Understanding the complex dynamics of social systems with diverse formal tools

Jordan Cambe

Numéro National de Thèse : 2019LYSEN043

THESE de DOCTORAT DE L'UNIVERSITE DE LYON
opérée par
l'Ecole Normale Supérieure de Lyon

Ecole Doctorale N° 52
Physique et Astrophysique de Lyon

Discipline : Physique

Soutenue publiquement le 26/09/2019, par :
Jordan CAMBE

Understanding the complex dynamics of social systems with diverse formal tools

-

Comprendre les dynamiques complexes des systèmes sociaux à l'aide de divers outils formels

---

Devant le jury composé de :

BARRAT, Alain	Professeur	CPT Marseille	Rapporteur
ROBARDET, Céline	Professeure	INSA de Lyon	Rapporteure
KARSAI, Márton	MCF HDR	ENS de Lyon	Examinateur
MASCOLO, Cecilia	Professeure	Université de Cambridge	Examinatrice
JENSEN, Pablo	Professeur	ENS de Lyon	Directeur de thèse
MERCKLÉ, Pierre	Professeur	Université Grenoble Alpes	Co-encadrant

Si nous prenons la nature pour guide
Nous ne nous égarerons jamais.
— Cicéron

To my family,

Acknowledgements

I would like to start this section by expressing how grateful I am to my supervisors, Pablo Jensen and Pierre Mercklé, for giving me guidance through my PhD. A PhD is a marathon with a fair amount of sprints and I am glad they have always been present to share their experience to me.

I would like to thank as well Professor Cecilia Mascolo and her team who welcomed me in Cambridge University. I have learnt so much from our collaboration, both personally and professionally, that words can hardly translate my gratitude to this amazing team.

Finally, I would like to thank my family, friends and colleagues for their constant support during my study and more particularly this PhD. They have been the invigorating breeze preventing me from ever getting out of breath.

Lyon, August 5th, 2019 J. C.

Abstract

For the past two decades, electronic devices have revolutionized the traceability of social phenomena. Social dynamics now leave numerical footprints, which can be analyzed to better understand collective behaviors. The development of large online social networks (like Facebook, Twitter and more generally mobile communications) and connected physical structures (like transportation networks and geolocalised social platforms) resulted in the emergence of large longitudinal datasets. These new datasets bring the opportunity to develop new methods to analyze temporal dynamics in and of these systems.

Nowadays, the plurality of data available requires to adapt and combine a plurality of existing methods in order to enlarge the global vision that one has on such complex systems. The purpose of this thesis is to explore the dynamics of social systems using three sets of tools: *network science*, *statistical physics modeling* and *machine learning*. This thesis starts by giving general definitions and some historical context on the methods mentioned above. After that, we show the complex dynamics induced by introducing an infinitesimal quantity of new agents to a Schelling-like model and discuss the limitations of statistical model simulation. The third chapter shows the added value of using longitudinal data. We study the behavior evolution of bike sharing system users and analyze the results of an unsupervised machine learning model aiming to classify users based on their profiles. The fourth chapter explores the differences between global and local methods for temporal community detection using scientometric networks. The last chapter merges complex network analysis and supervised machine learning in order to describe and predict the impact of new businesses on already established ones. We explore the temporal evolution of this impact and show the benefit of combining networks topology measures with machine learning algorithms.

Résumé

Au cours des deux dernières décennies les objets connectés ont révolutionné la traçabilité des phénomènes sociaux. Les trajectoires sociales laissent aujourd'hui des traces numériques, qui peuvent être analysées pour obtenir une compréhension plus profonde des comportements collectifs. L'essor de grands réseaux sociaux (comme Facebook, Twitter et plus généralement les réseaux de communication mobile) et d'infrastructures connectées (comme les réseaux de transports publiques et les plate-formes en ligne géolocalisées) ont permis la constitution de grands jeux de données temporelles. Ces nouveaux jeux de données nous donnent l'occasion de développer de nouvelles méthodes pour analyser les dynamiques temporelles *de* et *dans* ces systèmes.

De nos jours, la pluralité des données nécessite d'adapter et combiner une pluralité de méthodes déjà existantes pour élargir la vision globale que l'on a de ces systèmes complexes. Le but de cette thèse est d'explorer les dynamiques des systèmes sociaux au moyen de trois groupes d'outils : les réseaux complexes, la physique statistique et l'apprentissage automatique. Dans cette thèse je commencerai par donner quelques définitions générales et un contexte historique des méthodes mentionnées ci-dessus. Après quoi, nous montrerons la dynamique complexe d'un modèle de Schelling suite à l'introduction d'une quantité infinitésimale de nouveaux agents et discuterons des limites des modèles statistiques. Le troisième chapitre montre la valeur ajoutée de l'utilisation de jeux de données temporelles. Nous étudions l'évolution du comportement des utilisateurs d'un réseau de vélos en libre-service. Puis, nous analysons les résultats d'un algorithme d'apprentissage automatique non supervisé ayant pour but de classer les utilisateurs en fonction de leurs profils. Le quatrième chapitre explore les différences entre une méthode globale et une méthode locale de détection de communautés temporelles sur des réseaux scientométriques. Le dernier chapitre combine l'analyse de réseaux complexes et l'apprentissage automatique supervisé pour décrire et prédire l'impact de l'introduction de nouveaux commerces sur les commerces existants. Nous explorons l'évolution temporelle de l'impact et montrons le bénéfice de l'utilisation de mesures de topologies de réseaux avec des algorithmes d'apprentissage automatique.

Contents

List of Figures

List of Tables

1 Introduction

The amount of available data has been continuously increasing for the past two decades. Due to the dramatic increase of computer power, storage capacity and the ubiquity of connected devices (smartphones, Internet Of Things) large datasets have emerged with a fine-grained description of complex systems. I will for the rest of this thesis use the definition of complex systems from [Barrat et al., 2008a]:

> "complex systems consist of a large number of elements capable of interacting with each other and their environment in order to organize in specific emergent structures"

Such systems have fascinating properties regarding their dynamics. One can think of systems where the overall dynamics is more than the dynamics of its units. Whereas other systems can exhibit a steady state architecture and properties despite the continuous changes of its units. These examples illustrate the need to develop different approaches to think the emergent dynamics of complex systems.

Complex systems can be found in very wide range of areas: (1) in biology (proteins and genes interact to form and regulate cells activities, see [Barrat et al., 2008b, Kovács et al., 2019]); (2) in ecology (food webs can be represented as a complex network of species, see [Caldarelli et al., 2003]); (3) in finance (the world financial market is a network of banks and institutions trading assets [Schweitzer et al., 2009, Caccioli et al., 2014]); and more closely related to the subject of this thesis in sociology and economics. We will take a deeper look at these in section 1.1.3.

In parallel with this increase in available data, methods and fields of research have been developing accordingly to process these data and help understand the complex nature of real-world problems with data, see [Donoho, 2017, Liao et al., 2012, Sagiroglu and Sinanc, 2013].

It has been now around twenty years that the data journey has started leading to the relatively recent constitution of large longitudinal datasets taking into account one more dimension: *time*. Many systems have an inherent temporal component and there is nowadays a need for new tools in order to get an understanding of the temporal dynamics of such complex systems. In order to get a global understanding of complex systems problems, it is necessary to tackle them from different angles. Therefore, the necessity to accumulate methods from different fields of research and combine them is

critical. The aim of this thesis is to explore the plurality of approaches available and merge them into new methodologies.

The tools used in this thesis will focus on network science, machine learning, and statistical physics modeling. In the following paragraphs, I will present the fields of research I explored before developing my contributions.

1.1 An Overview of Network Science

1.1.1 Static Complex Networks

Let us first speak about what network science is and what are the motivations of this field of research. Network science aims to model interactions between different units of a system using graph theory. It was first used for information technology purposes - like the mapping the World Wide Web (WWW) in order to optimize search engines. Then, network sciences appeared to be efficient enough to model a huge amount of complex systems belonging to the field of biology, ecology, finance, and so on, as seen previously.

The principle of complex networks study is to understand the global system behavior by describing the interactions between its units. Formally a network can be represented as a graph G. This graph is composed of a set of nodes N and a set of edges E linking the nodes to each other. This formalism makes it easy to visualize and measure interactions between the units of the network as the interactions are coded in the topology of the graph. In order to understand the topology of the system, one can look at measures describing the state of connectivity of its units. Some of the most common measures are degree-related measures, clustering, centrality measures (e.g. closeness, betweenness, etc), modularity. We will come to a formal definition of most of these measures in the core of this thesis. More details on complex networks measures can be found in [Newman, 2010].

1.1.2 Temporal Complex Networks

As discussed earlier, the appearance of temporal data led to the evolution of network science to include the new area of temporal networks [Holme and Saramäki, 2012]. Some real-world networks are by nature temporal, these are networks where nodes can (dis)appear over time or where edges are not continuously active. For example, communication networks (i.e. e-mails, text messages, and phone calls) are temporal networks. Cities are temporal networks as well and one can study and visualize the flow of people in them [Roth et al., 2011]. A temporal network is a union of static network snapshots at each time t. Hence, the temporal network G_T is composed of the set of nodes $N_T = \bigcup_{t=1}^{T} N_t$ and set of edges $E_T = \bigcup_{t=1}^{T} E_t$. Considering a sequence of interactions, we can construct the corresponding temporal graph; or a static aggregated graph, where an edge between two nodes accounts for all the interactions which happened between the nodes during the time window of observation.

Using a temporal graph rather than an aggregated one can be useful when we need to take into account the order of interactions (like in the case of modeling spreading

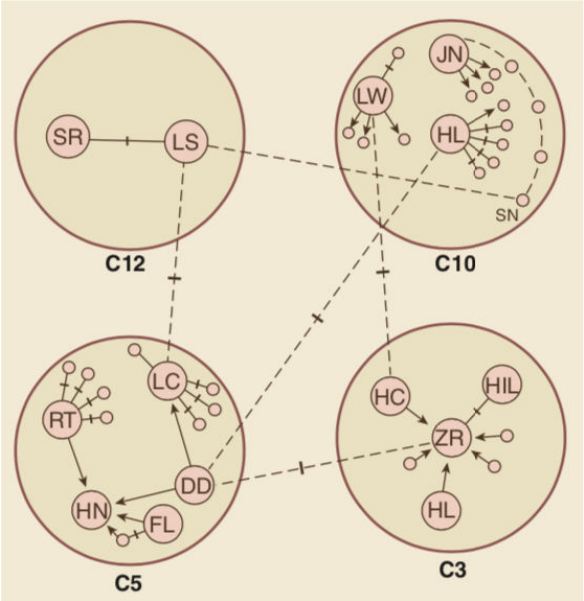


Figure 1.1 – Moreno's network of runaways, taken from [Borgatti et al., 2009]. Circles C12, C10, C5 and C3 represent the cottages in which the girls lived. Circles within the cottages represent girls and the 14 runaways are identified by their initials. Undirected edges between two girls represent feelings of mutual attraction, whereas directed edges represent one-way feelings of attraction. From this network, Moreno visualized the flow of social influence among the girls and argued that their location in the social network was determining when they ran away.

phenomena [Karsai and Perra, 2017]). In this case the temporal path matters as it does not result in the same people being informed/infected.

In the next section, I give some historical context of social and urban systems.

1.1.3 Networks in Social Sciences

In social science, networks are used to model interactions between people and collective behaviors. The use of social networks started in the early twentieth century when J. Moreno and H. Jennings worked on the epidemic of runaways at the Hudson School for Girls, NY, USA. They, for the first time, used *sociometry* to model relationships between the runaway girls [Moreno, 1934]. Figure 1.1 shows one of the first sociograms published by Moreno in 1934. They represent interactions between girls in the case of the Hudson School runaways mentioned earlier in this paragraph. This work falls within the movement of *social physics* initiated by A. Comte [Comte and Martineau, 1856]. This movement envisioned a description of sociology following natural sciences paradigms. Hence terms such as 'social atoms' and 'social gravitation' have emerged during that period.

Following the work of Moreno, *social networks* research continued growing. In the fifties, M. Kochen and de Sola Pool [de Sola Pool and Kochen, 1978] formulated the *small world* problem, which would then become the subject of the well-known *Milgram*

experiment in 1967 [Milgram, 1967] and finally was modeled by Watts and Strogatz in 1998 [Watts and Strogatz, 1998]. The *small-world phenomenon* is the observation that there exists a 'short path' connecting any two nodes in a network. Milgram, in his experiment, randomly sent packets to individuals living in two U.S. cities. Then these individuals needed to send the letter to a target person in Boston. If the randomly selected individuals did not know the target person, then they had to send the letter to someone who they thought would be more likely to know the target person. Every time a letter reached a new person, they must add their identity details to the letter, and repeat the process until reaching the target person. Over the 296 letters, 232 never reached destination [Milgram, 1967]. The remaining 64 letters eventually reached the target individual. The average path length, that is the number of intermediate people the letters went through, was around six. Later on, in 2008, small-world network properties were popularized in a Hollywood movie *Connected: The Power of Six Degrees* [Talas, 2008]. Nowadays, modern online social networks have usually smaller average path lengths: Facebook, 4.7 [Backstrom et al., 2012]; Twitter, 5.4 [Kwak et al., 2010]; MSN, 6.6 [Leskovec and Horvitz, 2008];

Back to the history of social networks, in the 60-70's research developed around community structures and their role in socio-economic levels of individuals [Bott, 1957] - notably with the theory of the influential strength of weak ties [Granovetter, 1973] and later the theory of social capital [Bourdieu, 1986, Putnam, 2000], acknowledging the relationships of people as one of their main resources.

After this short overview of social networks evolution, it is worth noting that for a long time the dichotomy between social science and computer science has constrained social scientists to limit their analysis to relatively small networks [Degenne and Forsé, 2004]. Fortunately, the development of large online social networks and connected devices led computer scientists to develop tools to treat and visualize large amounts of data. Following the foundations of the twentieth century, the last twenty years have seen the emergence of new social systems. I will divide datasets into two groups:

- Online social networks: like Facebook, Twitter and more generally mobile communications. These data allow to describe interactions between people and to study recurring structures in the society. For example in [Leo et al., 2016] the authors studied the socio-economic class structure of Mexican society using (1) mobile phone data, (2) credit and purchase data and (3) location data. In another work [Kovanen et al., 2013], the authors studied the correlations between demographics (gender, age) and differences in communication patterns.

- mobility data: geolocalised social platforms such as foursquare, transportation systems (e.g. bike sharing systems [Fishman et al., 2013], plane network [Neal, 2014], etc, enable to model the flow of people at various scales (city, country) and the interactions between flows and events, new policies and so on. For example, in [Zhou et al., 2017] the authors study the impact of cultural investment policies in London neighborhoods using government data (socio-economic variables, cultural expenditures) and foursquare data. Another article, [Faghih-Imani et al., 2014], explores the relationship between land use (restaurants, facilities, etc) and

the flow of Montreal's bike sharing system.

After this overview on complex networks and social systems, I would like to introduce another story of research on social systems which was built in parallel with traditional social sciences.

1.2 Statistical Physics Models

> Statistical mechanics is the art of turning the microscopic laws of physics into a description of Nature on a macroscopic scale.

This definition is taken from [Tong, 2011]. Indeed the whole idea behind statistical physics is to find a way to describe from the microscopic properties of system components (atoms, particles, etc) the macroscopic evolution of the system itself. As its name suggests statistical physics uses methods from probability theory and statistics. More specifically, if one knows the states of atoms in a box, can one infer the state of the box content as a whole?

With this kind of framework, it is easy to understand why some statistical physicists went from the study of matter to describing complex systems from other fields of research such as biology [Peyrard, 2004], society [Castellano et al., 2009a] and economics [Jovanovic and Schinckus, 2013]. Therefore, over the past century, terms such as *socio-physics* and *econophysics* have emerged. Scientists have developed statistical models to understand concepts and model social phenomena.

For example, the Schelling model is an agent-based model developed by Nobel prize winner Thomas Schelling. In the model, each agent belongs to one of two groups (let us say 'green' and 'red') and aims to reside in a neighborhood with at least a small amount of agents with same color (not necessarily a majority). Schelling showed that when having this kind of personal (microscopic) rule, the system would converge to a segregated (macroscopic) state where red agents and green agents do not live in the same neighborhoods, despite their personal preferences. This model, which oversimplifies society does not aim to describe any real world phenomenon. This counterintuitive result illustrates the difficulty to infer microscopic states of system components (personal preferences, homophily, racism) from the macroscopic state of a system (segregation).

In chapter 2, we present a variant of the Schelling model studying the effect of mixing populations in such dynamical models and the resulting chaotic convergence tendency.

Finally, after the small social networks analysis and the mathematical toy models of the past century, the recent avalanche of data brings the hope of someday being able to model and predict social systems' evolution with an accuracy better than ever. In the next section, I give an overview of the more recent field of *Machine Learning*.

1.3 Machine Learning

Machine learning algorithms are being developed to analyze and make sense of large amounts of data. They have progressively conquered various fields of application, such as urban planning, natural language processing (text and speech), computer vision, market segmentation, to only name a few.

In the following, I describe only two subareas of machine learning, *supervised* and *unsupervised* learning. In this context, a machine learning model is implemented to predict the class of a data point (classification) or the value of a data point (regression). Developing such a model always requires to split the given data into two disjoint subsets: (1) the training data which are used for training the model and (2) the testing data which are used for testing that the model can actually predict the outcomes correctly.

In the case of *supervised learning* the classes/values that we try to predict are known. Hence, we can use this information during the training to learn the combinations of explanatory variables which 'explain' the class. For example, in the article [Todorova and Noulas, 2019], the authors identify spatio-temporal patterns in ambulance call activity and assess the health risk of a geographic area. They, then, build a supervised learning model (Random Forest classifier) to predict for a given time which regions need an ambulance.

In *unsupervised learning*, the labels/values that we want to predict are not known. Therefore, unsupervised learning algorithms try to detect clusters in the data. In the case of customer segmentation, segments are constructed from customer data such as demographic characteristics, past purchase and product-use behaviors. See for example [Venkatesan, 2018], where the author uses K-means algorithm to discuss the importance of customer segmentation in marketing.

When implementing a machine learning model all the difficulty lies in being able to fit the training data with good accuracy (not underfitting) while still having a good generalization of outcomes (not overfitting), that is predicting accurately labels/values using test data. This is known as the *Bias-Variance Tradeoff*.

In chapter 3, we discuss the use of an unsupervised machine learning model to classify users of a bike sharing system using temporal data. In chapter 5, we implement a binary classification supervised machine learning model in order to predict the impact of a new business on other businesses in its neighborhood.

1.4 Contributions and Chapter Outline

The purpose of this thesis is to explore the dynamics of social systems. The rest of this thesis will start with a statistical model in chapter 2. This chapter will show the extreme sensitivity of Schelling-like models to population composition and discuss the importance of seeing such models as what they are: toy models that help shape concepts to better think our society. However, they should not be intended to model

society.

In the first part of chapter 3, we study the behaviors evolution of the Vélo'v Lyon's bike sharing system (BSS) users over 5 years. We, then, discuss the results of an unsupervised machine learning model aiming to classify users based on their use profile. Using a methodology similar to a previous work [Vogel et al., 2014] analyzing a 1-year dataset, we show this study overestimated one class density due to the lack of temporal component.

Chapter 4 explores temporal community detection using two scientometric networks. It describes the differences between local and global approaches. It also discusses ways to evaluate temporal community detection methods. More particularly, it compares Mutual Information measures to measures based on a bipartite graph representation of the partitions.

Chapter 5 merges complex network analysis and supervised machine learning in order to describe and predict interactions between new businesses and already established businesses in a neighborhood. This chapter combines practices from both complex networks modeling and machine learning. We show that using networks topology measures as explanatory variables increases the predictive power of the algorithm.

1.5 List of PhD Publications

During my PhD, I have worked on the following four papers, one has been published and three are currently under review. The chapters of this thesis are based on these articles.

[1] J. Cambe, K. D'Silva, A. Noulas, C. Mascolo, A. Waksman, (2019) 'Modelling Cooperation and Competition in Urban Retail Ecosystems with Complex Network Metrics'.

[2] J. Cambe, S. Grauwin, P. Flandrin, P. Jensen (2019) 'Exploring and comparing temporal clustering methods'.

[3] J. Cambe, P. Abry, J. Barnier, P. Borgnat, M. Vogel, P. Jensen, (2019) 'Evolutions of Individuals Use of Lyon's Bike Sharing System', pre-print: https://arxiv.org/abs/1803.11505

[4] P. Jensen, T. Matreux, J. Cambe, H. Larralde, E. Bertin, (2018) 'Giant Catalytic Effect of Altruists in Schelling's Segregation Model', Phys. Rev. Lett. 120, 208301, URL: https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.120.208301

2 Complex Dynamics From a Simple Social Model

2.1 Introduction

As mentioned earlier, simple social models can be useful to improve our intuitive, conceptualizations of social processes [Castellano et al., 2009b, Watts, 2011, Jensen, 2018]. For example, the segregation model proposed by Schelling [Schelling, 1971] helps understanding that the collective state reached by agents may well be different from what each of them seeks individually. Specifically, Schelling's model shows that even when all agents share a preference for a mixed city, the macroscopic stationary state may be segregated [Grauwin et al., 2009]. In this thesis, we show that introducing a vanishingly small concentration of altruist agents gives rise to a strongly non linear response.

Our model combines two important themes for many disciplines, including physics and economics: The large effects of small perturbations and the influence of altruistic behavior on coordination problems. On the first point, microscopic causes leading to macroscopic effects are well-known in physics. Chaos theory has shown that some dynamical systems are prone to an exponential increase of small perturbations [Eckmann and Ruelle, 1985], a topic of recurring interest in other fields, such as modeling of ecological competition [Hébert-Dufresne et al., 2017] or pattern formation [Cross and Hohenberg, 1993]. More related to this chapter, there are several examples of large effects arising from small changes in population composition. It has been shown that a small variation in the proportion of uninformed individuals may lead to strong changes in the way collective consensus is achieved by animal groups manipulated by an opinionated minority [Couzin et al., 2011]. In the minority game [Challet and Zhang, 1997], introducing a small proportion of fixed agents - i.e. agents that always choose the same option - induces a global change in the population behavior, leading to an increase of the overall gain [Liaw and Liu, 2005, Liaw, 2009]. In the voter model, a finite density of voters that never change opinion can prevent consensus to be reached [Mobilia et al., 2007].

On the second point, altruism is a major topic in evolutionary biology and economics [Fehr and Gachter, 2002, Boyd et al., 2003, Kirman and Teschl, 2010]. Many models have shown that pair interactions between selfish players lead to stationary states of low utility. They have introduced various types of altruistic behavior to in-

vestigate how it may lead to a better equilibrium: altruistic punishment [Fehr and Gachter, 2002, Boyd et al., 2003], inequity aversion [Hetzer and Sornette, 2013], fraternal attitudes [Szabo et al., 2013], agent mobility[Cong et al., 2017] ... Here, we use a simple definition of altruism (see below) and concentrate on the proportion of altruists needed to reach the social optimum. We show that, unexpectedly, an infinitesimal proportion of altruists can coordinate a large number of egoists and allow the whole system to reach the social optimum.

2.2 Description of the model

Our model represents the movement of a population of agents in a "city", which is divided into $Q \gg 1$ non overlapping blocks, also called neighborhoods. Each block is divided into H sites and has the capacity to accommodate H agents (one per site). Initially, a number of agents $N = QH\rho_0$ are distributed randomly over the blocks, leading to an average block density ρ_0 ($\rho_0 = 0.4$ throughout the chapter). All agents share the same utility function $u(\rho)$ that depends on the agents density ρ in the neighborhood where they are located. We choose a triangular utility (see Fig. 2.1): agents experience zero utility if they are alone ($\rho = 0$) or in full blocks ($\rho = 1$), and maximum utility $u = 1$ in half-filled blocks ($\rho = 0.5$). The collective utility U is defined as the sum of all agents' utilities, $U = H \sum_{q=1}^{Q} \rho_q u(\rho_q)$ and the average utility $\tilde{u}$ per agent is $\tilde{u} = U/N$.

Building upon past work on Schelling's segregation model [Grauwin et al., 2009], we now mix two types of agents: "egoists", who act to improve their own, individual, utility, and a fraction p of "altruists", who act to improve the collective utility. Thus, egoists have as objective function the variation of their individual utility Δu, while altruists consider the variation of the overall utility ΔU. The dynamics is the following: at each time step, an agent and a free site in another block are selected at random. The agent accepts to move to this new site only if its objective function *strictly* increases (note that the moving agent is taken into account to compute the density of the new block). Otherwise, it stays in its present block. Then, another agent and another empty site are chosen at random, and the same process is repeated until a stationary state is reached, i.e., until there are no possible moves for any agent.

2.3 Limiting cases: pure egoist or altruist populations

Authors in [Grauwin et al., 2009] computed analytically the stationary states of a *homogeneous* population of egoist or altruist agents. Altruists always reach the optimal state, given by half filled (or empty) blocks and an average pure altruist utility $\tilde{u}_A \simeq 1$. In contrast, a pure egoist population *collectively* maximizes not U but an effective free energy that we have called the *link* L. The link is given by the sum over all blocks q of a potential l_q: $L = \sum_q l_q$, where $l_q = \sum_{n_q=0}^{N_q} u(n_q/H)$, with $N_q = H\rho_q$ is the total number of agents in block q. In the large H limit,

$$l(\rho_q) \approx H \int_0^{\rho_q} u(\rho)\, d\rho. \tag{2.1}$$

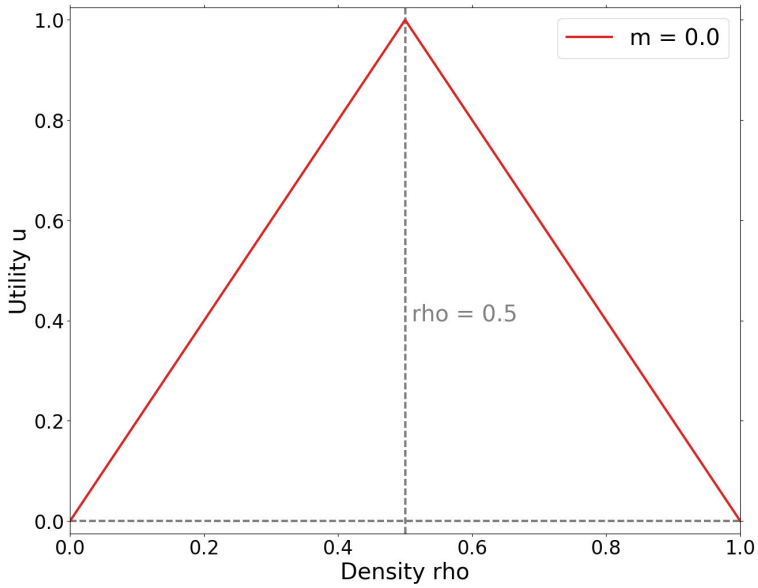
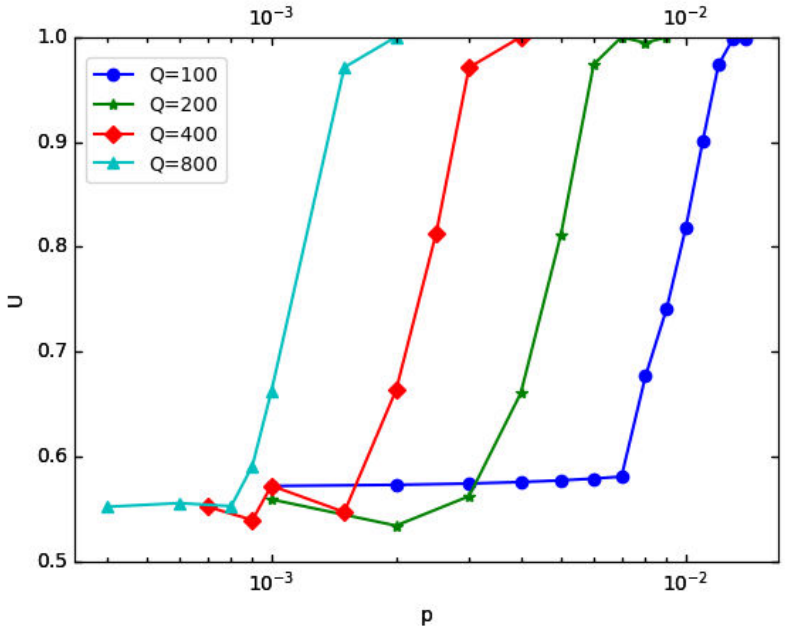


Figure 2.1 – Agent utility function: $u(\rho) = 2\rho$ for $\rho \leq 0.5$ and $u(\rho) = 2(1-\rho)$ for $\rho > 0.5$.

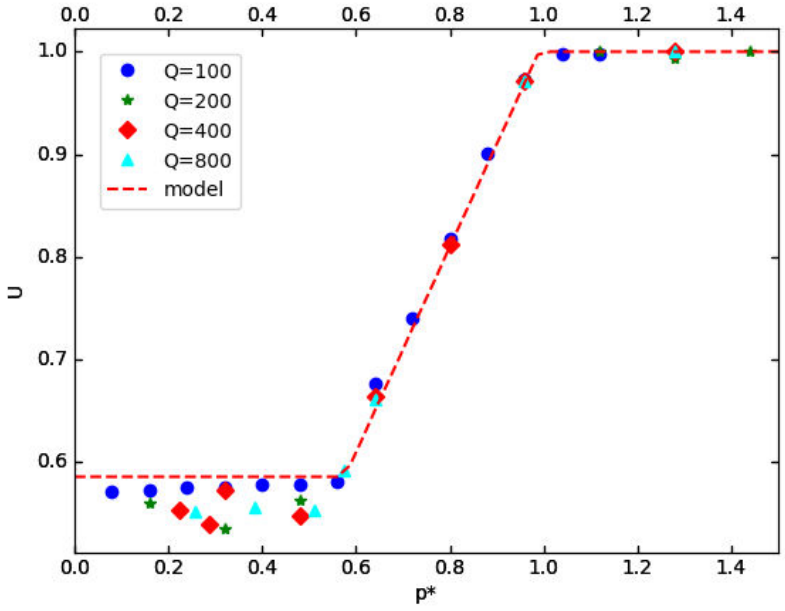
The link may be interpreted as the cumulative of the individual marginal utilities gained by agents, as they progressively enter the blocks from a reservoir of zero utility. Its key property is that, *for any move*, $\Delta L = \Delta u$. Since egoists move only when their individual Δu is positive, the stationary state is given by maximizing L over all possible densities $\{\rho_q\}$ of the blocks, from which no further $\Delta u > 0$ can be found. Analytical calculations [Grauwin et al., 2009] show that this stationary state corresponds to crowded neighborhoods, far above the state of maximum average utility given by $\rho_q = 1/2$. For the case studied in this chapter, the stationary density is given by $\rho_E = 1/\sqrt{2}$, leading to a pure egoist utility $\tilde{u}_E = 2(1-\rho_E) \simeq 0.586 \ll 1$. Numerical simulations have confirmed these results, though the existence of many metastable states around $\rho_E \simeq 0.7$ leads to fluctuations in the simulated final densities.

