population is less dense and per capita production therefore higher. Unfortunately, Hotelling never puts this argument in a shape suitable for a mathematical model. Skellam, in contrast, argues that animal populations move at random. Hence from a concentration area more animals are likely to move out than in from the surroundings. The argument is in fact better than Hotelling's, and it is by no means easy to straighten out and formalize what Hotelling had in mind.

Denote sustainable population s, taken as a constant over time and space. Actual population changes with time and over locations, so we have $p\,(x_1, x_2, t)$. The Malthusian growth (today we would say logistic) is simply $\frac{dp}{dt} = \gamma\,(s-p)\,p$, and it has a closed form solution $p = \frac{s}{1+ce^{-\gamma t}}$ easily obtainable through variable separation and integration by parts.

As for the dispersion, consider that space is different from time. Of course, geographical space has two dimensions, time only one, but this is not the point. More fundamental is that time has a forward direction. In contrast, even in a one dimensional space, left and right are equivalent. So, if we have to measure the difference of a variable, such as p, in a point with that in its neighbourhood, we have one difference to the right, measured by the derivative $\left.\frac{dp}{dx}\right|_+$, and another to the left, measured by the derivative $\left.\frac{dp}{dx}\right|_-$. If $\left.\frac{dp}{dx}\right|_+ > 0$ the value of p increases as we move to the right from the point, if $\left.\frac{dp}{dx}\right|_- > 0$, the value, however, decreases as we move to the left. To get the net effect for moving away from a point we have to form the difference $\left.\frac{dp}{dx}\right|_+ - \left.\frac{dp}{dx}\right|_-$. In the limit of infinitesimal moves this difference goes to the second derivative $\frac{d^2p}{dt^2}$. Hence the second derivative, which is independent of reversal of the directions, is the lowest of relevance for linear spatial processes.

Again, we can treat the two space dimensions separately in terms of a box and form

$$\nabla^2 p = \frac{\partial^2 p}{\partial x_1^2} + \frac{\partial^2 p}{\partial x_2^2} \tag{14}$$

This is the Laplacian, and it measures the difference of the value of a variable in a point with its values in all its neighbourhood. If it is positive there is a net increase if we leave the point, if negative we deal with a net decrease. This Laplacian turns up in all physical processes involving diffusion, such as the heat equation. Incidentally it is the composite of two differential operators we already introduced, the gradient and the divergence, i.e., $\nabla^2 p = \nabla \cdot \nabla p$. Taking the gradient makes a vector field out of the scalar field of p, then applying the divergence makes it a scalar field anew. Anyhow, there are again stringent explanations in terms of Gauss's or Green's theorem.

Using this operator both Hotelling and Skellam formulated the combined system

$$\frac{dp}{dt} = \gamma\,(s-p)\,p + \delta \nabla^2 p \tag{15}$$

Unfortunately it has no closed form solution. Equations of this type are currently much studied in connection to chemical applications, and go under the heading of reaction-diffusion equations.

3.1 Model with Production

To see something more of possible solutions, we end this section with a related system, proposed by the present author, which, though more complicated, in fact has a solution. It also adds some more realism to the growth part of the model.

Humans are not solely dependent on the subsistence means given by nature, but produce their own means of subsistence. Suppose we have a production function with increasing/decreasing returns.

$$q = 2\alpha\beta p^2 - \alpha p^3$$

Here α is a technical efficiency parameter, whereas β is a measure for "optimal scale", the point where production switches from increasing to decreasing returns.

Then rephrasing the growth part of the Hotelling-Skellam equation

$$\frac{dp}{dt} = \gamma\left(\frac{q}{p} - \omega\right)p + \delta\nabla^2 p$$

so that per capita production triggers birth, whereas ω is a given death rate, we obtain

$$\frac{dp}{dt} = \gamma\left(2\alpha\beta p - \alpha p^2 - \omega\right)p + \delta\nabla^2 p$$

Next define two new compound parameters

$$a = \beta - \sqrt{\beta^2 - \frac{\omega}{\alpha}}, \quad b = \beta + \sqrt{\beta^2 - \frac{\omega}{\alpha}}$$

and rescale the measurement units of time and space so that $\gamma = \delta = 1$. Then factorizing:

$$\frac{dp}{dt} = (b - p)(p - a)p + \nabla^2 p$$

This equation has a closed form solution:

$$p = \frac{b}{1 + e^{-\frac{b}{\sqrt{2}}\left(x_1 + x_2 + \sqrt{2}\left(\frac{b}{2} - a\right)\right)}} \qquad (16)$$

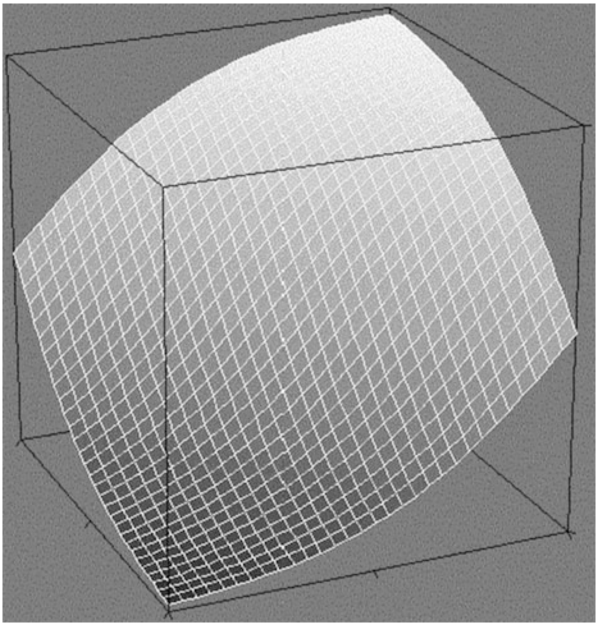


Fig. 6 Example of a stationary population wave with densely and sparsely populated areas. It can also be taken as a momentary picture of a moving wave that eventually goes to a homogenous state with either maximum sustainable or extinct population

For $a = \frac{b}{2}$ we get a standing wave as illustrated in Fig. 6 with regions of high and low population density. Low local population growth is compensated by immigration, and high growth by emigration. When the wave moves, population density can go to maximum or extinction.

