Universitat
de les Illes Balears

DOCTORAL THESIS
2024

AGING AND MEMORY EFFECTS IN SOCIAL AND ECONOMIC DYNAMICS

David Abella Bujalance

Universitat de les Illes Balears

DOCTORAL THESIS
2024
Doctoral programme in Physics

AGING AND MEMORY EFFECTS IN SOCIAL AND ECONOMIC DYNAMICS

David Abella Bujalance

Thesis Supervisor: José Javier Ramasco Sukia
Thesis Supervisor: Maxi San Miguel
Thesis Tutor: Cristóbal López Sánchez

Doctor by the Universitat de les Illes Balears

Supervisors:
José Javier Ramasco Sukia
Maxi San Miguel

A en Manuel Miranda
pel seu suport i ajuda
durant tots aquests anys.
Sempre estaràs amb mi.
i recordare sempre
el que em vas ensenyar.

Dr José Javier Ramasco of the Consejo Superior de Investigaciones Científicas (CSIC) and Dr Maxi San Miguel of the Universitat de les Illes Balears (UIB)

WE DECLARE:

That the thesis titles *Dynamics of social interactions*, presented by David Abella Bujalance to obtain a doctoral degree, has been completed under my supervision and meets the requirements to opt for an International Doctorate.

For all intents and purposes, I hereby sign this document.

Signature

Dr. José Javier Ramasco Sukia	Dr. Maxi San Miguel
Thesis Supervisor	Thesis Supervisor

Palma de Mallorca, June 2024

Acknowledgements

M'agradaria agrair aquesta tesi a totes les persones que m'han ajudat a fer-la possible. En primer lloc, vull agrair a la meva família, per tot el suport que m'han donat durant tots aquests anys. En especial, vull agrair a la meva mare, per tot el que ha fet per mi, i per tot el que ha hagut de patir per mi. També vull agrair a la meva parella, per tot el suport que m'ha donat, i per tot el que m'ha ajudat a tirar endavant. I finalment, vull agrair a tots els meus amics, per tot el suport que m'han donat, i per tots els bons moments que hem passat junts.

Des d'un primer moment, vull agrair a la meva directora, la professora Marta Arias, per haver-me donat l'oportunitat de fer aquest projecte, i per tot el suport que m'ha donat durant tot el projecte. També vull agrair al meu tutor, el professor Jordi Casas, per tot el suport que m'ha donat durant tot el projecte. I finalment, vull agrair a tots els professors que m'han ensenyat durant tots aquests anys, per tot el que m'han ensenyat, i per tot el que m'han ajudat a tirar endavant.

Tambe afegir que aquest projecte no hagués estat possible sense l'ajuda de tots els companys que han fet possible que aquest projecte sigui una realitat. Jo que soc un dels que ha fet possible que aquest projecte sigui una realitat, vull agrair a tots els companys que han fet possible que aquest projecte sigui una realitat, per tot el suport que m'han donat durant tot el projecte.

Resum

En els sistemes complexos distribuïts, els sistemes de memòria transaccional distribuïda (DTM) són una eina molt útil per a la programació concurrent. Aquests sistemes permeten als desenvolupadors de software escriure codi concurrent sense haver de preocupar-se per la gestió de la memòria compartida. A més, els DTM ofereixen una interfície molt senzilla per a la programació concurrent, ja que permeten als desenvolupadors de software escriure codi concurrent de forma semblant a com ho farien si el codi fos seqüencial. Tot i això, els DTM no són una eina perfecta, ja que tenen un rendiment molt inferior al de les estructures de dades distribuïdes. A més, els DTM no són capaços de gestionar estructures de dades distribuïdes de forma eficient. Per aquest motiu, els DTM no són una eina adequada per a la programació de sistemes distribuïts.

Resumen

En los sistemas complejos distribuidos, los sistemas de memoria transaccional distribuida (DTM) son una herramienta muy útil para la programación concurrente. Estos sistemas permiten a los desarrolladores de software escribir código concurrente sin tener que preocuparse por la gestión de la memoria compartida. Además, los DTM ofrecen una interfaz muy sencilla para la programación concurrente, ya que permiten a los desarrolladores de software escribir código concurrente de forma similar a como lo harían si el código fuera secuencial. Sin embargo, los DTM no son una herramienta perfecta, ya que tienen un rendimiento muy inferior al de las estructuras de datos distribuidas. Además, los DTM no son capaces de gestionar estructuras de datos distribuidas de forma eficiente. Por este motivo, los DTM no son una herramienta adecuada para la programación de sistemas distribuidos.

Abstract

In complex systems distributed transactional memory (DTM) systems are a very useful tool for concurrent programming. These systems allow software developers to write concurrent code without having to worry about managing shared memory. In addition, DTM systems offer a very simple interface for concurrent programming, as they allow software developers to write concurrent code in a similar way to how they would if the code were sequential. However, DTM systems are not a perfect tool, as they have a much lower performance than distributed data structures. In addition, DTM systems are not able to manage distributed data structures efficiently. For this reason, DTM systems are not a suitable tool for programming distributed systems.

Contents

List of Figures

List of publications

The list of articles detailed below, in chronological order by date of publication, form the basis of the present thesis.

1. David Abella, Maxi San Miguel, and José J. Ramasco. "Aging effects in Schelling segregation model". In: *Scientific Reports* 12.1 (Nov. 2022). ISSN: 2045-2322. DOI: 10.1038/s41598-022-23224-7. URL: http://dx.doi.org/10.1038/s41598-022-23224-7

2. David Abella, Maxi San Miguel, and José J. Ramasco. "Aging in binary-state models: The Threshold model for complex contagion". In: *Phys. Rev. E* 107 (2 Feb. 2023), page 024101. DOI: 10.1103/PhysRevE.107.024101. URL: https://link.aps.org/doi/10.1103/PhysRevE.107.024101

3. David Abella et al. "Ordering dynamics and aging in the symmetrical threshold model". In: *New Journal of Physics* 26.1 (Jan. 2024), page 013033. DOI: 10.1088/1367-2630/ad1ad4. URL: https://dx.doi.org/10.1088/1367-2630/ad1ad4

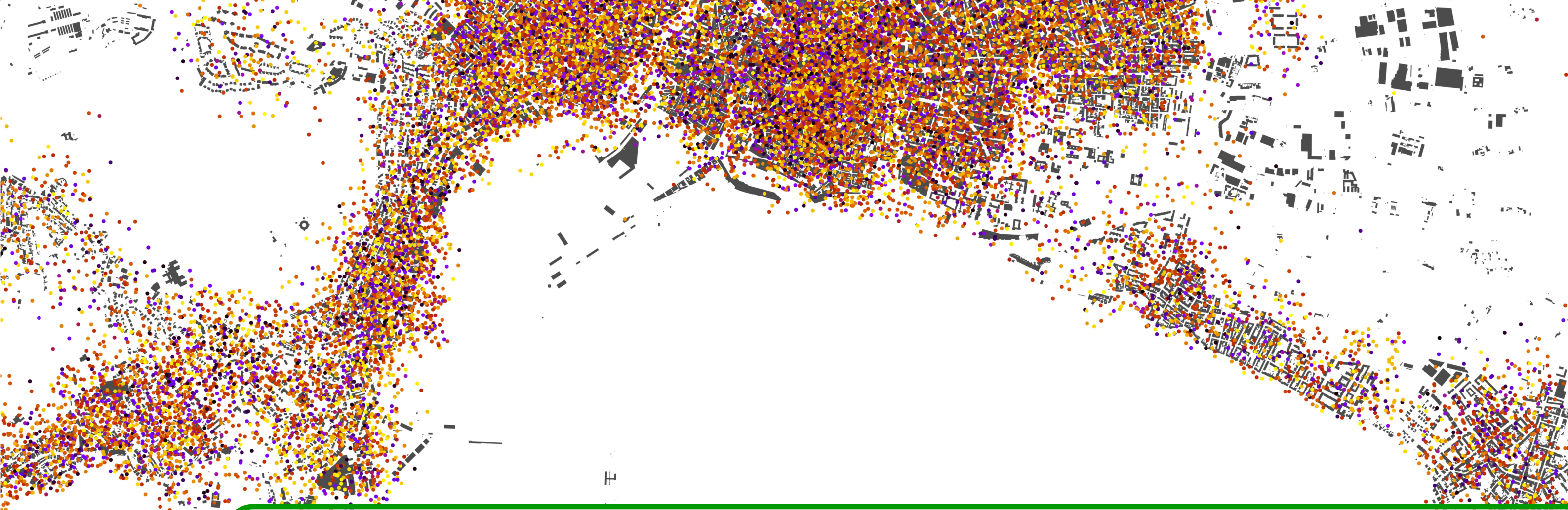
4. Idealista model for complex systems housing

5. Idealista spatial segmentation of the real state market

Other publications published during the PhD period are also included in the following list.

- David Abella, Giancarlo Franzese, and Javier Hernández-Rojas. "Many-Body Contributions in Water Nanoclusters". In: *ACS Nano* 17.3 (Jan. 2023), pages 1959–1964. ISSN: 1936-086X. DOI: 10.1021/acsnano.2c06077. URL: http://dx.doi.org/10.1021/acsnano.2c06077

- David Abella et al. "Unraveling higher-order dynamics in collaboration networks". In: *arXiv preprint arXiv:2306.17521* (2023)



1. Introduction and positioning

This thesis provides a general overview of the research that I have been developing since the beginning of my PhD studies in September, 2021. I could define myself as a curious, creative and open-minded person, following the so called *IFISC attitude*, which means that I am always willing to learn new methods and address new problems, even though they are not directly related to my field of expertise. That is why, through this thesis many topics will be covered, from the study of human behavior and social systems, to the study of complex systems and network theory........

1.1 Scientific Landscape

This thesis adress the study of human behavior and social systems from a *complex systems* perspective, which studies the emergence of collective phenomena that arise from the interactions of many individuals, and that cannot be understood by studying the behavior of individual agents in isolation (the so-called *reductionist* approach) (0). The study of collective phenomena has a long history in the natural sciences, specially in the branch of statistical physics (0). This branch traditionally studies the emergence of collective phenomena in physical systems, such as the phase transitions in magnetic materials via spin models (0), the turbulence in fluids (0), the synchronization in oscillatory systems (0), or percolation (0). However, in recent years, the study of complex systems has evolved into the study of emergent phenomena beyond physical systems, such as biological (0), ecological (0), economic (0), and social systems (0). Even though this branch of physics is relatively young, in 2021, the Nobel Prize in Physics was awarded to Syukuro Manabe, Klaus Hasselmann, and Giorgio Parisi for their contributions to the study of complex systems (0), giving a recognition to this field in the scientific community.

From the migration of birds (0) to the spreading of a fake news through social media (0), there are many examples of collective phenomena beyond the real of physics at which the study of complex systems can be applied. The cascade of failures in power grids (0), the spread of a disease in a population (0), the consensus in political elections (0), the emergence of social norms (0) are some examples of social collective behavior in which the global phenomena cannot be understood by looking at a single individual. Social and economic collective phenomena has been studied from a variety of perspectives (sociology, psychology, economics, political sciences...), which often relies on qualitative methods, such as interviews, surveys, or ethnographic studies (0). However, the study of social systems from a complex systems perspective aims to provide a quantitative framework to understand the collective behavior, based on methodologies from statistical mechanics and network theory (0, 0). Nevertheless, this approach needs for a large amount of detailed data to validate theories and develop models, which historically has been a limitation for the study of social systems. It is in fact surprising how other branches of physics, where the typical scale of the phenomena is very large, as astrophysics, or very small, as particle physics, do not suffer from a lack of data, while the study of social systems, where the typical scale is human, has been historically limited by the lack of data.

Thankfully, the digital revolution has changed this picture, allowing the storage of large amounts of data generated by human activities, such as social media, mobile phones, or online platforms. Nowadays, every two year, more human socio-economic data is produced than during all the preceding years of human history together (0). This data, often referred to as *Big data*, has opened a new era for the study of social systems at a large scale, together with a paradigm shift in the way we understand human behavior (0). Nevertheless, this new era comes with an awareness, as the use of big data for the study of human behavior raises important ethical and privacy concerns, which need to be addressed in order to ensure the responsible use of data for the study of social systems (0). Moreover, from the technical point of view, this huge amount of data needs for a set of computational and mathematical resources to be analyzed and modeled.From this demand, the field of *Computational Social Science* has emerged, with the aim to develop new methods to study human behavior (0). This branch of the complex systems science was born as a combination of methodologies borrowed from social sciences, such as sociology, psychology, or economics, with computational methods from computer science, such as machine learning, data mining, or network theory (0). This interdisciplinary approach has allowed to develop new methods for forecasting social phenomena and understanding the basic mechanism behind human interactions.