2.4 Mixing populations: qualitative picture

We now investigate how adding a fraction of altruists drives the system away from the frustrated pure egoist case to the optimal configuration observed in the pure altruist case. We find that, instead of a linear response, the system reaches the optimal state even at very low altruist concentrations ($p < 0.01$ in figure 2.2 a). To help understanding the origin of this strongly non-linear effect, the different panels of Fig. 2.3 illustrate the evolution of a small system ($H = 225$, $Q = 36$ and $p = 0.04$). Initially, altruists (yellow) and egoists (red) are distributed randomly in the blocks (a), which all have a density $\rho \simeq \rho_0 = 0.4$. Then, blocks with the lowest densities are depleted by both altruists and egoists that prefer districts with higher densities. At some point, when the block density increases, the behavior of the two kinds of agents diverge. Altruists "sacrifice" themselves and leave these high density blocks, moving to blocks with lower densities, as this increases the utility of their many (former) neighbors, leading to an



(a)



(b)

Figure 2.2 – Evolution of the average utility as a function of (a) the altruists' fraction p (note the log scale on the x-axis) and (b) the rescaled fraction $p^* = 2pQ\rho_0$. We take $H = 200$ and vary Q as shown. The fluctuations for low p^* values (before the transition) arise from metastable states in the pure egoist regime.

increase in global utility. On the other hand, egoists would loose individual utility by doing so, and therefore remain in these high density blocks which continue to feed on the remaining neighborhoods with $\rho < 1/2$. After a few iterations (Fig. 2.3b-c), selfish agents have gathered into "segregated" neighborhoods. This is the classical segregation observed in the pure egoist case [Grauwin et al., 2009], arising from the well studied amplification of density fluctuations. Note that all altruists have left the egoist blocks and gather into few blocks with lower densities (Fig. 2.3c) and then into a single neighborhood, whose density increases until it becomes attractive for egoist agents who "invade" it (Fig. 2.3d-e), while altruists leave it for other lower density blocks (Fig. 2.3e). The density of some of these new blocks then increases, allowing for successive egoist invasions (Fig. 2.3f-g). These migrations of egoist agents reduce the density of the overcrowded egoist blocks, increasing the overall utility. Eventually, the system reaches a stationary state in which no agent can move to increase its objective function (Fig. 2.3h).

2.5 Quantitative description

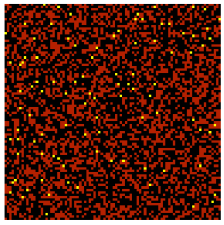
We now give a quantitative explanation of the decrease of egoist block densities and show that an altruist concentration $p \simeq 1/Q$ is sufficient to drive the system towards the optimal state, $\tilde{u} = 1$. To understand altruists' dynamics, it is useful to replace their dynamics by an equivalent egoist dynamics with a utility $u_{\text{altr}}(\rho)$ that differs from the original utility $u(\rho)$. An exact mapping can be done in the following way. As mentioned above, each altruist agent tries to maximize the global utility $U = H \sum_{q=1}^{Q} \rho_q u(\rho_q)$. In contrast, an egoist agent acts to maximize the link function $L = \sum_q \ell(\rho_q)$, with $\ell(\rho_q)$ given in Eq. (2.1). As a result, an altruist agent exactly behaves as an equivalent egoist agent with a utility function $u_{\text{altr}}(\rho)$ satisfying the relation

$$\rho u(\rho) = \int_0^\rho u_{\text{altr}}(\rho')\, d\rho' \tag{2.2}$$

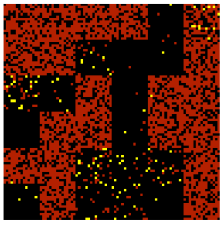
since the resulting function to be maximized is the same. Differentiating this last equation, one finds

$$u_{\text{altr}}(\rho) = \frac{\partial\big(\rho\, u(\rho)\big)}{\partial \rho} = \begin{cases} 4\rho, & \text{for } \rho \leq \frac{1}{2} \\ 2(1-2\rho), & \text{for } \rho > \frac{1}{2} \end{cases} \tag{2.3}$$

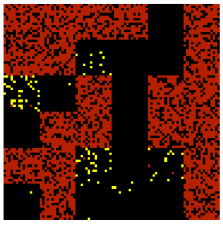
This effective utility function for altruists is plotted on Fig. 2.4. Note that this effective utility is *not* the one used to compute average or global utilities, it only helps understanding altruists' moves, since an altruist moves to a new block only if $u_{\text{altr}}(\rho)$ increases. Fig. 2.4 shows that altruists have a clear preference for blocks with densities just below $1/2$. The large discontinuity at $\rho = 1/2$ arises because at this density the original utility function $u(\rho)$ changes slope and starts to decrease. Then, an altruist moving from a block with $\rho < 1/2$ to a slightly more populated one with $\rho > 1/2$ induces a large decrease of total utility, since all its former neighbors loose utility (as the density of the initial block decreases) and so do its new neighbors, as the density of their block increases.



(a) t = 0



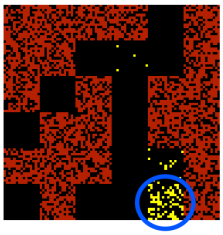
(b) t = 2



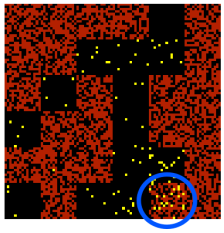
(c) t = 5

(d) t = 10

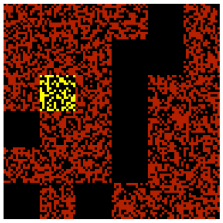
(e) t = 12



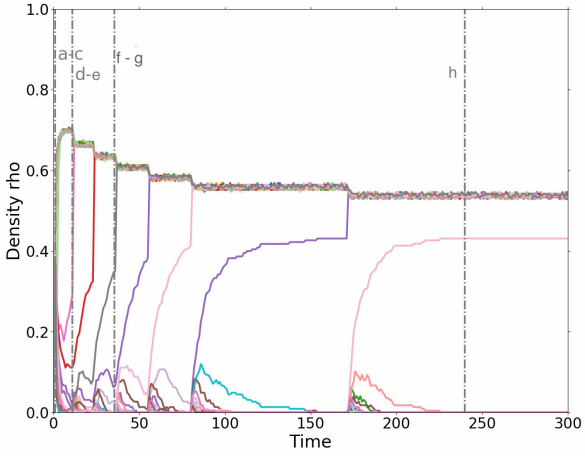
(f) t = 35



(g) t = 36



(h) t = 240



(i) Density profile

Figure 2.3 – Evolution of the city for $p = 0.03, Q = 36$ and $H = 225$. Panels (a-h) show the occupation of the different neighborhoods at different times. Egoists are represented in red, altruists in yellow, empty sites in black. (a) initial; (b) first steps; (c) usual segregation; (d-e): first invasion and altruist escape from the block surrounded in blue; (f-g): final invasion of the block surrounded in blue; (h): stationary state. In panel (i), each continuous line represents the evolution of the density of a single neighborhood. Vertical dashed lines show the times corresponding to panels (a-h).

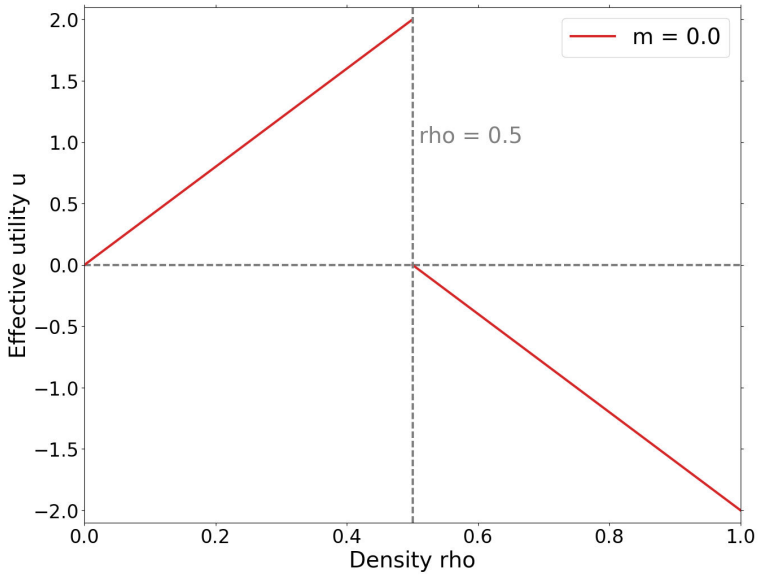


Figure 2.4 – *Effective* utility function of altruistic agents.

Fig. 2.2b suggests that the transition towards the optimal state is continuous and takes place at an altruist concentration $p \simeq 1/Q$ for all values of Q. This Q dependence is important, since in the thermodynamic limit ($Q \to \infty$), the transition would take place at $p \to 0$. We now derive this result in a simple way by computing analytically the evolution of the average utility as a function of the altruist concentration p. Let's start with very low altruist concentrations and assume that the initial dynamics is dominated by egoists, which form the usual Schelling's overcrowded blocks, as observed above (Fig. 2.3c) and in previous work [Grauwin et al., 2009]. Therefore, we take as starting point a city composed of n_E egoist blocks with uniform density $\rho_e = \rho_E > 1/2$, such that $\rho_E = (1-p)Q\rho_0/n_E$. Taking a uniform ρ_e value is justified because any density fluctuation for $\rho_e > 1/2$ is rapidly wiped out by the dynamics, as shown by the unique density of egoist blocks in Fig. 2.3i. Altruists can be initially somewhat scattered over the remaining blocks but, as their effective utility clearly shows (Fig. 2.4), they rapidly aggregate into a single block, leading to an altruist density $\rho_a = pQ\rho_0$ provided $\rho_a < 1/2$, or equivalently

$$p < p_{\text{high}} \equiv \frac{1}{2\rho_0 Q}\,. \tag{2.4}$$

The driving force for the transition are the relative values of agents' utilities in egoist and altruist blocks, respectively $u_e = u_E = 2 - 2\rho_E$ and $u_a = 2\rho_a$ since $\rho_a < 1/2$ and $\rho_E > 1/2$. For very low p values, ρ_a is small, leading to $u_e > u_a$ and the system remains in the usual frustrated Schelling egoist state $\tilde{u}(p) \simeq \tilde{u}_E$ which is essentially constant. When p reaches a value p_{low} such that $u(\rho_a + 1/H) > u(\rho_E)$, a first egoist can improve its utility by moving into the altruist block, whose density becomes $\rho_a + 1/H$ (Fig. 2.3d-f). This gives :

$$p_{\text{low}} \equiv \frac{1 - \rho_E - 1/H}{\rho_0 Q}\,, \tag{2.5}$$

The density of the invaded block rapidly increases (Fig. 2.3e) and eventually reaches 1/2. At this point, altruists' effective utility becomes negative, pushing them to leave

for other lower density blocks (Fig. 2.3f). As previously, altruists gather in another single block of identical density $\rho_a = pQ\rho_0$. The invasion has led to a slight decrease of the density of egoist blocks to $\rho_e < \rho_E$, and therefore to a slight increase of egoists' utility, $u_e = u(\rho_e) > u(\rho_E)$. Successive invasions of the block partially filled by altruists are possible until ρ_e decreases down to the value ρ_e^* such that $u(\rho_e^*) = u(\rho_a + 1/H)$. This leads to $\rho_e^* = 1 - pQ\rho_0 - 1/H$ ($\rho_e^* > 1/2$ as long as $p < p_{\rm high}$). The equality of utilities implies $\tilde{u}(p) = u(\rho_a) = 2pQ\rho_0 + 2/H$. When $p = p_{\rm high}$, the final (lowest) egoist density reaches the optimal value $\rho_e^* = 1/2$ and no further improvement in average utility is possible: $\tilde{u}(p) = 1$ (to simplify the discussion, we ignore here corrections of order $1/H$ that depend on the parity of H). This description remains valid for larger altruist concentrations, the only difference being that, at the end, the additional altruists form stable blocks with densities $\rho_a = 1/2$.

In summary, the evolution of the average utility $\tilde{u}$ follows:

$$\begin{cases} \tilde{u}(p) = 2 - 2\rho_E & \text{for } p \leq p_{\rm low} \\ \tilde{u}(p) = 2pQ\rho_0 + 2/H & \text{for } p_{\rm low} \leq p \leq p_{\rm high} \\ \tilde{u}(p) = 1 & \text{for } p \geq p_{\rm high} \end{cases} \tag{2.6}$$

Our analysis predicts that plotting $\tilde{u}$ as a function of the rescaled altruist proportion $p^* = p/p_{\rm high} = 2pQ\rho_0$ should lead to a universal transition starting at $p^* = 2 - 2\rho_E \simeq 0.586$ and ending at $p^* = 1$. Simulations perfectly confirm our calculations (Fig. 2.2b).

2.6 Discussion

Our model illustrates the complexity of the dynamics produced by two types of agents, even when they follow simple rules. Introducing altruists into a population dominated by egoists increases the average utility much more rapidly than expected from a linear projection. The interplay between the different behaviors leads to complex "catalytic" phenomena. By catalytic, we mean that altruists are not "consumed" once they coordinate egoists, and can continue to help egoists finding the optimal configuration indefinitely. The global utility increase *per* altruist can be computed easily: $\delta U_{\rm altr} \equiv (U(p) - U(p=0))/N_A \simeq (1 - 0.56)\rho_0 QH/(p\rho_0 QH) = 0.44/p$. When $p = 1/Q$, $\delta U_{\rm altr} \simeq 0.44Q$. Each altruist induces a utility change proportional to the system size, which becomes infinite for infinite systems.

Interestingly, while the stationary state of a system composed of a single type of agents (either egoists or altruists) can be mapped to an equilibrium state, this is no longer the case when including two types of agents, except if some restrictive conditions are met [Grauwin et al., 2009]. In a thermodynamic analogy, the utility function can be mapped (in the zero temperature limit considered here) to a chemical potential, as shown in [Lemoy et al., 2012], when a single type of agents is present. If a system with both egoist and altruist agents could be mapped to an equilibrium system, chemical potentials could be defined as $\mu_e(\rho_a, \rho_e) = u(\rho_a + \rho_e)$ and $\mu_a(\rho_a, \rho_e) = u_{\rm altr}(\rho_a + \rho_e)$. As chemical potentials derive from a free energy, their cross derivatives would be equal, $\partial\mu_e/\partial\rho_a = \partial\mu_a/\partial\rho_e$, leading to $u'(\rho) = u'_{\rm altr}(\rho)$. This equality is not satisfied as seen from Figs. 2.1 and 2.4, showing that the system reaches a non equilibrium steady state.

We are well aware that simple models do not allow to draw any rigorous conclusion about what is going on in the *real* world [Venturini et al., 2015, Ostrom, 2010, Jensen, 2018]. While Schelling's segregation model neatly shows that one cannot logically deduce individual racism from global segregation, it may well be that for some towns racism is one cause of segregation, for some others not; at any rate the reasons behind urban segregation are far more complex than those that any simple model can come up with. Simple models can be helpful to analyze some interesting phenomena, the origin of which may be obscured in more complicated realistic settings. Ours may help thinking about the effectiveness of coordination by an infinitesimal proportion of altruist agents, but it cannot be directly applied to real systems. Real agents do not behave like these virtual robots: they are able to put their actions into context, to anticipate the behavior of the others and moreover, they disagree about what is the social "optimum" [Jensen, 2018, Latour, 1988].

In this chapter we discussed the limitations of simple statistical physics models of society and the complex dynamics which can emerge from simple simulated models. In the next chapter, we explore the dynamics of a real world system using the dataset of Vélo'v users, a bike sharing system based in Lyon. We will illustrate the advantage of using temporal data and the difference between global evolution of the system and individual evolutions of its units.

3 Dynamics of Bike Sharing System Users

3.1 Introduction

Bike Sharing Systems (BSS) have been developing rapidly all over the world in the last decades, being now present in more than 500 cities. The number of studies of BSS has followed a similar pattern, focusing on 3 topics : quantifying BSS characteristics, describing users' socio-demographic profiles and evaluating its impacts on environment and public health.

The automatic recording of BSS activities has allowed a quantitative description of many BSS characteristics: Circadian and monthly activity patterns (see [Borgnat et al., 2011, Côme et al., 2014]), average speed ([Jensen et al., 2010]), patterns of bicycle flows over the cities (see [Côme et al., 2014, Jensen et al., 2010, Borgnat et al., 2011, Borgnat et al., 2013]) and influence of weather conditions ([Borgnat et al., 2011]). The knowledge derived from these studies, especially on bicycle flows between stations (see [Côme et al., 2014, Tran et al., 2015]) and the prediction of bike reallocation schedules ([Zhang et al., 2016]), can help the management of station balancing (see [Singla et al., 2015, Côme et al., 2014, Côme and Oukhellou, 2014]), one of the main financial challenges of BSS ([Yang et al., 2011]).

Socio-demographics profiles of BSS users differ generally from the overall cities demographics. Studies carried out in Europe and North America (see [Beecham and Wood, 2014, LDA-Consulting, 2012, Ogilvie and Goodman, 2012, Shaheen et al., 2012, Fuller et al., 2011, Raux et al., 2017]) have shown that users are more likely to be young, male, with a high level of education and living in the city center.

Finally, several studies have described the impact of BSS policies on environment and public health (see [Pucher and Buehler, 2012]). [Shaheen et al., 2010, Shaheen et al., 2011] have listed the benefits of BSS: Emission reductions, individual financial savings, physical activity benefits, reduced congestion and facilitation of multimodal transport connections. Yet, other studies question the real impact of BSS on some of the latter. Notably [Shaheen et al., 2012] showed the relatively low impact on people favorite mode of transportation. In particular [Fishman et al., 2013, Midgley, 2011, LDA-Consulting, 2012, Buttner et al., 2011] exhibited, for several cities in Europe and Canada the low substitution rates from car usage to BSS. Most BSS riders are indeed people who used to walk and take public transportation.

Among all the research axes cited, questions remain on the commitment of BSS subscribers in the long term. This is due to the lack of accurate trip datasets over long periods of time, as mentioned in [Fishman et al., 2013]. Some articles have tried to characterize travel behaviors using surveys, such as [Guo et al., 2017, Raux et al., 2017]. But the temporal evolution of users has never been deeply investigated. This is the reason why in this chapter we approach BSS travel behaviors and usage rates under the temporal angle. We address questions related to BSS sustainability, such as : how long do users remain active over the years ? Does their activity increase, decrease or remain stable? Do these trajectories depend on their level of activity? These questions are addressed using a five years long dataset covering about 150,000 long-term distinct users, among which 13,358 have stayed in the system for the whole period. We follow previous work on Lyon's BSS, *Vélo'v*, by [Vogel et al., 2014] which, using a single year dataset (2011), characterized users according to their intensity and frequency of uses at different time scales (day, week, month and year). This work found 9 classes of users, ranging from 'extreme users', that use Vélo'v twice a day on average to 'sunday cyclists', who only use the system a few week-ends per year. Using a single year dataset to classify users has however two main limitations. Firstly, there is no way to distinguish between two possible interpretations for a user that appears to be very active from September to December. This could correspond either to (a) someone arriving in town in September that remains very active for the months/years to come or (b) someone who for an unknown reason uses the system only in those months. The second limitation arises from the impossibility to test the stability of users' characteristics over years, which would allow to interpret them as real user properties. For example, do users classified in 2011 as 'sunday cyclists' retain this characteristic over the years? Have they only used Vélo'v in this way in 2011 or is this pattern a more personal - and stable - use of the system that lasts for longer periods?

The work presented in this chapter is one of the first giving a detailed description of how different segment of customers use Vélo'v over years. A recent article from [Jain et al., 2018] highlighted the lack of temporal analysis on BSS' users and importance of tracing longitudinal usage trends. Our study could open the way to the establishment of a methodology to dynamically assess impact of BSS policies and other transportation facilities on different segment of users.

After presenting our 5-years dataset in Section 3.2, we start by describing the overall system stability. We, then, show the heterogeneous individual trajectories masked by this overall system stability in Section 3.4. Finally, we compute in section 3.5 user classes using a similar approach to [Vogel et al., 2014] and break down individual evolutions from section 3.4 using the classes we computed.

3.2 Dataset

The Vélo'v program started in 2005 in Lyon, France. The Vélo'v network now has 340 stations, where roughly 4000 bicycles are available. The stations are in the street and can be accessed at anytime (24/7) for rental or return. More information about the history of Vélo'v and the deployment of stations can be found in [Borgnat et al., 2011]. The dataset used in this chapter records all bicycle trips from 2011/01 to 2015/12 for the Vélo'v system. It contains more than 38 million trips made by more than 3.8

million users. Each trip is documented with starting and ending times, duration, a user ID code and a tag describing the class of user (year-long subscriber, weekly or daily subscription, maintenance operation, etc). Data are filtered according to the process used in [Vogel et al., 2014], keeping only year-long users and eliminating any anomalies. This leads to a subset of the original population containing 147,354 long-term users. For each person, we count years from the first active day: For example, a user appearing in the records for the first time on March 16^{th}, 2011 will end the first adapted year on March 15^{th}, 2012. To avoid boundary artifacts for users that are active over several years, we stop recording trips at the anniversary date in 2015, even if there are recorded trips later in 2015.

Note that our elementary unit of analysis is therefore the 'person-year', i.e. the vector of 21 features for each user and each year. One person can therefore appear several times (up to 5) and change group from year to year. One could adopt a different point of view, using persons as the entities and computing a single vector for each of them, averaged over their whole period of activity. This would have two drawbacks: masking the single user trajectories over the years and comparing vectors computed over different periods (from 1 to 5 years). Comparing the third and fourth columns of Table 3.3 shows that using the 'person-year' or the 'person' as the basic entity leads to roughly the same proportions for the different classes. We then retain the 'person-year' description, which allows studying users' trajectories.

3.3 Overall evolution

We first analyze the global system evolution over the 5 years. Table 3.1 shows that there is a steady increase in the number of users and trips. However, the average number of trips per user remains remarkably stable around 92 trips/year, despite the large variability (standard deviation larger that the average). A similar general temporal trend is found in [Jain et al., 2018].

3.4 Individual evolutions

The overall system stationarity (slow increase of user numbers) hides a great variability at the individual level that can be uncovered only using long-term datasets at the individual level as ours. Every year, there is a strong user renewal, as the majority of users leave the system after their first year of activity and are replaced by a greater

Year	2011	2012	2013	2014	2015
Active Users	50,393	55,896	61,806	70,068	76,511
Trips	4,702,498	5,138,971	5,576,973	6,625,090	7,044,707
Trips per user	**93.3**	**91.9**	**90.2**	**94.5**	**92.0**
Median trips per user	**45**	**46**	**45**	**49**	**46**
Standard Deviation	**125.6**	**122.8**	**120.7**	**123.8**	**123.8**

Table 3.1 – Number of trips per active user.

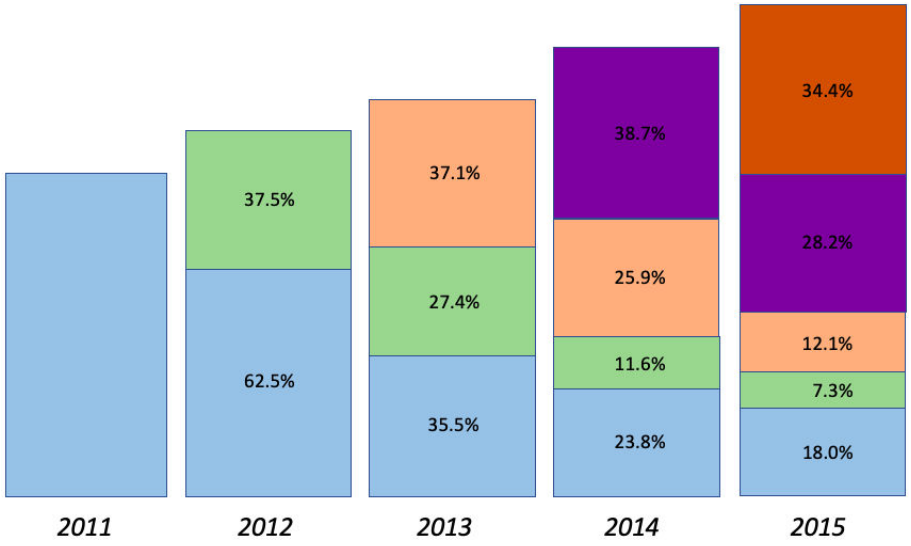


Figure 3.1 – Progressive renewal of users over the years. For each year, the box height represents the total number of users and the colors the year users have entered the system. For example, in 2015, 34.4% of users are new to the system, while 18% started in 2011.

number of new users. Figure 3.1 shows that every year the new users represent around 35% of the total. Then, they progressively leave the system, in a quite predictable way: They represent 26-28% of users the year after and 11-12% two years later. The only exception is the 2011 cohort, which by lack of data over the previous years, also includes users that entered the system *before* 2011.

3.4.1 Most users leave the system after one year

Analyzing user activity over calendar years as in Figure 3.1 is confusing, since users enter the system at any time during the year. To follow *individual* evolutions, we have to shift the different starting dates to a common origin using 'adapted' years as explained above.

Figure 3.2 shows that a large majority of users (60.8%, blue rectangle) quit after a single year of practice. These users are significantly younger than more loyal users (24 years old against 31), more likely women (51.1% of men compared to 59.1%) and less active: their median number of trips is 47, to be compared to 91. This low activity is mainly explained by a shorter time span of their activity (median close to 9 months instead of the whole year). This means that many of them stop using the system before the 12-month validity of their subscription, because they leave Lyon, buy a bike, change job...

Almost 20% of users stay in the system for 2 years (yellow rectangles in figure 3.2). Note that their activity is significantly lower than that of more loyal users, that will stay in the system for 3 or 4 years (89 trips against 100, p-value $< 2.2 * 10^{-16}$). In this case, this reduced activity cannot be explained by a shorter activity time span. These users are consistently less active over the whole year, a feature that allows to predict a higher chance of quitting the system the following year as we will check below. When

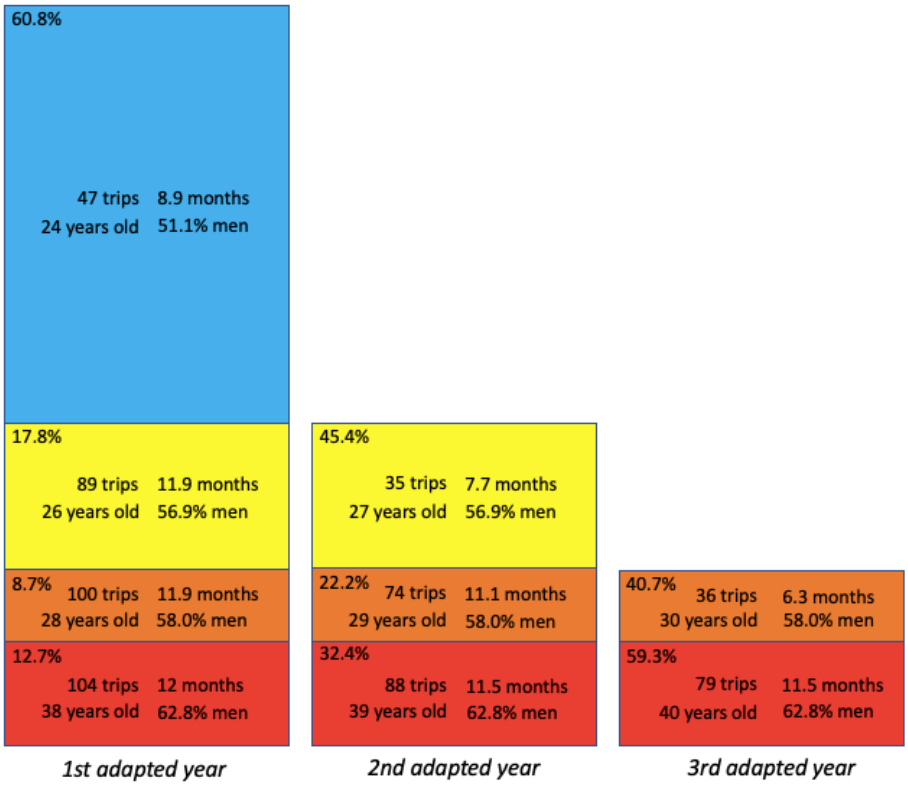


Figure 3.2 – Percentages of users leaving the system at the end of different adapted years. For each group of users is given the median number of trips per year, the median number of active months, the median age and the percentage of men. For example 'Blue users' stopped at the end of their first adapted year, after a median number of 8.9 months of activity. They represent 60.8% of first year users. They had a median number of trips during that year of 47, a median age of 24 years old and 51.1% were men. Yellow users stopped at the end of their second adapted year. They had a median number of trips during their first year of 89 and during their second year of 35. They stopped after a number of 7.7 active months during their second year.

these users reach their second (and last) active year, their activity becomes quite similar to the blue users, as their time span is reduced to 7.7 months and their activity much lower than in their first year (35 trips instead of 89).

Almost 9% of users stay in the system for 3 years (orange rectangles in figure 3.2). Again, their activity, even two years before leaving the system, is significantly lower than that of more loyal users (100 trips against 104, p-value $< 2.2 * 10^{-16}$). This activity progressively diminishes over the years, reaching a very low value on the third and final year (36 trips over 6.3 months).

Finally, 12.7% of users stay in the system for at least 4 years (red rectangles in figure 3.2). Their activity is consistently higher than average, and these users are older and more often men. Their activity also progressively diminishes over the years, a feature that we study in more detail below.