4 Diffusion of Business Cycles

Business cycle models such as the celebrated Samuelson 1939 model (Samuelson 1939) were usually set in discrete time. This was natural at the time because in the wake of Keyensian macroeconomics national accounting using periodized empirical data emerged. Before that national income was just an abstraction, defined as current interest on national wealth, itself defined as the discounted value of all expected future incomes for all agents belonging to the economy. This was the idea of, for instance, Erik Lindahl 1939 who departed from some Wicksellian ideas.

On the other hand economic growth as suggested by Sir Roy Harrod in 1948 was modelled in continuous time, and remained so. Growth remained a pure abstraction, as exponential growth can never be fitted to anything real. It is curious that the "balanced growth" rate defined by Harrod was considered unstable by himself. Something, his many followers forgot.

What happened if the actual growth rate deviated from the balanced rate, was, however, modelled by Phillips a few years later in 1954 (Phillips 1954). This is the model we will start out from. So we stick to continuous time, as it makes no harm in theoretical digression. Though we make some amendments. First, Phillips, like Harrod and Samuelson, proposed linear models. Harrod's was first order, producing just exponential growth, though unstable. Phillips and Samuelson proposed second order equations, producing cycles. But also those were linear and so they could not produce sustained bounded oscillation. Either all movement died out, or the oscillations grew in amplitude beyond any limit—even in the negative.

In 1950 Sir John Hicks (Hicks 1950) suggested bounds to the oscillations, the "floor" and the "ceiling", for the Samuelson model making it nonlinear. This was never done for the Phillips model, only in 1989 the present author suggested an accelerator with a cubic nonlinearity (Puu 1989). We will keep to this case.

Another change is that we put the model in a spatial context, which was never done for any of the models discussed above. As a matter of fact Metzler (1949) proposed interregional trade models where local income triggered imports and exterior income triggered exports. Export surplus would thus be proportional to the difference of income in the surrounding of a location and in the location itself. In continuous space this seems to be a perfect case for using the Laplacian, discussed above, in fact much better than in the Hotelling model for population dispersal. So, we also reformulate the Phillips model for an open economy with interregional exports and imports.

Recall some simple facts from national (or, rather, regional) accounting. Income denoted Y, is generated by investment and consumption expenditures, $Y = I + C$. It is disposed on consumption C and a residual saving S, so $Y = S + C$, and, $S = I$. According to legend in capital theory, in the primitive society, saving, i.e., abstaining from immediate consumption, and investment represented the same act. In modern society this is no longer so, the producing sector invests in capital, whereas the consumers provide space for it through abstaining from consumption. The equality $S = I$ has been considered as an equilibrium condition, or an accounting identity, in which case savings also include unintentional saving due to shortage of goods, and investments include unintended inventory investments of unsold goods, as these by definition are considered investments.

As for the determinants of the items, a simple assumption always used in macroeconomics takes consumption as a given fraction c of income, $C = cY$, so $S = (1 - c)\,Y$. This is called multiplier theory, because if we take investments as given, $Y = cY + I$ can be solved for $Y = \frac{1}{1-c}I$. As $c < 1$, the multiplier $\frac{1}{1-c} > 1$ "multiplies up" investments or any increment of those.

As for investments, one needs to consider that they are defined as *changes* in capital stock K. As goods that go in the income are produced, one needs some simple assumption about production and the services of capital that are used in the process. The simplest of such assumptions is, again, to assume proportionality, i.e., to put $K = aY$. This, however, concerns the capital stock needed—it is by no means sure that actual capital stock at any time equals this. As investments are the change of capital stock, $I = a\triangle Y$, where $\triangle Y$ denotes the increment of income. In discrete time we would measure the increment as a difference of incomes in two subsequent periods, in continuous time we would use the time derivative, $I = a\frac{dY}{dt}$. As investments are part in the formation of income, but themselves depend on the rate of change of income, this "principle of acceleration" tends to accelerate changes in the system. Hence the name.

These elements were what Harrod used $I = S$, or $a\frac{dY}{dt} = (1-c)\,Y$, which provided his first order differential equation with its simple solution $Y = Ae^{\frac{1-c}{a}t}$ for the (unstable) balanced growth rate.

To produce an oscillatory system in discrete time, Samuelson used delayed action. In continuous time Phillips used adaptation, which amounts to the same.

4.1 The Phillips Model

First, income is not always equal to the sum of investments and consumption—it increases in proportion to the difference of this sum and actual income, $\frac{dY}{dt} \propto I + C - Y = I - cY - Y$. For simplicity we delete the adaptation coefficient and put

$$\frac{dY}{dt} = I - (1-c)\,Y \tag{17}$$

As for investments, Phillips assumes that investments are not always equal to what they should be $a\frac{dY}{dt}$, they are only adjusted to the extent actual investments fall short of this value, i.e.,

$$\frac{dI}{dt} = a\frac{dY}{dt} - I \tag{18}$$

One can now differentiate (17) once more, substitute from the (18) for $\frac{dI}{dt}$, and use (17) as it stands to eliminate I, so getting the reduced form:

$$\frac{d^2Y}{dt^2} - (a+c-2)\,\frac{dY}{dt} + (1-c)\,Y = 0 \tag{19}$$

with solution

$$Y = Ae^{\alpha t}\cos(\omega t + \varphi)$$

where

$$\alpha = \frac{1}{2}(a+c)\,, \quad \omega = \frac{1}{2}\sqrt{4a-(a+c)^2}$$

and A, φ are arbitrary, determined by initial conditions.