One can differentiate two main approaches to build a representation of the reality from the data source. The first one is to focus on the prediction and forecasting of a certain social phenomena, such as the spread of a disease or the price of a stock. In this approach, the data is seen as a necessary input to our methodology to make quantitative predictions (0). However, in this approach, the mechanisms behind the phenomena are often hidden in the data, and the model is seen as a black box that provides accurate predictions (0). In this context, the use of machine learning (0) and deep learning (0) models are very popular, as they are able to capture complex patterns in the data. The second approach is to focus on the understanding of the mechanisms behind the phenomena. In this approach, the data is seen as a problem to be understood, an observation from which we can extract qualitative behaviors and patterns (0). In this context, the aim is to develop very simple models that are able to reproduce the main features of the data, and to extract the basic mechanisms behind the phenomena.

Following the later approach, network science has a critical role in the study of socio-economic systems, as it provides a natural framework to study the interactions between individuals. A network, or graph, is a mathematical representation of a set of nodes (individuals) connected by links (interactions), which allows to study the structure of the interactions and the dynamics of the system. The study of networks has a long history in the natural sciences, from the neurons network in the brain (0) to food webs in an ecosystem (0, 0, 0). However, in recent years, new data sources lead to the discovery that complex properties and heterogeneities, present in most social systems, need for a topological description in terms of a complex network (**CITA EN KARSAIS**). A complex network can be defined is a network that exhibits non-trivial topological properties, which we will explain later on this thesis. These properties are often found in social networks, such as a social media (0), the collaboration network of scientists (0, 0), or international conflicts (0, 0). In particular, the study of information spreading as a dynamical system on networks has allowed to understand how information spreads through a social system and how consensus emerges.

Contagion of information has been a topic of interest for many social scientists. Eearly theoretical frameworks,influenced by psychological and sociological theories, show how individuals in a crowd lose their sense of self and are more susceptible to the ideas and emotions of the crowd (0). Social imitation of behaviors and ideas was proposed as a mechanism for social change, facilitated by close contact and communication among individuals (0). Similarly, peer pressure was also proposed another possible mechanism, where individuals are influenced by their peers to adopt certain behaviors or ideas (0, 0). However, until now these theoretical dissertations were not supported by quantitative results with real data analysis. For example, the spread of innovations of innovations (0), the diffusion of information (0), or the spread of diseases (0) are some examples of social phenomena where we can test if the traditional theoretical frameworks are able to reproduce the data.

Beyond the network structure and the contagion dynamics, there is a third ingredient that is critical for the study of social systems: the human temporal interactions. Human interactions exhibit complex activity patterns that are difficult to understand and to model, since there are a lot of mechanisms that drive the human behavior. For example, the human activity patterns can be driven by the circadian rhythms (**roenneberg-2012**), bursty interactions (0), cascades (0), periodic commuting behaviors (0), recurring patterns in online behavior (0), etc. All this effects need for a proper characterization. MIRAR LO QUE DIUS MARTON.

1.2 Challenges of Computational Social Science

- The study of human behavior and social systems is a complex problem that requires the use of computational methods to study human behavior and social systems.

- There are some challenges that are unique to the study of human behavior and social systems, and that are not present in the study of physical systems.

- The main problem is the data availability, and the fact that the data that is generated by human activities is not always available for study.

- Notice that the data sources typically used for the study of human behavior does not come from controlled experiments, but from the digital traces that are generated by human activities.

- The second problem is the data analysis, and the fact that the data that is generated by human activities is not always easy to analyze.

- The data source to analyze usually is a piece of a larger dataset, so we need to be careful to avoid biases in the analysis driven by the data size.

- Temporal windows are also a problem, because when we analyze the dynamics of a system, we need to be careful to avoid biases in the analysis driven by the temporal window.

- The third problem is the modeling, and the fact that the data that is generated by human activities is not always easy to model.

- Deterministic models are not always useful to model human behavior, and we need to use stochastic models to model human behavior.

- Also, mechanistic models and data driven models is something that we need to consider when we model human behavior.

- Another possibility is to use agent-based models to model human behavior. With the advent of computational methods in the latter half of the 20th century, researchers gained powerful tools to simulate and analyze complex social systems. Agent-based modeling (ABM) emerged as a particularly influential approach, enabling scientists to create and study systems of interacting agents (individuals or collective entities) and observe emergent behaviors from simple rules of interaction.

- Computational social science has many applications, and it is being used to study human behavior and social systems.

- Sociotechnical systems, social networks, and human dynamics are some of the applications of computational social science.

- fake news detection, information spreading, and social influence are some of the applications of computational social science.

1.3 Terminology and general concepts

1.3.1 Complex networks

As I understood it, what makes a network complex is having a short path length, such that average distance between two nodes is relatively small, high levels of clustering, such that nodes tend to form triangles, and a degree distribution with a fat tail, such that there are a few nodes with a very high degree. These properties are often found in social networks, such as a social media, the collaboration network of scientists, or international conflicts.

- In this section, we introduce some terminology and general concepts that are used in the study of human behavior and social systems.

Complex networks, interface density, and community structure are some of the concepts that are used in the study of human behavior and social systems.

1.3.2 Models

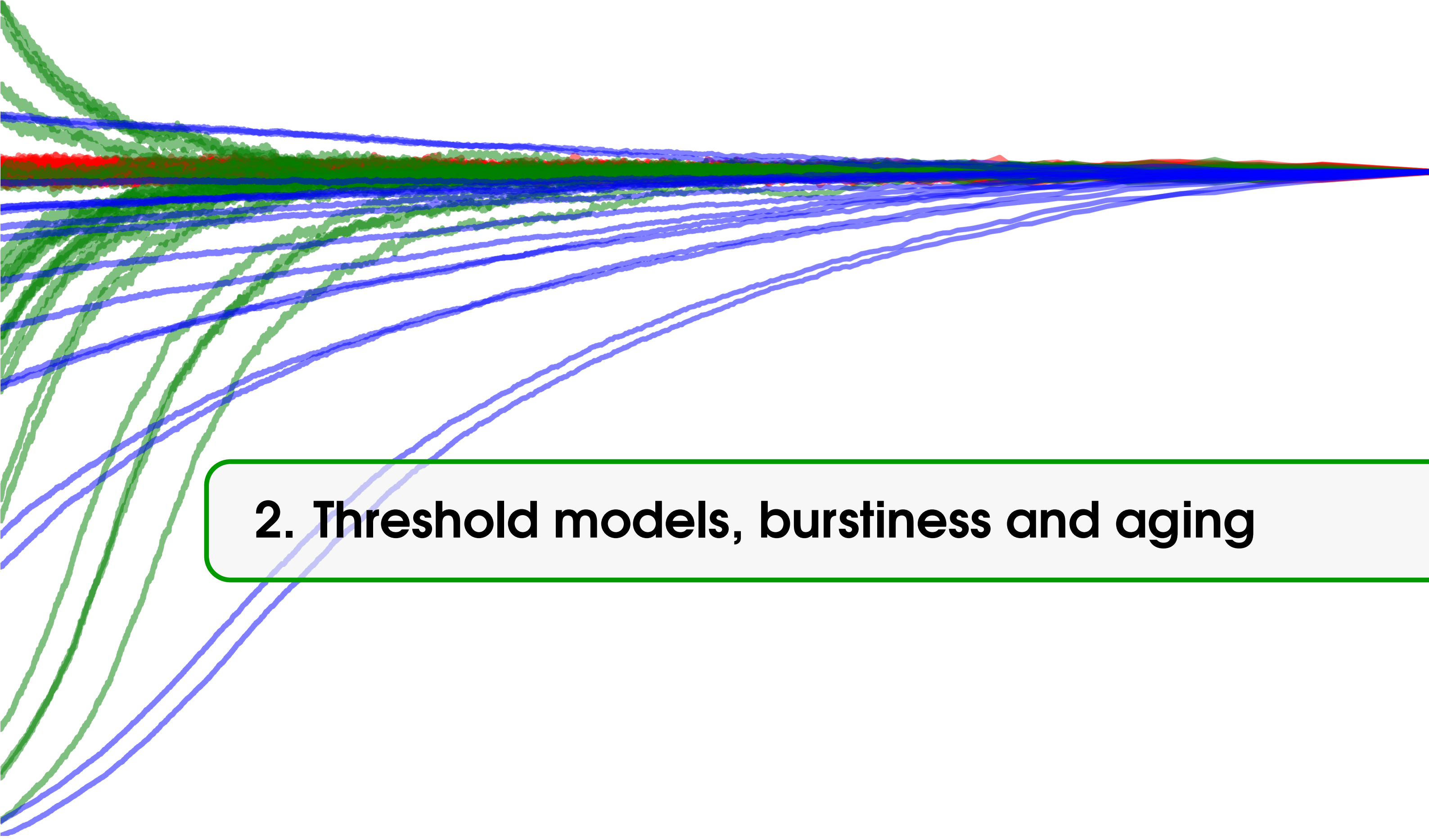
binary state models, random networks, configuration models, and preferential attachment are some of the models that are used in the study of human behavior and social systems.

1.4 Datasets

- We used the idealista dataset

- The strong point of the idealista dataset is that it contains information about the real estate market in Spain, and that it is a large dataset that contains information about the real estate market in Spain.

- The missing point of the idealista dataset is that it contains information about the real estate market in Spain, and that it is a large dataset that contains information about the real estate market in Spain.



2. Threshold models, burstiness and aging

I Aging in threshold models

3. Aging effects in the Sakoda-Schelling segregation model

4. Aging in binary state dynamics: The Approximate Master Equation

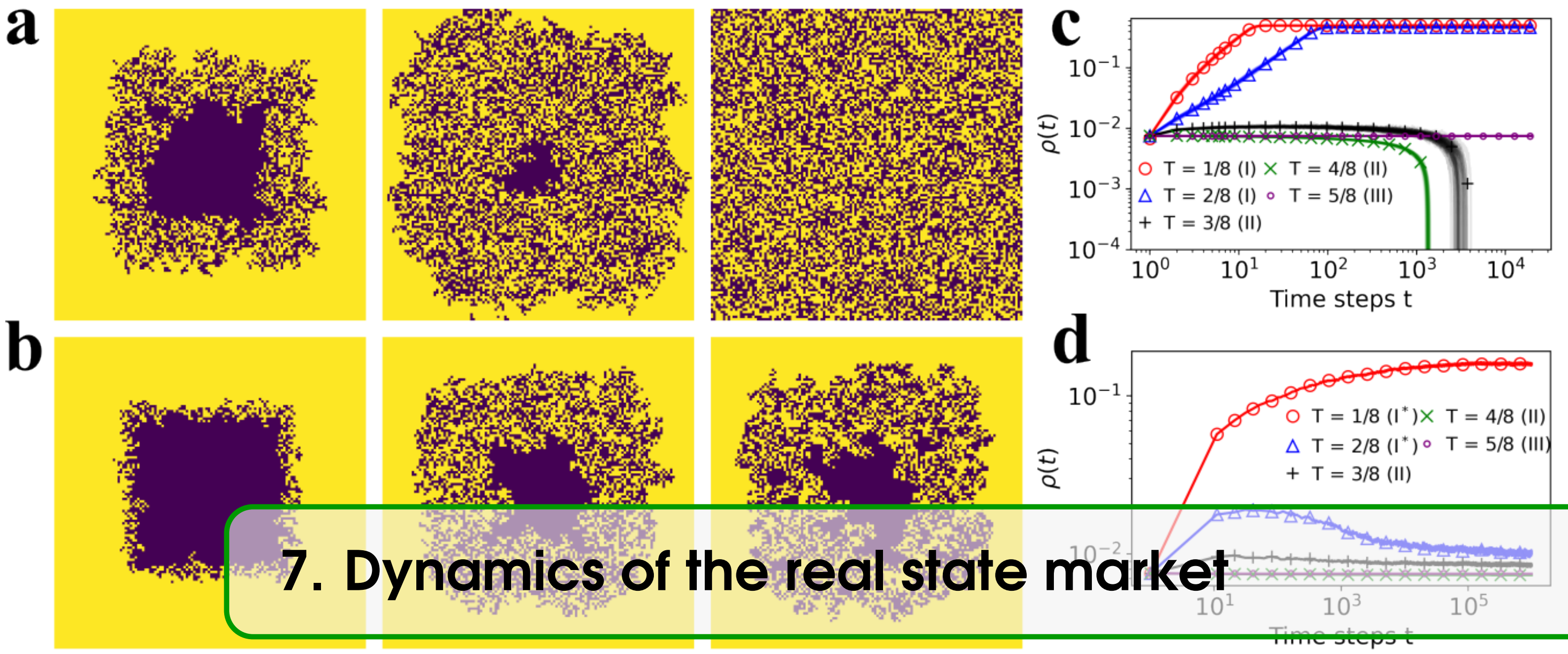
5. Impact of Aging in the Granovetter-Watts model