The most striking result is the high proportion (60.8%) of users that quit after a single year of practice (called 'leavers' hereafter). To the best of our knowledge, this surprising figure was previously unknown. It is worth noting however that this figure

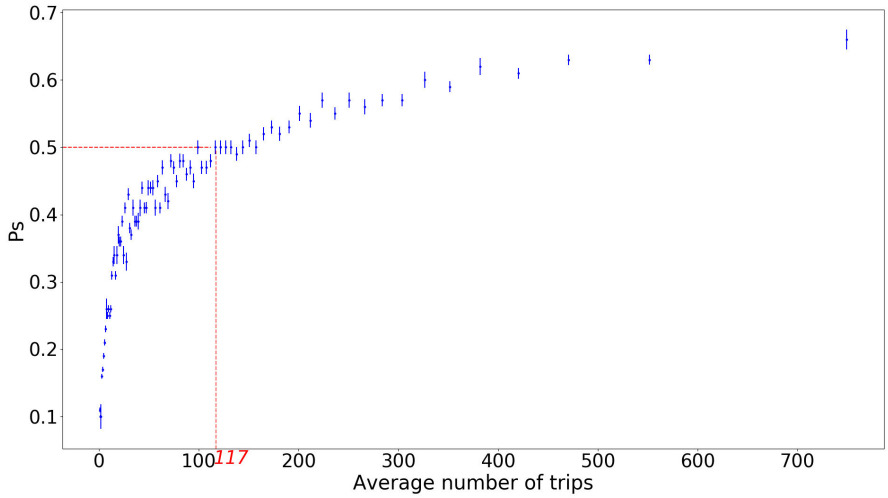


Figure 3.3 – Probability to stay in the system at the end of an adapted year (P_s) as a function of the average number of trips during that year. When activity reaches 110-120 trips per year, users are more likely to stay ($P_s \geq 0.5$). Each point represents an average of P_s over 2500 person-years.

might be slightly overestimated. The reason is that users are identified through the ID of different cards, the most common being Velo'v own card (30.3% of the users), public transportation card (Tecely, 59.7%) and train card (Oura, 5.2%). The point is that the Tecely cards have to be renewed every 5 years. In some (uncontrolled) cases, this leads to a change of ID, which our analysis interprets as if the user had left the system and another had entered it. To estimate the proportion of incorrectly labeled exits from the system, we note that only 46.6% of Velo'v cards users give up after one year, the corresponding figure being 61.3% for Tecely users. As Velo'v cards do not go through the yearly renewal process, this percentage could represent a lower bound on the 'leavers' proportion, if we assume that the proportion of leavers does not depend on the card used, which seems unlikely as using the specific Velo'v card suggests a higher loyalty. To obtain another estimation, we may assume that all renewed Tecely cards (20% per year) change their ID. This would mean that the 61.3% figure is an overestimation of the real figure (61.3 - 20)/0.8 = 51.6%. These estimates converge to a proportion of leavers between 50 and 55%.

We noted above that the loyalty of users was correlated to their activity. Figure 3.3 shows the general trend over all users. It confirms that the higher the intensity of use, the higher the probability P_s to stay in the system. This result can help predicting users' loyalty.

3.4.2 Long-term users are older, more likely men and more urban than average

Let us now focus on the most loyal users, the 25,963 users that have been active for at least 3 years, which we now call 'long-term' users. Comparing them to those that leave after a single year reveals interesting facts about their specific characteristics.

Ages	N	% men	% long-term users	% men in long-term users
13-22	11,231	54.8	25.2	60.5
23-32	18,081	54.5	31.6	59.3
33-42	9,572	62.6	49.8	65.7
43-52	6,413	57.6	57.3	58.8
53-62	3,774	55.9	59.6	57.0
63-72	1,194	65.1	64.7	67.4

Table 3.2 – Description of long-term users characteristics

They are older (median age 35, against 24, p-value $< 2.2 * 10^{-16}$), more likely men (men proportion 62.9% against 49.9%, p-value $< 2.2 * 10^{-16}$) and live within the Lyon-Villeurbanne urban area (85.3% against 81.7%, p-value $< 2.2 * 10^{-16}$)

Table 3.2 shows the proportions of long-term users for different 10-years slices. Clearly, loyalty steeply increases with age, from 25.2% for 13-22 years old users up to 64.7% for 63-72 years old users. Men are over-represented among the long-term users for all ages, but the difference is highly significant among younger users. It would be interesting to understand why there are (comparatively) so few young women among the most loyal BSS users.

Finally, we study how long-term users change their activity over the years. For each user, we computed the percentage of change in the number of trips per year from one year to another. Figure 3.4 shows that only one quarter (26.5%) maintain their number of trips within a ±20% range. Roughly two-thirds (61.8%) users lower their activity, almost halving it (median decrease 42.3%). The remaining third increases its activity (median increase 42.6%). The median evolution of long term users is a decrease of activity by 16.3%.

We now apply a k-mean algorithm to the set of person-years to describe to finer level the evolution of users, we discuss the number of clusters and compare our findings to the results of [Vogel et al., 2014], in which the authors used the same methodology on a 1-year dataset.

3.5 Classes of users

In this section, we compute users classes on our 5 years dataset, using a similar approach to [Vogel et al., 2014], and offer a brief description of them. Computing these classes allows us to compare our results to the one found in [Vogel et al., 2014]. This comparison illustrates the effect of temporality on users classification accuracy.

3.5.1 Computing users classes

From this dataset, we compute the same 21 normalized features characterizing the activity as in [Vogel et al., 2014]. For each person, these features quantify the intensity and regularity of use over the year (14 features) and the week (7 features).

- *trips week*, averaged number of trips made per week, calculated over the weeks

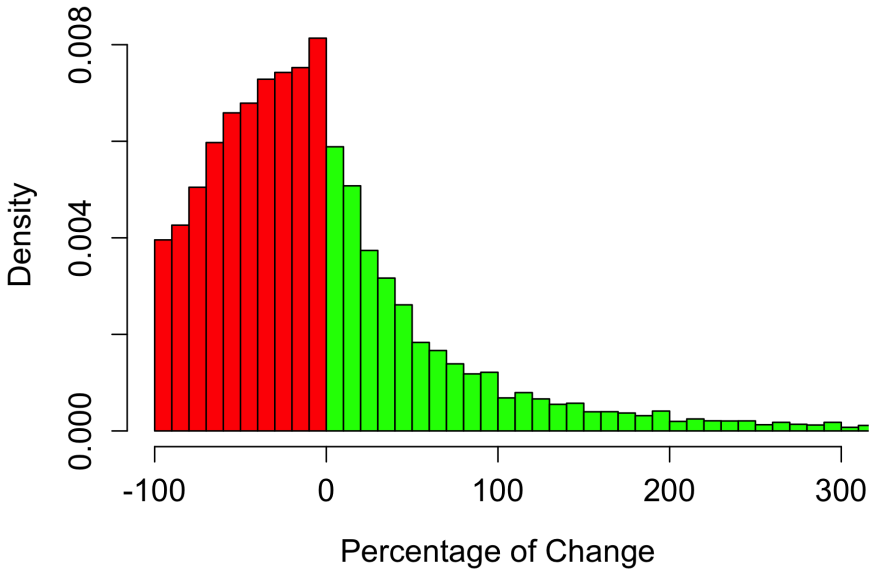


Figure 3.4 – Density distribution of percentage of change in the number of trips per year from one year to another. Percentages are computed for each user that remained active for at least 3 years.

during which users traveled at least once, and normalised dividing by 1.5 times the interquartile range of the distribution for all users (equal to the difference between the lower and upper quartile of the distribution).

- *trips day*$1-5$, number of trips per week day. Days are ranked from one to five, *day 1* being the day with the highest number of trips and *day 5* the day with the lowest number of trips. *trips saturday*, average number of trips made on Saturdays. *trips sunday*, average number of trips made on Sundays. These seven features are normalized over a total sum unity over the week.

- *trips year*, total number of trips made over the adapted year, normalized dividing by 1.5 times the interquartile range of the distribution for all users (i.e. the difference between the lower and upper quartile of the distribution).

- *trips month*$1-12$, number of trips per month, normalized to a total sum unity, months are ranked from one to twelve, *month 1* being the month with the highest number of trips and *month 12* the month with the lowest number of trips.

As explained in [Vogel et al., 2014], there is some correlation and redundancy between intensity and regularity features. Nevertheless, a simple K-means clustering method (see, e.g., [MacKay, 2003]) is used, coupled with statistical appraisal and careful analysis of the results, as our main intent is to create and interpret a relevant typology, not to find well-defined, pre-existing, classes. Also, we prefer to use the original variables instead of PCA axes, because this makes the interpretation of the obtained classes straightforward.

To choose the number of clusters, we studied as [Vogel et al., 2014] the partitions obtained when choosing from three to twelve clusters. However, as shown on figure

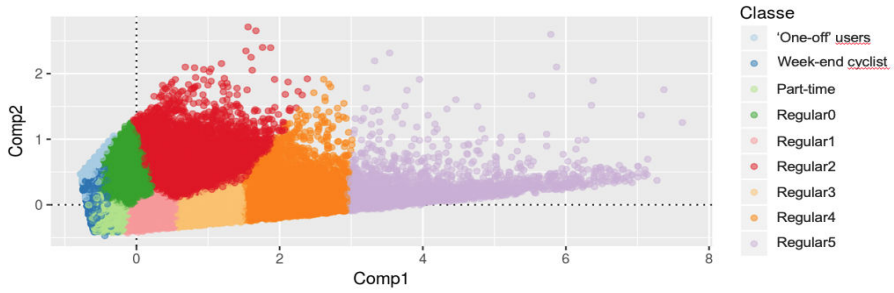


Figure 3.5 – Visualization of the users-year on the two main axis given by principal component analysis (PCA)

3.5, there is a continuum distribution of users-year and therefore no clear separation to split the clusters. Moreover we also computed the *Akaike Information Criteria* (AIC) for the different number of clusters k. The AIC is defined as follow:

$$AIC(k) = -2L(k) + 2q(k) \tag{3.1}$$

Where $-L(k)$ is the negative maximum likelihood for k clusters and $q(k)$ is the number of parameters of the model with k clusters. Hence, the AIC represents the trade-off between a model which clusters well the data (given by the likelihood) and the complexity of the model. The optimal number of clusters k is in theory $argmin_k[-2L(k)+2q(k)]$. However after plotting $AIC(k)$ (Figure 3.6), we observe that there is no clear knee showing an optimal value.

Hence, we chose nine clusters for two reasons. Firstly, we found that as in [Vogel et al., 2014], this number of clusters represents the best compromise between a reduced but rich enough description (differentiating for example week-end users from week users), without introducing uninformative clusters. Secondly, this number allows a simple comparison of our results with the study from 2014.

A detailed description of the nine classes is given in Table 3.3 and Figure 3.7.

3.5.2 User Classes

The nine classes correspond to different profiles of use. There are 6% of 'one-off' users, who make on average only 3 trips per year, generally the same month and then disappear from the database. Another almost 12% of users are mainly active in week-ends, either for shopping (Saturdays) or leisure (Sundays) (second line of Table 3.3). The last 6 lines of Table 3.3 represent users that show a regular activity over the year and differ mainly by their intensity of use, from twice a month (regular0 class, gathering 27% of users) to nearly twice a day (regular5, 1% of users). The part-time class is quite peculiar: we will show below that it can be interpreted as the class where users end up for the last year of activity.

3.6 Evolution of user classes

This section answers to the question: what are the differences between stable users in terms of practice, i.e. what is the dynamics of stable users within the classes?

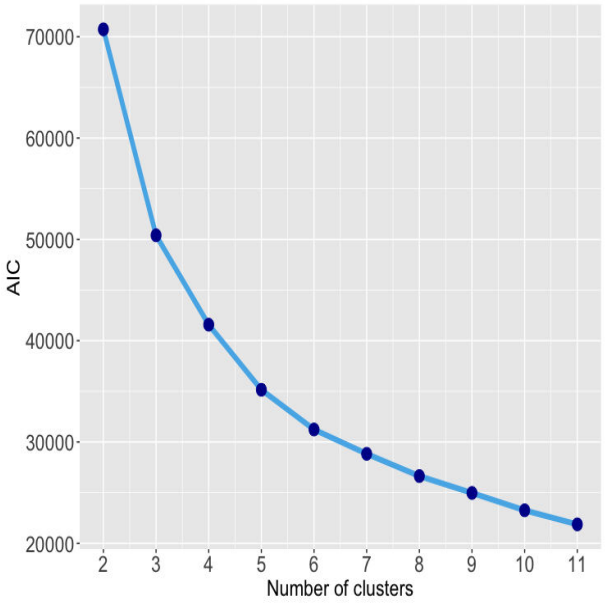


Figure 3.6 – AIC

Figures 3.8b and 3.8a present these 'transfer matrices' for different years for each class.

These matrices give key informations about the evolution of stable users in the system, it shows:

- the overall stability of classes (diagonal terms). The high values found around the matrix diagonals (light colors in Figures 3.8b and 3.8a) show that many users remain in the same class over several years : the probability to stay in the same class is higher than any other probability (except leaving). As discussed briefly below, the second highest probability corresponds to a shift to the neighboring class with lower activity. The matrix also shows that this 'class fidelity' is correlated to the intensity of use: For example, on the first year $P_{I5\to I5} = 30.2\% > P_{I4\to I4} = 22.8\% > P_{I3\to I3} = 17.5\% \ldots$ This intensity is therefore a good predictor of future behavior : Staying in the same class or, as discovered earlier (Figure 3.3), leaving the system.

- the decrease of activity for users that remain active (asymmetry of the non diagonal terms). We observe the upper parts contain higher values than the lower parts. As the lines in the graph are ordered by intensity, this means that users have a high probability of reducing their activity from one year to another. This joins results from Figure 3.4.

- in each year, a high proportion of users leave the system (last column). However the probability to leave the system decreases continuously as the classes increase in intensity of use.

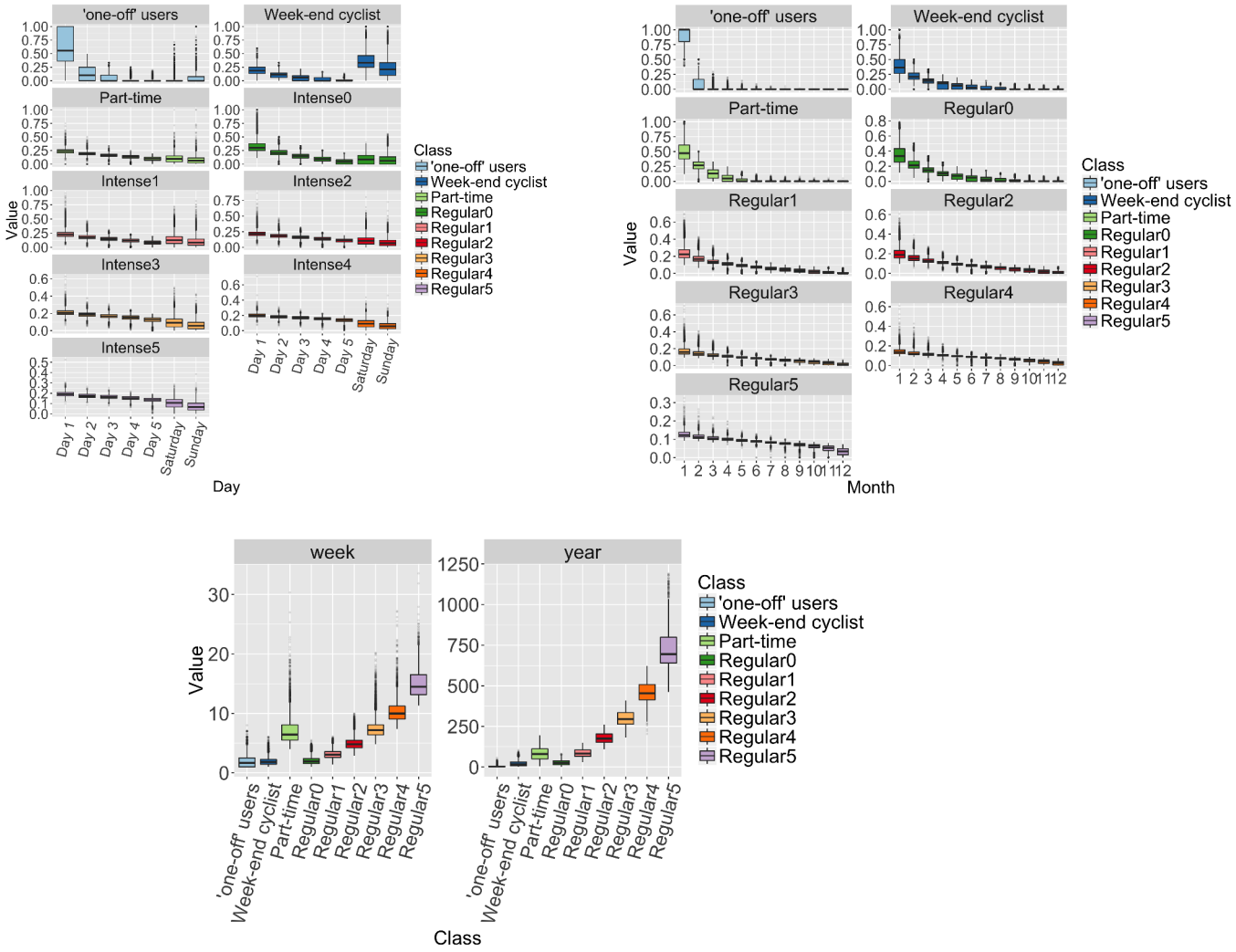
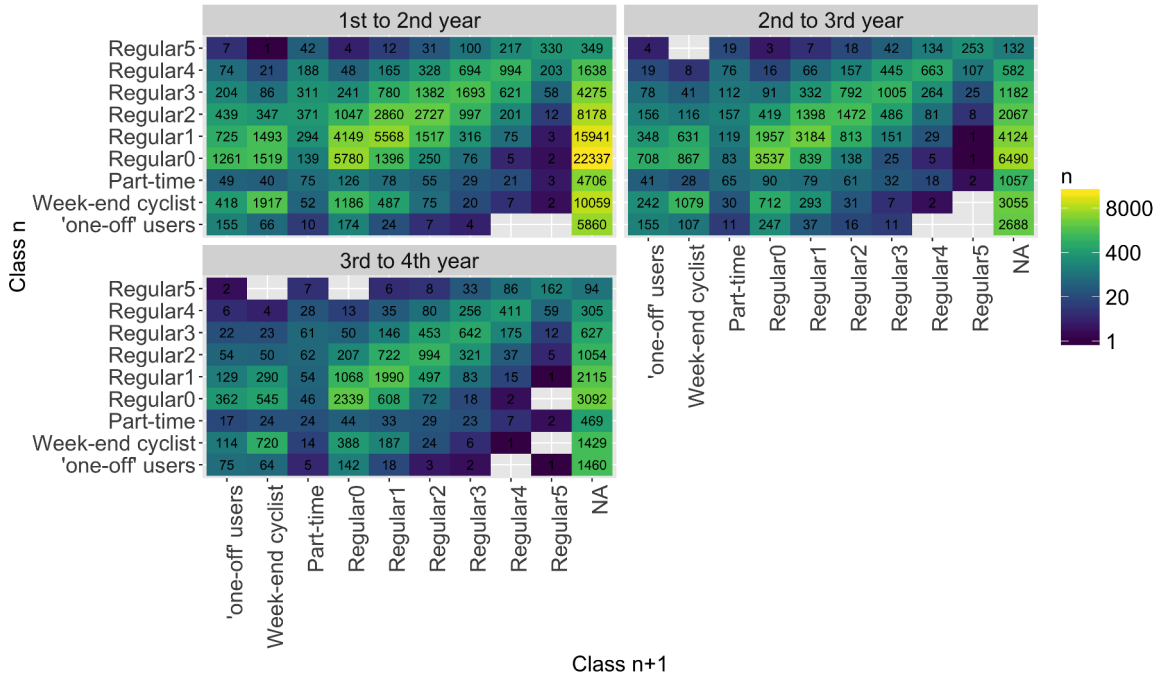
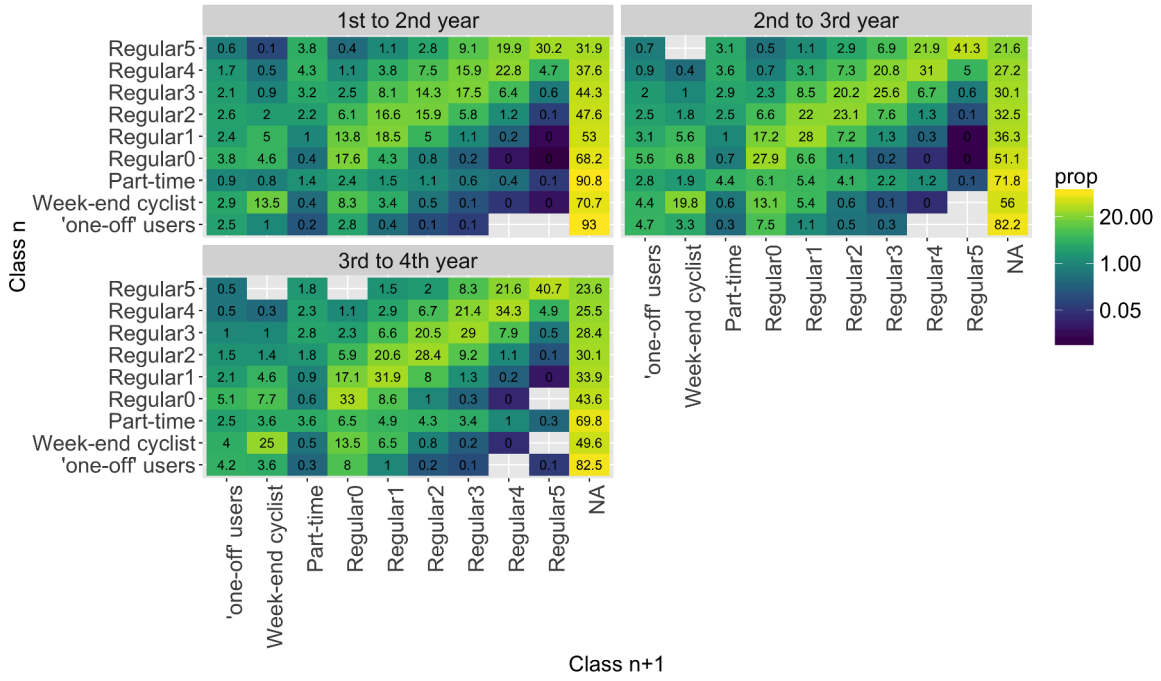


Figure 3.7 – Boxplots of the behavioral patterns at different time scales of the 9 classes. (a) number of uses per year (right) and per *active* week (left), for each class. A week is considered 'active' for a user whenever he/she takes a bicycle at least once. (b) normalized number of uses for each day of the week and for each class. Week days range from one to five in decreasing order of activity for each user. Saturday and Sunday are computed separately as users' activity is different on week-ends. (c) normalized number of uses for each month of the year and for each class.



(a)



(b)

Figure 3.8 – Transfer matrices from year n (lines) to year n+1 (columns). (a) Absolute number of users. Reading: (first line, first table) : 217 users that belonged to the regular-5 class in their first year became regular-4 users in their second year; 330 remained regular-5 and 349 were no longer active; (second line, third table): 411 users that belonged to the regular-4 class in their third year remained regular-4 users in their fourth year; 59 became regular-5 and 305 left the system. (b) Percentage of users. Reading : (first line, first table) : 19.9% users that belonged to the regular-5 class in their first year became regular-4 users in their second year; 30.2% remained regular-5 and 31.9% were no longer active; (second line, third table): 34.3% users that belonged to the regular-4 class in their third year remained regular-4 users in their fourth year; 4.9% became regular-5 and 25.5% left the system.

Class	# person-year	freq	1^{st}-year freq	#trips/year
'one-off' users	12164	5.8	5.2	3.0
Week-end cyclist	24313	11.6	11.8	17
Part-time	7639	3.6	4.3	80
Regular0	56849	27.1	27.1	25
Regular1	51446	24.5	24.9	83
Regular2	29225	13.9	14.2	175
Regular3	17183	8.2	8.0	295
Regular4	8444	4.0	3.6	454
Regular5	2361	1.1	0.9	695

Table 3.3 – Description of user classes found by the k-means for the 21 features. Note that since the entity is a 'person-year', these counts do not directly represent proportions of individuals, because users that stay in the system for long periods are over-represented. However, the comparison with the proportions obtained for year one (third column), which correspond to real users, shows that this effect is relatively weak. *#trips/year* is the median number of trips in a year for each class.

Length of use and activity

Let us now break down the length use, first mentioned in 3.2. The increase of use from classes regular0 to regular5 arises from the combination of two factors: both the number of active weeks (figure 3.9) and the number of trips per active week increase. For example, regular5 users are almost 30 times more active over the year than regular0 users (695 instead of 25 trips, see Table 3.3), because they are at the same time more often active (50 active weeks instead of 12) and more active during these weeks (14 trips per week instead of 2). Note that the decrease of the number of active weeks is not due to a simple seasonal effect: Figure 3.10) shows that active weeks of regular0 users span a median period of nearly 40 weeks (between the first and the last trip of the year).

3.6.1 Comparing the 5-years and 1-year classes

When comparing the classification obtained here to that found over a single year [Vogel et al., 2014], we note many similarities and a major difference. As in [Vogel et al., 2014], a 'one-off' and a 'week-end' class are found, with similar proportions, as well as six 'regular' classes differing mainly by their intensity of use. The major difference is the 'part-time' class, that represents 3.6% of users, instead of 29% for the single year classification (summing their 'intensive and part-time' and 'irregular' classes). This means that those two 1-year classes mostly gathered users that have in fact a *regular* behavior appearing to be 'part-time' because they are observed over a too short period of time. For example, a user starting in September 2011 will appear active only for (at most) 4 months, even if they keep the same activity over the subsequent (unobserved) year. Figures 3.8b and 3.8a show that the 'part-time' class gathers users that massively leave the system at the end of the year (nearly 90% the first year and 70% the second). Year after year, the 'part-time' class is filled again by users coming from all (previous year) classes, as shown by the numbers in the 'part-time' column in Figures 3.8b and

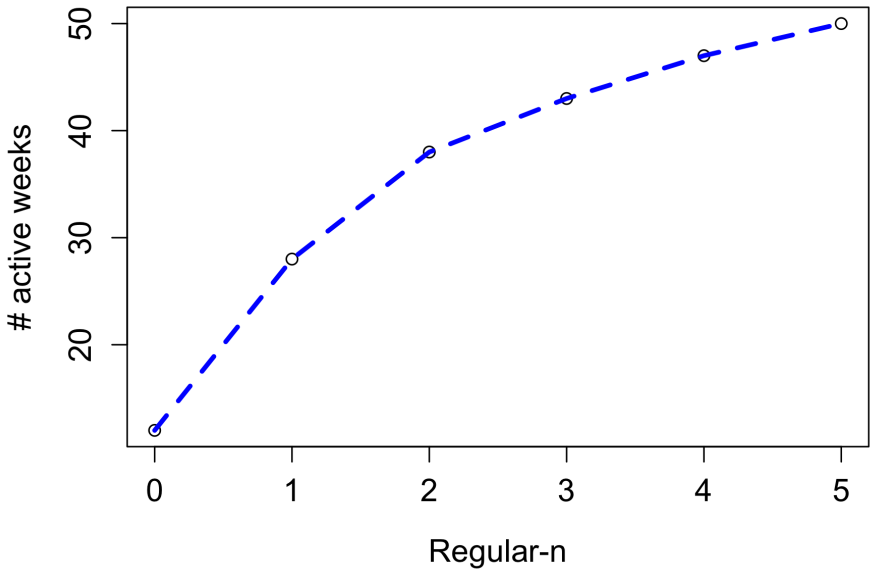


Figure 3.9 – Number of active weeks per year for each class of regular users. An active week is defined as a week where at least a trip took place.

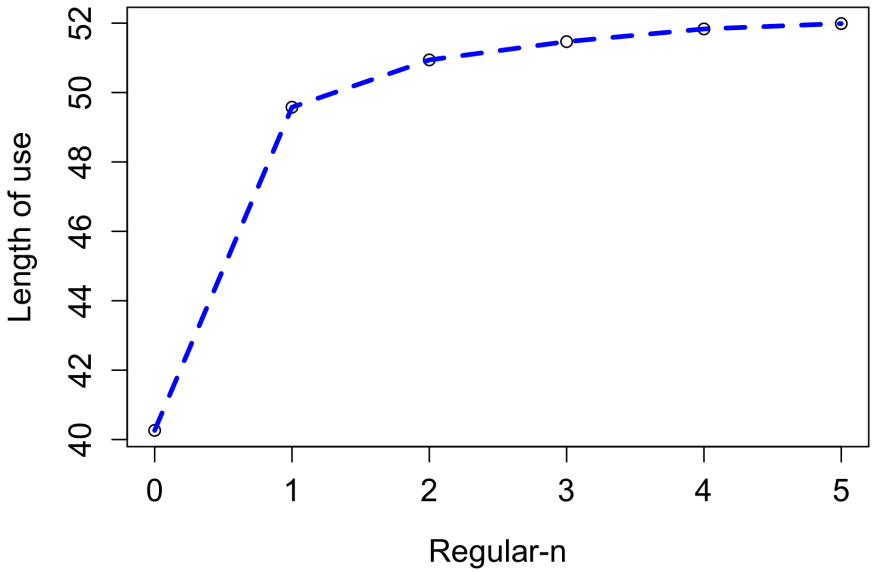


Figure 3.10 – Length of use per adapted year for each class of regular users. The length of use is the time difference between the first active week and the last active week of the adapted year.

3.8a. Therefore, this class does not represent a stable behavior of a class of users - people being active every year only 3 months - but the class where users end up for their last year of activity.

3.7 Discussion

As seen in the introduction, there was a lack of in-depth studies on the temporal evolution of long-term bicycle usage, mainly due to the lack of long-term datasets. Thus, we studied the temporal evolutions of year-long Vélo'v users thanks to a unique dataset spanning over 5 years. After adapting the data, we extended the method from [Vogel et al., 2014] to characterize temporal patterns and we showed that using a 5-year database corrects the 1-year classification by avoiding the overestimation of part-time users (from 29% to 3.6%). This indicates that the seasonal effect in bike usage is much smaller than expected.