This is the Phillips model of 1948 (Phillips 1954). But it is still linear so it can only produce oscillations that either die out or explode.

4.2 Cubic Nonlinearity

Now, suppose we replace the linear accelerator $I = a\frac{dY}{dt}$ by $I = a\frac{dY}{dt} - a\left(\frac{dY}{dt}\right)^3$, as suggested by the present author in 1989 (Puu 1989). The argument runs along the lines outlined by Sir John Hicks for the discrete time model. Investments can become negative, i.e., disinvestments, which happens when income decreases so fast that there is no point in replacing all the depreciating capital. However, there is a lower limit, a floor, to the negative disinvestment. If income during a depression decreases so fast that capital stock could be reduced by more than its natural depreciation, then nobody would actively destroy working capital. It is better to leave it idle in expectation of the next upswing. Likewise, there is a ceiling; if income increases so fast that manpower or raw materials become short there is no point in building up excess capacity in terms of capital. Hicks imposes the upper limit, the ceiling on income and not on investment, but it seems even better to apply it to investment as it is an issue for decisions, whereas income is just an outcome of the process.

This accounts for the limits. There is also a good argument for the actual bending in the opposite of the cubic. The public sector would always act contracyclically in terms of policy. Further, there are always long run investment projects in infrastructure and the like, both in the private and the public sector which could better be carried out in depression when resources are underutilized and therefore cheap than in an overheated prosperity period.

With this linear/cubic accelerator the Phillips equation is converted to:

$$\frac{d^2Y}{dt^2}\,(1-c)\,Y-(a+c-2)\,\frac{dY}{dt}+a\left(\frac{dY}{dt}\right)^3=0$$

It has no longer any simple closed form solution, but it closely relates to physical models much studied by Rayleigh (1894).

4.3 Interregional Trade Added

We will not make any study of the nonlinear equation before we also introduced space. Therefore we first have to reformulate the income equation so that it suits an open economy with exports and imports. It now reads $Y = I + C + X - M$, where X denotes exports and M imports. Metzler (1949) suggested the intuitively reasonable assumption that exports are proportional to income in the exterior, imports to income in the interior. We already have the Laplacian as measure for interregional income differences, so, assuming a constant propensity to import, we get the new building block $X - M = m\nabla^2 Y$. As this export surplus enters quite as investments, we can assume exactly the same delayed adaptation mechanism, and so finally arrive at the equation

$$\frac{d^2Y}{dt^2} + (1-c)\,Y - (a+c-2)\,\frac{dY}{dt} - a\left(\frac{dY}{dt}\right)^3 - m\nabla^2 Y = 0 \tag{20}$$

where

$$\nabla^2 Y = \frac{\partial^2 Y}{\partial x_1^2} + \frac{\partial^2 Y}{\partial x_2^2}$$

It does not seem so easy to solve, but intuitively we can find an obvious solution for $Y(x_1, x_2, t)$. If the sides of this surface at any location and any time is flat, then $\nabla^2 Y = 0$. Further, if income grows or decreases at constant rates always and everywhere, also $\frac{d^2Y}{dt^2} = 0$. Finally, if the constant growth rate $\frac{dY}{dt}$ is zero, or equals $\pm\sqrt{\frac{a+c-2}{a}}$ then the PDE is identically satisfied everywhere and at all times. It remains to piece the assumptions

$$\frac{d^2Y}{dt^2} = 0, \quad \nabla^2 Y = 0, \quad \frac{dY}{dt} = \begin{cases} -\sqrt{\frac{a+c-2}{a}} \\ 0 \\ +\sqrt{\frac{a+c-2}{a}} \end{cases} \tag{21}$$

together.

The result is shown in Fig. 7. The oscillating waves are pyramids that are erected, erased, reversed and then erected again, over and over in a never ending process. The picture shows four pyramids on a square, two in the process of erection, two in excavation at a frozen point of time. This definitely is a solution, at least for the case where boundary conditions prescribe that income is always at rest on the edges of the square. But the same solution could give us only one, four as in the picture, or sixteen small pyramids.

The question is if they are attractive. All the subdivisions coexist, and numerics indicate that the system goes to one pattern of the type according to initial conditions. If we start with a smooth sinusoid wave with sixteen extrema, then the pattern seems to be attracted to the pyramid solution with sixteen bases, if we start with just one smooth wave, we get just one pyramid.

Myerscough (1973, 1975) investigated the problem for the simpler case of one space dimension, modelling oscillations in overhead lines subject to a strong side wind. He could actually prove that such saw-toothed oscillating patterns were attractors. However, we just chose the square region for convenience, which by no means

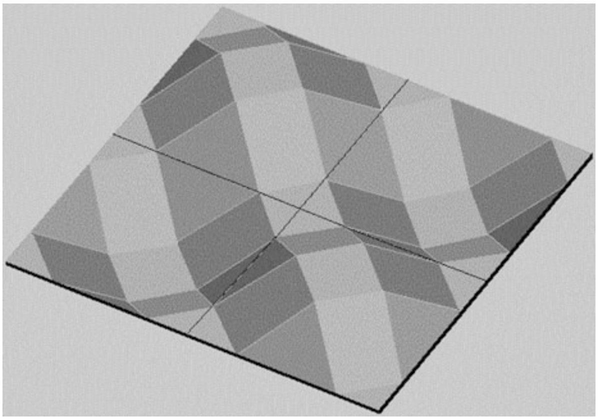


Fig. 7 Momentary picture of pyramid shaped waves

is as evident as fixing a line to two posts, so creating a fixed interval in 1D. A good reminder that for PDE the boundary conditions are essential parts of the problem.