6A. Symmetrical Threshold model: Ordering dynamics

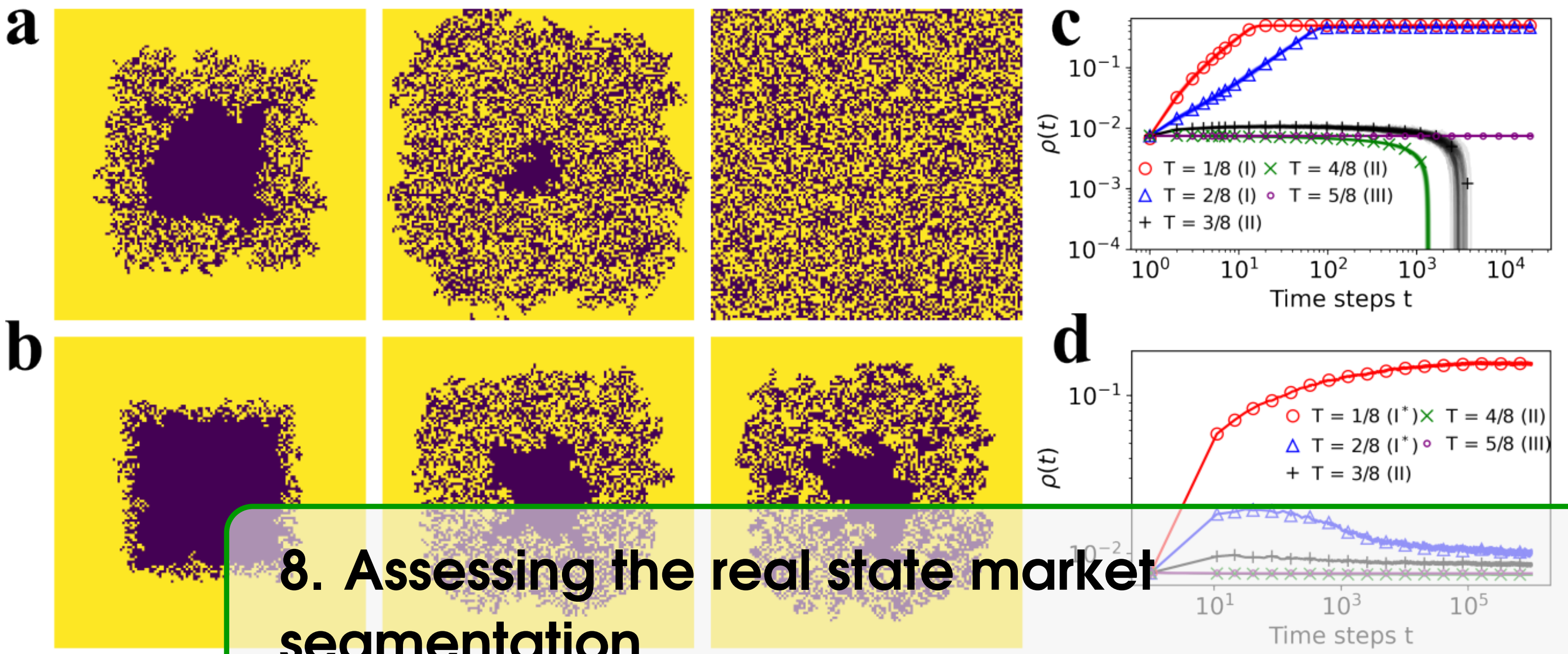
6B. Symmetrical Threshold model: Aging implications

II

Real estate market dynamics



7. Dynamics of the real state market



8. Assessing the real state market segmentation

The results in this chapter are published as:

David Abella et al. "Ordering dynamics and aging in the symmetrical threshold model". In: *New Journal of Physics* 26.1 (Jan. 2024), page 013033. DOI: 10.1088/1367-2630/ad1ad4. URL: https://dx.doi.org/10.1088/1367-2630/ad1ad4

The real estate market shows an inherent connection to space. Real estate agencies unevenly operate and specialize across space, price and type of properties, thereby segmenting the market into submarkets. We introduce here a methodology based on multipartite networks to detect the spatial segmentation emerging from data on housing online listings. Considering the spatial information of the listings, we build a bipartite network that connects agencies and spatial units. This bipartite network is projected into a network of spatial units, whose connections account for similarities in the agency ecosystem. We then apply clustering methods to this network to segment markets into spatially-coherent regions, which are found to be robust across different clustering detection algorithms, discretization of space and spatial scales, and across countries with case studies in France and Spain. This methodology addresses the long-standing issue of housing market segmentation, relevant in disciplines such as urban studies and spatial economics, and with implications for policymaking.

8.1 Introduction

The spatial dimension of housing markets is a crucial aspect for urban studies and planning. Understanding the spatial segmentation of the housing market into submarkets (**morawakage2022housing**, **bourassa2003housing**) has important implications for real estate valuation and investment decisions, which together affect urban development and social equity (**bourassa2003housing**). Spatial segmentation is the product of many factors such as residential location and the proximity to amenities (**bourassa2003housing**), differences in housing stock (**keskin2017defining**), price levels (**goodman1998housing**), and consumer preferences (**leishman2013predictive**).

The spatial division of the real estate market has been studied from different perspectives and with different methods in the literature. Some studies have examined the spatial segmentation of the urban housing market focusing on neighborhood correlations of housing prices (**palm1978spatial**), the spatial effects of urban public policies on housing values (**baumont2009spatial**), the neighborhood quality and accessibility effects on housing prices (**dubin1992spatial**), while others have determined if a specific property market is spatially segmented into submarkets, and whether accounting for the existence of submarkets improves the accuracy of price modeling (**keskin2017defining**, **usman2021priori**). This is especially important for hedonic pricing models that seek to incorporate spatial autocorrelation and heterogeneity (**usman2021priori**, **paez2009recent**, **bitter2007incorporating**, **case2004modeling**). Ref. (**hu2022NovelApproach**) distinguishes two main approaches for spatial segmentation: using pre-defined geographical boundaries based on *a priori* knowledge, such as local administrative boundaries or expert areas used by market stakeholders, or relying on clustering methods to infer patterns from the structure of the data. For the latter, popular statistical approaches to divide space into submarkets are principal component analysis and hierarchical clustering (**goodman1998housing**, **bourassa1999defining**, **bourassa2003housing**).

The digitization of the housing market (**raehousing2024**) provides untapped research opportunities for data-driven studies of market segmentation. With property portals being nowadays the dominant way to create and access market information, online listings constitute a new type of data to study housing markets (**sawyer1999ict**, **boeing2017new**, **boulay2021moving**). Scholars studied the spatio-temporal distribution of housing prices (**yao2018mapping**, **adolfsen_segmentation_2022**), revealed the persistence of spatial inequalities in the housing information landscape (**boeing2020online**), predicted the social profile of neighborhoods (**delmelle2021language**), or detected the segmentation of the market from online search patterns (**rae2015online**). Aside price, pictures or textual descriptions, a listing includes a critical piece of information: the identity of the marketing agency that has posted the listing on the portal. As such, listings constitute digital traces (**salganikbit2017**) of the work performed by real estate agencies when acquiring, selling or marketing on property portals. It is therefore possible to reconstruct, for each agency, its own portfolio of listings, whose volume and location patterns result from and reflect the heterogeneous practices and market shares of real estate agencies. By informing on *who sells where*, listings offer new ways

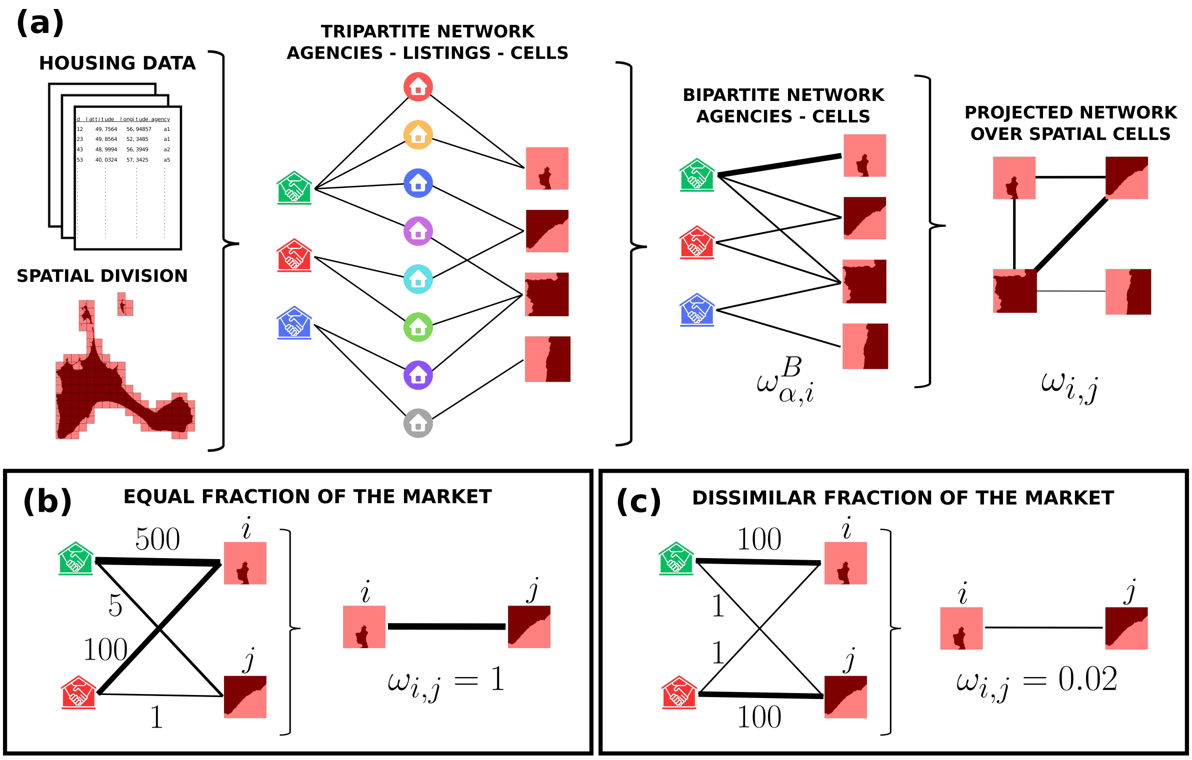


Figure 8.1: (a) A tripartite network is constructed between real estate agencies, listings, and spatial units obtained from geolocalized housing data and the division of space in regular grid cells. In this network, each listing is connected to its real estate agency and the spatial cell where it is located. This simple tripartite network is contracted into a bipartite network linking agencies and cells, where the link weight $\omega^B_{\alpha,i}$ corresponds to the number of listings the agency α has in the spatial cell i. Finally, the network is then projected over the cells to form a weighed network of spatial units, where the weight $\omega_{i,j}$ of the link between cells i and j quantifies how much they are similar in the market – $\omega_{i,j}$ is properly defined by Equation (8.2). Two simple examples of the projection process are shown below: with **(b)** equal and **(c)** complementary listings distributions for the agencies in the cells.

to examine how real estate agencies unevenly operate and specialize across space, thereby segmenting the market into submarkets (**palm1976RealEstate**).

There is ample evidence underlining how real estate agencies influence market segmentation by determining housing prices, sorting homebuyers into different market channels, and specializing in certain types of neighborhoods and market segments (**palm1976RealEstate**, **palm1978spatial**, **keskin2017defining**, **bonneval2017agents**, **besbris2017investigating**). Furthermore, it has been shown that the definition of submarkets based on agencies is far superior to other segmentation techniques (**leishman2013PredictivePerformance**).