Also, users' yearly activity was organized in three main classes: 6% of 'one-off' users, with only 3 trips per year; 12% of 'week-end' users and the rest presenting a regularly distributed activity over the year. We divided the latter class into 6 groups with considerable differences in numbers of trips (from twice a month to twice a day). From these classes, we studied the evolution of activity for longer times and found two main trajectories: About 60% of the users stay in the system for one year at most and show a low median activity (47 trips); the remaining 40% of users are more active (median activity of 96 trips in their first year) and remain continuously active for several years (mean time = 2.9 years).

This high proportion of leavers can be explained by many reasons : moving to other towns, buying one's own bike, finding the service unsatisfactory... It would be interesting to ascertain the relative proportions of each. Note that considering the number of trips (and not of users) leads to a more stable picture, as the 40% of stable users perform around 60% of the annual trips. Their activity is relatively stable, slightly decreasing over the years. We showed that this long-term behavior strongly depends on the user initial class, as fidelity rapidly increases with the number of trips observed the first year.

On the socio-demographic point of view, stable users are generally older than average users (30 to 40 years old) and live closer to the city center. This result agrees with previous articles socio-demographically characterizing BSS users population (see [Beecham and Wood, 2014, LDA-Consulting, 2012, Ogilvie and Goodman, 2012, Shaheen et al., 2012, Fuller et al., 2011, Raux et al., 2017]).

A comprehensive analysis of the reason of evolution patterns would be an insightful extension of this study. Though data about the motivations of users are not yet available, a survey would help addressing questions on the motivations behind the patterns such as why users leave the system. Such answer would complement our description.

This analysis is a first step to understanding usage trends and monitoring the evolution of users' segments which could help authorities to design more adapted systems.

This chapter describes for the first time in great detail the evolution of customers of a large BSS and suggest further work on important policy issues which we cannot

address for lack of appropriate data. So far, one study performed a similar analysis to ours. [Jain et al., 2018] found similar demographic characteristics to long-term users and general system's usage (slowly increasing over years). The development of temporal study on the evolution of usage and commitment for other BSS in the world would be of high profitability, in order to compare the evolution of trends and socio-demographics in the world and releasing BSS more adapted to users' segments.

In this chapter, we studied individual dynamics of users using mobility data. Now, after individual dynamics, we would like to study collective dynamics. In the next chapter, we explore temporal community detection on scientometric networks. More particularly, we investigate the differences between two approaches for generating automatic history of scientific research.

4 Dynamics of Scientific Research Communities

4.1 Introduction

Networks are a convenient way to represent real-world complex systems, such as social interactions [Newman, 2003, Barabási et al., 2002, Onnela et al., 2007], metabolic interactions [Boccaletti et al., 2006, Barabási and Oltvai, 2004], the Internet/world wide web [Pastor-Satorras and Vespignani, 2001, Barabási and Albert, 1999], transportation systems [Dall'Asta et al., 2006, Barthélemy, 2011], etc. For several systems it is interesting to find and describe areas of the network which are more densely connected, i.e. the communities of the network. In 20 years of complex networks history extensive work was conducted on community detection in static -non evolving- networks, see [Newman, 2006, Blondel et al., 2008, Fortunato and Barthélemy, 2007] and the review [Fortunato and Hric, 2016] for an overview on community detection in static graphs.

However, many networks have a temporal dimension and need a dynamic mesoscopic description at risk of non-negligible information losses if studied as static networks. Therefore the description of large temporal graphs has been a hot topic of research for the last decade, see the excellent reviews [Holme and Saramäki, 2012] and [Holme, 2015] for a complete description of temporal networks. Most recently the detection of dynamic communities, that is communities on temporal networks, has become one of the main interests in network science, as temporal networks require to adapt the methods of static community detection. So far no consensual method was found and around 60 methods have been proposed to try to detect dynamic communities evolving with temporal networks. A total of 4 published reviews try to classify and summarize them [Aynaud et al., 2013], [Hartmann et al., 2016], [Masuda and Lambiotte, 2016] and [Rossetti and Cazabet, 2018].

In the most recent one ([Rossetti and Cazabet, 2018]), these methods are classified into 3 main categories: (a) *instant optimal*, (b) *temporal trade-off* and (c) *cross-time*. In this chapter we propose a method inspired from [Morini et al., 2017], which falls into category (a). These methods aim to detect clusters at different times t, i.e. for many snapshots of the temporal network. As these clusters are only dependent on the state of the network at time t, it is then necessary to match the communities at different t with some similarity measures, e.g. Jaccard based [Morini et al., 2017, Lorenz et al., 2018, Greene et al., 2010], core-node [Wang et al., 2008]. Methods in

category (b) define clusters at t depending on current and past states of the network. Clusters are incrementally temporally smooth. However such methods are subject to drift as clusters are added up to each other locally. There is no compromise between temporal smoothness and 'optimal' partition at time t, see for example [Rossetti et al., 2017, Guo et al., 2014, Görke et al., 2010, Görke et al., 2013]. Finally, in category (c) clusters at t depend on both past and future states of the network, see [Duan et al., 2009, Mucha et al., 2010, Matias and Miele, 2016, Ghasemian et al., 2016]. Clusters are completely temporally smooth and not subject to drift, but they do not respect causality as communities at t are determined using network's information at $t + n$, i.e. communities at time t can change depending on what comes next, which makes communities temporally unstable. This limitation also makes these methods difficult to use on-the-fly, as it requires full knowledge of the history.

This chapter has two goals : (1) present tools to describe community dynamics in history of science, such as : the description of the emergence and death of scientific disciplines, the description of merge and split of two disciplines; (2) illustrate the difference between global optimization (category (c) methods) and local temporal optimization (category (a) methods). We describe a method in-between which returns a mesoscopic description of the network, which we call: temporal streams. The main difficulties for meta-community detection methods are to find the right temporal smoothing and to quantify the 'stability' of communities. It is difficult to distinguish if changes between snapshots are due to structural evolution of the community or algorithm instability, as static community detection methods used at each time t can find different communities for a same topology (see [Rossetti and Cazabet, 2018] for a complete description of pros and cons of each clustering category). In this chapter we propose an algorithm which aims to find balance between temporal inertia (smoothness) and 'optimal' community at particular time t. We compare this method to the most basic algorithm which runs a Louvain method on the aggregated network. The latter can be assimilated to a category (c) method in [Rossetti and Cazabet, 2018]. We introduce the platform *BiblioMaps* that we used for visualizing dynamic communities. We then describe the methods we used to analyze differences between partitions: mutual information (MI) measures and bipartite network (BN) representation. We see that MI based measures are interesting but give a limited amount of information on how different two partitions are, whereas bipartite network representation allows to see how streams split between partitions. We used the methods on two bibliographic datasets: (1) the scientific publications of the ENS Lyon and (2) publications related to the field of wavelets. We find that the basic aggregated method finds partitions which present similarities with our method, but differs in cases where optimal partitioning at time t is preferred over smoothness of the history.

4.2 Methods

We start by presenting the two building blocks used in the algorithms we want to compare: how we define and partition a Bibliographic Coupling (BC) network and how we match clusters from successive time period to create streams (meta-clusters).

4.2.1 Bibliographic Coupling partitioning

Given a set of publications on a given period, a Bibliographic Coupling (BC) network can be defined based on the relative overlap between the references of each pair of publications. More specifically, we compute Kessler's similarities $\omega_{ij} = R_{ij}/\sqrt{R_i R_j}$, where R_{ij} is the number of shared references between publications i and j and R_i is the number of references of publication i. In the BC network, each publication corresponds to a node and two publications i and j share a link of weight ω_{ij}. If they don't share any reference, they are not linked ($\omega_{ij} = 0$); if they have an identical set of references, their connexion has a maximal weight ($\omega_{ij} = 1$). In this chapter, we considered that the link between two publications is only meaningful if they share at least two references and we impose $\omega_{ij} = 0$ if they share only one reference.

We use weighted links to reinforce the dense (in terms of links per publication) regions of the BC networks. This reinforcement facilitates the partition of the network into meaningful groups of cohesive publications, or communities. We measure the quality of the partition with the *modularity* Q (eq. 4.1), a quantity that roughly compares the weight of the edges inside the communities to the expected weight of these edges if the network were randomly produced:

$$Q = \frac{1}{2\Omega} \sum_{i,j} \left[\omega_{i,j} - \frac{\omega_i \omega_j}{2\Omega} \right] \delta\left(c_i, c_j\right), \tag{4.1}$$

where $\omega_i = \sum_j \omega_{ij}$ is the sum of the weights of the edges linked to node i, c_i and c_j are the communities containing respectively nodes i and j, δ is the Kronecker function ($\delta(u, v)$ is 1 if $u = v$ and 0 otherwise) and $\Omega = \frac{1}{2}\sum_{ij} \omega_{ij}$ is the total weight of edges. We compute the graph partition using the efficient heuristic algorithm presented in [Blondel et al., 2008].

4.2.2 Matching communities from successive time periods

Given the sets of communities $\{C_1^t, ...C_{k_t}^t\}$ in each time windows t, the problem at hand is to identify a set of relevant historical communities, or *s*treams, that correspond to a chain of communities from successive time periods (at most one per period). In order to decide which community of a given period should be added to a chain of communities from previous periods, we need to use some measure to assess the similarity between communities from different time periods. In the dynamical communities literature, one method often use is the Jaccard index, based on a proportion of shared nodes between clusters of successive and overlapping periods (see e.g. [Claveau and Gingras, 2016, Morini et al., 2017]). One drawback from this method is that because of the use of overlapping periods, there is no bijection between the publications and the streams (a given publications can be part of several streams).

Here, we take advantage of the BC nature of our network, which ensures that links can exist between nodes from different time periods (publications from different periods can have common references). We can thus define a similarity measure between two

clusters C_a and C_b from different periods either by the total sum of the links between pairs of publications from these clusters $\Omega_{a,b} = \sum_{i \in C_a, j \in C_b} \omega_{i,j}$ or by a normalized version of this sum $\omega_{a,b} = \Omega_{a,b}/|C_a||C_b|$, which is comprised between 0 and 1. While these two measures can appear quite intuitive, each of them has some drawbacks as well: using $\Omega_{a,b}$ may sometime bias the construction of the streams by linking two "large" (in terms of publications) but dissimilar top clusters. On the opposite, using $\omega_{a,b}$ may sometime put too much emphasis on one cluster (e.g. a strong similarity between clusters of very different size over a second-best similarity between cluster of similar sized).

To be coherent with our construction of clusters from each time period by maximizing the modularities within each time period, we propose here to use a modularity-based concept on the set of clusters from 2 successive time periods. We, thus, use as similarity measure the quantity $\delta Q = \Omega_{a,b} - \Omega_a \Omega_b / 2\Omega_{A,B}$ which corresponds to an increase in the modularity of the BC network built from the two periods A and B.

Matching Algorithm

Only compare pairs of communities (a, b) with a minimum similarity $\omega_{a,b} > \Theta = 10^{-6}$.

Define the best match of each cluster by the one maximizing δQ.

for *each temporal window* **do**

  Define the **predecessor** of each cluster as its best match from the previous time period.

  Define the **successor** of each cluster as its best match from the next time period.

end

Two clusters are said to be *paired* if they are predecessors / successors of each other.

If a cluster is not the successor of its predecessor, we have a *split*.

If a cluster is not the predecessor of its successor, we have a *merge*.

Streams are defined as chains of paired clusters.

4.2.3 Different algorithms used to define historical streams

We investigated the results of four algorithms to define historical communities, or *s*treams, on a given dataset of publications over a long period, which we cut into successive periods of ΔT years. Two methods have a temporally global approach (GA, GPA), and two methods have a temporally local approach (BMLA, BCLA). In fact results between (GA and GPA) and (BMLA and BCLA) are very close (see Annex B). For this reason and for the sake of illustration, we only present the GA and BCLA methods. The two other methods (GPA, BMLA) and their results are shown in Annex B.

Global Algorithm (GA)

Building the global BC network by taking into account all the publications in the dataset, we simply define the streams as the communities maximizing the global mod-

ularity found by running the Louvain algorithm. Since we are working in a single (large) time period, this approach does not yield any dynamical events such as splitting / merging of communities, but it provides a simple reference.

Best-Combination Local Algorithm (BCLA)

The BCLA is an incrementation of the work from [Morini et al., 2017]. On each time period, we run N independent runs (we used $N = 100$) of the Louvain algorithm. Because of the noise inherent to the Louvain algorithm, choosing the best-modularity partitions chosen in each time period are not necessarily the ones that best match each other across successive time periods. We propose the BCLA to optimize the inter-period combination.

BCLA Algorithm

Compute the Bibliographic Coupling Graph ;
Split the dataset into temporal windows Δt ;
for *each of the* $N = 100$ *partitions of the first period and each of the* $N = 100$ *partitions of the second period* **do**
  Run the matching algorithm to define the 2-periods streams;
end
Among the $N * N$ defined streams, select the ones maximizing the modularity of the BC network on the first 2 periods. ;
Define the "best combination" partitions of the first 2 periods as those corresponding to those streams;
for *each pair of successive temporal windows A and B, starting from the second one* **do**
  for *each of the* $N = 100$ *partitions of the period B* **do**
    Run the matching algorithm between these partition and the "best combination" partition of period A (known from a previous step) to define 2-periods streams;
  end
  Among the N defined streams, select the ones maximizing the modularity of the BC network on periods A and B;
  Define the "best combination" partition of period B as the one corresponding to those streams
end

Note that maximizing a global indicator over the T periods with N runs would take too long as there would be N^T possibilities to explore. For this reason, we first choose the best combination between the first two periods (N^2 checks) and then we choose the "best match" one period at a time ($N(T-2)$ checks).

This algorithm returns temporal streams which we call *BCLA-streams*. These streams still maximize the modularity at each time t while using some cross-time information to improve the global modularity.

Choosing the value of the period T is a trade-off. It needs to be long enough so that communities within each period have enough articles to give meaning to the communities. But it also needs to be small enough compared to the total dataset duration

Dataset	Type	Period	N	N_{BC}	ρ_{links}	$< d >$	$< w >$	Q_{GA}
Wavelets	Thematic	1963-2012	6,582	5,568	0.0065	35.98	0.000719	0.677
ENS-Lyon	Institution	1988-2017	16,679	14,389	0.0019	27.04	0.000175	0.919

Table 4.1 – Statistics on datasets investigated in the chapter. *Type* is the type of organization data come from. *Period* is the period over which spans the dataset. N is the number of publications in the dataset. N_{BC} is the number of articles in the BC table. ρ_{links} is the density of links in the BC network. $< d >= (N_{BC} - 1) * \rho_{links}$, it indicates the average number of publications a given publication shares references with. $< w >$ indicates the average link weight. Q is the modularity of the network using global partioning (GA).

to see dynamics. After trying different periods, we chose, in the following, a period $T = 5$ years.

Akin to [Claveau and Gingras, 2016], these methods take advantage of local information to partition the data.

4.2.4 BiblioTools / BiblioMaps

All the datasets were extracted from the ISI Web of Knowledge Core Collection database[1]. The bibliographic records were parsed and analyzed using Bibliotools, a Python-based open-source software and the historical streams figures were generated using the web-based visualisation platform BiblioMaps. Bibliotools and its extension BiblioMaps were developed by one of us and are available online[2]. They were also used and presented in previous studies [Grauwin and Jensen, 2011, Lund et al., 2017, Grauwin and Sperano, 2018].

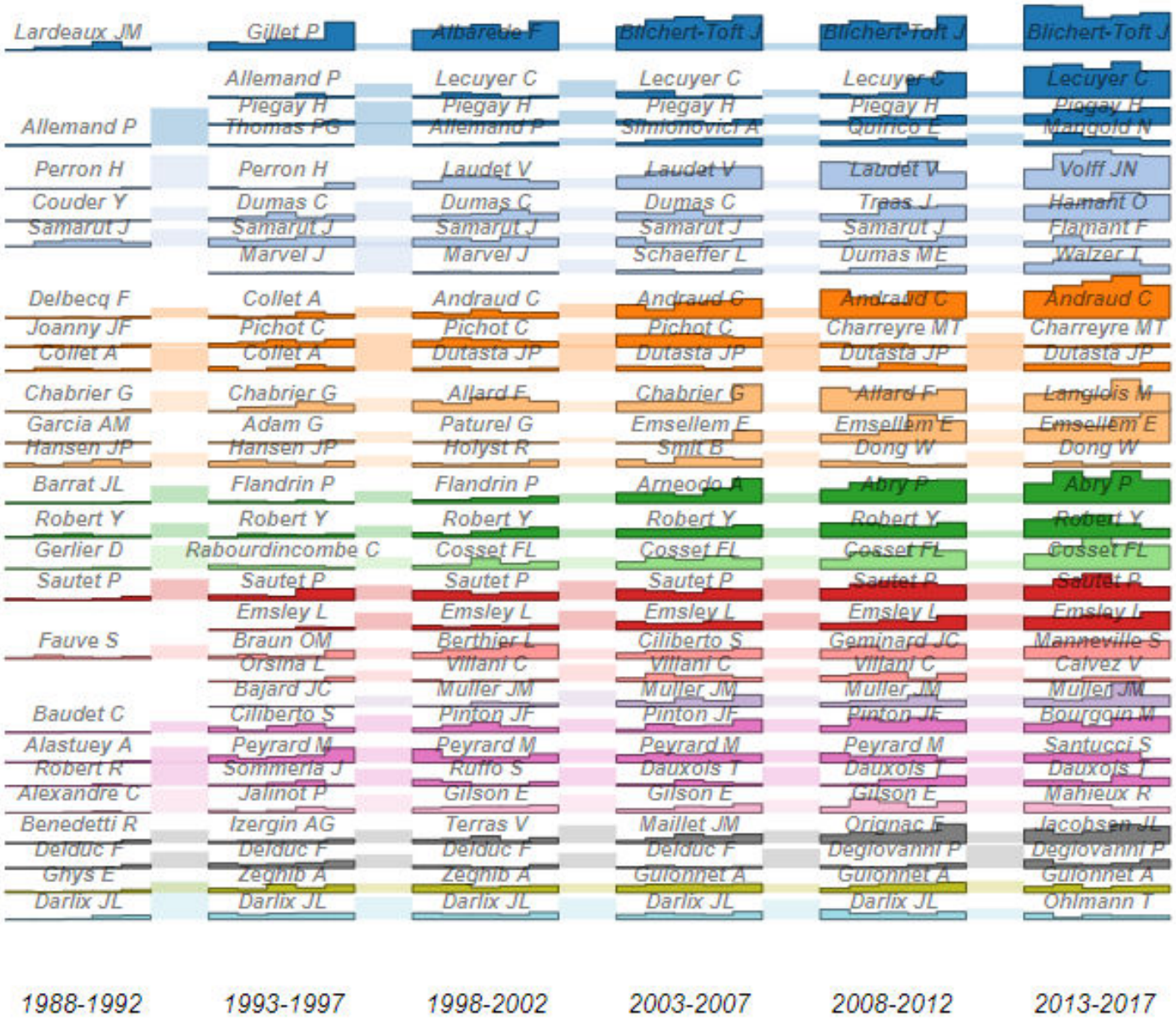
4.3 Datasets

In this section, we present specificities of each dataset and the motivations to use them. Key informations are summarized in table 4.1.

4.3.1 ENS-Lyon Publications Dataset

The ENS-Lyon Publications Dataset contains all publications produced by researchers affiliated to the *École Normale Supérieure de Lyon* in natural science fields. It spans from 1988 to 2017 and contains 16,679 publications. The ENS-Lyon is an institution, hence this dataset exhibits publications well classified into academic departments with a relatively low level of interdisciplinarity, see [Grauwin and Jensen, 2011] for a first study on this dataset. In this chapter, we compare our temporal clustering methods to a temporal partition of articles following the laboratories of the ENS-Lyon. We call it our reference partition (P_{REF}).

[1] *http://apps.isiknowledge.com/*
[2] *http://www.sebastian-grauwin.com/bibliomaps/*



(a) Historical streams computed from the ENS Lyon natural sciences publications. Streams were determined using the global algorithm (GA).

(b) Historical streams computed from the ENS Lyon natural sciences publications. Streams were computed using our local method (BCLA).

Figure 4.1 – Labels on each stream correspond to the most frequent author name in that stream during a given period. Streams with the same color have close research topics (here the proximity of streams to each other is computed from the weight of BC network links between clusters of a same period). Bar height is proportional to the number of publications in a given year. Links between streams show the streams that are preceding/following each other.

4.3.2 Wavelets Publications Dataset

The Wavelets Publications Dataset contains all publications related to wavelets and spans from 1910 to 2012 (however the period before 1960 contains only a few publications). This dataset contains 6,582 publications, corresponding to all the publications of a list of 83 key actors in the field of wavelets selected by expert advice and bibliographic searches. See [Morini et al., 2017] where the dataset was initially presented. The study of this dataset represents a difficult task because it emerged from the collaboration of several research fields, constituted by many entangled subfields. Based on the knowledge of field's expert, Patrick Flandrin from the physics laboratory of the ENS Lyon, we built manually a temporal partition drawing the history of wavelets. We refer to this partition as P_{REF} and compare our automatically generated partitions to this partition of reference. We acknowledge that this partition is not an absolute ground truth as it relies on the subjectivity of an expert. However, it gives an approximation of the field dynamics.

4.4 Results

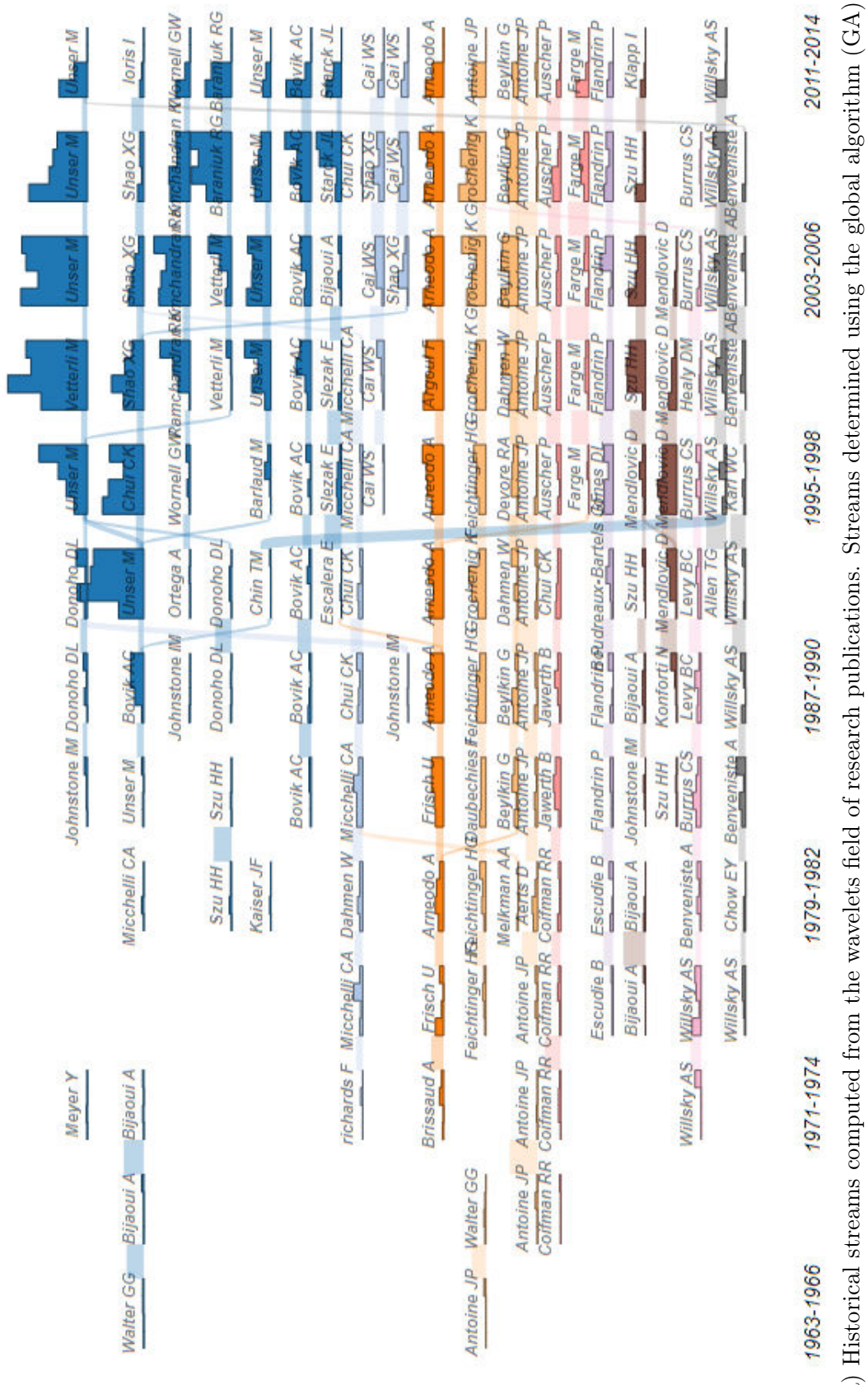
As illustrated in Figures 4.1 and 4.2, the global method (a) does not include much dynamics in the historical streams it returns. GA-streams are flat in Figure 4.1a and have only a few splits/merges in 4.2a. On the opposite, BCLA-streams give a dynamical history of research for both datasets. In Figure 4.1b many streams remain flat due to the low multidisciplinarity of research between the different departments of the institution. However, we can see some splits corresponding to teams splitting to focus on different research topics (like streams 'Blichert-Toft/Lecuyer'). Similarly, many splits and merges occur in 4.1b.

As we observe these differences in dynamics, we compare the partitions for each dataset using two sets of measures to better describe their dynamics: (1) mutual information based measures and (2) measures from a bipartite network representation of the partitions. Then, we show some examples of major differences in partitioning that we spotted using bipartite network representation.

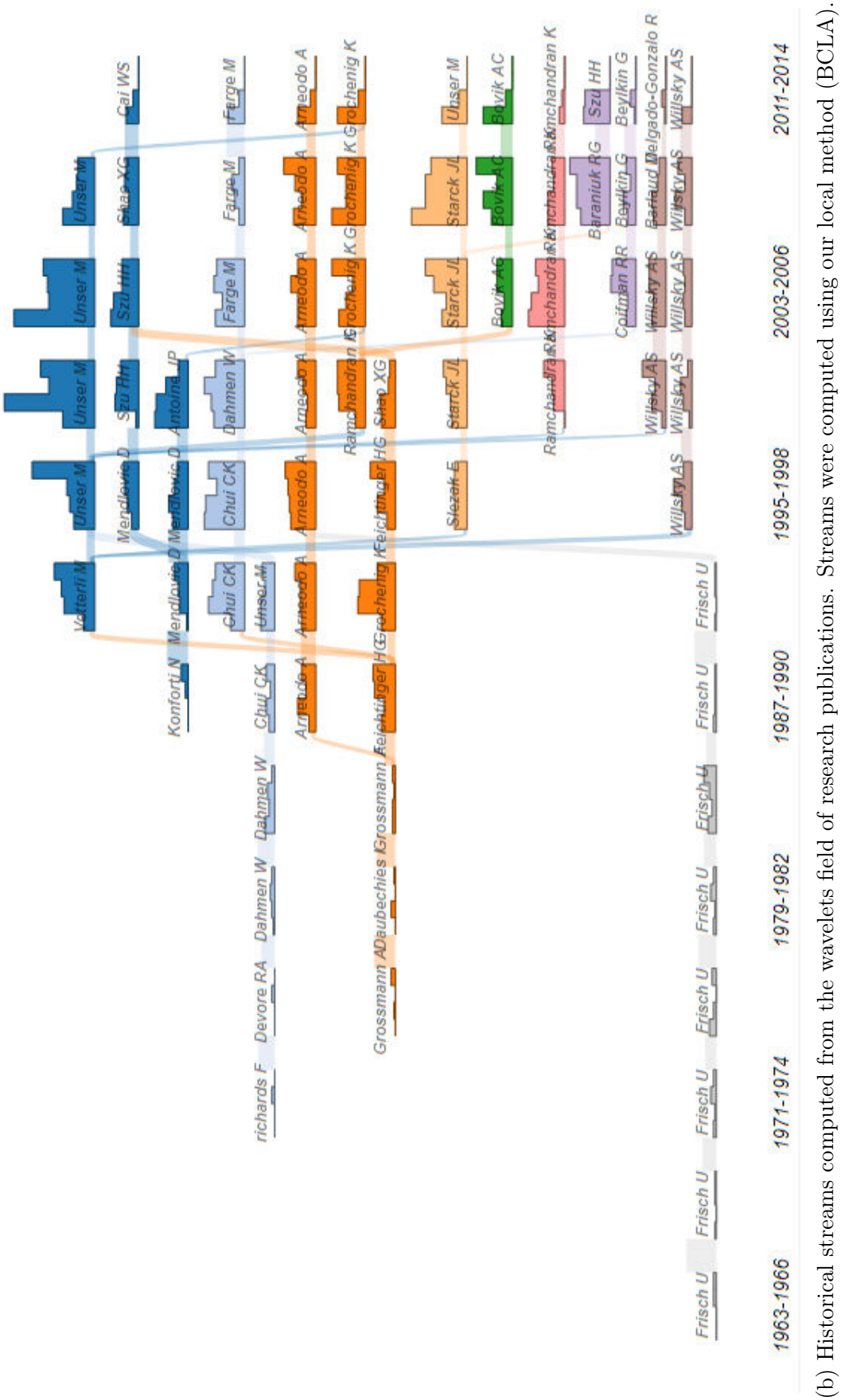
4.4.1 Normalized Mutual Information

The mutual information (MI) is a widely used measure for comparing community detection algorithms. It is defined as a measure of the statistical independence between two random variables (see eq. 4.2). In other words, if $H(P_X)$ is the entropy associated with partition X and $H(P_Y)$ is the entropy associated with partition Y -the entropy is a measure of how partitioned is our network, the more communities (here temporal streams), the higher the entropy- then $MI(P_X, P_Y)$ represents the overlap of the two partitions. It gives an answer to the question: how much do I know about the partition P_X when the partition P_Y is given? Note that the mutual information is a symmetrical measure, that is $MI(P_X, P_Y) = MI(P_Y, P_X)$. See [Wagner and Wagner, 2007, Kvålseth, 2017] for deeper description on mutual information.