5 Market Areas

The shape of market areas has interested economists and geographers alike. An ideal shape would be circular, but circles cannot tessellate the plane without overlapping or empty corners, Among regular shapes only equilateral triangles, squares, and hexagons can. Walter Christaller in 1933 (Christaller 1933) observed that market areas in Bavaria had approximately hexagonal shapes, and August Lösch in 1940 (Lösch 1940) presented a thorough analysis for nested hexagonal market areas. It is, by the way, curious that we still speak of the Christaller-Lösch theory as if they had collaborated. In truth Christaller got the printing of his little thesis sponsored by SS Reichsführer Heinrich Himmler, whereas Lösch refused oath of fidelity to the Führer, and therefore was denied the “venia legendi” (permit to teach). Lösch lived on the verge of starvation, and died from a trivial infection by the end of the war, whereas Christaller immediately upon the Soviet victory joined the communist party.

5.1 Hexagonality

In fact, the hexagon among the three tessellations is the most compact. It has been known to solve the isoperimetric problem, having the shortest boundary for a cell of given area. Charles Darwin was astonished at the inherited wisdom of the bees who built hive cells of hexagonal crosssection (rather than square or triangular), and took it as the ultimate proof for survival of the fittest, as natural selection had chosen those bees who could best economize with wax and labour.

The case for market areas is a bit different (and more complicated) than the isoperimetric problem. It concerns average or total distance to all points of the enclosure from its centre. The total distance from the centre to all points of an n-gon of unit area is

$$\frac{1 + \cot\frac{\pi}{n} \ln\tan\left(\frac{\pi}{4} + \frac{\pi}{2n}\right)}{3\sqrt{n \cos\frac{\pi}{n} \sin\frac{\pi}{n}}} \tag{22}$$

We get 0.4036 for the triangle, 0.3826 for the square, and 0.3761 for the hexagon. So, the hexagon is indeed the most economical shape, but savings of transport distance (=cost) are only 1.43 % as compared to the squares. Any change of location structure involves considerable relocation costs, so, if market areas indeed are hexagonal, can this tiny gain on transportation cost explain a transit from squares to hexagons?

Hexagons abound in all kinds of natural structures. Among the most curious cases is the skeleton of Aulonia Hexagona described by d'Arcy Wentworth-Thompson in his remarkable book 1917 (Thompson 1917), a spherical radiolarian seemingly made up of hexagonal cells. However, it is known that for mathematical reasons a sphere cannot be paved by hexagons, as this would result in a wrong topological genus number. Yet one has to look very carefully at the radiolarian to discover a few pentagons. As an Israeli geographer once put it; "there must be at least one pentagon on the sphere".

Hence, one must find a better reason for the abundance of hexagons than optimality of shape. In the case of market areas, we can easily offer one.

A tessellation has not only the characteristic of shape, but also of the number of cells meeting in a corner: Three for hexagons, four for squares, six for triangles. Maybe this characteristic is more relevant than shape?

5.2 Launhardt Funnels

Therefore, consider how Wilhelm Launhardt in 1885 once explained the emergence of market areas. Referring to Fig. 8, consider the 2D geographical plane. Three firms are located. The vertical sticks have the height of the mill prices charged. Local delivered prices are these mill prices accrued by transportation costs. If transportation costs are independent of location and the same for all competitors, then delivered prices generate surfaces that take conical shapes. One speaks of Launhardt "funnels". Consumers choose the lowest price represented by the lowest surface at their location.

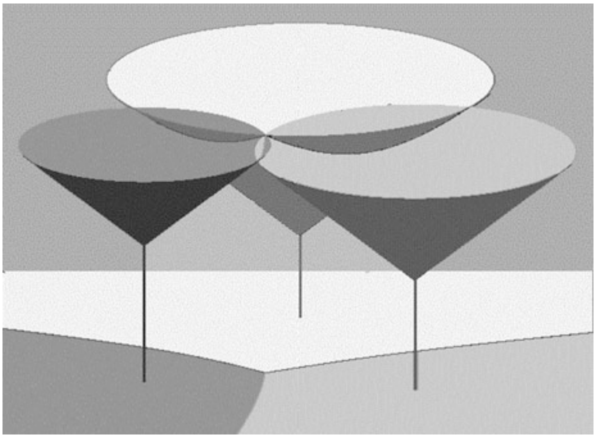


Fig. 8 Three Launhardt "funnels" for delivered price, with the intersection curves projected down on the base plane to provide market boundaries

Two such price surfaces intersect along a curve, which defines the market boundary, separating market areas. As there are three firms in the picture, we get three such boundary curves, which intersect in a point.

The crucial question is, can we add a fourth firm so that its delivered price surface intersects the same point as the three already existent? For an answer, sure we can add one more firm, but the intersection is extremely unlikely. Two surfaces intersect along a curve, three in a point, but the intersection of four would be an exceptional and unlikely occurrence.