This work introduces a new method to identify the housing market segmentation using geospatial data, complex network analysis techniques, and taking as a basis the local ecosystem of real estate agencies. We build a network structure based on two factors: the *presence* of an agency within a particular area, and the relative *influence* of an agency in this area, determined by the agency's proportional share of all listings located in the area. Our methodology is applied to the residential property market in 3 Spanish provinces and 3 French urban areas, for which we have a rich, high resolution dataset sourced from property portals. We find that the market in those regions is divided into a hierarchy of subregions. We test the robustness of our results against different community detection algorithms, scales, and administrative boundaries in different countries.

8.2 Materials and methods

8.2.1 Data description

For Spain, we analyze listings published on the portal `Idealista.com` (**idealista**). The dataset covers a 2-year time period, from January 2017 to December 2018 and it comprises a comprehensive collection of online listings georeferenced with their (lat, long) coordinates in the Spanish provinces of Balearic Islands, Barcelona, and Madrid. These listings were posted by more than $50,000$ real estate agencies, each identified with its unique id. There are about one million listings for sales, and over $800,000$ for rentals.

French listings were obtained from the portal `SeLoger.com` (**SeLoger**). The dataset includes all listings posted in the country over a 6-month period from July to December 2019 - representing over 2 million sale listings. Geographical information is only available at the administrative and census levels, such as ZIP codes ("*code postal*"), municipalities ("*communes*"), and census tracts ("*IRIS*"), the finest and basic scale for sub-municipal information in France. We focus on three major urban areas: Paris, Marseilles and Toulouse.

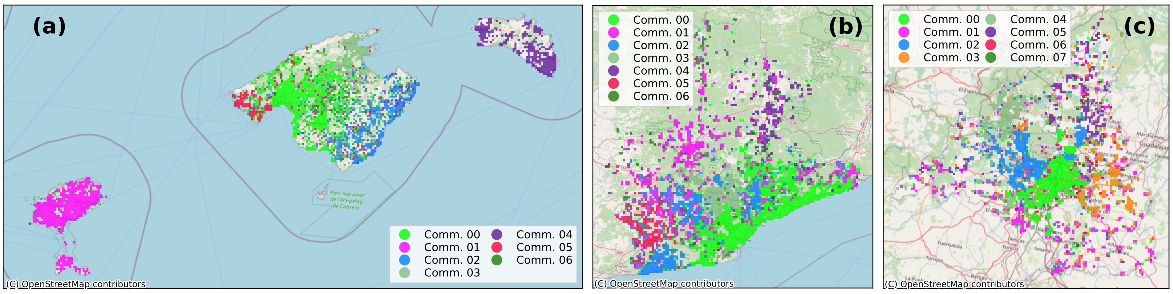
For both datasets, we focus on houses and apartments, and do not consider farms or rural parcels.

8.2.2 Building a network

We begin by discretizing the space into spatial units (square grid cells, municipalities, districts, postal codes, census-tracts, etc). This allows us to label each listing according to the spatial unit it falls into, along with the agency that posted this listing. By doing so, we build a tripartite relation between agencies, listings, and spatial units. Based on this structure, we can build a weighted bipartite network that connects agencies and spatial units, where the link weight $\omega^B_{\alpha,i}$ accounts for the number of listings posted by agency α that are located in the spatial unit i. The resulting network contains all the information about the spatial characteristics of the housing market.

Bipartite networks can be projected to create networks with a single type of nodes (**newman2001structure**, **newman2001scientific**, **zhou2007bipartite**). In our case, we project it to build a new weighted network connecting spatial units (see Fig. **??**(a) for schematic representation, taking as an example the discretization of space with square grid cells). Let us assume that we have N spatial units and N^a real estate agencies. The set of all agencies operating in the entire area is $\{\alpha\}$, while the

Figure 8.2: Communities from the projected network for the three Spanish provinces studied: Balearic Islands **(a)**, Barcelona **(b)**, and Madrid **(c)**. The spatial cells are $1\ \mathrm{km}$ square cells. The communities shown are detected using the Louvain algorithm with a consensus clustering of 1000 realizations. The underground map data is rendered from OpenStreetMap under ODbL.



subset operating in the spatial unit i is denoted by $\{\alpha\}_i$. The fraction of listings in i that belong to a certain agency α is

$$f_{\alpha,i} = \frac{\omega^B_{\alpha,i}}{\sum_{\gamma\in\{\alpha\}_i}\omega^B_{\gamma,i}}, \tag{8.1}$$

where the index γ runs over all the agencies operating in i. In the projected network, we define the influence weight between two spatial units i and j as

$$\omega_{i,j} = \frac{\sum_{\gamma\in\{\alpha\}_{ij}} f_{\gamma,i} f_{\gamma,j}}{\frac{1}{2}\left[\sum_{\gamma\in\{\alpha\}_i} f^2_{\gamma,i} + \sum_{\beta\in\{\alpha\}_j} f^2_{\beta,j}\right]}, \tag{8.2}$$

where $\{\alpha\}_{ij} \equiv \{\alpha\}_i \cap \{\alpha\}_j$ is the subset of agencies operating in i and j. The weight $\omega_{i,j} = 1$ if the agencies operating in i and j are the same, and cover an equal fraction of the market in both spatial units. If the market distribution is similar, but not equal, the weight will deviate from 1. Reciprocally, if no common agency is found across the two spatial units, the weight is zero and there is no link between them. Fig. **??**(b) and **??**(c) show examples of the influence weights between two spatial units with equal distribution of the listings in (b), for which $\omega_{i,j} = 1$, and a complementary distribution in (c) with a value of $\omega_{i,j} = 0.02$. Note that our influence weight is related to the participation ratio introduced by Derrida *et al.* in (**Derrida_1987**).

The projected network is thus built with the spatial units as nodes, which are connected with links weighted according to Equation 8.2. A group of spatial units strongly connected between them implies that they share a common ecosystem of agencies, that operate with a similar market share in these units. Searching for clusters in this weighted spatial network should therefore inform us on the spatial segmentation of the housing market, the clusters corresponding to submarkets. In the network literature, such clusters are commonly referred to as communities, with numerous methods proposed to detect them (**fortunato2010community**). We use several classic community detection algorithms (**newman2004finding**, **infomap**, **Louvain**, **Louvain-Leiden**, **OSLOM**) that account for network weights, including Louvain (**Louvain-Leiden**), Infomap (**infomap**), and OSLOM (**OSLOM**). These algorithms enable us to classify the spatial units into communities. Since these algorithms are stochastic, we perform several realizations of each method, and perform consensus clustering (**lancichinetti2012consensus**) for higher stability.

8.3 Results

8.3.1 Segmenting the market according to agencies' operations

We start by analyzing the spatial segmentation that arises from the data geolocated in the Balearic Islands, Barcelona, and Madrid using 1 km-sided square cells. Fig. 8.2 presents the

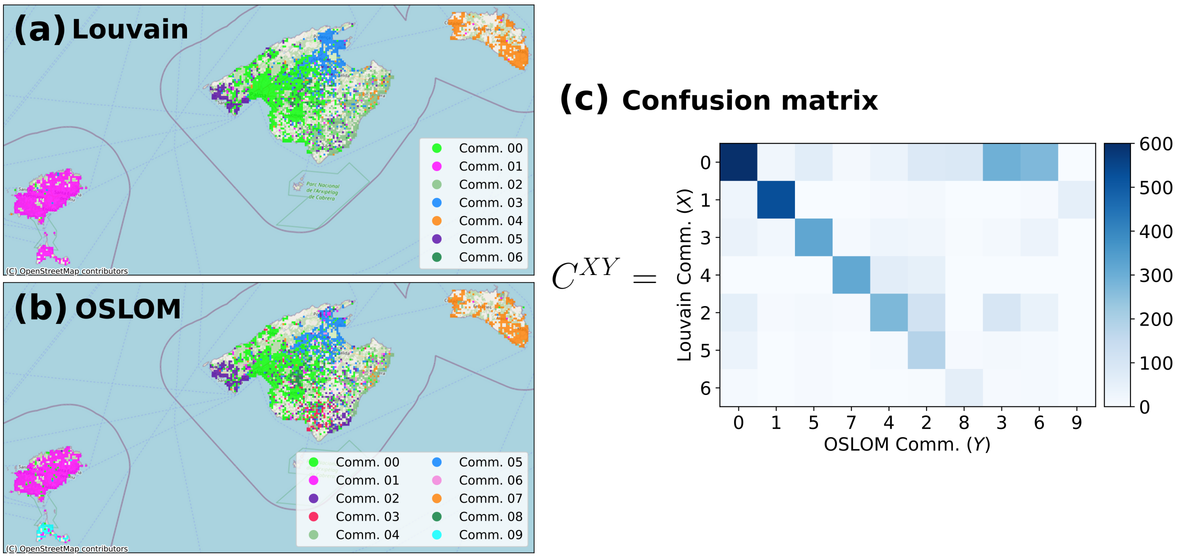


Figure 8.3: Partition result of the community detection methods at the Balearic Islands using Louvain algorithm **(a)** and OSLOM method **(b)** $1\mathrm{km}$ square cells. The confusion matrix C^{XY} of the two partitions **(c)**, is ordered according to the maximum overlap. The underground map data is rendered from OpenStreetMap, under ODbL.

communities listed according to their size, from larger to smaller. Even though our methodology does not consider spatial proximity, we observe spatial segmentation in adjacent regions with few exceptions. For the Balearic Islands, we observe that spatial constraints, such as insular nature of the environment, affect the segmentation of the housing market: while the same submarket covers Minorca or Ibiza-Formentera, Majorca is divided into four different ones. It is noteworthy that the submarkets that emerge in all these three provinces are slightly larger than municipalities.

To study the robustness of identified submarkets in each of the three provinces, we run several community detection algorithms, and compare the communities obtained across realizations of different algorithms. We define as a network partition the classification of the cells in communities, $X = \{x_0, x_1, \cdots, x_{|X|-1}\}$, where each community x_i is a set of cells. The partition X has $|X|$ communities in this notation. Every cell must be in at least one community, but in some clustering methods a cell may belong to several. In order to compare two partitions X and Y, we compute a confusion matrix C^{XY} in which each element is defined as

$$C_{ij}^{XY} = |x_i \cap y_j|, \tag{8.3}$$

where x_i and y_j are communities in the partitions X and Y, respectively, and $|.|$ stands for the cardinal (number of elements) of a set. An element C_{ij}^{XY} can be zero if there is no overlap between the communities, and it can be large if the two communities coincide across the partitions. We reorder then the elements of the matrix C^{XY} to have the largest values in the pseudo-diagonal. Note that C^{XY} is not necessarily a squared matrix because the number of communities in each partition may differ. This process is essentially the identification of the communities in one partition that correspond to the communities in the other. This is a statistical match, given that the cells of a community in X may be distributed in several communities in Y. As shown in Fig. **??**, if the partitions between the two methods are similar, we must observe a strong pseudo-diagonal in the confusion matrix. The sum of the elements of this pseudo-diagonal is the number of cells clustered in the same way in the two partitions. To compute a measure of the agreement between two partitions, we use the fraction $H(X,Y)$ (**girvan2002community**,