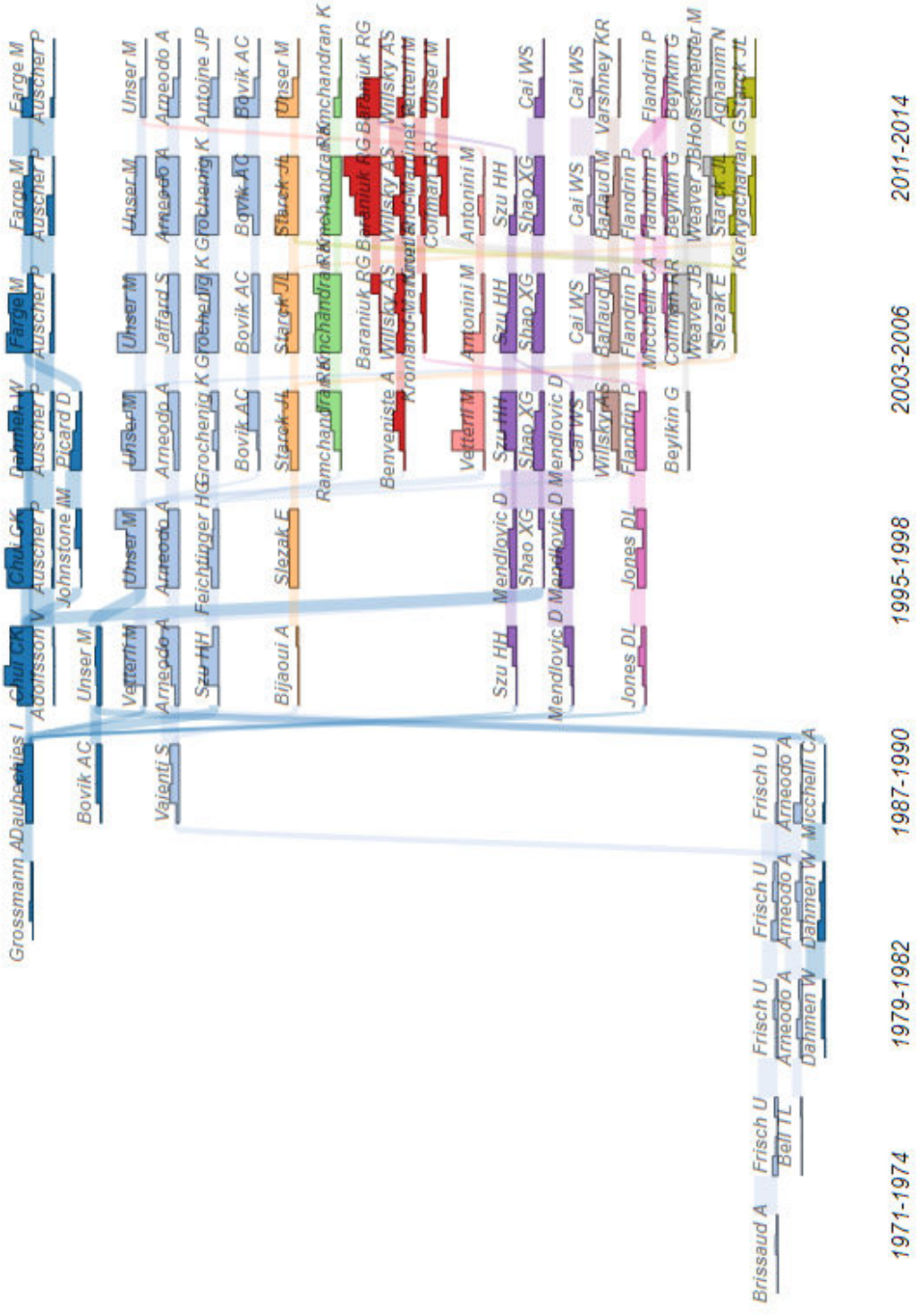
$$MI(P_X, P_Y) = H(P_Y) - H(P_Y|P_X) = MI(P_Y, P_X) \tag{4.2}$$



(a) Historical streams computed from the wavelets field of research publications. Streams determined using the global algorithm (GA).



(b) Historical streams computed from the wavelets field of research publications. Streams were computed using our local method (BCLA).



(c) Historical streams computed from the wavelets field of research publications. Streams are the reference streams determined manually.

Figure 4.2 – Labels on each stream correspond to the most frequent author name in that stream during a given period. Streams with the same color have close research topics (here the proximity of streams to each other is computed from the weight of BC network links between clusters of a same period). Bar height is proportional to the number of publications in a given year. Links between streams show the streams that are preceding/following each other.

MI is defined on $[0, +\infty]$, therefore it is difficult to make sense of it without an upper-bound. There exists different ways to normalize the mutual information. The idea is to take into account the entropies of the partitions we consider, so that it is possible to gauge the proportion of mutual information between the partitions. Normalizing by the entropy of one of the partition, e.g. $H(P_X)$ (see eq. 4.3) measures how much of the partition P_X is included in the partition P_Y. We call this normalized mutual information NMI_X. If it reaches its maximum value 1, then it is possible to retrieve all the information -all the partition- of P_X from the partition P_Y. However this measure does not take into account the size of the other partition, P_Y. A partition P_Y where each node would be its own community would make NMI_X equals to 1 even though both partitions are very different. This measure then needs to be combined with at least another NMI which takes into account the relative size of both partitions (see eq. 4.4). Here the mutual information is normalized by $\sqrt{H(P_X) * H(P_Y)}$, which shows how much of the two entropies overlap on a scale between 0 and 1. This expresses how similar are the partitions. It is equal to 1 when the partitions are the same. Moreover, this last NMI is symmetrical, so it takes into account both retrieval of P_X from P_Y and retrieval of P_Y from P_X.

$$NMI_X(P_X, P_Y) = \frac{MI(P_X, P_Y)}{H(P_X)} \tag{4.3}$$

$$NMI(P_X, P_Y) = \frac{MI(P_X, P_Y)}{\sqrt{H(P_X) * H(P_Y)}} = NMI(P_Y, P_X) \tag{4.4}$$

Entropies, MI and NMIs computed on the 3 datasets are given in table 4.2.

Mutual information based measures can give a value of similarity between two partitions. However, it is not straightforward to analyze and it does not allow to track where the (dis)similarity comes from. To allow in depth comparison, we represent pairs of partitions as bipartite networks. We present our results in the next section.

4.4.2 Bipartite Network of streams

To track and quantify differences between partitions X and Y, we compute a bipartite network where the $n_X^i \in N_X$ are the first kind of nodes. They represent the streams $s_X^i \in P_X$ (hence $|N_X| = |P_X|$). It follows that the second kind of nodes $n_Y^j \in N_Y$ represent the streams $s_Y^j \in P_Y$. A weighted directed edge is drawn between n_X^i and n_Y^j only if their corresponding streams s_{GA}^i and s_{BCLA}^j share articles. For a given pair of nodes (n_X^i,n_Y^j) the weights of the two edges between them (one in each direction) are defined in eq.4.5. We quantify differences between streams of two partitions from this graph, quantities are given in table 4.3.

$$\begin{cases} w_{n_X^i \to n_Y^j} = \dfrac{|s_X^i \cap s_Y^j|}{|s_X^i|} \\ w_{n_Y^j \to n_X^i} = \dfrac{|s_Y^j \cap s_X^i|}{|s_Y^j|} \end{cases} \tag{4.5}$$

Measures	ENS-Lyon	Wavelets
$\|P_{GA}\|$	57	27
$\|P_{BCLA}\|$	97	36
$\|P_{REF}\|$	17	36
$H(P_{GA})$	3.63	2.87
$H(P_{BCLA})$	4.05	3.04
$H(P_{REF})$	2.37	3.18
$MI(GA, REF)$	1.93	2.03
$MI(BCLA, REF)$	1.93	2.49
$MI(GA, BCLA)$	3.10	1.90
$NMI_{GA}(GA, REF)$	0.53	0.73
$NMI_{REF}(GA, REF)$	0.82	0.64
$NMI(GA, REF)$	0.66	0.68
$NMI_{BCLA}(BCLA, REF)$	0.48	0.84
$NMI_{REF}(BCLA, REF)$	0.81	0.80
$NMI(BCLA, REF)$	0.63	0.82
$NMI_{GA}(GA, BCLA)$	0.86	0.67
$NMI_{BCLA}(GA, BCLA)$	0.77	0.62
$NMI(GA, BCLA)$	0.81	0.64

Table 4.2 – $|P_X|$ is the number of streams in partition X. $H(P_X)$ is the entropy of partition X. $MI(P_X, P_Y)$ is the mutual information between the partitions X and Y. NMI_X is the mutual information MI normalized by $H(P_X)$. $NMI(P_X, P_Y)$ is the symmetrical normalized mutual information (normalized by $\sqrt{H(X) * H(Y)}$).

Figure 4.3 shows a part of the bipartite network between P_{GA} (left) and P_{BCLA} (right) on the ENS Lyon publications dataset. The part of the network is centered on nine streams from P_{GA} equivalent to 17 streams from P_{BCLA}.

4.4.3 Results on ENS-Lyon Dataset

From Table 4.2, the first thing to notice on this dataset is the very different number of streams of each partition. Our global method GA has 57 streams whereas our local method $BCLA$ contains 97 streams. The reference partition contains 17 streams, which are the 17 laboratories of the ENS-lyon in natural sciences. The high values of $NMI_{REF}(GA, REF)$ (0.82) and $NMI_{REF}(BCLA, REF)$ (0.81) suggests that the extra streams in both P_{GA} and P_{BCLA} are mostly hierarchical subdivisions of the laboratory streams from P_{REF}. A partition being a subdivision of another does not result in a decrease of MI between them. The MI decreases only if communities of a partition need to be mixed to become communities of another. These results argue that P_{GA} and P_{BCLA} are both the same description as P_{REF} at different scales. Similarly, the high value of $NMI(GA, BCLA)$ (0.81) suggests that P_{BCLA} and P_{GA} share mostly the same information.

The measures from Table 4.3 confirm the previous analysis. $\overline{1^{st}E}(GA, REF)$ shows that streams from P_{GA} share on average $86 \pm 17\%$ of their articles with a stream from P_{REF} and an average of 3.37 ± 1.76 streams from P_{GA} are needed to retrieve 80% of streams from P_{REF}. Similar observations can be made for P_{BCLA}. Moreover, $\overline{Sum_{80}}(GA, BCLA)$ shows that it takes on average two streams from P_{BCLA} to reach 80% of streams from P_{GA}. This is illustrated in Figure 4.3

These observations confirm that here streams from P_{REF} are not mix of different streams parts from P_{GA} (P_{BCLA}). But they are unions of (almost) entire streams. In this case, GA and BCLA yield almost the same partition, only different in scale.

Examples of major differences

Here, we illustrate our analysis with BN representations. We show the hierarchical subdivisions between P_{GA} and P_{BCLA} for the ENS Lyon dataset. In figure 4.3, we see that P_{GA} streams (in red) can be almost entirely retrieved from the union of a few *complete* P_{BCLA} streams (in blue). Sometimes P_{BCLA} streams share articles with more than one P_{GA} stream. However, for all P_{BCLA} streams sharing articles with more than one P_{GA} streams, the proportion of shared articles with the most similar article is more than five times higher on average than the other proportions of shared articles with the other P_{GA} streams. This illustrates that P_{GA} differs from P_{BCLA} mostly by scale.

4.4.4 Results on Wavelets Dataset

Describing the history of the wavelet research field is a complicated task as it is born from the collaboration of multiple fields and sub-fields. The values from table 4.2 point that, even though partitions have closer number of streams (27 for P_{GA} and 36 for P_{BCLA}) they are significant differences between our local and global method. In this case, $NMI(BCLA, REF)$ is significantly higher than $NMI(GA, REF)$ (0.82 vs.

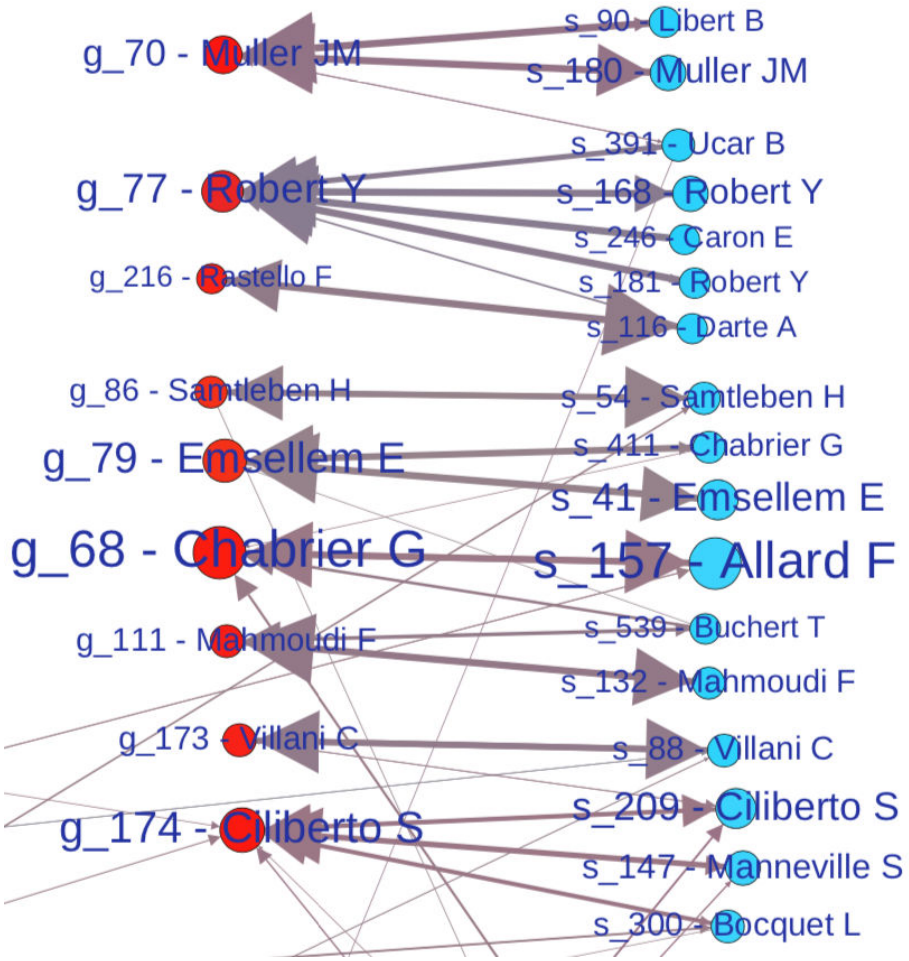


Figure 4.3 – Part of the bipartite network representation of ENS Lyon dataset. This network shows the links between temporal communities from P_{GA} (left in red) and P_{BCLA} (right in blue). On each node is given the stream ID and the most frequent author name of the temporal community. Size of nodes accounts for the size of the streams, each stream contains at least 20 articles.

Measures	ENS-Lyon	Wavelets
$\overline{1^{st}E}(GA, REF)$	0.86 ± 0.17	0.75 ± 0.20
	0.49 ± 0.20	0.81 ± 0.17
$\overline{Sum_{80}}(GA, REF)$	1.26 ± 0.54	1.88 ± 0.93
	3.37 ± 1.76	1.5 ± 0.73
$\overline{1^{st}E}(BCLA, REF)$	0.89 ± 0.14	0.87 ± 0.17
	0.49 ± 0.26	0.87 ± 0.15
$\overline{Sum_{80}}(BCLA, REF)$	1.23 ± 0.44	1.26 ± 0.50
	4.87 ± 3.35	1.31 ± 0.57
$\overline{1^{st}E}(GA, BCLA)$	0.74 ± 0.23	0.72 ± 0.23
	0.85 ± 0.16	0.83 ± 0.19
$\overline{Sum_{80}}(GA, BCLA)$	1.96 ± 1.14	1.88 ± 0.96
	1.34 ± 0.51	1.61 ± 0.83

Table 4.3 – In this table each cell contains two lines. Each measure $M(X, Y)$ is made on edges. The first line correspond to M measured on edges from n_X to n_Y and the second line corresponds to M being measured on edges from n_Y to n_X. So, the first row in $\overline{1^{st}E}(X, Y)$ is the average proportion of articles n_X shares with n_Y $\pm$ its standard deviation. The second row is the average proportion of articles n_Y shares with n_X $\pm$ its standard deviation. For instance, for the ENS-Lyon, this means that streams of P_{GA} share on average 86% of their articles with their most similar stream in P_{REF}, whereas streams from P_{REF} only share on average 49% of their articles with their most similar stream in P_{GA}. $\overline{Sum_{80}}(X, Y)$ is the average number of streams from P_Y it takes to retrieve 80% of the streams' articles from P_X. For example in the case of the Wavelet Dataset, on average 1.88 ± 0.96 streams from P_{BCLA} are needed to retrieve 80% of a stream from P_{GA}.

0.68). Moreover $NMI(GA, BCLA)$ is rather low (0.64) which suggests that differences here are not only due to the difference of scales. We visualize some of these differences in the section 4.4.4.

From Table 4.3 we see that most similar streams between P_{GA} and P_{REF} share $75\% \pm 20\%$ of articles on average, whereas it is $87\% \pm 17\%$ of articles shared between P_{BCLA} and P_{REF}.

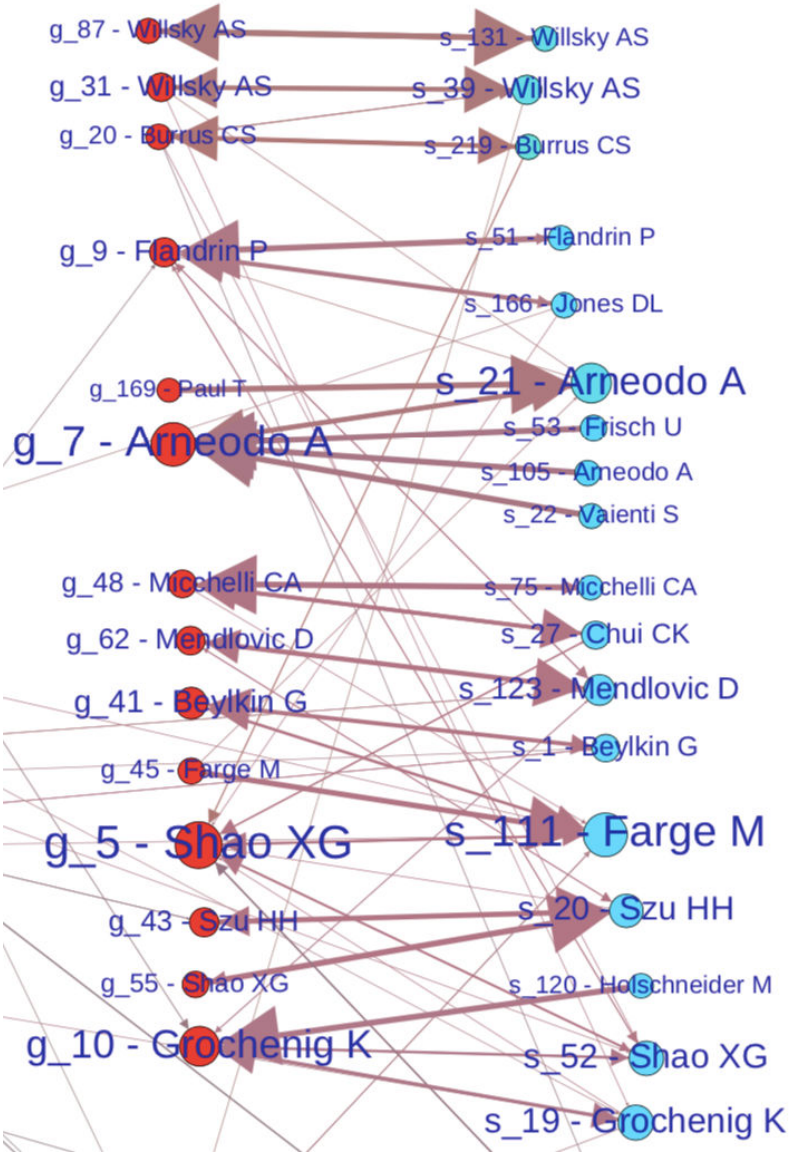
Examples of major differences

We now show some major differences between P_{GA} and P_{BCLA} for the wavelets dataset. From Figure 4.4a, we can see two kind of differences between partitions: scale differences (e.g. g_9 with s_42 and s_9) like in the ENS Lyon case; and differences where P_{GA} streams are mixed to retrieve P_{BCLA} streams, such as the group of streams around g_5 and g_31. Interestingly, g_7 cumulates scale and mixing differences.

If we now look at the BN representation of the same P_{BCLA} streams with corresponding P_{REF} streams (Figure 4.4b), we see that our P_{BCLA} description is closer to the description in P_{REF}. There are more 'stream-to-stream' equivalences, represented by the double arrow on each side of the edge linking streams. Note that, though P_{BCLA} is closer to P_{REF}, there are still scale differences (e.g. s_21, s_111) and mixing dif-

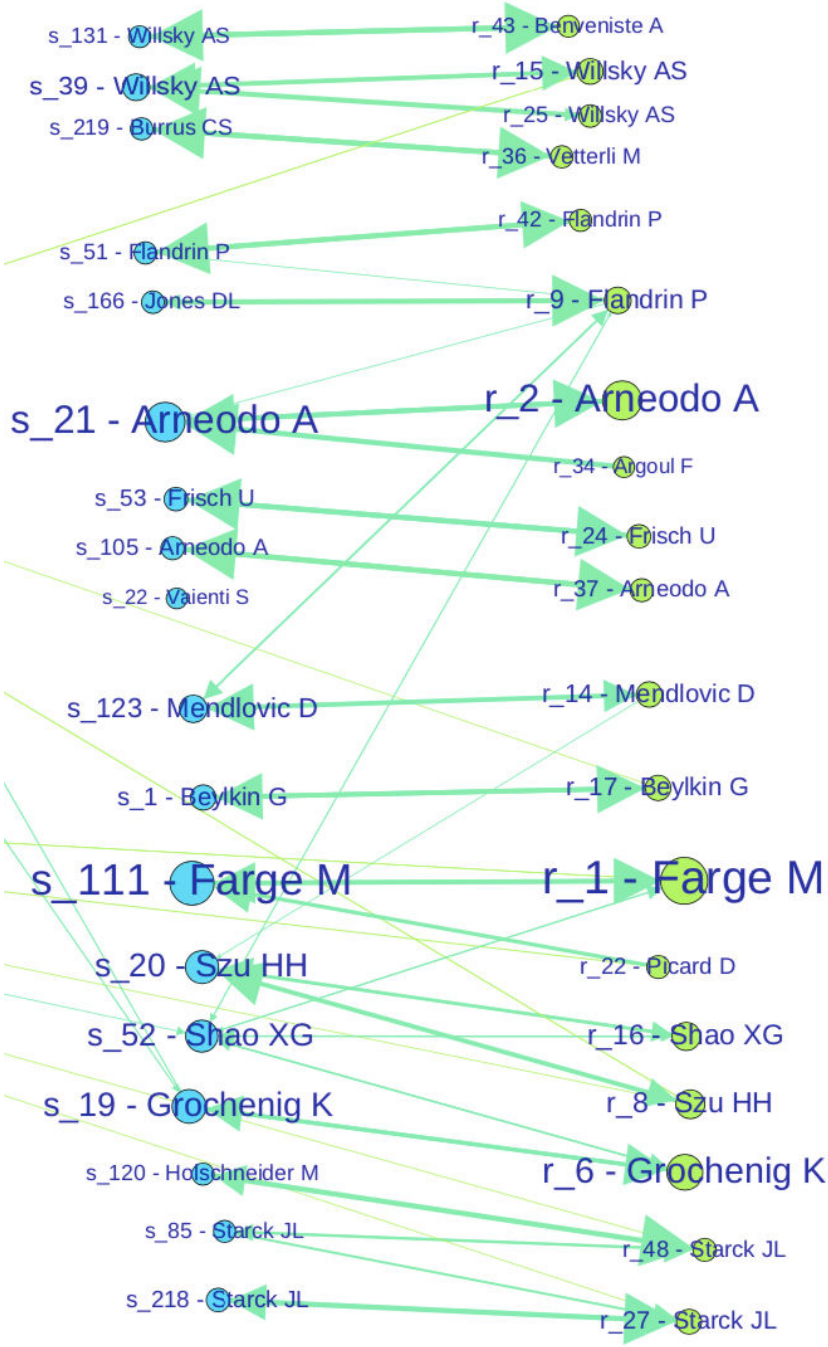
ferences (e.g. s_85, s_52).

It is also worth noting the difference between s_21 and s_53. The global method (GA) merges both into one stream g_7 (see Fig. 4.4a), However, P_{REF} and P_{BCLA} partitioned them separately (respectively (s_21, s_53) and (r_2 and r_24). If we look at Figure 4.2b, we see that these streams do not belong to the same time period. s_53 corresponds to one of the early works on wavelets, from 1963 to 1994. The second one, s_21 came after (1987 - 2014) working on different problems.



(a) Part of the bipartite network representation of Wavelets dataset. This network shows the links between temporal communities from P_{GA} (left in red) and P_{BCLA} (right in blue). On each node is given the stream ID and the most frequent author name of the temporal community. Size of nodes accounts for the size of the streams, each stream contains at least 20 articles.

From these examples we saw that our $BCLA$ method takes into account better the complexity in the dynamics than global methods which simply merged these streams into one in P_{GA}.



(b) Part of the bipartite network representation of Wavelets dataset. This network shows the links between temporal communities from P_{BCLA} (left in blue) and P_{REF} (right in green). On each node is given the stream ID and the most frequent author name of the temporal community. Size of nodes accounts for the size of the streams, each stream contains at least 20 articles.

Figure 4.4 – Part of the bipartite network representation of Wavelets dataset.

4.5 Discussion

In our method, we first analyzed mutual information based measures. We saw that MI based measures are interesting to quantify *how* different two partitions are. But they are limited when it comes to describing these differences. For this reason it was necessary to visualize the partitions. The bipartite network representation allowed us to see how streams split between partitions. We saw with some examples that global methods could do as well as local methods to automatically generate history from datasets with a clear structure (high modularity), like in the case of ENS-Lyon data. But it reached limitations when working on datasets where streams are entangled as illustrated in Section 4.4.4. On the opposite, our $BCLA$ could better take into account the complexity in the dynamics and led to a partition closer to the reference, manually drawn by the field expert.

Now, for the last chapter of this thesis, in the fashion of a grand finale, we will combine complex networks, mobility data of customers, and supervised machine learning. More precisely, we will look at how mobility data can enable us to study the impact of new businesses on already established ones.

5 Dynamics of Retail Environments

Keywords: Urban mobility; Spatio-temporal patterns; Predictive modeling

5.1 Introduction

Location is known to be highly influential in the success of a new business opening in a city. Where a business is positioned across the urban plane not only determines its reach by clienteles of relevant demographics, but more critically, it determines its exposure to a local ecosystem of businesses who strive to increase their own share in a local market. The types of businesses and brands that are present in an urban neighborhood in particular has been shown [Jensen, 2006] to play a vital role in determining whether a new retail facility will grow and blossom, or instead whether it will become a sterile investment and eventually close. Competition is nonetheless only one determinant in retail success. How a local business establishes a *cooperative network* with other places in its vicinity has been shown to also play a decisive role in its sales growth [Daggitt et al., 2016]. Local businesses can complement each other by exchanging customer flows with regards to activities that succeed each other (e.g. going to a bar after dining at a restaurant), or through the formation of urban enclaves of similar local businesses that give rise to characteristic identities that then become recognisable by urban dwellers. A classic example of the latter is the presence of many Chinese restaurants in a Chinatown [Zhou, 2010].

It is therefore natural to hypothesize that the rise of a business lies on the complex interplay between cooperation and competition that manifests in a local area. Measuring these cooperative and competitive forces in a city remains, however, a major challenge. Today's cities change rapidly driven by urban migration and phenomena such as gentrification as well as large urban development projects, which can lead to shops opening and closing at increasing rates. Already in 2011, the *fail rate* of restaurants in certain cities, such as New York, was as high as 80% [1] with some businesses closing in only a matter of months. A similar picture has been reported recently for high street retailers in the United Kingdom with part of the crisis being also attributed

[1] *https://www.businessinsider.com/new-york-restaurants-fail-rate-2011-8*

to the increasing dominance of online retailers [2].

Data generated in location technology platforms by mobile users who navigate the city provides a unique opportunity to respond to the aforementioned challenges. In addition to providing quick updates, in almost real time, on the places that open (or close) in cities - thus accurately reflecting the set of local businesses in a given area - they offer a view on urban mobility flows between areas and places at fine spatial and temporal scales. The ability to describe these two dimensions of urban activity - places and mobility - paves the way for measuring the impact, either positive or negative, that retail facilities have on each another. In this chapter, we harness this opportunity, building on a longitudinal dataset by Foursquare that describes mobility interactions between places in 26 cities around the world. Our contribution are summarized in more detail in the following:

- **Detecting patterns of cooperation in urban activity networks:** We model businesses in a city as a connected network of nodes belonging to different activity types. We examine the properties of these networks spatially and temporally. With respect to a null network model, we observe higher clustering coefficient, higher modularity and lower closeness centrality scores which are indicators of strong tendencies for local businesses to cluster and form collaborative communities that exchange customer flows. In numerical terms, the modularity of urban activity networks is ≈ 0.6 relative to the corresponding null models with ≈ 0.15.

- **Measuring the impact of new businesses using spatio-temporal metrics:** We next model the impact of a new business opening in a given area. Previous work [Daggitt et al., 2016] conducted a preliminary analysis of homogeneous impact of new businesses. However, this work was limited as it did not isolate numerous contributing factors such as whether multiple new venues have opened when attributing the impact to a given business. We develop a methodology more robust to bias through the use of spatio-temporal filters. Moreover our approach generalizes to heterogeneous interactions between venue types present in an urban system by considering impact measurements between different venue categories (e.g. measuring the impact of a restaurant to a bar). Notably, we observe that the opening of a Fast Food Restaurant near other Fast Food Restaurants results in the most significant competition, with a median decline in customer flows of 21% over 6 months. We also show how this competitive ranking can be created for heterogeneous categories.

- **Predicting optimal retail environment using urban dynamics and network topology measures:** Lastly, we built a supervised learning model to predict the impact of multiple new venues on an existing venue. We incorporated additional network metrics into our model to consider the urban environment of a given venue and its capacity to operate cooperatively with other places in vicinity. We observe significant heterogeneity between categories where some have more predictable trends while others have greater variations. In light of this

[2] *https://www.theguardian.com/cities/ng-interactive/2019/jan/30/high-street-crisis-town-centres-lose-8-of-shops-in-five-years*

heterogeneity across venue types we tailor supervised learning models by training them in a manner that reflects the idiosyncrasies of its category. Despite the inherent difficulty of the prediction task, due to the multi-factorial nature of dynamic and complex interactions in retail ecosystems, our results suggest that incorporating complex signals in predictive machine learning frameworks can offer meaningful insight in real world application scenarios. For certain business categories, AUC scores above 0.7 are consistently attained, whereas complex network metrics consistently boost the performance of classifiers offering a clear advantage over baseline methods.