All this is regulated formally by the principle of transversality. It states that the sum of dimensions of the intersecting manifolds must equal the sum of the dimensions of the intersection manifold and the embedding space. For example:

$2 + 2 = 1 + 3$, two surfaces intersect along a curve in space
$2 + 1 = 0 + 3$, a third surface intersects this curve in a point
$2 + 0 = ? + 3$, no way for a fourth surface to intersect a point

So, recall that hexagons intersect three by three, squares four by four. Accordingly it is the meeting of three market areas in a point, not four, that is transverse, or structurally stable. If a hexagonal tessellation is subject to some disturbance, is perturbed, then it remains hexagonal. For a square tessellation, on the contrary, every tiny perturbation makes a meeting of four areas split into two disjoint intersections of three. It is therefore structurally unstable.

On a map of national states, we only see four country incidence where the map has been drawn by an administrator with pen and ruler, never where national state boundaries are results of wars and negotiations. Then the number of countries meeting is always three. Of course, the shape of national states is not exactly hexagonal, as they have so different sizes, and as there are natural barriers in terms of oceans, but the principle applies.

6 Topology of the Beckmann Model

Now that we got acquainted with topological issues, it is time to return to the Beckmann model, because structural stability and transversality have something to say also about this case, though the results are quite different from those for the case of market areas, in favour of square rather than hexagonal shapes. It depends on what we apply structural stability to. For market areas it was to the incidence of market areas, in the Beckmann case we apply it to the flow structure.

A model such as Beckmann's, by its very elegance and generality tends to be a final statement, quite as the case of general economic equilibrium. Examples can be supplied, but it is difficult to say anything more.

6.1 Structural Stability

Fortunately, the present author came across some most useful results due to Mexican mathematician M.M. Peixoto dating from 1977 (Peixoto 1977). His remarkable contribution makes it possible to characterize structurally stable flows associated with ordinary differential equations, such as those of the Beckmann model.

Recall that once we solved the partial differential equation for $\lambda\left(x_1, x_2\right)$, we could derive the ordinary differential equations for the transportation routes $x_1\left(s\right), x_2\left(s\right)$, where s is just an arc length parameter, defined as $s = \int\sqrt{dx_1^2 + dx_2^2}$. They were stated as

$$\left.\begin{array}{c}\frac{dx_1}{ds}\\ \frac{dx_2}{ds}\end{array}\right\} = \left\{\begin{array}{c}\frac{\partial\lambda}{\partial x_1}\\ \frac{\partial\lambda}{\partial x_2}\end{array}\right.$$

or $\frac{dx_1}{ds} = \frac{\partial\lambda(x_1,x_2)}{\partial x_1}$, $\frac{dx_2}{ds} = \frac{\partial\lambda(x_1,x_2)}{\partial x_2}$. As the right hand sides are known, they are a pair of coupled ordinary differential equations which can be solved provided they are simple enough.

We will not try to actually solve them, but to get some information about the global portrait of these flow lines.

By the theorem for existence and uniqueness of solutions to ODE, there is one and only one solution curve through each point, unless both right hand sides vanish. When they do we have singular points, something awkward in older differential calculus, but the most interesting points in the present context. After all, regular solutions only provide non-intersecting flow lines, which are topologically equivalent to parallel straight lines, and therefore trivial.

We get a global picture, a phase portrait, of the flow of trade through only considering these singular points and the trajectories joining them.

Now Peixoto's remarkable theorem of 1977 (Peixoto 1977) states that, for a structurally stable flow,

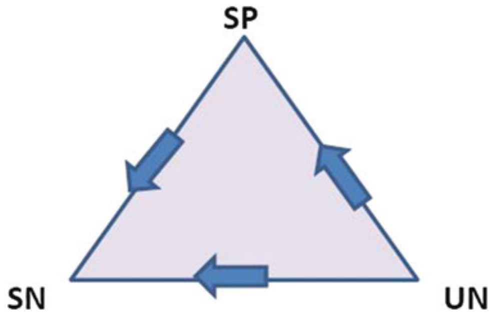


Fig. 9 Elementary triangle in a stable flow with stable node, unstable node and saddle point as vertices, and the connecting trajectories with flow directions indicated. Note that to the stable node there is only inflow, and from the unstable node only outflow. The saddle has one ingoing and one outgoing trajectory

(i) There are only a finite number of isolated singular points.
(ii) These are stable nodes SN, unstable nodes UN, or saddle points SP.
(iii) There are no heteroclinic (or homoclinic) saddle connections, i.e. such that go out from the unstable manifold of a saddle and in to the stable manifold of another (or the same) saddle.

So, consider the smallest possible building blocks of the phase portrait spoken of, a triangle with three corners and three connecting edges. They can only look as in Fig. 9. This is because there cannot be two nodes of the same type, as in a triangle they would be connected by an arc, and on this arc the direction of flow could not be oriented in a consistent way. There cannot be two saddles either, because Peixoto's theorem forbids this. Remains only one possibility, a singular point of each kind, SN, UN, and SP.

If we like in the case of market areas want organize regular tessellation elements, we can organize the elementary triangles cyclically, six for a triangle (Fig. 10), eight for a square (Fig. 11), and twelve for a hexagon (Fig. 12).

Any of these can be fitted together to form a regular tessellation. We should, however, note one fact. Using the triangle and the hexagon result in the *same* type of flow pattern. There are twice as many hexagons as triangles or the other way around. So, in fact, these result in only one case, a mixed triangular-hexagonal tessellation. Also the number of unstable nodes is twice or half the number of stable nodes. However, a pure hexagonal tessellation does not occur, as it is an unstable structure and violates Peixoto's conditions.