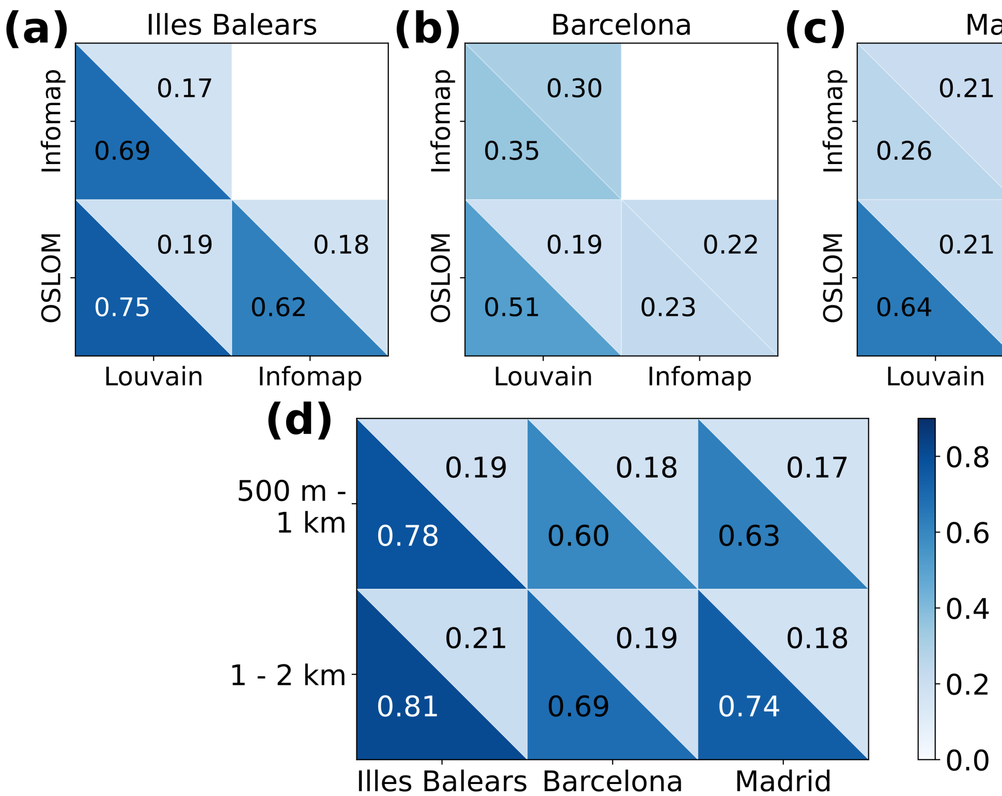


Figure 8.4: Agreement across the different community detection methods for the network in Balearic Islands **(a)**, in the province of Barcelona **(b)** and of Madrid **(c)**. The metric used to compute the agreement between method partitions is $H(X,Y)$, shown in the lower triangles for each pair of methods, denoted by X and Y. The upper triangles display the value $H(X_r,Y_r)$, being X_r and Y_r the partitions randomized (preserving the communities size). In **(d)**, comparison of partitions obtained with the Louvain method for networks generated with different cell sizes: 500 m-sided vs 1 km-sided cells (top row), and 1 km-sided vs 2 km-sided cells (bottom row).

hric2014community) defined as

$$H(X,Y) = \sum_{i=0}^{\min(|X|,|Y|)-1} \frac{C_{ii}^{XY}}{N}, \tag{8.4}$$

where the matrix C^{XY} is ordered to maximize the pseudo-diagonal, and N is the total number of cells. $H(X,Y)$ is a metric commonly used in the literature to compute the accuracy between community detection algorithms (**danon2005comparing**, **duch2005community**, **li2008quantitative**, **darst2014improving**, **chen2015deep**, **saoud2016community**, **wang2017mitigation**, **fortunato2016community**), its value is bounded in the interval $(0,1]$, but it has the downside that $H(X,Y)$ depends on the size of the communities. To determine if the value of $H(X,Y)$ is significant, it is necessary to compare it with a randomized version of the partitions, $H(X_r,Y_r)$, in which the cells are reshuffled at random across the communities of each partition respecting the community sizes.

Figure **??**(a-c) compares the three community detection algorithms (Louvain, OSLOM, and Infomap) used for different provinces. In all cases, the agreement between the communities detected from the real partition is higher than that of the randomized communities. The OSLOM-Louvain comparison exhibits the highest agreement, which is significant in all provinces. In the Balearic Islands, a robust and statistically significant agreement is evident among all methods. However, when examining Barcelona and Madrid, Infomap detects a large community probably due to the high density of the network, and this does not compare well with the other methods which detect more communities. In fact, the value of $H(X,Y)$ approaches the one of the randomized model. This issue is absent in the Balearic Islands, where the network has a stronger intrinsic spatial division into different islands.

So far, we have focused on the results for the networks built with 1 km-sided square cells. It is, nevertheless, important to check whether the results may vary depending on the scale of the unit cells. We thus recalculate the networks taking as basis square cells of side 500 m and 2 km and compute the communities using the Louvain method with consensus clustering. The cells of the different scales have been delimited to keep spatial coherence: four 500 m cells form one of the 1 km cells used in the previous figures, and four 1 km cells aggregate to form a 2 km cell. This hierarchical structure allows us to compare communities at various levels because we can identify the cells across scales. For example, if a 2 km cell belongs to a community, then the four 1 km cells composing it share the same community label. In parallel, we also run the community detection algorithm in the network composed of 1 km cells, and then we can use the confusion matrix and $H(X,Y)$ to compare the partitions at these two scales using 1 km cells. Note that the calculation of $H(X,Y)$ requires the same number of basic units in the two partitions. Figure **??**(d) shows the results of this analysis, where we use 1 km-sided cells as a reference for comparison with the other scales. In all cases, we notice a consistently high and statistically significant level of agreement. This demonstrates that our methodology generates communities that remain robust across the three spatial scales.

8.3.2 Comparison with networks obtained from administrative boundaries

In this section, we examine how incorporating administrative spatial boundaries to build networks impacts the detection of communities. In many cases, the geographical information for listings is only available at the level of existing administrative boundaries and statistical units, which are by design more heterogeneous than square cells.

We aggregate listings into administrative and statistical spatial units to determine if the emergent submarkets are stable and consistent when comparing with the ones observed with the networks built with square cells. In this case, we consider municipalities and census tracts as they are the most common administrative divisions applied to spatial statistics.

Fig. **??** shows the communities found in the province of Madrid. We observe clear differences between the results obtained using census tracts (Fig. **??**(a)-(b) and using municipalities (Fig. **??**(d)-(e). The results for census tracts are characterized by a large community that covers almost all the territory and the agreement between methods is not significant. In contrast, the

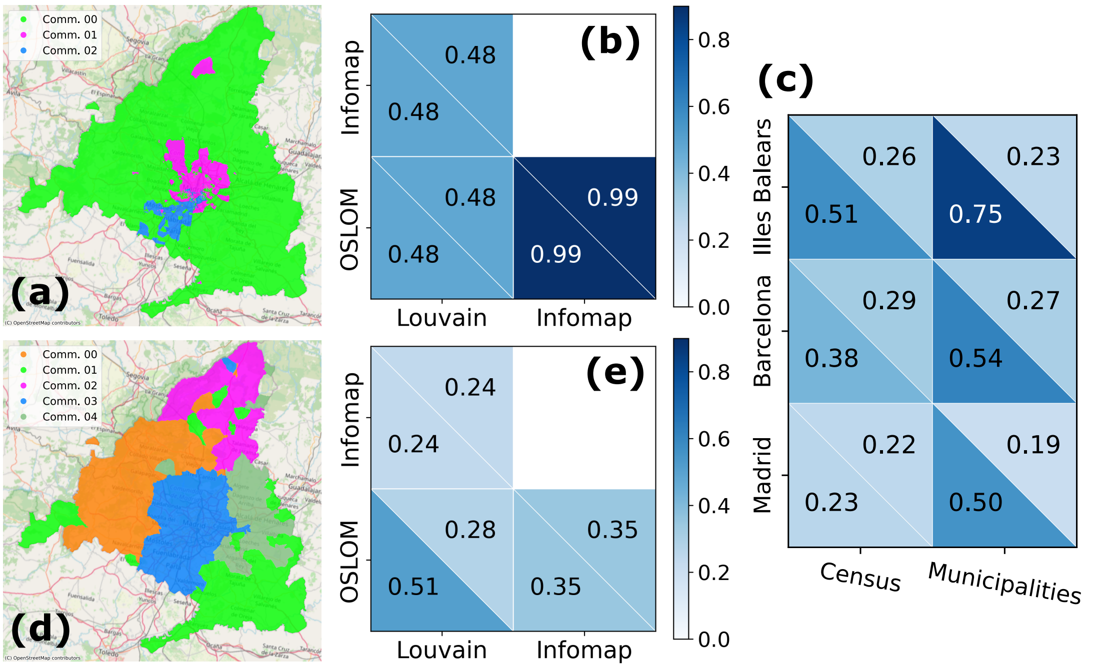


Figure 8.5: Communities detected using census areas **(a)** and municipalities **(d)** as spatial units to build the network in Madrid. The clustering method employed is the Louvain algorithm. The agreement across the different methods for the census **(b)** and municipalities **(e)**. **(f)** shows the communities' agreement between 1 km cells and administrative boundaries networks for all Spanish provinces. The agreement in (b)-(c)-(e) is computed using $H(X,Y)$ (lower triangles) compared with the value randomizing the communities (upper triangles). The underground map data is rendered by OpenStreetMap, under ODbL.

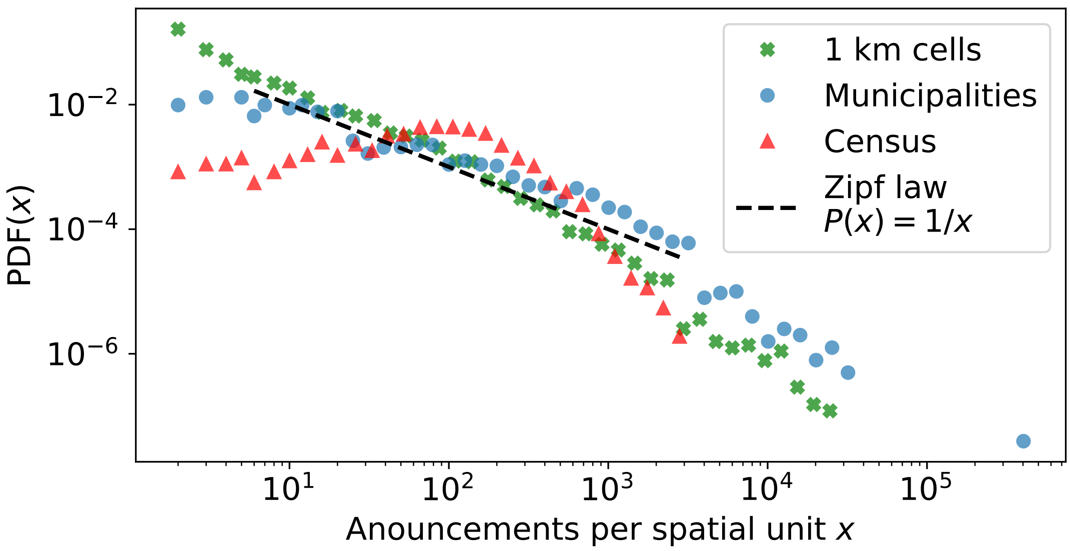


Figure 8.6: Each spatial unit is shown by a different color and marker: green crosses (1 km-sided cells), blue circles (municipalities), and red triangles (census). The dashed black line shows the slope of a Zipf law distribution.

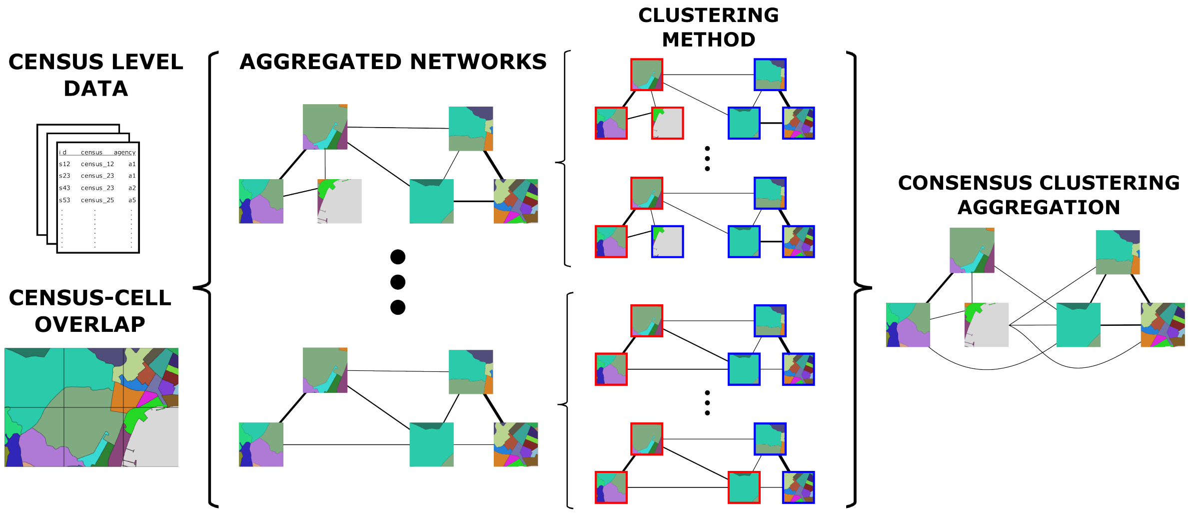


Figure 8.7: From a census-level listing and the spatial division of the census in square cells, we generate an ensemble of networks. In this ensemble, each listing within a census tract is associated to a cell with a probability based on the overlapping area between the census and the cell. For each of these cell networks, we run a community detection algorithm multiple times. The next step involves combining the results from these partitioned networks through consensus clustering, resulting in an aggregated network.

results using municipalities have a good and significant OSLOM-Louvain agreement. Keeping Louvain as the reference method, we compare the partitions of the networks originated from $1\,\mathrm{km}$, census tracts, and municipalities in Fig. **??**(c). The communities in the networks using cells and municipalities show significant agreement, while those based on census tracts show non-significant values in Barcelona and Madrid.