Our results are especially important in a digital age with shifting customer preferences as physical business are forced to adapt to remain competitive. Our methodology can enable a better understanding of interactions within local retail ecosystems. Modern data and methods, such as those employed in the present chapter, not only can allow for monitoring these phenomena at scale, but also offer novel opportunities for retail facility owners to assess the risk of opening a new venture through location-based analytics. Similar methods can be applied beyond the scope of the retail sector we study here, namely for urban planning and innovation e.g. by assessing the impact of opening transport hubs, leisure and social centers or health and sanitation facilities in city neighborhoods.

5.2 Related Work

Understanding retail ecosystems and determining the optimal location for a business to open have for long been questions in operations research and spatial economics [Ghosh and Craig, 1983, Eiselt and Laporte, 1989]. Compared to modern approaches, these methods were characterized by static datasets informing on population distribution across geographies, tracked through census surveys and the extraction of retail catchment areas through spatial optimization methods [Applebaum, 1966]. Gravity models on population location and mobility later became a common approach for site placement of new brands [Gibson and Pullen, 1972].

The availability of spatio-temporally granular urban datasets and the popularization of spatial analysis methods in the past decade led to a new generation of approaches to quantify retail success in cities. In this line, network-based approaches have been proposed to understand the retail survival of local businesses through quality assessment on the interactions of urban activities locally [Jensen, 2006]. In addition to networks of places, street network analysis emerged as an alternative medium to understand customer flows in cities, with various network centrality being proposed as a proxy to understand urban economic activities [Crucitti et al., 2006, Porta et al., 2012].

More recently, machine learning and optimization methods have been introduced to solve location optimization problems in the urban domain, focusing not only on retail store optimization [Karamshuk et al., 2013] but also real estate ranking [Fu et al., 2014] amongst other applications. Location technology platforms such as Foursquare opened the window of opportunity for customer mobility patterns to be studied at fine spatio-temporal scales [D'Silva et al., 2018b, D'Silva et al., 2018a] and moreover,

semantic annotations on places presented direct knowledge on the types of urban activities that emerge geographically and led to works that allowed for the tracking and comparing of urban growth patterns at global scale [Daggitt et al., 2016]. Closer to the spirit of the present chapter from a modelling perspective, the authors in [Hidalgo and Castanẽr, 2015] study co-location patterns of urban activities in Boston and subsequently recommend areas where certain types of activities may be missing.

5.3 Dataset Description

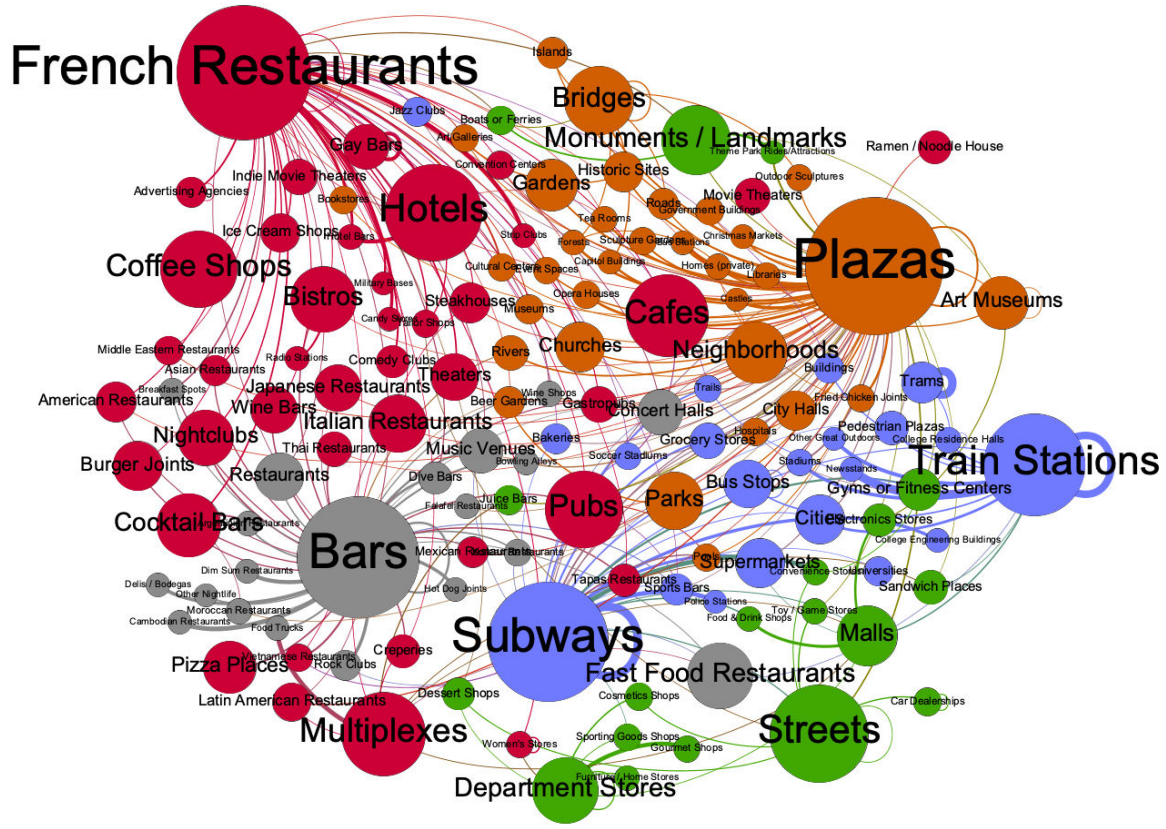
Within the last decade, Online Location-based Services have experienced a surge in popularity, attracting hundreds of millions of users worldwide. These systems have created troves of data which describe, at a fine spatio-temporal granularity, the ways in which users visit different businesses and areas of a city. We hypothesize these data can be used to build a predictive model of the impact of a new venue on the surrounding businesses. To this end, we utilize data from Foursquare, a location technology platform with a consumer application that allows users to check into different locations. As of August 2015, Foursquare had more than 50 million active users and more than 10 billion check-ins [VentureBeat, 2015].

The basis of our analysis is a longitudinal dataset from 26 cities that spans three years, from 2011 to 2013, and included over 80 million checkins. We aggregate data from the 10 most represented cities in North America[3] and the 16 most represented cities in Europe[4]. A summary of our data is described in Table 5.1.

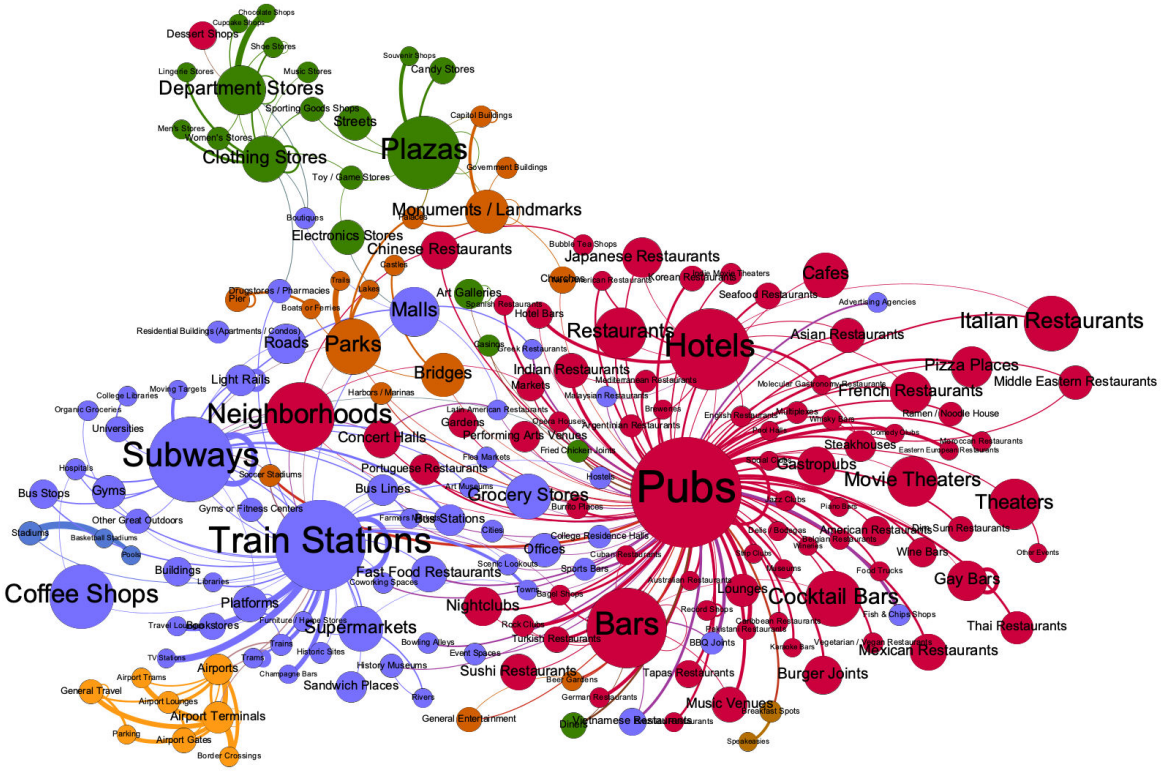
For each venue, we have the following information: geographic coordinates, specific and general category, creation date, total number of check-ins, and number of unique visitors. The specific and general categories fall within Foursquare's API of hierarchical categories. A full list of the categories can be found by querying the Foursquare API [Foursquare, 2018]. The dataset also contains a list of transitions within a given city. A transition is defined as a pair of check-ins by an anonymous user to two different venues within the span of three hours. It is identified by a start time, end time, source venue, and destination venue. We consider the set of venues V in a city. A venue $v \in V$ is represented with a tuple $< loc, date, category >$ where loc is the geographic coordinates of the venue, $date$ is its creation date, and $category$ is the specific category of the venue. The creation date, $date$, for a given venue refers to the date it was added to the Foursquare platform. Prior work by Daggitt et al. [Daggitt et al., 2016] showed that across all cities when examining the number of venues added per month, the last 20% of venues were new venues rather than existing venues added to the database for the first time. We apply this methodology to all cities and define new venues as those that fall within the last 20% of venues added to Foursquare for that city.

[3]North American cities are: Austin, Boston, Dallas, San Francisco, New York City, Houston, Las Vegas, Los Angeles, Toronto, Washington.

[4]European cities are: Amsterdam, Antwerpen, Barcelona, Berlin, Brussels, Budapest, Copenhagen, Gent, Helsinki, Kiev, Madrid, Milano, Paris, Prague, Riga, London



(a) Paris



(b) London

Figure 5.1 – Network visualization of categories during the evening in Paris (a) and London (b). Different colors represent different Louvain communities [Blondel et al., 2008]. We see clusters of travel and transport (blue), nightlife (red), and shopping activities (green).

5.4 Urban Activity Networks

We begin by examining transitions between Foursquare venues of different category types that we refer to as urban activities. While we are considering a mix of categories users check in in the city, our focus from an analysis and modelling point of view will be focusing on urban activities corresponding to retail establishments (e.g. restaurants).

5.4.1 Visualizing Mobility Interactions

To visualize an urban activity network, we create a graph G_i for each city i, where the set of nodes N_{cat} is the set of business categories defined previously in Section 5.3. In this network, business categories are linked by weighted directed edges $e_{s \to d}$. A directed link is created from the source category c_s to the destination category c_d if at least one transition happens during the time window we consider (e.g. weekend, weekday, or a period of hours during a day). Thus, the weight of each edge is proportional to the total number of transitions from the source category to the destination category for the particular time period of interest for each city. The weights are then normalized by the total number of check-ins that occurred at c_d. Therefore, the weight can be interpreted as the percentage of customers of c_d who come from c_s. To eliminate insignificant links, we filter out edges that have less than 50 transitions total. We examine two time intervals of interest: morning AM (6am-12pm) and evening PM (6pm-12am).

In Figure 5.1 we visualize the network in the evening for two cities, London and Paris. The colors represent different communities, obtained using the Louvain community detection algorithm [Blondel et al., 2008]. Further, the size of nodes is proportional to their degree. This visualization, as one example, describes similarities and variations in the structure of urban activities in different cities. We observe an underlying common structure for the two cities, even though cultural distinctions can also be noted. We have observed a similar pattern across many cities. In terms of similarity in network structure, we see a shopping cluster (green) centered around Department Stores; a cluster for travel and transport (blue) centered around categories such as Train Stations and Subways; a leisure cluster (light brown) centered around Plazas and containing outdoor categories (e.g. Parks, Gardens, Soccer Stadiums). On the other hand, differences in network structure become also apparent. We note for instance how recreation activities in the evenings differs across the two cities. London has a considerably large nightlife cluster (red) centered around pubs from which a number of different nightlife categories unfold (e.g. Nightclubs, restaurants of different types, Theater). Paris is more segregated and contains two nightlife clusters: one

Region	# venues	# new venues	# transitions
North America	94,094	29,552	43,200,432
Europe	101,101	40,275	44,600,446
Total	**195,195**	**69,827**	**87,800,878**

Table 5.1 – Foursquare dataset description for 10 North American cities and 16 European cities.

cluster around French Restaurants (red) linked to Coffee Shops, Theaters, Nightclubs; and another cluster (gray) centered around Bars which contains Food Trucks, Fast Food Restaurants, and Music Venues. This dichotomy translates to the presence of two classes of customers each of which adheres to different types of activity sequences during nighttime. Another observation is regarding variations in network structure over time: the Coffee Shop category in London is separated from the nightlife cluster, which may indicate different kind of customer behaviors between daytime and evening. Interestingly, we also see associations emerging between types of businesses. Taking Paris as an example, French Restaurants interact a lot with Coffee Shops and Nightclubs and so do Bars with Food Trucks. In both cities, Coffee Shops are drawing crowds from Subways, Toy Stores with Electronics Stores and Sport Stores.

Overall, these results suggest strong structural characteristics in urban activity networks where different categories of places form interaction patterns of cooperation, where mobile users move from one to the other. Competition on the other hand manifests in a more implicit manner in the network in two ways: first, retail facilities that are grouped in the same node (e.g. Bars) have to share customers that have been previously performing a different activity (e.g. going to a Restaurant) and second, through activities that do not share an edge in the network and as a result they do not interact with one another in terms of mobility patterns.

5.4.2 Network Properties

We next quantify the structure of these networks in terms of different network properties considering also different time intervals. For our two cities of comparison, we list the network metrics in Table 5.2 and enlist those next:

- *the average clustering coefficient*, $< C >$, is the tendency of categories to form triangles, that is to gather locally into fully connected groups. It varies between 0 and 1 with higher values implying a higher number of triangles in the network (see [Newman, 2010] for more details).

- *the average closeness centrality*, $< C_c >$, is the average length of the shortest path between the nodes and here accounts for the tendency of categories to be close to each in terms of shortest paths [Newman, 2010]. It varies between 0 and 1 where a higher closeness centrality score for a node suggests higher proximity to other nodes in the network.

	London				Paris			
	AM	Random AM	PM	Random PM	AM	Random AM	PM	Random PM
# of nodes	204	204	208	208	180	180	149	149
# of edges	2271	2271	3055	3055	2160	2160	2184	2184
$< C >$	0.657	0.304	0.645	0.301	0.744	0.134	0.645	0.173
$< C_c >$	0.313	0.552	0.402	0.541	0.352	0.513	0.434	0.551
Q	0.583	0.150	0.608	0.131	0.588	0.148	0.641	0.161

Table 5.2 – Network metrics for London and Paris for during the morning AM (6am - 12pm) and evening PM (6pm-12am). These metrics are compared to an Barabási-Albert model (Random).

- *the modularity*, Q, is a well established metric indicating how well defined communities are within the network [Blondel et al., 2008]. Modularity values fall within the range $[-1, 1]$, with greater positive values indicating greater presence of community structure.

We compare our network metrics to a random baseline which maintains the degree distribution, a Barabási-Albert network [Albert and Barabási, 2003], in Table 5.2. The comparison with the null model provides an indication of how significant empirical observations are with respect to the random case. First we note that for all three metrics the real networks are very different to the corresponding null models. In general, high clustering coefficient and modularity together with a lower closeness centrality scores point to the tendency of local businesses to form significantly tight clusters that are well isolated from one another. Furthermore, these networks properties vary for different period of the day and are subtly different from city to city. We see for instance closeness centrality being higher in the evening relative to morning hours whereas and the average clustering being lower in the evening. This means that categories are less locally connected to each other during evening hours. This could be due to the fact that users have more well planned series of movements in the morning following daily commuting routines. For instance, users are more likely to travel directly from a train station to an office in the morning on a weekday while they may instead wander from a pub to a bar in the evenings.

Looking closer at the network modularity scores presented in Table 5.2 we note a clear partitioning of different categories into communities with scores around 0.6 for both cities compared to much smaller values ≈ 0.15 for the null model. Modularity values increase in the evening in both cities. This translates to a stronger community structure, suggesting customers may be less likely to experience activities in different category communities than in the morning and confirm the dichotomies from previous sections such as *Coffee Shops vs. Pubs* in London and *Bars vs. French Restaurants* in Paris). Finally, the similarity in terms of network properties values between the two cities, as well as the prominent community structure in both suggest that the hypothesis that the organization of the retail business ecosystem is similar across cities is a plausible one. This is true to a certain degree nonetheless, as variations are also noted due to apparent cultural differences.

5.5 Measuring Impact

In this section we examine the impact of new businesses on other establishments within their vicinity. Previous work [Daggitt et al., 2016] has considered the *homogeneous* impact of a new business opening, that is the impact that venue categories have on categories of the same type (e.g. the impact of a new Coffee Shop on another Coffee Shop). In this section we generalize the impact metric to heterogeneous mixes of categories, and develop a methodology more robust to spatio-temporal bias effects.

5.5.1 Spatio-temporal Scope of Impact

To measure the impact of a new venue opening in an area we need to define its geographic scope. We define the spatial neighborhood of a venue as the set of venues

that are located within the radius r_s. Formally, we define the spatial neighbourhood of a venue v_n as:

$$SN(v_n, r_s) = \{v_e \in V : dist(v_n, v_e) < r_s \wedge v_n \neq v_e\} \tag{5.1}$$

where V is the set of venues in the city and $dist(v_n, v_e)$ is the Euclidean distance between venue v_n and v_e.

Further we introduce a similar notion across the temporal dimension. In particular, given a temporal radius r_t, then two businesses opening within r_t are considered to be *temporal neighbors*. Formally, we define the temporal neighbors of a venue v_n as:

$$TN(v_n, r_t) = \{v_e \in V : t\_dist(v_n, v_e) < r_t \wedge v_n \neq v_e\} \tag{5.2}$$

where t_dist is the difference in the number of months between the creation date of v_n and v_e. Finally, we define W_T as the temporal window of observation, i.e. the total period in months over which we measure the number of check-ins at a given business.

5.5.2 Measuring Impact

Defining the impact formula

We base our impact metric $I_{v_n}(v_e, t_{v_n})$ of a new venue v_n on an existing venue v_e on the metric introduced in [Daggitt et al., 2016]. This is defined as the normalized number of transitions for an existing venue in a specific time period prior to and after a new venue opens in its spatial neighborhood. We define the number of check-ins during the posterior time interval as follows:

$$C^{post}_{v_n}(v_e, t_{v_n}, \Delta t) = \sum_{d=t_{v_n}}^{t_{v_n}+\Delta t-1} n_{v_e}(d, d+1) \tag{5.3}$$

This calculates the sum of check-ins at venue v_e, n_{v_e}, after the opening of v_n at t_{v_n} and over a period of $\Delta t = W_T/2$ months. We define the number of checkins during the prior time interval as follows:

$$C^{prior}_{v_n}(v_e, t_{v_n}, \Delta t) = \sum_{d=t_{v_n}}^{t_{v_n}-\Delta t-1} n_{v_e}(d-1, d) \tag{5.4}$$

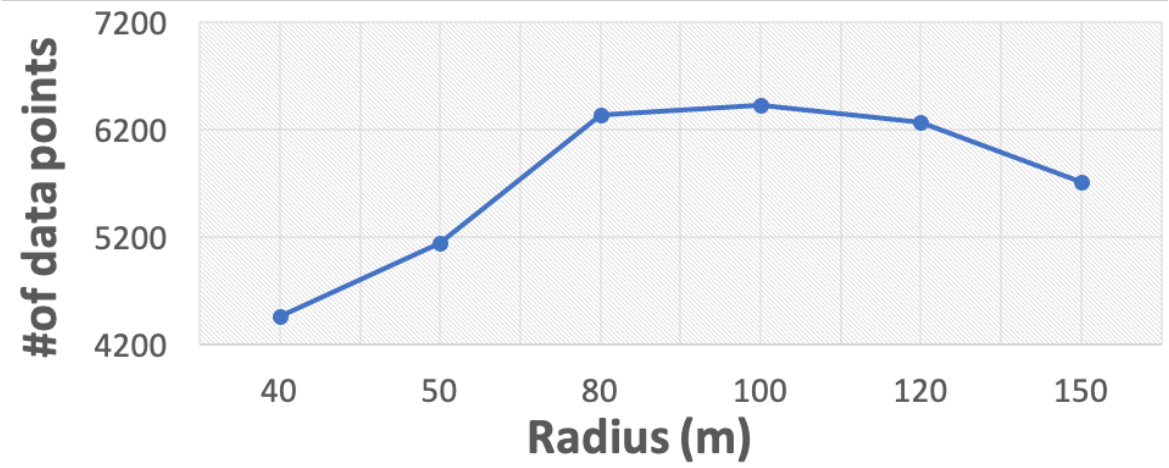
Similarly, this calculates the number of check-ins over the period of Δt months before the opening of v_n. Formally, the impact metric is defined as follows:

$$I_{v_n}(v_e, t_{v_n}, \Delta t) = \frac{C^{post}_{v_n}(v_e, t_{v_n}, \Delta t)/N_{cat}(t_{v_n}, t_{v_n} + \Delta t)}{C^{prior}_{v_n}(v_e, t_{v_n}, \Delta t)/N_{cat}(t_{v_n}, t_{v_n} - \Delta t)} \tag{5.5}$$

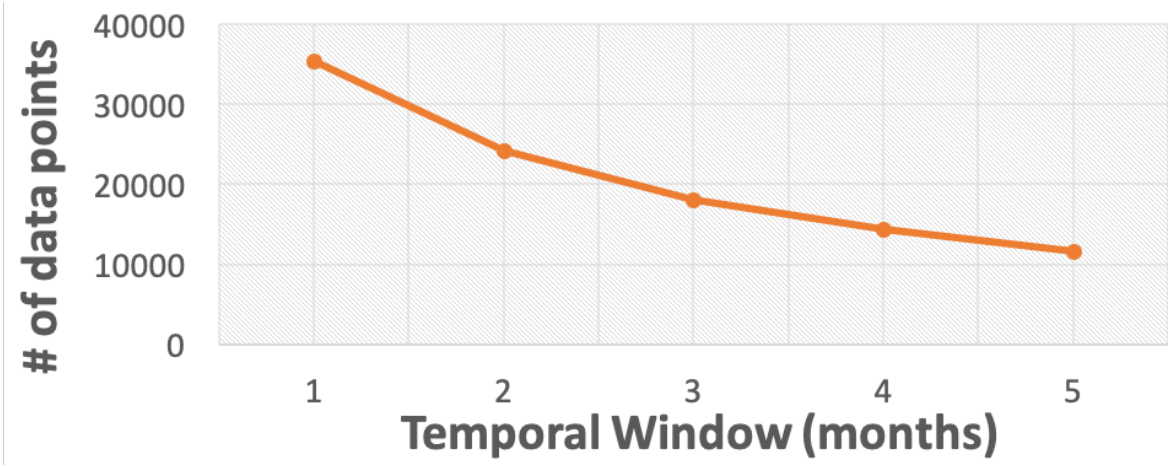
$N_{cat}(t_{v_n}, t_{v_n} + \Delta t)$ (and respectively $N_{cat}(t_{v_n}, t_{v_n} - \Delta t)$) is the total number of check-ins to all businesses of v_e's category during the period $[t_{v_n}, t_{v_n} + \Delta t]$ (respectively $[t_{v_n}, t_{v_n} - \Delta t]$). This normalizing factor takes into account both the background trend of the category and the potential season effect which can occur in the market. This factor is calculated at a per city level.

Category	Median Impact	% Businesses $I < 0$	# Businesses
Fast Food Restaurants	0.79	67.9	28
Bakeries	0.81	73.1	26
Pizza Places	0.81	69.3	39
Coffee Shops	0.84	66.4	202
Sandwich Places	0.84	62.0	63

Table 5.3 – Ranking of the categories which have the strongest homogeneous negative impact.



(a) Effect of the size of the spatial radius on the number of data points when r_t is set to 3 months. An optimal value of r_s is drawn for 100 m.



(b) Effect of the size of the temporal radius on the number of data points when r_s is set to 100m. The number of data points continues to decrease for $r_t \geqslant 1$ months.

Figure 5.2 – Tuning the spatial and temporal parameters of our model.

Impact metric interpretation

Above, we define $N_{cat}(t_{v_n}, t_{v_n} + \Delta t)$ as the total number of check-ins to all businesses of a specific category. We define the *market share* of a given venue as the total number of check-ins to that venue as a percentage of N_{cat}. Using the impact metric defined in Section 5.5.2 we can calculate the percentage of market share gained or lost by venue v_e after the opening of venue v_n. For example, if v_n is a clothing store and v_e is a sandwich shop $I_{v_n}(v_e, t_{v_n}, \Delta t) = 1.2$ means the market share of venue v_e (the sandwich shop) increased by 20% over the Δt months following the opening of v_n (the clothing store). Conversely, a number below 1 means the market share of the venue v_e decreased after the opening of venue v_n, suggesting the new venue was a competitor. For the example above if $I_{v_n}(v_e, t_{v_n}, \Delta t) = 0.81$ this would correspond to the market share of venue v_e (the sandwich shop) decreasing by 19% over the Δt months following the opening of v_n (the clothing store).

5.5.3 Tuning Spatial and Temporal Windows

A major challenge in analyzing correlation in a complex system is to implement a methodology which can limit the effect of potential hidden variables which may bias the observed correlation.

For our model, our methodology must isolate the impact of *one business* in constantly changing neighborhoods. For this reason, as a first analysis, we set r_t to $\Delta t = W_T/2$, that is half of the window of observation, and we only choose a pair (v_n, v_e) if the new business v_n has no temporal neighbors (defined above in Equation 5.2). In other words, if we set $\Delta t = W_T/2 = 3$ months, we only measure the impact of v_n on v_e if v_e was open at least 3 months before v_n and if no other business opened in the spatial neighborhood of v_e during $[t_{v_n} - \Delta t, t_{v_n} + \Delta t]$. We apply this form of filtering to isolate the confounding effect of multiple businesses opening next to each other. This naturally introduces a trade-off with regards to the data points considered for our measurement.

The number of data points resulting from our approach depends on two parameters. The spatial radius r_s and the temporal radius r_t. As shown in Figure 5.2(a) where r_t is fixed at 3 months, when the spatial radius is small the probability of a new business nearby opening is low. Hence the number of data points remains small as well. As the spatial radius increases so does the number of data points until a maximum is reached after which it decreases again. This likely due to the fact that after a certain distance threshold the probability of finding only one business opening within r_s decreases as well. Based on these results we set parameter r_s to 100 meters for our subsequent analysis. We apply the same reasoning for tuning the temporal radius r_t setting r_s to 100m. However, tuning r_t is more challenging. Our hypothesis is that after a time period, $t_{v_n} + r_t$, the positive (or negative) impact of a new business will become stable within the scope of a neighborhood. After this period of time, v_n will be considered as part of the baseline of the neighborhood, when examining the impact of future new venues. Our hypothesis is that as r_t increases, the less likely we are to be considering other new shops that are interfering with the impact we measure. Figure 5.2(b) shows that the number of data points steadily decreases from $r_t = \Delta t > 1$ month onwards. This implies that the optimal value for r_t in terms of data points is smaller than

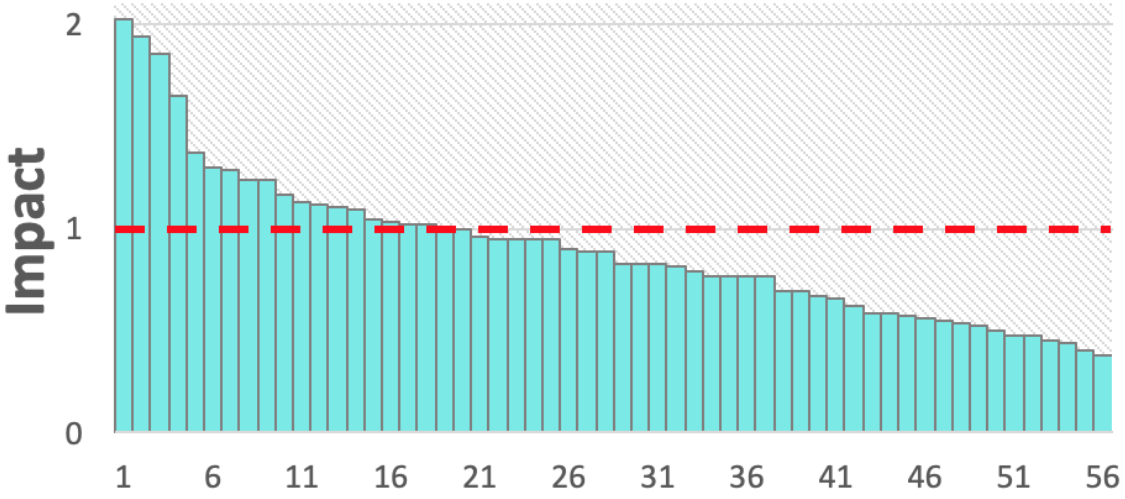


Figure 5.3 – Plot of the impact measure of Coffee Shops on Burger Joints. Each column is a Burger Joint for which only a Coffee Shop opened in its neighborhood during a period $\Delta t = 3$ months. We see that 37/56 joints are over $I = 1$, that is 66% of Burger Joints have been impacted negatively by the opening of a Coffee Shop.