The other stable type of tessellation is the square, where the numbers of stable and unstable nodes are equal. We display this case in Fig. 13. In the picture we also filled in a couple of typical regular (non-intersecting) trajectories for the trade flow, and the orthogonal curves marking the constant price contours. The empty circles represent one kind of nodes, the filled the other. The saddles are not marked, but are located at the corners of the square patterns, towards which the trajectories seem to be attracted.

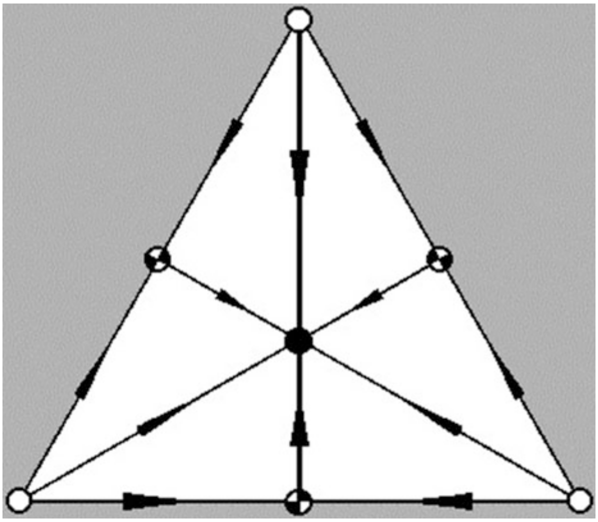


Fig. 10 Six elementary triangles arranged cyclically to create a triangular tessellation element

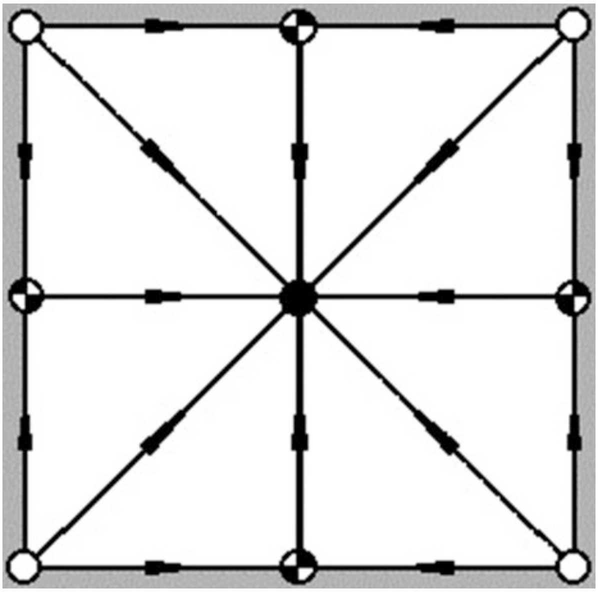


Fig. 11 Eight elementary triangles arranged cyclically to create a square tessellation element

6.2 Structure Change

To finish, as there are two stable regular tessellations, one may want to work out the transitions between them. This can be done using the elliptic umblic catastrophe as defined by René Thom 1972 (Thom 1975). We describe the transitions in two

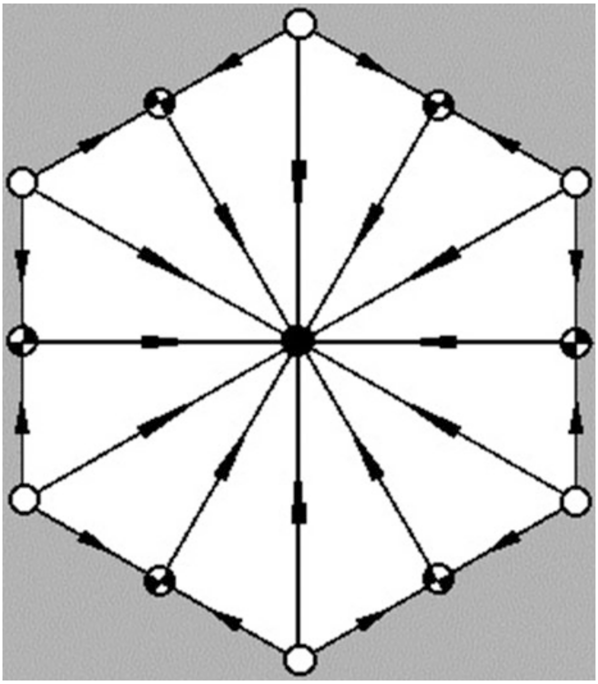


Fig. 12 Twelve elementary triangles arranged cyclically to create a hexagonal tessellation element

pictures, Fig. 14 displaying the catastrophe surface in parameter space, a, b, c as they appear in Thom's "canonical" form:

$$\lambda = x_1^3 - 3x_1x_2^2 + ax_1 + bx_2 + c\left(x_1^2 + x_2^2\right) \tag{23}$$

Figure 15 shows in five small medallions the types of change we encounter in x_1, x_2 phase space as we cross the catastrophe surface. The thin "waist" in parameter space and the central medallion represent just the leading terms $x_1^3 - 3x_1x_2^2$, a, b, c all being zero. This corresponds to the pure hexagonal pattern, which is the most unstable case we can find. Passing this point in the horizontal direction we see the transition from one triangular/hexagonal case to the other hexagonal/triangular case (passing the extremely unstable "monkey saddle" pattern). Passing it in the vertical, we see a transformation of one square tessellation to another.