While the distribution of listings per spatial unit in the other cases follows a heterogeneous distribution, well-described by a Zipf law, the one for census tracts follows a more homogeneous distribution (see Fig. **??**). This effect is a consequence of how the census tracts are built, forcing the population in each unit to be similar by a heterogeneous selection of the space included in each unit. This distribution is directly translated into the network weights and thus impacts the spatial segmentation method.

8.3.3 Recovering the submarkets from census level data

Multiple datasets, such as our French data, are avilable at census level. To maintain the broad applicability of our spatial segmentation methodology, we have devised a data aggregative method to recover the results obtained at the cell and municipality levels. This technique enables us to restore the Zipf law pattern using data gathered at the census level and to find similar segmentation results regardless of the basic spatial units.

We start with listings at a census scale, such that each listing is associated to an agency and a census tract. The first step is to divide the space into square cells, as we did in Section 8.2. The cells intersect with the census tracts. We then associate each listing to a cell with a probability proportional to the overlapping area between the listing census tract and the cell. This process is repeated for all the listings to reconstruct a tripartite network of agencies-listings-cells, from which we can follow the methodology explained to reach a cell-cell network and a segmentation in submarkets (communities). We observe that in the final networks the Zipf law distribution of listings per cell is recovered.

Since the assignation of listings to cells is stochastic, the projected network is different each time the process is repeated. To avoid uncertainty, we construct an ensemble of these networks. For each network, we run the community detection algorithm multiple times. Once our cells are labeled with a community, we perform consensus clustering to aggregate all partitions from

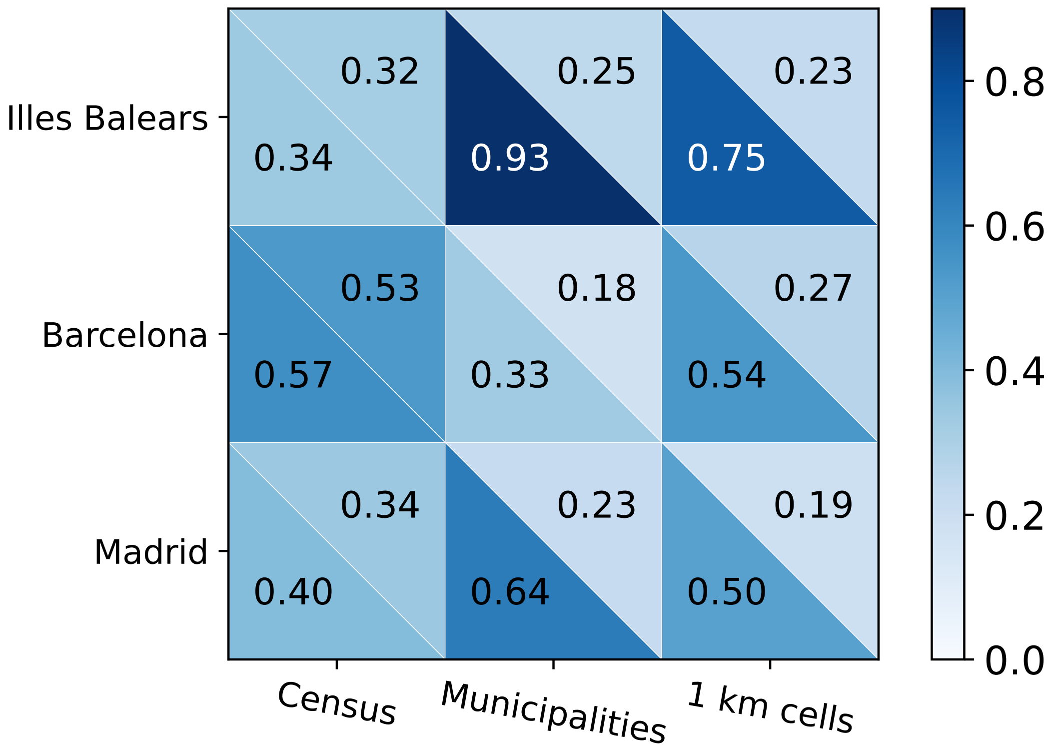


Figure 8.8: Each column shows the agreement between the communities of the 1 km aggregated cells networks (from census data) and the networks obtained from the other spatial units: Census, Municipalities, and 1 km cells from the original latitude longitude coordinate data. Each row shows the results for each province: Balearic Islands, Barcelona and Madrid. The agreement is computed via the fraction of correctly detected cells $H(X,Y)$ (lower triangles) compared with the value randomizing the communities (upper triangles).

all aggregated networks of our ensemble into a single consensus aggregated network. We represented this process in detail on Fig. **??**.

To verify the results of the aggregative method, we perform a comparison of the submarkets obtained out of different networks. Starting with our Spanish data, where the listings are geolocated using exact coordinates, we build networks at the level of 1 km cells, census tracts and municipalities. We then apply the method to aggregate the census tracts to the cells. This gives us a fourth family of networks, which we call aggregated cells network. We then run community detection methods and compare them across the networks, taking as a basis the partition obtained from the network of aggregated cells (see Fig. **??**). For all cases, the agreement exhibited by partitions of the aggregated cells network and the original cells or the municipalities is very high (and significant compared to the randomized communities). Therefore, by reconstructing the network with the aggregative method, we recover the original communities at the cell and municipality levels and avoid the issues caused by the natural spatial heterogeneity of census tracts.

8.3.4 Comparison across countries

In this section, we investigate whether the emergent spatial segmentation revealed by our method is a unique feature of the Spanish market, or can be understood as a more general phenomenon across geographical contexts. To this end, we use listing data for three major French urban areas, namely, Marseilles, Paris, and Toulouse. Since we do not have exact coordinates for the listings, which are only located at a census tract level, we have to employ

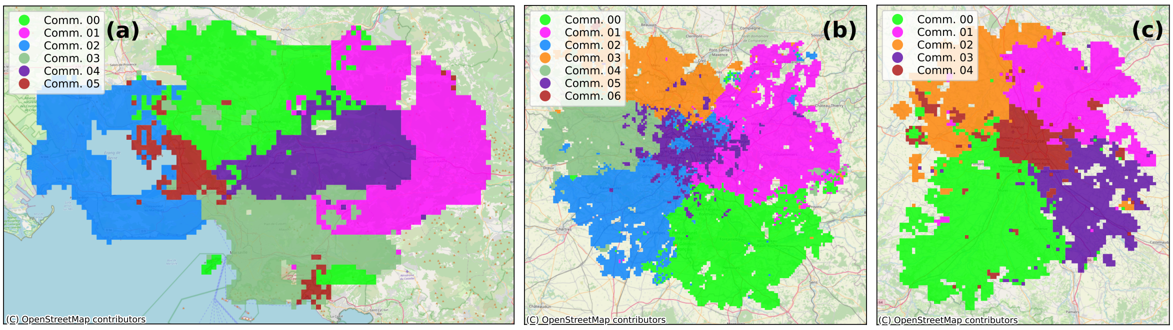


Figure 8.9: Communities detected at the stochastic projected network for the 3 French FUA studied: Marseilles-Aix en Provence **(a)**, Paris **(b)**, and Toulouse **(c)**. The communities shown are detected using the Louvain algorithm with a consensus clustering of 200 clustering method realizations for each of the 100 stochastic networks generated in the IRIS to cell aggregative process. The underground map data is rendered by OpenStreetMap, under ODbL.

the stochastic aggregative technique described in the previous section to obtain the networks at the cell level or to aggregate the data at the municipality (*commune*) level (since the census tracts can be grouped within each commune).

Communities emerge in these French urban areas at aggregated cell level as well (see Fig. **??**). The communities are contiguous in space, similar to the ones observed in Spain, suggesting that listings (as a source of information on listed properties and agencies) allow us to study the spatial segmentation of the housing market through a data-driven, bottom-up method that foregrounds the practices of key market intermediaries.

We repeat the exercise of comparing networks built from different spatial divisions. If France exhibits the same structures found in the Spanish dataset, we would expect the communities found from the aggregated cells and municipality networks to coincide, being the ones from the network of IRIS level very different. Fig. **??** displays the agreement between the communities using aggregated cells and administrative divisions (IRIS and communes). All values of the agreement are significant when compared with the randomized communities, but the largest agreement is found between aggregated cells and communes in all places, echoing results with the Spanish data. This indicates that our aggregative method is a general tool to compute a robust spatial segmentation of the housing market.

8.4 Conclusions

In this study, we present a new method for analyzing the spatial segmentation of housing markets through the activity of real estate agencies, using online listings to extract information on the location of both the property and the marketing agency. We apply this method to analyze comprehensive datasets of geolocated listings in two different countries: Spain and France.

Our methodology is based on dividing space into spatial units, to construct a tripartite network between listings, real estate agencies, and spatial units. We project the network, taking into account the presence and influence of real estate agencies. To divide our projected networks, we use different classic community detection algorithms that account for network weights, such as Louvain, Infomap, and OSLOM. Our methodology generates a spatial segmentation into regions that happen to be spatially connected and larger than municipalities. This segmentation into submarkets remains robust across different community detection algorithms, scales, and administrative boundaries across different countries.

We discovered a limitation of our method when the spatial units exhibit a highly heterogeneous area distribution, and the Zipf law of the distribution of listings per spatial unit is not fulfilled, as in the case of census tracts. To overcome this limitation and extend our methodology to

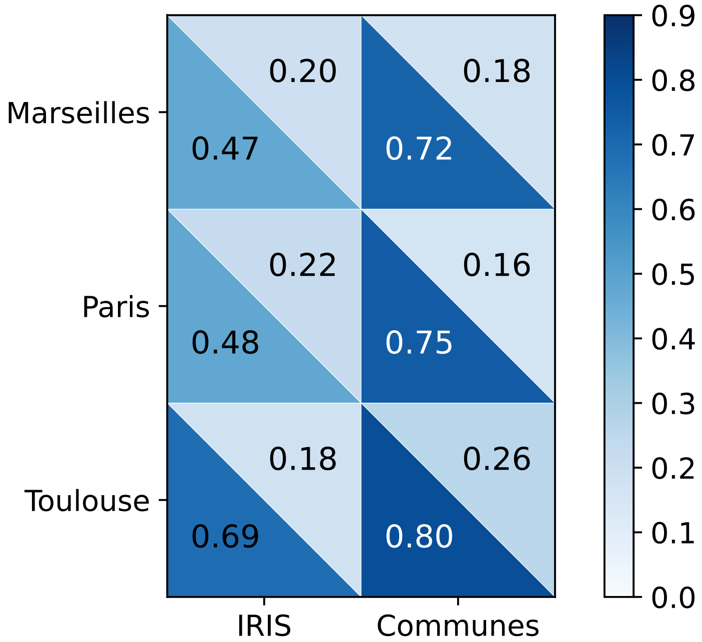


Figure 8.10: Each column shows the agreement between the 1 km aggregated cells and the French political spatial units: IRIS and Communes. Each row shows the results for each FUA: Marseilles-Aix en Provence, Paris, and Toulouse. The agreement is computed via the fraction of correctly detected cells (lower triangles) compared with the value randomizing the communities (upper triangles).

heterogeneous-level data, we developed a method to create an aggregated network via stochastic reconstruction and consensus clustering aggregation. This methodology exhibits good accuracy when compared with the communities from the original high-precision data.

To summarize, we have developed a new methodology that uses listings data to evaluate the spatial segmentation of housing markets into spatially-coherent submarkets. This methodology is generally applicable to different datasets of geolocated listings to infer the submarkets that emerge from the uneven presence and influence of real estate agencies across space. The market-based supra-municipal communities that emerge from the data are found to be robust. Future research should investigate how identifying the submarkets created by market intermediaries can inform policymaking and improve price modeling.