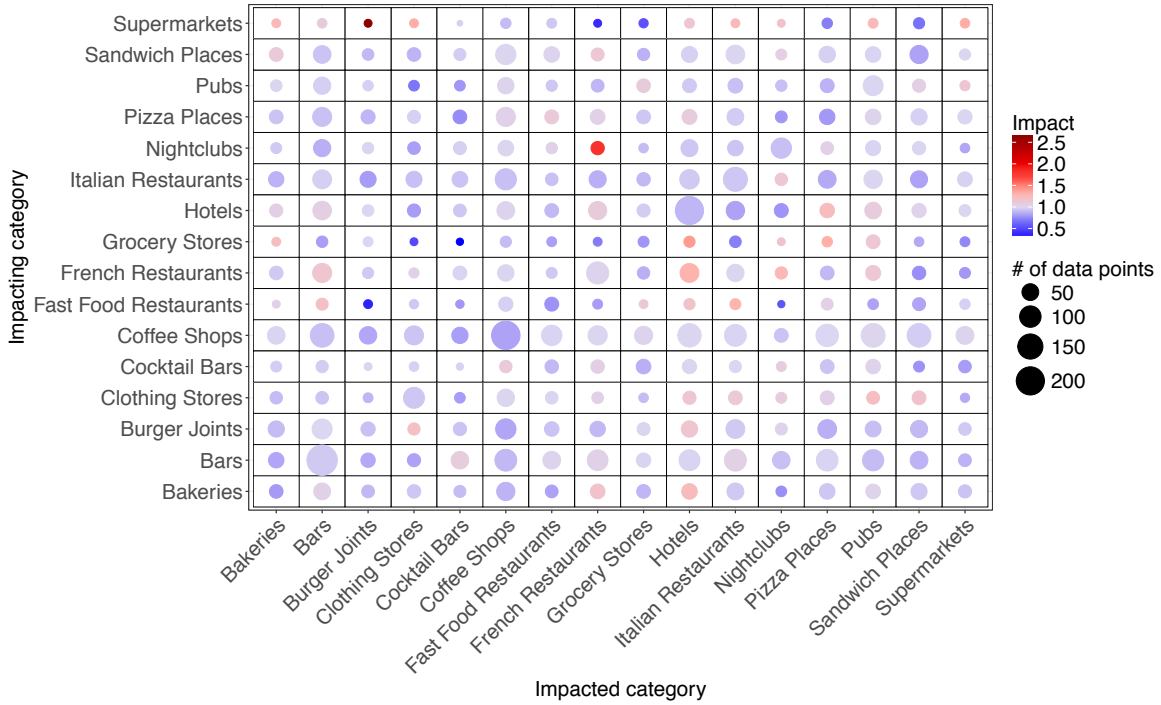


Figure 5.4 – Matrix of median impact of categories on each other. The size of the circles represent the number of pairs of businesses which were used (e.g. Coffee Shops - Coffee Shops is 200). The color varies according to the median impact value of the pair, dark blue corresponds to impact below 1 (representing competition), grey represents neutral median impact equal to 1, and dark red translates to high median impact (representing cooperation). Read: French Restaurants (9th row) have a median impact on Bars (2nd column) of 1.2 and less than 100 pairs were counted.

one month. However, as suggested earlier increasing r_t limits the risk to consider previously new venues that opened just before the time window of interest in the impact measure. For this reason we choose a value of r_t longer than one month. For the analysis described below, we set r_t to three months.

5.5.4 Measuring Impact on Retail Activity

We perform the analysis using the aforementioned setup on the 16 most popular retail categories and aggregated data from the 26 cities listed above in Section 5.3. We aggregated data from these cities based on observed similarities in consumer trends as discussed in Section 5.4. This enabled us to build a set of $10,238$ businesses. The top 16 retail categories examined were as follows: Bakeries, Bars, Burger Joints, Clothing Stores, Cocktail Bars, Coffee Shops, Fast Food Restaurants, French Restaurants, Grocery Stores, Hotels, Italian Restaurants, Nightclubs, Pizza Places, Pubs, Sandwich Places, and Supermarkets.

Figure 5.3 shows the impact measured for each Burger Joint where a Coffee Shop opened within its spatial neighborhood. We observe a heterogeneous distribution of impact scores. However 66.1% of the 56 burger joints in this configuration were impacted negatively. The median impact is 0.85 which corresponds to a 15% *decrease in demand to Burger Joint when a Coffee Shop opens within 100 meters.* In this analysis, due to the skewness of the impact measures distribution, we use the median and the percentage of positively (or negatively) impacted businesses as key numbers to represent the directed impact of pairs of categories. Furthermore, we note the structure of Burger Joints in Figure 5.3. It is composed of two parts: a high impact peak followed by a linear decline. Interestingly, this structure is consistent across all pairs of categories, generalized with a peak of $\sim 10\%$ of high impacts shops ($I > 1.5$), followed by a linear decline of impact ($\sim 80 - 90\%$ of shops), and succeeded with an optional $\sim 10\%$ discontinuously negatively impacted shops, representing competition in the area ($0 < I < 0.5$).

In Figure 5.4 we show the matrix of median impact of our top 16 retail categories on each other. A few observations are worth noting. First, impact scores on the diagonal tend to be negative. This observation shows and quantifies the direct competitive effect of businesses belonging to the same category. Second, most impact scores are negative suggesting the general tendency for retail establishments to compete. Third, some cooperative pairs are noticeable (e.g. Hotels on Pizza Place: 70% of positive impact, median impact: 1.2; Nightclubs on French Restaurants: 67% of positive impact, median impact: 1.8). We also note that certain network links seen in Figure 5.1 between categories can be found as positive impact pairs in the matrix, such as Nightclubs and French Restaurants and Fast Food Restaurants and Bars.

Building a competitive ranking

Using the aforementioned results we list a ranking of homogeneous competition in Table 5.3. This ranking shows the top categories in terms of negative impact, when a business of the same category opens within 100 meters. The list suggests that fast food restaurants feature the strongest homogeneous negative impact (-21%), closely followed by other businesses where people may visit for food or a snack, namely

Bakeries, Pizza Places, Coffee Shops, Sandwich places. Note also the high probability of negative impact for Bakeries with a 73.1% chance to be negatively impacted.

Our methodology can also be applied to heterogeneous pairs of businesses to determine which new business type would create the greatest competition or cooperation for an existing business. Using Coffee Shops as an example, we saw in Table 5.3 that a new Coffee Shop creates a median impact of $I = 0.84$. We similarly see in Figure 5.4 a competitive effect from new Burger Joints, which result in a median impact of $I = 0.85$. A competitive ranking of this type can be established for all categories to better understand trends in the impact of new businesses.

To our knowledge the methodology described above is novel, and despite the limitations that we discuss in more detail in Section 6, it allows for the principled quantification of the impact that new businesses have in the retail ecosystem they operate. We exploit this formulation in the context of a prediction task in Section 5.6 where impact scores are modeled as target labels and the goal becomes to predict the impact of new businesses opening on their local environment. From an economic perspective these results highlight the most competitive types of business categories that operate in cities and moreover the highlight cases of categories where cooperation in terms of customer flows is more likely to emerge.

5.5.5 Takeaways

The methodology we have developed in Section 5.5.2 can enable one to quantify competitive and cooperative behaviors between pairs of categories. Further, it highlights general trends, such as competition between food-related categories and cooperation between certain clusters of categories. However, reducing the complexity has two limitations. It limits the number of data as it filters out neighborhoods with more than one business opening within W_T. It also over-simplifies the problem. Often, for a given pair of categories, there is heterogeneity in the impact measure for those two categories (e.g. there is a range in the impact of a new Coffee Shop on a Bakery). This suggests there are complex interactions between a new venue and its environment and there may be additional factors to consider. In the following sections, we study the impact of combinations of multiple new businesses coupled with the network topology of their environment. This methodology leads to insights which can help determine the optimal location for a new venue.

In Section 5.6, we implicitly considered a more complex interaction setup by taking into account the combinations of multiple new businesses, modeled as features, and coupled those with network topological characteristics of the local environment. This methodology led to insights which can help determine the optimal location for a new venue of a particular category.

5.6 Predicting New Business Impact

In Section 5.5, we explored the impact of the opening of a single business on its neighborhood. Next, we investigate in the form of a prediction task the cumulative impact of multiple shops opening within the same spatial and temporal neighborhoods, as defined in the previous section.

Category	Coffee Shops	Bars	Italian Restaurants	Hotels	French Restaurants	Bakeries
Train/test size	1956/218	2275/253	1629/181	2129/237	1631/181	1526/170
AUC (Business Features Baseline)	0.611	0.603	0.618	0.621	0.701	0.771
AUC (Network Features Baseline)	0.549	0.551	0.564	0.557	0.591	0.584
AUC (All Features)	**0.627**	**0.642**	**0.658**	**0.702**	**0.742**	**0.861**

Table 5.4 – AUC Scores for a subset of categories using a Gradient Boosting model.

5.6.1 Prediction Task

To illustrate our methodology, let us consider the example of French Restaurants. To understand the impact of new venues on French Restaurants in Europe and North America, we consider all new venues that opened within the spatial radius r_s of a French restaurant within a given period r_t. We then examine how the demand of that given French restaurant changed after the opening of the new venues. We aim to predict whether the opening of those new venues will have a positive or negative impact on the French restaurant given network features about the venue, and a count of the types of venues that opened within r_t. We consider this prediction task for all retail categories aiming to understand how each category is impacted by new venues opening nearby.

We model the impact of new venues as a binary classification task where the impact on the existing venue is the dependent variable and the features described below in Section 5.6.2 are independent variables. We further represent a positive impact label as 1 and a negative impact label with a 0. Considering a supervised learning methodogy, our goal then becomes to learn an association of the input feature vector $\mathbf{x}$ with a binary label y. We experiment with a number of supervised learning algorithms described in the following paragraphs. All network features were min-max normalised. We split our dataset into training and test sets with the training set consisting of 80% of the data and the test set consisting of the remaining 20%. We perform 5-fold cross-validation to pick the best performing model and report the subsequent accuracy of prediction. Finally, we also sub-sample our dataset performing our predictions on a balanced dataset of positive and negative classes.

5.6.2 Extracting features

Business features

When a new venue v_n opens at time t_{v_n} within a given spatial neighborhood $SN(v_n, r_s)$ we count all other businesses which opened within the temporal neighborhood $TN(v_n, r_t)$ of v_n and in the same spatial neighborhood $SN(v_n, r_s)$. For each existing venue v_e, the counts of each category of venue within its spatial and temporal neighborhood is encoded as a feature. We refer to these features as *Business features*.

Network features

As described above, we utilize network topology measures at the venue neighborhood level considering each venue $v_e \in SN(v_n, r_s)$ where each venue is a node in the network. Edges in the network represent the number of transitions from the source venue to the destination venue. These features are described as follows. The *in-degree* of a venue v_e

is the number of edges coming from the other venues to v_e. The *out-degree* of a venue v_e is the number of edges coming from v_e to the other venues. The *degree* of v_e is the total number of edges between v_e and the other venues. The *closeness centrality* of a venue v_e is the average of its shortest paths to all the other venues of the network. The *diversity* of a venue v_e is defined by the Shannon equitability index from information theory which calculates the variety of the neighbors of venue v_e [Sheldon, 1969]. These features give a representation of how well a given venue is connected to the rest of the city network and also describe the distribution of its customers.

5.6.3 Evaluation

In this section, we report our findings on the predictive ability of the features as well as the supervised learning models we employ in the prediction task. We compare the predictions against different baselines:

- Network connectivity features, as described in section 5.6.2.
- Business features, as described in section 5.6.2.

We first discuss how our combined model outperforms both baselines, then show the predictive capabilities of different categories, and lastly discuss the value of network metrics in our prediction task.

Model Selection: We explored a number of different models, including Logistic Regression, Gradient Boosting, Support Vector Machines, Random Forests, and Neural Networks. As described above, we train our models to predict the impact of new businesses on a given type of retail category. When aggregating our data across all categories and working to predict the impact of a new venue on any type of existing venue category, this resulted in models with low predictive power. As discussed previously in Section 5.5, this suggests there is significant heterogeneity of behavior across different category types and therefore a training strategy tailored for each category would be more effective. To take this into account, we segregated our data by category and built a separate model for each type of category. We individually modelled the ten retail categories with the largest total number of data points. These categories were as follows: Bars, Hotels, Coffee Shops, Italian Restaurants, French Restaurants, Bakeries, Clothing Stores, Pizza Places, Nightclubs, and Grocery Stores. To compare our supervised learning models, we calculated the average AUC scores across all ten categories using our different feature baselines.

Our combined models, using both business and network features, had a resulting AUC of 0.611 for Logistic Regression, 0.687 for Gradient Boosting, 0.631 for Support Vector Machines, 0.640 for Random Forest, and 0.667 for Neural Networks. These results suggest that venue-specific metrics can support the prediction of the impact of a new venue. Across all categories, we saw that Gradient Boosting was the most robust model for our task with the highest average AUC across all ten categories. As such, we use Gradient Boosting to further analyze our models below.

Variations across Categories We next examine the predictability of different category types. Table 5.4 shows the predictability of six of the ten categories trained using a Gradient Boosting model. The AUCs of the remaining categories are as follows: Clothing Stores (AUC: 0.664, train/test size: 1362/152), Pizza Places (AUC:

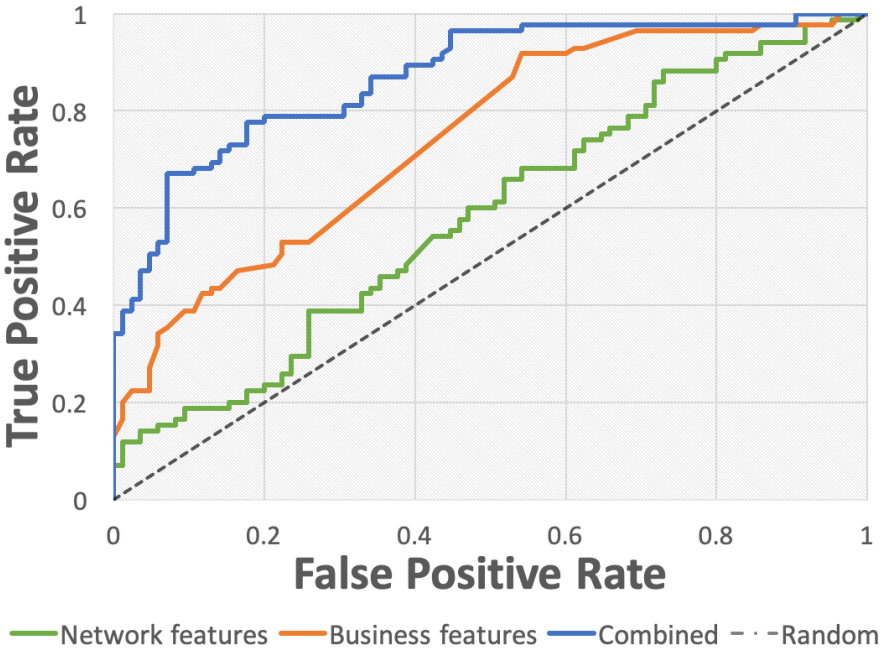


Figure 5.5 – ROC Curves of Bakeries for the performance of each class of features and for the combined model.

0.682, train/test size: 1287/143), Nightclubs (AUC: 0.686, train/test size: 622/70), and Grocery Stores (AUC: 0.690, train/test size: 581/65). We see a range in the predictability of different venue types: Bakeries have the highest AUC of 0.861 and Bars have the lowest AUC of 0.627. This suggests that certain activities adapt to a more heterogeneous set of environment, making them more challenging to predict. These more challenging categories tend to be those that are ubiquitous and general, such as Bars and Coffee Shops.

The Value of Network Metrics: To understand the influence of different feature classes on our results, we run our Gradient Boosting model on each class separately and in combination. We report our observations in Figure 5.5 and note that for Bakeries the category features alone reach AUC of ≈ 0.77, the network features alone reach AUC of ≈ 0.60 and combined model leads to a higher overall AUC of ≈ 0.85. These results, which were consistent across all retail categories, suggest using network features in lieu of venue features alone supports our prediction task. We see that the two feature classes together have the best prediction results, suggesting that the impact of new venues is driven by a number of forces, including the network connectivity, which modulate the nature of their interactions within their neighborhood.

5.7 Discussion And Future Work

Our methodology highlights the power of complex networks measures in building machine learning prediction models. This is especially valuable for systems in which interactions between agents must be taken into account. We began in Section 5.4 of

this chapter by detecting patterns of cooperation in urban networks and quantifying similarities and differences in network structure across cities. Next, in Section 5.5 we measured the impact of new businesses isolating cases in which exactly one new venue opened within a given region. Using urban network topology measures and our business impact metrics, we developed a machine learning model to predict the optimal location for a new business in Section 5.6. The novelty of our approach in methodological terms stems from the use of complex networks measures, combined with machine learning methods to tackle modern societal challenges. These results can support policy makers, business owners, and urban planners as they have the potential to pave the way for the development of sophisticated models describing urban neighborhoods and help determine the optimal conditions for establishing a new venue.

One of the limitations of the work presented here is that we only examine the impact of retail venues. However, this analysis can be more broadly applied to venues of other categories as well. The present study sets the frame for further general studies. As one example of future work, one could examine the impact of new bus stops or transport hubs on those venues within their proximity. Additionally, in this chapter we conduct a preliminary examination of the variations in network trends across cities worldwide. Future work could expand upon this to explore the duality between general network trends and cultural consumer idiosyncrasies across cities. Our analysis could also be further expanded by considering temporal network analysis, examining the variations in features across different time intervals of interest.

6 Conclusion

The explosion of connected devices all around the world has been changing the economical landscape by revolutionizing the way people interact, consume and move. These actions leave traces, footprints are left by some users for most of their actions and social dynamics emerge.

These traces are an opportunity to better understand collective behaviors. Nowadays, the plurality of data available is such that it requires adapting and combining a plurality of existing methods in order to enlarge the global vision one has on a given system. This thesis has been the opportunity for me to investigate some of the hottest topics in data analysis. From network science to machine learning, I have explored different ways to model and analyze data.

We illustrated the complexity of dynamics which can emerge from simple statistical models of society, even when models follow simplistic rules. We showed that simple models can be helpful to analyze social phenomena while keeping in mind that they do not allow for drawing any rigorous conclusion about *the real world*.

I, then, continued my exploration of complex systems modeling working on real world data - bike sharing system users data. We worked on describing temporal dynamics of long-term bike sharing system customers. We were the first to conduct such kind of analysis on *individual* dynamics. We highlighted a low seasonal effect in bike usage and two main trajectories of users: the majority of users ($\sim$ 60%) leaves the system after at most one year; and the minority ($\sim$ 40%) remains in the system for several years (average $\sim$ 3 years). We described the usage profile and socio-demographic profile of these two classes of users. Then, we used unsupervised learning (K-means algorithm) to generate segments of customers and we compared results to a static baseline from [Vogel et al., 2014].

Moving forward in the analysis of temporal data and collective dynamics, we explored tools to describe and visualize differences between temporal partitions. We used them to illustrate essential differences between global methods and local methods of temporal complex networks clustering.

Finally, using mobility data from Foursquare, we worked on describing business cooperation and competition in urban networks. This project enabled us to work on metrics to measure the impact of new businesses on their neighborhood and build supervised machine learning models aiming at predicting optimal location for new businesses. Further, we could illustrate the power of complex networks measures in

building machine learning models.

In a context of smart cities emergence, where technology is used with the goal to make transportation and urban planning more efficient, the topics of research presented in this thesis are of particular interest. They aim at a better understanding of customer usages and mobility at different layers of a city (public transportation, people's customer profiles, etc). Such research can benefit policy makers, business owners, and urban planners as understanding customers segments and trends can help design more adapted systems.

The work presented in this thesis can be expanded in many regards. In particular on the dynamics of retail businesses, we have conducted preliminary work on business trajectories and the evolution of impact over months. It shows there exists some significant trends depending on business categories. Taking into account more temporality in the design of impact metrics and machine learning features (through temporal networks metrics, opening hours of businesses, etc) together with exploring the variations in cooperation networks across cities worldwide (i.e. impact of cultural trends vs. general consumer trends) could help refine our understanding of retail environments. These are research projects of interest and I hope to continue contributing to them in the future.

Bibliography

[Albert and Barabási, 2003] Albert, R. and Barabási, A.-L. (2003). Statistical mechanics of complex networks. *Rev. Mod. Phys.*

[Applebaum, 1966] Applebaum, W. (1966). Methods for determining store trade areas, market penetration, and potential sales. *Journal of marketing Research*, pages 127–141.

[Aynaud et al., 2013] Aynaud, T., Fleury, E., Guillaume, J.-L., and Wang, Q. (2013). Communities in evolving networks: Definitions, detection, and analysis techniques. In Mukherjee, A., Choudhury, M., Peruani, F., Ganguly, N., and Mitra, B., editors, *Dynamics On and Of Complex Networks, Volume 2: Applications to Time-Varying Dynamical Systems*, pages 159–200. Springer New York, New York, NY.

[Backstrom et al., 2012] Backstrom, L., Boldi, P., Rosa, M., Ugander, J., and Vigna, S. (2012). Four degrees of separation. In *Proceedings of the 4th Annual ACM Web Science Conference*, WebSci '12, pages 33–42, New York, NY, USA. ACM.

[Barabási and Albert, 1999] Barabási, A. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.

[Barabási et al., 2002] Barabási, A., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., and Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311(3):590 – 614.

[Barabási and Oltvai, 2004] Barabási, A.-L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5:101–113.

[Barrat et al., 2008a] Barrat, A., Barthélemy, M., and Vespignani, A. (2008a). *Dynamical Processes on Complex Networks*. Cambridge University Press.

[Barrat et al., 2008b] Barrat, A., Barthélemy, M., and Vespignani, A. (2008b). *Dynamical Processes on Complex Networks*. Cambridge University Press.

[Barthélemy, 2011] Barthélemy, M. (2011). Spatial networks. *Physics Reports*, 499(1):1 – 101.

[Beecham and Wood, 2014] Beecham, R. and Wood, J. (2014). Exploring gendered cycling behaviours within a large-scale behavioural data-set. *Transp. Plan. and Technol.*, 37(1):83–97.

[Blondel et al., 2008] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

[Boccaletti et al., 2006] Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., and Hwang, D.-U. (2006). Complex networks: Structure and dynamics. *Physics Reports*, 424(4):175 – 308.

[Borgatti et al., 2009] Borgatti, S. P., Mehra, A., Brass, D. J., and Labianca, G. (2009). Network analysis in the social sciences. *Science*, 323(5916):892–895.

[Borgnat et al., 2011] Borgnat, P., Abry, P., Flandrin, P., Robardet, C., Rouquier, J.-B., and Fleury, E. (2011). Shared bicycles in a city: a signal processing and data analysis perspective. *Advances in Complex Syst.*, 14(03):415–438.

[Borgnat et al., 2013] Borgnat, P., Robardet, C., Abry, P., Flandrin, P., Rouquier, J.-B., and Tremblay, N. (2013). *A Dynamical Network View of Lyon's Vélo'v Shared Bicycle System*, pages 267–284. Springer New York, New York, NY.

[Bott, 1957] Bott, E. (1957). *Family and Social Network*. London: Tavistock Publications.

[Bourdieu, 1986] Bourdieu, P. (1986). *The forms of capital*. J. Richardson (Ed.) Handbook of Theory and Research for the Sociology of Education (New York, Greenwood).

[Boyd et al., 2003] Boyd, R., Gintis, H., Bowles, S., and Richerson, P. (2003). The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences*, 100(6):3531–3535.

[Buttner et al., 2011] Buttner, J., Mlasowsky, H., Birkholz, T., Groper, D., Fernandez, A., Emberger, G., and Banfi, M. (2011). Optimising bike sharing in european cities: A handbook. *Intelligent Energy Europe program (IEE)*.

[Caccioli et al., 2014] Caccioli, F., Shrestha, M., Moore, C., and Farmer, J. D. (2014). Stability analysis of financial contagion due to overlapping portfolios. *Journal of Banking & Finance*, 46:233 – 245.

[Caldarelli et al., 2003] Caldarelli, G., Garlaschelli, D., and Pietronero, L. (2003). Food web structure and the evolution of complex networks. In Pastor-Satorras, R., Rubi, M., and Diaz-Guilera, A., editors, *Statistical Mechanics of Complex Networks*, pages 148–166. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Castellano et al., 2009a] Castellano, C., Fortunato, S., and Loreto, V. (2009a). Statistical physics of social dynamics. *Rev. Mod. Phys.*, 81:591–646.

[Castellano et al., 2009b] Castellano, C., Fortunato, S., and Loreto, V. (2009b). Statistical physics of social dynamics. *Rev Mod Phys*, 81:591.

[Challet and Zhang, 1997] Challet, D. and Zhang, Y. (1997). Emergence of cooperation and organization in an evolutionary game. *Physica A*, 246(3):407 – 418.

[Claveau and Gingras, 2016] Claveau, F. and Gingras, Y. (2016). Macrodynamics of Economics: A Bibliometric History. *History of Political Economy*, 48(4):551–592.

[Côme and Oukhellou, 2014] Côme, E. and Oukhellou, L. (2014). Model-based count series clustering for bike sharing system usage mining: A case study with the vÉlib’ system of paris. *ACM Trans. Intell. Syst. Technol.*, 5(3):39:1–39:21.

[Côme et al., 2014] Côme, E., Randriamanamihaga, N., Oukhellou, L., and Aknin, P. (2014). Spatio-temporal Analysis of Dynamic Origin-Destination Data Using Latent Dirichlet Allocation: Application to Vélib' Bike Sharing System of Paris. In *TRB 93rd Annual meeting*, page 19p, France. Transportation Research Board.

[Comte and Martineau, 1856] Comte, A. and Martineau, H. (1856). *Social physics: From the Positive philosophy of Auguste Comte*. New York: C. Blanchard.

[Cong et al., 2017] Cong, R., Zhao, Q., Li, K., and Wang, L. (2017). Individual mobility promotes punishment in evolutionary public goods games. *Scientific Reports*, 7:14015.

[Couzin et al., 2011] Couzin, I. D., Ioannou, C. C., Demirel, G., Gross, T., Torney, C. J., Hartnett, A., Conradt, L., Levin, S. A., and Leonard, N. E. (2011). Uninformed individuals promote democratic consensus in animal groups. *Science*, 334(6062):1578–1580.

[Cross and Hohenberg, 1993] Cross, M. and Hohenberg, P. (1993). Pattern formation outside of equilibrium. *Rev Mod Phys*, 65:851–1112.

[Crucitti et al., 2006] Crucitti, P., Latora, V., and Porta, S. (2006). Centrality measures in spatial networks of urban streets. *Physical Review E*, 73(3):036125.

[Daggitt et al., 2016] Daggitt, M. L., Noulas, A., Shaw, B., and Mascolo, C. (2016). Tracking urban activity growth globally with big location data. *Royal Society Open Science*, 3(4):150688.

[Dall'Asta et al., 2006] Dall'Asta, L., Barrat, A., Barthélemy, M., and Vespignani, A. (2006). Vulnerability of weighted networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2006(04):P04006.

[de Sola Pool and Kochen, 1978] de Sola Pool, I. and Kochen, M. (1978). Contacts and influence. *Social Networks*, 1(1):5 – 51.

[Degenne and Forsé, 2004] Degenne, A. and Forsé, M. (2004). *Les réseaux sociaux*. A. Colin.

[Donoho, 2017] Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4):745–766.

[D'Silva et al., 2018a] D'Silva, K., Jayarajah, K., Noulas, A., Mascolo, C., and Misra, A. (2018a). The role of urban mobility in retail business survival. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(3):100:1–100:22.

[D'Silva et al., 2018b] D'Silva, K., Noulas, A., Musolesi, M., Mascolo, C., and Sklar, M. (2018b). Predicting the temporal activity patterns of new venues. *EPJ Data Science*, 7(1):13.

[Duan et al., 2009] Duan, D., Li, Y., Jin, Y., and Lu, Z. (2009). Community mining on dynamic weighted directed graphs. In *Proceedings of the 1st ACM International Workshop on Complex Networks Meet Information & Knowledge Management*, CNIKM '09, pages 11–18, New York, NY, USA. ACM.

[Eckmann and Ruelle, 1985] Eckmann, J. and Ruelle, D. (1985). Ergodic theory of chaos and strange attractors. *Rev Mod Phys*, 57:617–656.

[Eiselt and Laporte, 1989] Eiselt, H. A. and Laporte, G. (1989). Competitive spatial models. *European Journal of Operational Research*, 39(3):231–242.

[Faghih-Imani et al., 2014] Faghih-Imani, A., Eluru, N., El-Geneidy, A. M., Rabbat, M., and Haq, U. (2014). How land-use and urban form impact bicycle flows: evidence from the bicycle-sharing system (bixi) in montreal. *Journal of Transport Geography*, 41:306 – 314.

[Fehr and Gachter, 2002] Fehr, E. and Gachter, S. (2002). Altruistic punishment in humans. *Nature*, 415:137.

[Fishman et al., 2013] Fishman, E., Washington, S., and Haworth, N. (2013). Bike share: A synthesis of the literature. *Transport Reviews*, 33(2):148–165.

[Fortunato and Barthélemy, 2007] Fortunato, S. and Barthélemy, M. (2007). Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41.

[Fortunato and Hric, 2016] Fortunato, S. and Hric, D. (2016). Community detection in networks: A user guide. *Physics Reports*, 659:1 – 44. Community detection in networks: A user guide.

[Foursquare, 2018] Foursquare (2018). Get details of a venue. https://developer.foursquare.com/docs/api/venues/details. [Online; Last accessed 09-May-2018].

[Fu et al., 2014] Fu, Y., Ge, Y., Zheng, Y., Yao, Z., Liu, Y., Xiong, H., and Yuan, J. (2014). Sparse real estate ranking with online user reviews and offline moving behaviors. In *Data Mining (ICDM), 2014 IEEE International Conference on*, pages 120–129. IEEE.

[Fuller et al., 2011] Fuller, D., Gauvin, L., Kestens, Y., Daniel, M., Fournier, M., Morency, P., and Drouin, L. (2011). Use of a new public bicycle share program in montreal, canada. *American Journal of Preventive Medicine*, 41(1):80 – 83.