Things may relate better to the preceding discussion if we put the leading terms of the catastrophe in a global version such as

$$\lambda = \frac{1}{2}\sin\left(x_1 + \sqrt{3}x_2\right)\sin\left(x_1 - \sqrt{3}x_2\right)\sin\left(2x_1\right) \tag{24}$$

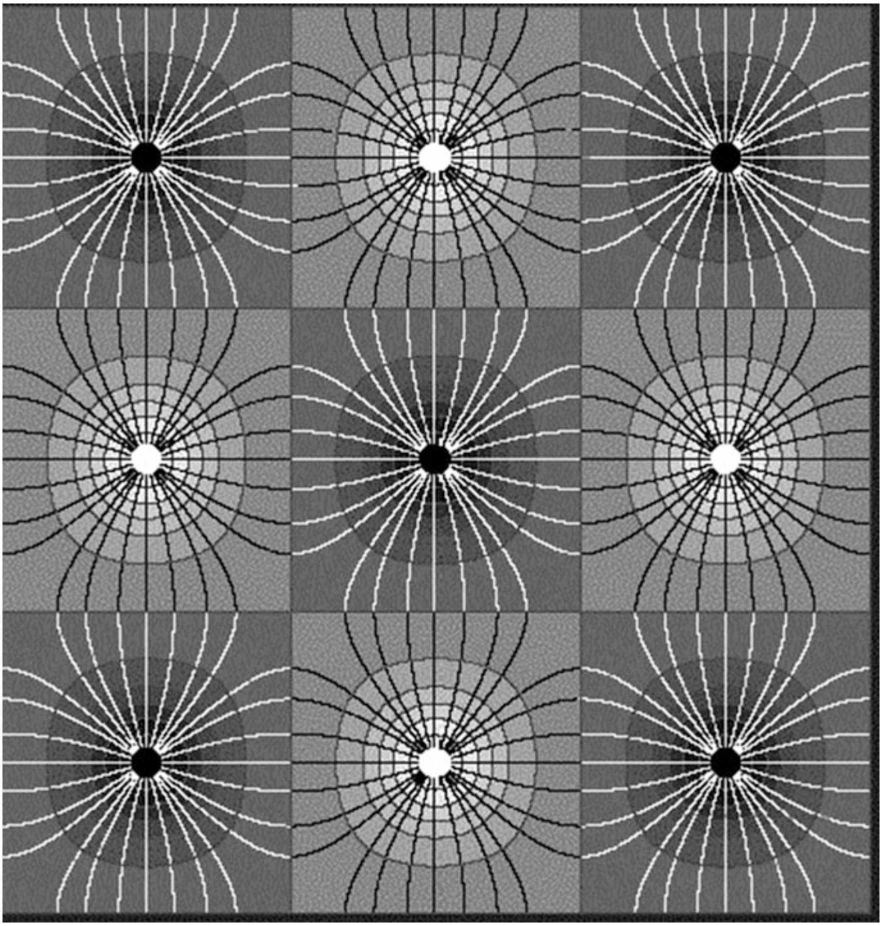


Fig. 13 Piece of a global square flow/price contour pattern. The price curves are almost *circular* around the nodes, but deform to *squares* when the saddle points are approached. The orthogonal flow lines start out and end almost radially, but they have intermediate parts that are attracted to the saddle points

of which Thom's normal form is a Taylor expansion. At every zero for λ in x_1, x_2 phase space, the Taylor series is

$$\lambda = x_1^3 - 3x_1x_2^2 + o\,(5)$$

We do not need to be concerned about the higher order terms. The perturbations, of first and second order multiplied by perturbation parameters a, b, c are essential. They can be obtained by other periodic functions producing first and second order terms with appropriate local Taylor expansions, but we chose to not mess things up through writing these down.

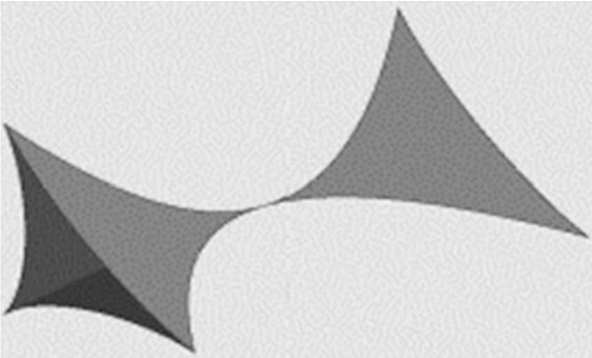


Fig. 14 Elliptic umblic catastrophe surface in parameter space

Fig. 15 The corresponding different flow patterns in phase space. Unstable monkey saddle in the *middle*, surrounded by square tessellation medallions *up* and *down*, hexagonal/triangular *right* and *left*

The global changes in phase space obtained through variations of a, b, c resulting in the vertical and horizontal passages described, can now be displayed in Figs. 16 and 17.

Note that all these characterisations are topological, they apply to all distorted cases obtained by stretching without cutting or gluing. Further, note that the examples use regular patterns; of course they can be combined. A very detailed account of structural stability applied to the stable patterns for the Beckmann model and their transitions can be found in Puu (1993).