The projected networks at the spatial resolution of 1 km cell, census-tract, and municipality are available at Zenodo (**zenodo-2024**) and Github (**Abella-github-2024**). These links also include the code and additional code to perform the stochastic aggregative method from generic census data.

Bibliography

Bibliography

[0] David Abella, Giancarlo Franzese, and Javier Hernández-Rojas. "Many-Body Contributions in Water Nanoclusters". In: *ACS Nano* 17.3 (Jan. 2023), pages 1959–1964. ISSN: 1936-086X. DOI: 10.1021/acsnano.2c06077. URL: http://dx.doi.org/10.1021/acsnano.2c06077 (cited on page 17).

[0] David Abella, Maxi San Miguel, and José J. Ramasco. "Aging effects in Schelling segregation model". In: *Scientific Reports* 12.1 (Nov. 2022). ISSN: 2045-2322. DOI: 10.1038/s41598-022-23224-7. URL: http://dx.doi.org/10.1038/s41598-022-23224-7 (cited on page 17).

[0] David Abella, Maxi San Miguel, and José J. Ramasco. "Aging in binary-state models: The Threshold model for complex contagion". In: *Phys. Rev. E* 107 (2 Feb. 2023), page 024101. DOI: 10.1103/PhysRevE.107.024101. URL: https://link.aps.org/doi/10.1103/PhysRevE.107.024101 (cited on page 17).

[0] David Abella et al. "Unraveling higher-order dynamics in collaboration networks". In: *arXiv preprint arXiv:2306.17521* (2023) (cited on page 17).

[0] David Abella et al. "Ordering dynamics and aging in the symmetrical threshold model". In: *New Journal of Physics* 26.1 (Jan. 2024), page 013033. DOI: 10.1088/1367-2630/ad1ad4. URL: https://dx.doi.org/10.1088/1367-2630/ad1ad4 (cited on pages 17, 23).

[0] Jeffrey J. Anderson and Arend Lijphart. "Patterns of Democracy: Government Forms and Performance in Thirty-Six Countries". In: *CrossRef Listing of Deleted DOIs* 30.2 (Jan. 2000), page 117. DOI: 10.2307/3331092. URL: https://doi.org/10.2307/3331092 (cited on page 1).

[0] Philip W Anderson. "More Is Different: Broken symmetry and the nature of the hierarchical structure of science." In: *Science* 177.4047 (1972), pages 393–396 (cited on page 1).

[0] Roy M Anderson and Robert M May. *Infectious diseases of humans: dynamics and control*. Oxford university press, 1991 (cited on pages 1, 3).

[0] W. Brian Arthur. *Increasing returns and path dependence in the economy*. Jan. 1994. DOI: 10.3998/mpub.10029. URL: https://doi.org/10.3998/mpub.10029 (cited on page 1).

[0] Robert Axelrod. *The Complexity of Cooperation: Agent-Based models of competition and collaboration*. Dec. 1997. DOI: 10.1515/9781400822300. URL: https://doi.org/10.1515/9781400822300 (cited on page 2).

[0] Albert-László Barabási. "The origin of bursts and heavy tails in human dynamics". In: *Nature* 435.7039 (May 2005), pages 207–211. DOI: 10.1038/nature03459. URL: https://nature.com/articles/nature03459 (cited on page 3).

[0] Albert-László Barabási. "Network science". In: *Philosophical transactions - Royal Society. Mathematical, Physical and engineering sciences/Philosophical transactions - Royal Society. Mathematical, physical and engineering sciences* 371.1987 (Mar. 2013), page 20120375. DOI: 10.1098/rsta.2012.0375. URL: https://doi.org/10.1098/rsta.2012.0375 (cited on page 1).

[0] Ugo Bastolla et al. "The architecture of mutualistic networks minimizes competition and increases biodiversity". In: *Nature* 458.7241 (Apr. 2009), pages 1018–1020. DOI: 10.1038/nature07950. URL: https://doi.org/10.1038/nature07950 (cited on page 2).

[0] Danah Boyd and Kate Crawford. "CRITICAL QUESTIONS FOR BIG DATA". In: *Information, communication & society* 15.5 (June 2012), pages 662–679. DOI: 10.1080/1369118x.2012.678878. URL: https://doi.org/10.1080/1369118x.2012.678878 (cited on page 2).

[0] B. Bradford Brown, Donna Rae Clasen, and Sue Ann Eicher. “Perceptions of peer pressure, peer conformity dispositions, and self-reported behavior among adolescents.” In: *Developmental psychology* 22.4 (July 1986), pages 521–530. DOI: 10.1037/0012-1649.22.4.521. URL: https://doi.org/10.1037/0012-1649.22.4.521 (cited on page 3).

[0] Alan Bryman. *Social research methods*. Feb. 2010, pages 157–184. DOI: 10.4324/9780203381175\{_}chapter\{_}9. URL: https://doi.org/10.4324/9780203381175_chapter_9 (cited on page 1).

[0] Claudio Castellano, Miguel A. Muñoz, and Romualdo Pastor-Satorras. “Nonlinearq-Voter model”. In: *Physical Review E* 80.4 (2009). DOI: 10.1103/physreve.80.041129 (cited on page 1).

[0] Hanshuang Chen et al. “Non-Markovian majority-vote model”. In: *Physical Review E* 102.6 (2020). DOI: 10.1103/physreve.102.062311 (cited on page 43).

[0] Fernando Diaz-Diaz, Paolo Bartesaghi, and Ernesto Estrada. “Network theory meets history. Local balance in global international relations”. In: *arXiv preprint arXiv:2303.03774* (2023) (cited on page 2).

[0] Ian Dobson et al. “Complex systems analysis of series of blackouts: Cascading failure, critical points, and self-organization”. In: *Chaos* 17.2 (June 2007). DOI: 10.1063/1.2737822. URL: https://doi.org/10.1063/1.2737822 (cited on page 1).

[0] Robin Dunbar et al. “The structure of online social networks mirrors those in the offline world”. In: *Social networks* 43 (Oct. 2015), pages 39–47. DOI: 10.1016/j.socnet.2015.04.005. URL: https://doi.org/10.1016/j.socnet.2015.04.005 (cited on page 2).

[0] Jane Elith and John R. Leathwick. “Species Distribution Models: Ecological explanation and prediction across space and time”. In: *Annual review of ecology, evolution, and systematics* 40.1 (Dec. 2009), pages 677–697. DOI: 10.1146/annurev.ecolsys.110308.120159. URL: https://doi.org/10.1146/annurev.ecolsys.110308.120159 (cited on page 2).

[0] Robert C. Ellickson. “The Evolution of Social Norms: A Perspective from the Legal Academy”. In: *Social Science Research Network* (Jan. 1999). DOI: 10.2139/ssrn.191392. URL: https://doi.org/10.2139/ssrn.191392 (cited on page 1).

[0] Uriel Frisch. *Turbulence: the legacy of AN Kolmogorov*. Cambridge university press, 1995 (cited on page 1).

[0] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. “Understanding individual human mobility patterns”. In: *nature* 453.7196 (2008), pages 779–782 (cited on page 3).

[0] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. Nov. 2016. URL: https://dl.acm.org/citation.cfm?id=3086952 (cited on page 2).

[0] Mark Granovetter. “Threshold Models of Collective Behavior”. In: *American Journal of Sociology* 83.6 (1978), pages 1420–1443. DOI: 10.1086/226707 (cited on page 3).

[0] Emilie M. Hafner-Burton, Miles Kahler, and Alexander H. Montgomery. “Network Analysis for International Relations”. In: *International organization* 63.3 (July 2009), pages 559–592. DOI: 10.1017/s0020818309090195. URL: https://doi.org/10.1017/s0020818309090195 (cited on page 2).

[0] Thomas C. Ings et al. “Review: Ecological networks – beyond food webs”. In: *Journal of animal ecology* 78.1 (Dec. 2008), pages 253–269. DOI: 10.1111/j.1365-2656.2008.01460.x. URL: https://doi.org/10.1111/j.1365-2656.2008.01460.x (cited on page 2).

[0] Rosabeth Moss Kanter and Herbert Blumer. “Symbolic interactionism: perspective and method.” In: *American sociological review* 36.2 (Apr. 1971), page 333. DOI: 10.2307/2094060. URL: https://doi.org/10.2307/2094060 (cited on page 3).

[0] Márton Karsai. “Computational human dynamics”. In: *arXiv preprint arXiv:1907.07475* (2019) (cited on page 2).

[0] David Lazer et al. “Computational Social Science”. In: *Science* 323.5915 (Feb. 2009), pages 721–723. DOI: 10.1126/science.1167742. URL: https://doi.org/10.1126/science.1167742 (cited on page 2).

[0] David Lazer et al. “Computational Social Science”. In: *Science* 323.5915 (Feb. 2009), pages 721–723. DOI: 10.1126/science.1167742. URL: https://pubmed.ncbi.nlm.nih.gov/19197046/ (cited on page 3).

[0] Gustave Le Bon. *The Crowd a Study of the Popular Mind*. BEYOND BOOKS HUB, 2023 (cited on page 2).

[0] James Manyika. *Big data: The next frontier for innovation, competition, and productivity*. May 2011. URL: http://abesit.in/wp-content/uploads/2014/07/big-data-frontier.pdf (cited on page 2).

[0] Robert M. May. *Stability and complexity in model ecosystems*. Dec. 2001. DOI: 10.1515/9780691206912. URL: https://doi.org/10.1515/9780691206912 (cited on page 1).

[0] Kevin Murphy. *Machine Learning : A Probabilistic Perspective*. Aug. 2012. URL: http://cds.cern.ch/record/1981503 (cited on page 2).

[0] M. E. J. Newman. "The structure of scientific collaboration networks". In: *Proceedings of the National Academy of Sciences of the United States of America* 98.2 (Jan. 2001), pages 404–409. DOI: 10.1073/pnas.98.2.404. URL: https://doi.org/10.1073/pnas.98.2.404 (cited on page 2).

[0] Mark Newman. *Networks: An Introduction*. Oxford University Press, Mar. 2010. ISBN: 9780199206650. DOI: 10.1093/acprof:oso/9780199206650.001.0001. URL: https://doi.org/10.1093/acprof:oso/9780199206650.001.0001 (cited on page 1).

[0] Lars Onsager. "Crystal Statistics. I. A Two-Dimensional Model with an Order-Disorder Transition". In: *Physical review* 65.3-4 (Feb. 1944), pages 117–149. DOI: 10.1103/physrev.65.117. URL: https://doi.org/10.1103/physrev.65.117 (cited on page 1).

[0] Arkady Pikovsky et al. "A universal concept in nonlinear sciences". In: *Self* 2 (2001), page 3 (cited on page 1).

[0] Ilya Prigogine and Gregoire Nicolis. "Self-organization in Non equilibrium Systems". In: *Non-equilibrium system* (1977) (cited on page 1).

[0] Filippo Radicchi, Santo Fortunato, and Claudio Castellano. "Universality of citation distributions: Toward an objective measure of scientific impact". In: *Proceedings of the National Academy of Sciences of the United States of America* 105.45 (Nov. 2008), pages 17268–17272. DOI: 10.1073/pnas.0806977105. URL: https://doi.org/10.1073/pnas.0806977105 (cited on page 2).

[0] John P. Roche and John H. Rappole. "The Ecology of Migrant Birds: A Neotropical Perspective." In: *Ecology* 78.1 (Jan. 1997), page 328. DOI: 10.2307/2266005. URL: https://doi.org/10.2307/2266005 (cited on page 1).

[0] Everett M Rogers, Arvind Singhal, and Margaret M Quinlan. "Diffusion of innovations". In: *An integrated approach to communication theory and research*. Routledge, 2014, pages 432–448 (cited on page 3).

[0] Cynthia Rudin. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". In: *Nature machine intelligence* 1.5 (May 2019), pages 206–215. DOI: 10.1038/s42256-019-0048-x. URL: https://doi.org/10.1038/s42256-019-0048-x (cited on page 2).

[0] Olaf Sporns and Jonathan D. Zwi. "The small world of the cerebral cortex". In: *Neuroinformatics* 2.2 (Jan. 2004), pages 145–162. DOI: 10.1385/ni:2:2:145. URL: https://doi.org/10.1385/ni:2:2:145 (cited on page 2).

[0] H Eugene Stanley. *Phase transitions and critical phenomena*. Volume 7. Clarendon Press, Oxford, 1971 (cited on page 1).

[0] Dietrich Stauffer and Amnon Aharony. *Introduction to percolation Theory*. Jan. 1985. DOI: 10.4324/9780203211595. URL: https://doi.org/10.4324/9780203211595 (cited on page 1).