[Ghasemian et al., 2016] Ghasemian, A., Zhang, P., Clauset, A., Moore, C., and Peel, L. (2016). Detectability thresholds and optimal algorithms for community structure in dynamic networks. *Phys. Rev. X*, 6:031005.

[Ghosh and Craig, 1983] Ghosh, A. and Craig, C. S. (1983). Formulating retail location strategy in a changing environment. *The Journal of Marketing*, pages 56–68.

[Gibson and Pullen, 1972] Gibson, M. and Pullen, M. (1972). Retail turnover in the east midlands: A regional application of a gravity model. *Regional Studies*, 6(2):183–196.

[Görke et al., 2013] Görke, R., Maillard, P., Schumm, A., Staudt, C., and Wagner, D. (2013). Dynamic graph clustering combining modularity and smoothness. *J. Exp. Algorithmics*, 18:1.5:1.1–1.5:1.29.

[Görke et al., 2010] Görke, R., Maillard, P., Staudt, C., and Wagner, D. (2010). Modularity-driven clustering of dynamic graphs. In Festa, P., editor, *Experimental Algorithms*, pages 436–448, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Granovetter, 1973] Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1380.

[Grauwin et al., 2009] Grauwin, S., Bertin, E., Lemoy, R., and Jensen, P. (2009). Competition between collective and individual dynamics. *Proceedings of the National Academy of Sciences*, 106(49):20622–20626.

[Grauwin and Jensen, 2011] Grauwin, S. and Jensen, P. (2011). Mapping scientific institutions. *Scientometrics*, 89(3):943.

[Grauwin and Sperano, 2018] Grauwin, S. and Sperano, I. (2018). Bibliomaps-a software to create web-based interactive maps of science: The case of ux map. *Proceedings of the Association for Information Science and Technology*, 55(1):815–816.

[Greene et al., 2010] Greene, D., Doyle, D., and Cunningham, P. (2010). Tracking the evolution of communities in dynamic social networks. In *2010 International Conference on Advances in Social Networks Analysis and Mining*, pages 176–183.

[Guo et al., 2014] Guo, C., Wang, J., and Zhang, Z. (2014). Evolutionary community structure discovery in dynamic weighted networks. *Physica A: Statistical Mechanics and its Applications*, 413:565 – 576.

[Guo et al., 2017] Guo, Y., Zhou, J., Wu, Y., and Li, Z. (2017). Identifying the factors affecting bike-sharing usage and degree of satisfaction in ningbo, china. *Plos One*.

[Hartmann et al., 2016] Hartmann, T., Kappes, A., and Wagner, D. (2016). Clustering evolving networks. In Kliemann, L. and Sanders, P., editors, *Algorithm Engineering: Selected Results and Surveys*, pages 280–329. Springer International Publishing, Cham.

[Hébert-Dufresne et al., 2017] Hébert-Dufresne, L., Allard, A., Noël, P., Young, J., and Libby, E. (2017). Strategic tradeoffs in competitor dynamics on adaptive networks. *Scientific Reports*, 7:7576.

[Hetzer and Sornette, 2013] Hetzer, M. and Sornette, D. (2013). An evolutionary model of cooperation, fairness and altruistic punishment in public good games. *PLOS ONE*, 8(11):1–13.

[Hidalgo and Castanẽr, 2015] Hidalgo, C. A. and Castanẽr, E. E. (2015). Do we need another coffee house? the amenity space and the evolution of neighborhoods. *ArXiv*.

[Holme, 2015] Holme, P. (2015). Modern temporal network theory: a colloquium. *The European Physical Journal B*, 88(9):234.

[Holme and Saramäki, 2012] Holme, P. and Saramäki, J. (2012). Temporal networks. *Physics Reports*, 519(3):97 – 125. Temporal Networks.

[Jain et al., 2018] Jain, T., Wang, X., Rose, G., and Johnson, M. (2018). Does the role of a bicycle share system in a city change over time? a longitudinal analysis of casual users and long-term subscribers. *Journal of Transport Geography*, 71:45 – 57.

[Jensen, 2006] Jensen, P. (2006). Network-based predictions of retail store commercial categories and optimal locations. *Phys. Rev. E*, 74:035101.

[Jensen, 2018] Jensen, P. (2018). *Pourquoi la societe ne se laisse pas mettre en equations*. Seuil (Paris).

[Jensen et al., 2010] Jensen, P., Rouquier, J.-B., Ovtracht, N., and Robardet, C. (2010). Characterizing the speed and path of shared bicycle use in lyon. *Transp. Res. Part D: Transp. Environ.*, 15:522–524.

[Jovanovic and Schinckus, 2013] Jovanovic, F. and Schinckus, C. (2013). The Emergence of Econophysics: A New Approach in Modern Financial Theory. *History of Political Economy*, 45(3):443–474.

[Karamshuk et al., 2013] Karamshuk, D., Noulas, A., Scellato, S., Nicosia, V., and Mascolo, C. (2013). Geo-spotting: mining online location-based services for optimal retail store placement. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 793–801. ACM.

[Karsai and Perra, 2017] Karsai, M. and Perra, N. (2017). Control strategies of contagion processes in time-varying networks. In Masuda, N. and Holme, P., editors, *Temporal Network Epidemiology*, pages 179–197. Springer Singapore, Singapore.

[Kirman and Teschl, 2010] Kirman, A. and Teschl, M. (2010). Selfish or selfless? the role of empathy in economics. *Phil Trans Roy Soc B*, 365(1538):303–317.

[Kovanen et al., 2013] Kovanen, L., Kaski, K., Kertész, J., and Saramäki, J. (2013). Temporal motifs reveal homophily, gender-specific patterns, and group talk in call sequences. *Proceedings of the National Academy of Sciences*, 110(45):18070–18075.

[Kovács et al., 2019] Kovács, I. A., Luck, K., Spirohn, K., Wang, Y., Pollis, C., Schlabach, S., Bianand, W., Kim, D.-K., Kishore, N., Hao, T., Calderwood, M. A., Vidal, M., and Barabási, A.-L. (2019). Network-based prediction of protein interactions. *Nature Communications*, 10(1240):2041–1723.

[Kvålseth, 2017] Kvålseth, T. O. (2017). On normalized mutual information: Measure derivations and properties. *Entropy*, 19(11):631.

[Kwak et al., 2010] Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 591–600, New York, NY, USA. ACM.

[Latour, 1988] Latour, B. (1988). *Politics of explanation, in Knowledge and Reflexivity, ed. by Steve Woolgar*. SAGE.

[LDA-Consulting, 2012] LDA-Consulting (2012). Capital bikeshare 2011 member survey report. *Washington, DC: LDA Consulting*.

[Lemoy et al., 2012] Lemoy, R., Bertin, E., and Jensen, P. (2012). Socio-economic utility and chemical potential. *European Physical Letters*, 93:38002.

[Leo et al., 2016] Leo, Y., Fleury, E., Alvarez-Hamelin, J. I., Sarraute, C., and Karsai, M. (2016). Socioeconomic correlations and stratification in social-communication networks. *Journal of The Royal Society Interface*, 13(125):20160598.

[Leskovec and Horvitz, 2008] Leskovec, J. and Horvitz, E. (2008). Planetary-scale views on a large instant-messaging network. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 915–924, New York, NY, USA. ACM.

[Liao et al., 2012] Liao, S.-H., Chu, P.-H., and Hsiao, P.-Y. (2012). Data mining techniques and applications – a decade review from 2000 to 2011. *Expert Systems with Applications*, 39(12):11303 – 11311.

[Liaw, 2009] Liaw, S. (2009). Maximise global gain in the minority game. *Research Letters in Physics*, 2009.

[Liaw and Liu, 2005] Liaw, S. and Liu, C. (2005). The quasi-periodic time sequence of the population in minority game. *Physica A*, 351(2):571 – 579.

[Lorenz et al., 2018] Lorenz, P., Wolf, F., Braun, J., Djurdjevac Conrad, N., and Hövel, P. (2018). Capturing the dynamics of hashtag-communities. In Cherifi, C., Cherifi, H., Karsai, M., and Musolesi, M., editors, *Complex Networks & Their Applications VI*, pages 401–413, Cham. Springer International Publishing.

[Lund et al., 2017] Lund, K., Jeong, H., Grauwin, S., and Jensen, P. (2017). Une carte scientométrique de la recherche en éducation vue par la base de données internationales scopus. *Les Sciences de l'education-Pour l'Ere nouvelle*, 50(1):67–84.

[MacKay, 2003] MacKay, D. J. C. (2003). Information theory, inference, and learning algorithms. *IEEE Transactions on Information Theory*, 50:2544–2545.

[Masuda and Lambiotte, 2016] Masuda, N. and Lambiotte, R. (2016). *A Guide to Temporal Networks*. WORLD SCIENTIFIC (EUROPE).

[Matias and Miele, 2016] Matias, C. and Miele, V. (2016). Statistical clustering of temporal networks through a dynamic stochastic block model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1119–1141.

[Midgley, 2011] Midgley, P. (2011). Bicycle-sharing schemes: Enhancing sustainable mobility in urban areas. *New York: United Nations*.

[Milgram, 1967] Milgram, S. (1967). The small-world problem. *Psychology Today*, 1:60–67.

[Mobilia et al., 2007] Mobilia, M., Petersen, A., and Redner, S. (2007). On the role of zealotry in the voter model. *J Stat Mech*, 2007(08):P08029.

[Moreno, 1934] Moreno, J. L. (1934). *Who shall survive?: A new approach to the problem of human interrelations*. Washington, DC, US: Nervous and Mental Disease Publishing Co.

[Morini et al., 2017] Morini, M., Flandrin, P., Fleury, E., Venturini, T., and Jensen, P. (2017). Revealing evolutions in dynamical networks. working paper or preprint.

[Mucha et al., 2010] Mucha, P. J., Richardson, T., Macon, K., Porter, M. A., and Onnela, J.-P. (2010). Community structure in time-dependent, multiscale, and multiplex networks. *Science*, 328(5980):876–878.

[Neal, 2014] Neal, Z. (2014). The devil is in the details: Differences in air traffic networks by scale, species, and season. *Social Networks*, 38:63–73.

[Newman, 2003] Newman, M. (2003). The structure and function of complex networks. *SIAM Review*, 45(2):167–256.

[Newman, 2010] Newman, M. (2010). *Networks: An Introduction*. Oxford University Press, Inc., New York, NY, USA.

[Newman, 2006] Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582.

[Ogilvie and Goodman, 2012] Ogilvie, F. and Goodman, A. (2012). Inequalities in usage of a public bicycle sharing scheme: Socio-demographic predictors of uptake and usage of the london (uk) cycle hire scheme. *Preventive Medicine*, 55(1):40 – 45.

[Onnela et al., 2007] Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., de Menezes, M. A., Kaski, K., Barabási, A.-L., and Kertész, J. (2007). Analysis of a large-scale weighted network of one-to-one human communication. *New Journal of Physics*, 9(6):179.

[Ostrom, 2010] Ostrom, E. (2010). Beyond markets and states: Polycentric governance of complex economic systems. *American Economic Review*, 100(3):641–72.

[Pastor-Satorras and Vespignani, 2001] Pastor-Satorras, R. and Vespignani, A. (2001). Epidemic spreading in scale-free networks. *Phys. Rev. Lett.*, 86:3200–3203.

[Peyrard, 2004] Peyrard, M. (2004). Nonlinear dynamics and statistical physics of DNA. *Nonlinearity*, 17(2):R1–R40.

[Porta et al., 2012] Porta, S., Latora, V., Wang, F., Rueda, S., Strano, E., Scellato, S., Cardillo, A., Belli, E., Cardenas, F., Cormenzana, B., et al. (2012). Street centrality and the location of economic activities in barcelona. *Urban Studies*, 49(7):1471–1488.

[Pucher and Buehler, 2012] Pucher, J. and Buehler, R. (2012). *City cycling*. Cambridge, MA: MIT Press.

[Putnam, 2000] Putnam, R. D. (2000). *Bowling Alone: The Collapse and Revival of American Community*. Simon & Schuster, New York.

[Raux et al., 2017] Raux, C., Zoubir, A., and Geyik, M. (2017). Who are bike sharing schemes members and do they travel differently? the case of lyon's "velo'v" scheme. *Transp. Res. Part A*, 106:350–363.

[Rossetti and Cazabet, 2018] Rossetti, G. and Cazabet, R. (2018). Community discovery in dynamic networks: A survey. *ACM Comput. Surv.*, 51(2):35:1–35:37.

[Rossetti et al., 2017] Rossetti, G., Pappalardo, L., Pedreschi, D., and Giannotti, F. (2017). Tiles: an online algorithm for community discovery in dynamic social networks. *Machine Learning*, 106(8):1213–1241.

[Roth et al., 2011] Roth, C., Kang, S. M., Batty, M., and Barthélemy, M. (2011). Structure of urban movements: Polycentric activity and entangled hierarchical flows. *PLOS ONE*, 6(1):1–8.

[Sagiroglu and Sinanc, 2013] Sagiroglu, S. and Sinanc, D. (2013). Big data: A review. In *2013 International Conference on Collaboration Technologies and Systems (CTS)*, pages 42–47.

[Schelling, 1971] Schelling, T. (1971). Dynamic models of segregation. *Journal of Mathematical Sociology*, 1:143–186.

[Schweitzer et al., 2009] Schweitzer, F., Fagiolo, G., Sornette, D., Vega-Redondo, F., Vespignani, A., and White, D. R. (2009). Economic networks: The new challenges. *Science*, 325(5939):422–425.

[Shaheen et al., 2010] Shaheen, S., Guzman, S., and Zhang, H. (2010). Bikesharing in europe, the americas, and asia: Past, present, and future. *Transportation Research Record: Journal of the Transportation Research Board*, 2143:159 – 167.

[Shaheen et al., 2012] Shaheen, S., Martin, E., Cohen, A., and Finson, R. (2012). Public bikesharing in north america: Early operator and user understanding. *San Jose, CA: Mineta Transportation Institute*.

[Shaheen et al., 2011] Shaheen, S., Zhang, H., Martin, E., and Guzman, S. (2011). Hangzhou public bicycle: understanding early adoption and behavioral response to bikesharing in hangzhou, china. *Transportation Research Record*, 2247:34 – 41.

[Sheldon, 1969] Sheldon, A. L. (1969). Equitability indices: Dependence on the species count. *Ecology*, 50(3):466–467.

[Singla et al., 2015] Singla, A., Santoni, M., Bartók, G., Mukerji, P., Meenen, M., and Krause, A. (2015). Incentivizing users for balancing bike sharing systems. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 723–729. AAAI Press.

[Szabo et al., 2013] Szabo, G., Szolnoki, A., and Czako, L. (2013). Coexistence of fraternity and egoism for spatial social dilemmas. *Journal of Theoretical Biology*, 317:126 – 132.

[Talas, 2008] Talas, A. (2008). Connected: The power of six degrees.

[Todorova and Noulas, 2019] Todorova, G. and Noulas, A. (2019). Exploiting population activity dynamics to predict urban epidemiological incidence. *CoRR*, abs/1902.10260.

[Tong, 2011] Tong, D. (2011). Course on statistical physics.

[Tran et al., 2015] Tran, T.-D., Ovtracht, N., and Faivre d'Arcier, B. (2015). Modeling bike sharing system using built environment factors. *Procedia CIRP*, 30(Supplement C):293 – 298. 7th Industrial Product-Service Systems Conference - PSS, industry transformation for sustainability and business.

[Venkatesan, 2018] Venkatesan, R. (2018). *Cluster Analysis for Segmentation*. UVA-M-0748. Darden Case.

[VentureBeat, 2015] VentureBeat (2015). Foursquare by the numbers. https://goo.gl/Vi1UUf.

[Venturini et al., 2015] Venturini, T., Jensen, P., and Latour, B. (2015). Fill in the gap: A new alliance for social and natural sciences. *JASSS*, 18(2):11.

[Vogel et al., 2014] Vogel, M., Hamon, R., Lozenguez, G., Merchez, L., Abry, P., Barnier, J., Borgnat, P., Flandrin, P., Mallon, I., and Robardet, C. (2014). From bicycle sharing system movements to users: a typology of vélo'v cyclists in lyon based on large- scale behavioural dataset. *J. Transp. Geogr.*

[Wagner and Wagner, 2007] Wagner, S. and Wagner, D. (2007). Comparing clusterings - an overview. Technical Report 4, Karlsruhe.

[Wang et al., 2008] Wang, Y., Wu, B., and Pei, X. (2008). Commtracker: A core-based algorithm of tracking community evolution. In Tang, C., Ling, C. X., Zhou, X., Cercone, N. J., and Li, X., editors, *Advanced Data Mining and Applications*, pages 229–240, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Watts, 2011] Watts, D. (2011). *Everything is Obvious*. Crown Business.

[Watts and Strogatz, 1998] Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393:60–67.

[Yang et al., 2011] Yang, T., Haixiao, P., and Qing, S. (2011). Bike-sharing systems in beijing, shanghai, and hangzhou and their impact on travel behavior. *Transportation Research Board 90th Annual Meeting*. Paper presented at the Transportation Research Board Annual Meeting, Washington, DC.

[Zhang et al., 2016] Zhang, J., Pan, X., Li, M., and Philip, S. (2016). Bicycle-sharing system analysis and trip prediction. *Mobile Data Management (MDM), 17th IEEE International Conference*, 1:174–179.

[Zhou, 2010] Zhou, M. (2010). *Chinatown: The socioeconomic potential of an urban enclave*. Temple University Press.

[Zhou et al., 2017] Zhou, X., Hristova, D., Noulas, A., Mascolo, C., and Sklar, M. (2017). Cultural investment and urban socio-economic development: a geosocial network approach. *Royal Society Open Science*, 4(9):170413.

Appendices

A

Résumé long

Au cours des deux dernières décennies les objets connectés ont révolutionné la traçabilité des phénomènes sociaux. Les trajectoires sociales laissent aujourd'hui des traces numériques, qui peuvent être analysées pour obtenir une compréhension plus profonde des comportements collectifs. L'essor de grands réseaux sociaux (comme Facebook, Twitter et plus généralement les réseaux de communication mobile) et d'infrastructures connectées (comme les réseaux de transports publiques et les plate-formes en ligne géolocalisées) ont permis la constitution de grands jeux de données temporelles. Ces nouveaux jeux de données nous donnent l'occasion de développer de nouvelles méthodes pour analyser les dynamiques temporelles *de* et *dans* ces systèmes.

De nos jours, la pluralité des données nécessite d'adapter et combiner une pluralité de méthodes déjà existantes pour élargir la vision globale que l'on a de ces systèmes complexes. Le but de cette thèse est d'explorer les dynamiques des systèmes sociaux au moyen de trois groupes d'outils : les réseaux complexes, la physique statistique et l'apprentissage automatique. Cette thèse commence par donner quelques définitions générales et un contexte historique des méthodes mentionnées ci-dessus.

Après quoi, nous montrons la dynamique complexe d'un modèle de Schelling suite à l'introduction d'une quantité infinitésimale de nouveaux agents, dits altruistes. Nous trouvons que cette quatité infinitésimale induit un effect catalytique sur l'utilité globale du système. Ces altruistes permettent au système d'échapper à l'état sous-optimal normalement atteint lorsqu'il est seulement constitué d'agents égoïstes. À partir de cet exemple, nous concluons le chapitre en discutant des limites des modèles statistiques de société.

Le troisième chapitre montre la valeur ajoutée de l'utilisation de jeux de données temporelles. Nous étudions l'évolution du comportement des utilisateurs d'un réseau de vélos en libre-service. Ces réseaux se sont développés rapidement à l'échelle mondiale et peu de données temporelles individuelles sont disponibles. Ce chapitre donne une première description détaillée de l'évolution temporelle de 120827 utilisateurs annuels répartis sur 5 ans. Nous montrons que l'apparente stabilité générale du système est constituée d'une distribution hétérogène de trajectoires individuelles. Les utilisateurs suivent principalement deux trajectoires : environ 50 à 55% quittent le système après au plus un an. Ces utilisateurs ont une activité médiane basse (env. 40 trajets) ; les autres 45% correspondent aux utilisateurs plus actifs (activité médiane de 91 trajets

la première année) qui restent actifs sur plusieurs années (en moyenne 2.9 ans). Ces utilisateurs réduisent généralement leur activité progressivement (diminution médiane de -16.3%). Nous montrons que les hommes, d'âge moyen, et vivant en centre-ville sont sur-représentés parmi cette dernière classe d'utilisateurs. Enfin, nous analysons les résultats d'un algorithme d'apprentissage automatique non supervisé ayant pour but de classer les utilisateurs en fonction de leurs profils.

Le quatrième chapitre explore les différences entre une méthode globale et une méthode locale de détection de communautés temporelles sur des réseaux scientométriques. La description de réseaux temporels et la détection de communautés dynamiques sont des sujets de recherche en pleine croissance depuis une décennenie. Cependant, il n'y a pas encore de réponse unanime à ces questions dû à la complexité de la tâche. Les communautés statiques ne sont pas des objets bien définis et l'ajout de la composante temporelle ne fait qu'ajouter à la difficulté. Dans ce chapitre, nous proposons de comparer une méthode basique de partionnement global (Global Algorithm (GA)) et une méthode intuitive de partitionnement temporel (Best-Combination Local Algorithm (BCLA)), qui a pour but de trouver un compromis entre *partitionnement optimal* au temps t et *continuité temporel*. Nous testons ces algorithmes sur deux jeux de données bibliographiques. Afin de faciliter la visualisation des dynamiques temporelles nous introduisons la plate-forme *BiblioMaps*. Nous montrons que les deux algorithmes ne présentent que très peu de différences sur le jeu de données aux dynamiques simples, avec peu d'enchevêtrement entre trajectoires. En revanche, pour le jeu de données avec une dynamique plus complexe, la méthode locale permet une description plus fine des trajectoires.

Le dernier chapitre combine l'analyse de réseaux complexes et l'apprentissage automatique supervisé pour décrire et prédire l'impact de l'introduction de nouveaux commerces sur des commerces existants.

Comprendre l'impact d'un nouveau commerce sur son écosystème local est une tâche difficile de par sa nature multi-facteurs. Des études précédentes ont examiné le rôle collaboratif ou compétitif de commerces de même type (i.e. l'impact d'une nouvelle librairie sur les librairies existantes) ce qui limitaient leur horizon. Pour mieux mesurer les performances des commerces dans une ville moderne, il est nécessaire de considérer plusieurs facteurs interagissant de façon synchrone. Ce chapitre étudie les intéractions multi-facteurs se produisant dans des villes pour examiner l'impact de nouveaux commerces. En utilisant un jeu de données longitudinal venant de Foursquare, une réseau social géolocalisé, nous modélisons l'impact de nouveaux commerces pour 26 villes de grandes tailles dans le monde. Nous représentons les villes comme des réseaux de commerces et quantifions leur structure et dynamique temporelle. Nous relevons une forte structure en communautés dans ces réseaux ce qui souligne les relations de coopération et compétition propres aux écosystèmes de commerces locaux. Ensuite, nous mettons en place une métrique capturant les impacts de premier ordre d'un nouveau commerce sur son écosystème locale en tenant compte des intéractions homogènes et hétérogènes entre commerces. Finalement, nous construisons un modèle d'apprentissage automatique supervisé pour prédire l'impact d'un nouveau commerce sur son écosystème commercial local. À l'aide de deux classes de variables explicatives, tenant en compte la présence de différents types de commerces et le réseau de coopérativité de ceux-ci, le modèle atteint une aire sous la courbe jusqu'à 80% pour certaines

catégories de commerces. Notre approche souligne la puissance de l'utilisation de mesures de topologies de réseaux complexes dans le développement de modèle prédictifs d'apprentissage automatique. Cette méthodologie et resultats pourraient assister les autorités urbaines et les propriétaires de commerces dans le développement de modèles pour décrire et prédire des changement dans l'environnement urbain.

B

Dynamics of Scientific Research Communities

We investigated four temporal community detection methods, two global and two local methods. However, as measures from GA and GPA are very close and measures from BMLA and BCLA are also very close, we only presented the GA and BCLA methods in the core of this thesis (see Chapter 4). The two other methods (GPA and BMLA) and their measures are described below.

B.1 Global Projected Algorithm (GPA)

Here, we want to include some dynamics into our global algorithm. So, as for our global method, we first run the Louvain algorithm on the total BC network. It results a set of GA-streams. Then, we define successive BC networks by taking into account articles sharing at least two references with articles within their own publication period. We, now, have a subset of the total BC network (BC network articles minus the articles not sharing two references within their own period). This subset is on average 7.8% smaller than the total BC network. For each time period, we define *local* communities by grouping together the publications that are in the same GA-streams. We now have a set of local projected communities in each period. Finally, we compute historical streams by applying our matching algorithm to the projected communities. This approach allows the emergence of dynamical events.

B.2 Best-Modularity Local Algorithm (BMLA)

The BMLA is an incrementation of the work from [Morini et al., 2017]. On each time period, we run N independent runs (we used $N = 100$) of the Louvain algorithm. Because of the noise inherent to the Louvain algorithm, these partitions may be a bit different, while having close values of modularity. BMLA historical streams are defined by applying the matching algorithm to the partitions with the best modularity in each time period.

BMLA Algorithm

Compute the Bibliographic Coupling Graph ;
Split the dataset into temporal windows Δt ;
for *each temporal window* **do**
    run $N = 100$ Louvain algorithm on the instant network;
    select the instant partition with the highest modularity Q;
end
Match the most similar communities between successive temporal windows ;
Link the paired communities along time;

This algorithm returns temporal streams which we call *BMLA-streams*. These streams maximize the modularity at each time t without considering the global modularity of the whole system.

B.3 Comparing All Algorithms

Table B.1 and Table B.2 show there is very little difference between the local algorithms and between the global algorithms, for all measures on both datasets.

Measures	ENS-Lyon	Wavelets
$\|P_{GA}\|$	57	27
$\|P_{GPA}\|$	54	30
$\|P_{BCLA}\|$	97	36
$\|P_{BMLA}\|$	103	40
$\|P_{REF}\|$	17	36
$H(P_{GA})$	3.63	2.87
$H(P_{GPA})$	3.63	2.94
$H(P_{BCLA})$	4.05	3.04
$H(P_{BMLA})$	4.04	3.17
$H(P_{REF})$	2.37	3.18
$MI(GA, REF)$	1.93	2.03
$MI(GPA, REF)$	1.94	2.09
$MI(BCLA, REF)$	1.93	2.49
$MI(BMLA, REF)$	1.94	2.47
$MI(GA, BCLA)$	3.10	1.90
$NMI_{GA}(GA, REF)$	0.53	0.73
$NMI_{REF}(GA, REF)$	0.82	0.64
$NMI(GA, REF)$	0.66	0.68
$NMI_{GPA}(GPA, REF)$	0.54	0.74
$NMI_{REF}(GPA, REF)$	0.82	0.66
$NMI(GPA, REF)$	0.67	0.70
$NMI_{BCLA}(BCLA, REF)$	0.48	0.84
$NMI_{REF}(BCLA, REF)$	0.81	0.80
$NMI(BCLA, REF)$	0.63	0.82
$NMI_{BMLA}(BMLA, REF)$	0.48	0.78
$NMI_{REF}(BMLA, REF)$	0.82	0.80
$NMI(BMLA, REF)$	0.63	0.79
$NMI_{GA}(GA, BCLA)$	0.86	0.67
$NMI_{BCLA}(GA, BCLA)$	0.77	0.62
$NMI(GA, BCLA)$	0.81	0.64

Table B.1 – Similarly to Table 4.2, $|P_X|$ is the number of streams in partition X. $H(P_X)$ is the entropy of partition X. $MI(P_X, P_Y)$ is the mutual information between the partitions X and Y. NMI_X is the mutual information MI normalized by $H(P_X)$. $NMI(P_X, P_Y)$ is the symmetrical normalized mutual information (normalized by $\sqrt{H(X) * H(Y)}$).

Measures	ENS-Lyon	Wavelets
$\overline{1^{st}E}(GA, REF)$	0.86 ± 0.17	0.75 ± 0.20
	0.49 ± 0.20	0.81 ± 0.17
$\overline{Sum_{80}}(GA, REF)$	1.26 ± 0.54	1.88 ± 0.93
	3.37 ± 1.76	1.5 ± 0.73
$\overline{1^{st}E}(GPA, REF)$	0.87 ± 0.16	0.78 ± 0.19
	0.54 ± 0.23	0.83 ± 0.17
$\overline{Sum_{80}}(GPA, REF)$	1.24 ± 0.5	1.65 ± 0.84
	3.12 ± 1.61	1.47 ± 0.72
$\overline{1^{st}E}(BCLA, REF)$	0.89 ± 0.14	0.87 ± 0.17
	0.49 ± 0.26	0.87 ± 0.15
$\overline{Sum_{80}}(BCLA, REF)$	1.23 ± 0.44	1.26 ± 0.50
	4.87 ± 3.35	1.31 ± 0.57
$\overline{1^{st}E}(BMLA, REF)$	0.89 ± 0.14	0.85 ± 0.19
	0.49 ± 0.25	0.84 ± 0.17
$\overline{Sum_{80}}(BMLA, REF)$	1.23 ± 0.44	1.34 ± 0.63
	5.0 ± 3.60	1.37 ± 0.59
$\overline{1^{st}E}(GA, BCLA)$	0.74 ± 0.23	0.72 ± 0.23
	0.85 ± 0.16	0.83 ± 0.19
$\overline{Sum_{80}}(GA, BCLA)$	1.96 ± 1.14	1.88 ± 0.96
	1.34 ± 0.51	1.61 ± 0.83

Table B.2 – Similarly to Table 4.3, In this table each cell contains two lines. Each measure $M(X, Y)$ is made on edges. The first line correspond to M measured on edges from n_X to n_Y and the second line corresponds to M being measured on edges from n_Y to n_X. So, the first row in $\overline{1^{st}E}(X, Y)$ is the average proportion of articles n_X shares with $n_Y \pm$ its standard deviation. The second row is the average proportion of articles n_Y shares with $n_X \pm$ its standard deviation. For instance, for the ENS-Lyon, this means that streams of P_{GA} share on average 86% of their articles with their most similar stream in P_{REF}, whereas streams from P_{REF} only share on average 49% of their articles with their most similar stream in P_{GA}. $\overline{Sum_{80}}(X, Y)$ is the average number of streams from P_Y it takes to retrieve 80% of the streams' articles from P_X. For example in the case of the Wavelet Dataset, on average 1.88 ± 0.96 streams from P_{BCLA} are needed to retrieve 80% of a stream from P_{GA}.