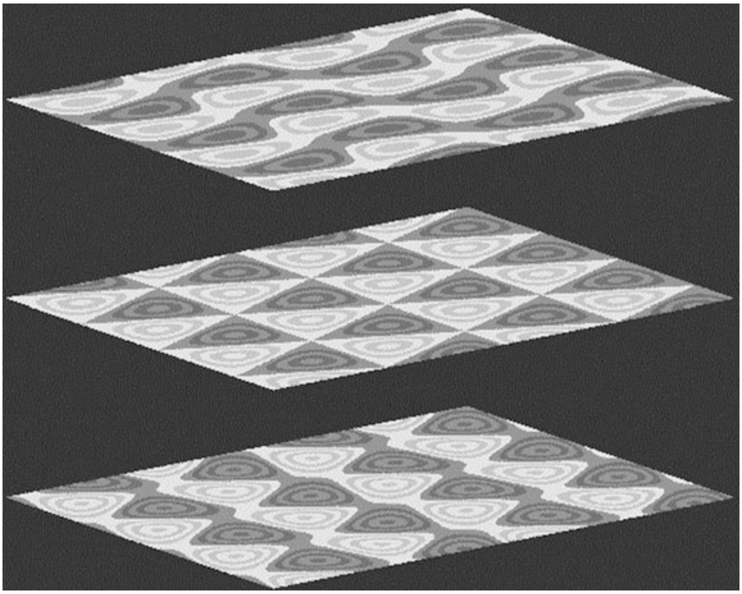


Fig. 16 Global transformation of square tessellation through the monkey saddle case

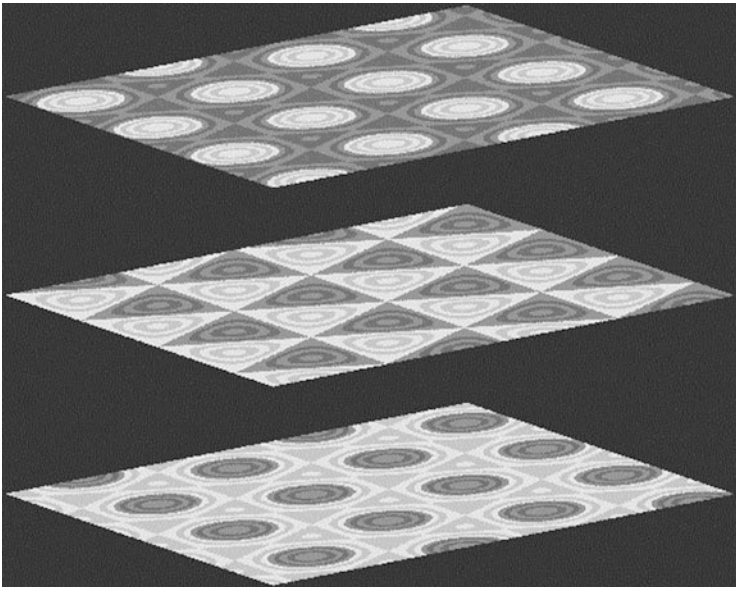


Fig. 17 Global hexagonal/triangular tessellation passing through the monkey saddle case

Acknowledgements This work has been supported by the grants MTM2014-52920-P from Ministerio de Economía y Competitividad (Spain), and COST Action IS1104 "The EU in the new complex geography of economic systems: models, tools and policy evaluation". Finally "este trabajo es resultado del proyecto de investigación 19294/PI/14 financiado por la Fundación Séneca-Agencia de Ciencia y Tecnología de la Región de Murcia en el marco del PCTIRM 2011–2014".

References

Beckmann MJ (1952) A continuous model of transportation. Econometrica 20:642–660
Beckmann MJ (1976) Equilibrium and stability in a continuous space market. Oper Res Veifahren 14:48–63
Beckmann MJ, Puu T (1985) Spatial economics. North-Holland, Amsterdam
Beckmann MJ, Puu T (1990) Spatial structures. Springer, Berlin
Christaller W (1933) Die zentralen Orte in Süddeutschland. Fischer, Jena
Harrod RF (1948) Towards a dynamic economics. Macmillan, London
Hicks JR (1950) A contribution to the theory of the trade cycle. Oxford University Press, Oxford
Hotelling H (1921) A mathematical theory of migration. MA thesis, University of Washington. Reprinted in Environ Plan A 10:1223–1239
Launhardt W (1885) Mathematische Begründung der Volkswirtschaftslehre. Teubner, Leipzig
Lindahl E (1939) Studies in the theory of money and capital. Allen and Unwin Ltd., London
Lösch A (1940) Die räumliche Ordnung der WirtschaJt (English translation: The Economics of Location, Yale University Press 1954). Fischer, Jena
Metzler LA (1949) Tariffs, the terms of trade, and the distribution of national income. J Polit Econ 57:1–29
Myerscough CJ (1973) A simple model of the growth of wind-induced oscillations in overhead lines. J Sound Vib 28:699–713
Myerscough CJ (1975) Further studies of the growth of wind-induced oscillations in overhead lines. J Sound Vib 39:503–517
Peixoto MM (1977) Generic properties of ordinary differential equations. MAA Stud Math 14:52–92
Phillips AW (1954) Stabilization policy in a closed economy. Econ J 64:290–323
Puu T (1989) Nonlinear economic dynamics. Springer, Berlin
Puu T (1993) Pattern formation in spatial economics. Chaos Solitons Fractals 3:99–129
Puu T (1997/2002) Mathematical location and land-use theory. Springer, Berlin
Rayleigh L (1894) The theory of sound. MacMillan, London
Samuelson PA (1939) Interactions between the multiplier analysis and the principle of acceleration. Rev Econ Stat 21:75–81
Samuelson PA (1952) Spatial price equilibrium and linear programming. Am Econ Rev 42:283–303
Schumpeter JA (1954) History of economic analysis. Oxford University Press, Oxford
Skellam JG (1951) Random dispersal in theoretical populations. Biometrika 38:196–218
Thom R (1975) Structural stability and morphogenesis. Benjamin, Reading, MA
Thompson DW (1917) On growth and form. Cambridge University Press, Cambridge. Dover reprint of 2nd edn., 1942
von Thunen H (1826) Der isolierte Staat in Beziehung auf LandwirtschaJt und Nationalokonomie. Perthes, Hamburg
Wilson A (1970) Entropy in urban and regional modelling. Pion Ltd., London