[0] *The Nobel Prize in Physics 2021*. URL: https://www.nobelprize.org/prizes/physics/2021/press-release/ (cited on page 1).

[0] Thomas W. Valente. "Social network thresholds in the diffusion of innovations". In: *Social Networks* 18.1 (1996), pages 69–89. ISSN: 0378-8733. DOI: https://doi.org/10.1016/0378-8733(95)00256-1. URL: https://www.sciencedirect.com/science/article/pii/0378873395002561 (cited on page 3).

[0] Soroush Vosoughi, Deb Roy, and Sinan Aral. "The spread of true and false news online". In: *Science* 359.6380 (Mar. 2018), pages 1146–1151. DOI: 10.1126/science.aap9559. URL: https://doi.org/10.1126/science.aap9559 (cited on page 1).

[0] D. J. Watts. “A simple model of global cascades on random networks”. In: *Proceedings of the National Academy of Sciences* 99.9 (2002), pages 5766–5771. DOI: 10.1073/pnas.082090499 (cited on page 3).

[0] Duncan J. Watts. “A twenty-first century science”. In: *Nature* 445.7127 (Jan. 2007), page 489. DOI: 10.1038/445489a. URL: https://doi.org/10.1038/445489a (cited on page 2).

A. Vacancy density effect on the Schelling model dynamics

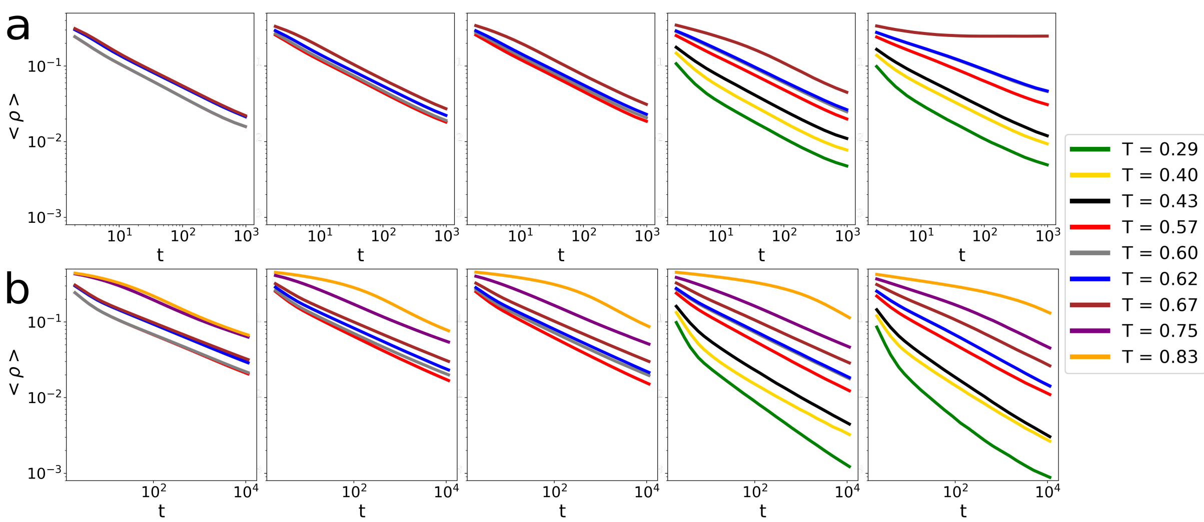


Figure A.1: Average interface density $\langle\rho(t)\rangle$ as a function of time steps for different values of the tolerance parameter T fr the Schelling model (**a**) and the version with aging (**b**). The different plots show a the evolution at a different value of the vacancy density, increasing from left to right $\rho_v = 0.005$, 0.15, 0.2, 0.3 and 0.45. Average performed over 10^3 realisations with system size $100\ x\ 100$.

Since we restrain ourselves to the region $\rho_v < 0.5$, the increase/decrease of the number of vacancies does not change dramatically the behaviour. Above this value, we approach the segregated-dilute transition ($\rho_v \sim 0.62$). Nevertheless, it is worth to mention a few features we observe on the coarsening dynamics. Essentially, when we set a higher vacancy density, the number of agents which see vacancies at their surroundings increases. This results in a family of similar power-law decays towards the segregated state for every meaningful value of T (see Fig. A.1).

Moreover, a higher ρ_v allows us to study the coarsening phenomena for lower values of T according to the phase diagram for the original Schelling model. For those particular cases, when the aging is introduced, we observe a power law decay faster than without aging (Fig. A.1b). Therefore, the aging effect accelerates segregation in this region of the phase diagram, contrary as for lower values of ρ_v. This acceleration is not caused by reaching the 2-clusters state in less time. Since there is a large presence of vacancies, aging causes a formation of vacancy

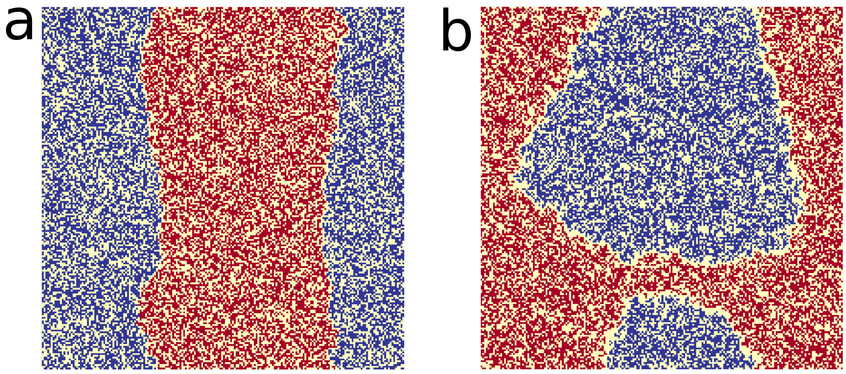


Figure A.2: Snapshots of the system at the final segregated state (after 10^6 MC steps) for the Schelling model (**a**) and the version with (**b**). System size $200\ x\ 200$ with $\rho_v = 0.45$ and $T = 0.29$.

clusters at the interface. Fig. A.2 shows the final segregated state with and without aging. This spontaneous behaviour is result of the low tolerance combined with the persistence of clusters (once formed) due to aging effect and the large number of vacancies that allows the possibility of the formation of clusters at the interface.

In order to quantify this vacancy cluster formation, we define a measure inspired in the segregation coefficient:

$$s_v = \frac{1}{(L^2\rho_v)^2}\sum_{\{c\}} n_c^2 \tag{A.1}$$

where c is the size of a vacancy cluster and n_c is the number of clusters with size c. The sample average of s_v after reaching equilibrium is called the cluster coefficient of vacancies $\langle s_v \rangle$.

The results of this measure as a function of ρ_v for a few values of T are represented in Fig.A.3 for the Schelling model with and without aging. We observe an increasing dependence of $\langle s_v \rangle$ with ρ_v for both models, but the effect reducing tolerance changes dramatically the behaviour for the case with aging, highlighting the vacancy cluster formation.

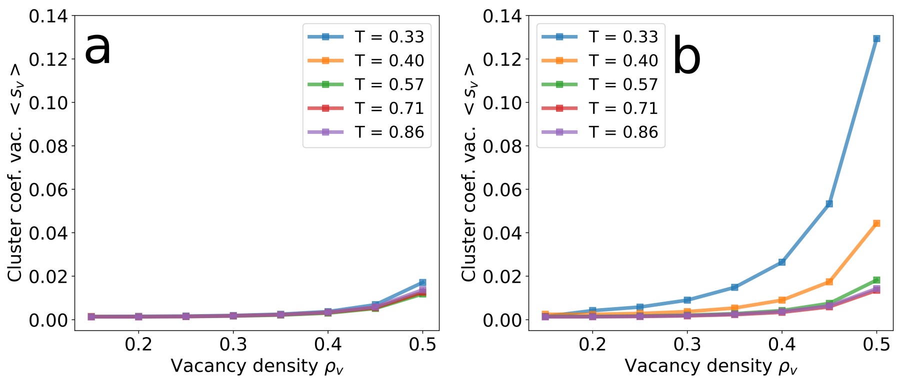


Figure A.3: Cluster coefficient of vacancies as a function of the vacancy density ρ_v for the Schelling model (**a**) and the version with (**b**) for different values of the tolerance T.

B. Heterogeneous mean-field taking into account aging (HMFA)

Setting the time derivatives to 0 in Eqs. (**??**), we obtain the relations for the stationary state:

$$x_{k,0}^{\pm} = \sum_{j=0}^{\infty} x_{k,j}^{\mp}\, \omega_{k,j}^{\mp}, \qquad x_{k,j}^{\pm} = x_{k,j-1}^{\pm}\,(1-\omega_{k,j-1}^{\pm}) \qquad j>0, \tag{B.1}$$

from where we extract the stationary condition $x_{k,0}^{-} = x_{k,0}^{+}$, as in Ref. (0). Notice that by setting $p_A(j) = 1$ and summing over all ages j, we recover the HMF approximation (Eq. **??**) for the model without aging. Defining $x_j^{\pm}(t)$ as the fraction of agents in state ± 1 with age j:

$$x_j^{\pm} = \sum_k p_k\, x_{k,j}^{\pm}, \tag{B.2}$$

and using the degree distribution of a complete graph $p_k = \delta(k-N+1)$ (where $\delta(\cdot)$ is the Dirac delta), we sum over the variable k and rewrite Eq. (B.1) in terms of $x_j^{\pm}$:

$$x_0^{\pm} = \sum_{j=0}^{\infty} x_j^{\mp}\, \omega_j^{\mp}, \qquad x_j^{\pm} = x_{j-1}^{\pm}\,(1-\omega_{j-1}^{\pm}) \qquad j>0, \tag{B.3}$$

where $\omega_j^{\pm} \equiv \omega_{N-1,j}^{\pm}$. Note that the stationary condition $x_0^{-} = x_0^{+}$ remains valid after summing over the degree variable. We compute the solution $x_j^{\pm}$ recursively as a function of $x_0^{\pm}$:

$$x_j^{\pm} = x_0^{\pm}\, F_j^{\pm} \qquad \text{where} \qquad F_j^{\pm} = \prod_{a=0}^{j-1} (1-\omega_a^{\pm}), \tag{B.4}$$

and summing all j,

$$x^{\pm} = x_0^{\pm}\, F^{\pm} \qquad \text{where} \qquad F^{\pm} = 1 + \sum_{j=1}^{\infty} F_j^{\pm}. \tag{B.5}$$

Using the stationary condition $x_0^{-} = x_0^{+}$, we reach:

$$\frac{x^{+}}{x^{-}} = \frac{F^{+}}{F^{-}}. \tag{B.6}$$

Notice that, for the complete graph, $\tilde{x}^{+} = x$, $\tilde{x}^{-} = 1-x$. Therefore, $F^{\pm}$ is a function of the variable $x^{\mp}$ ($F^{+} = F(1-x)$). Thus, we rewrite the previous expression just in terms of the variable x:

$$\frac{x}{1-x} = \frac{F(1-x)}{F(x)}. \tag{B.7}$$

C. Internal time recursive relation in Phase $\mathbf{I}/\mathbf{I}^*$

In Phase I and I^*, the exceeding threshold condition ($m/k > T$) is full-filled for almost all agents in the system. Thus, agents will change their state and reset the internal time once activated. For the original model, all agents are activated once in a time step on average, but for the model with aging, the activation probability plays an important role. We consider here a set of N agents that are activated randomly with an activation probability $p_A(j)$ and, once activated, they reset their internal time. Being $n_i(t)$ the fraction of agents with internal time i at the time step t, we build a recursive relation for the previously described dynamics in terms of variables i and t:

$$n_1(t) = \sum_{i=1}^{t-1} p_A(i) n_i(t-1) \qquad n_i(t) = (1 - p_A(i-1)) n_{i-1}(t-1) \qquad i > 1. \tag{C.1}$$

This recursion relation can be solved numerically from the initial condition ($n_1(0) = 1$, $n_i(0) = 0$ for $i > 1$). To obtain the mean internal time at time t, we just need to compute the following:

$$\bar{\tau}(t) = \sum_{i=1}^{t} i n_i(t). \tag{C.2}$$

The solution from this recursive relation describes the mean internal time dynamics with great agreement with the numerical simulations performed at Phase I (for the complete graph) and Phase I^* (for the Erdős-Rényi and Moore lattice).