Modelling the social dynamics of contagion and discovery using dynamical processes on complex networks

by

Iacopo Iacopini

A thesis submitted to the University of London for the degree of
Doctor of Philosophy

School of Mathematical Sciences
Queen Mary, University of London
United Kingdom

January 2021

To Bibi e Iommetta

Abstract

Complex networks have been successfully used to describe the social structure on top of which many real-world social processes take place. In this thesis, I focus on the development of network models that aim at capturing the fundamental mechanisms behind the dynamics of adoption of ideas, behaviours, or items.

I start considering the transmission of a single idea from one individual to another, in an epidemic-like fashion. Recent evidence has shown that mechanisms of complex contagion can effectively capture the fundamental rules of social reinforcement and peer pressure proper of social systems. Along this line, I propose a model of complex recovery in which the social influence mechanism acts on the recovery rule rather than on the infection one, leading to explosive behaviours. Yet, in human communication, interactions can occur in groups. I thus expand the pairwise representation given by graphs using simplicial complexes instead. I develop a model of simplicial contagion, showing how the inclusion of these higher-order interactions can dramatically alter the spreading dynamics.

I then consider an individual and model the dynamics of discovery as paths of sequential adoptions, with the first visit of an idea representing a novelty. Starting from the empirically observed dynamics of correlated novelties, according to which one discovery leads to another, I develop a model of biased random walks in which the exploration of the interlinked space of possible discoveries has the byproduct of influencing also the strengths of their connections. Balancing exploration and exploitation, the model reproduces the basic footprints of real-world innovation processes. Nevertheless, people do not live and work in isolation, and social ties can shape their behaviours. Thus, I consider interacting discovery processes to investigate how social interactions contribute to the collective emergence of new ideas and teamwork, and explorers can exploit opportunities coming from their social contacts.

Acknowledgments

(Almost) any person holding a PhD would tell you that it is a hell of a journey. Anxiety, self-doubts, sleepless periods, stress, depression, and so on. Well, that was not my case. Sure, motivation is a roller coaster, but it has been a fantastic journey. This is only because of the people I encountered along the way.

First, I have to thank my supervisor, the one and only Prof. Vito Latora. From the very first Skype call –I was still in Switzerland at that time– he started bombarding me with ideas, and he never stopped. I think it took me no more than three minutes to realize what an amazing opportunity I had been given. Vito is notoriously an exceptional scientist, but with me he has been way more than that. He gave me trust, support when I needed the most, and –probably even more importantly– he didn't give me support when I (wrongly) thought I needed some. This mechanism, correctly balanced at what I might call the *sweet spot of the good supervisor*, drastically shaped me towards becoming a more independent young researcher. I can't be more proud to have become a Latoriano. I have to equally thank the complex and dynamical systems groups at QMUL, starting from Prof. Christian Beck, who gave me the chance to start in the first place and then let me bridge the two groups. A huge thank you to all the other members of the family with whom I had the pleasure to interact with in the office, in the corridors, or at the SCR bar (damn, I miss the beers after the seminar with shrimp crisps that do not have shrimps inside): Enzo, Ginestra and Lucas (indubitably the dream team of complex systems), Fede, Valerio, Moreno, Jaia, Oliver and Owen (la vecchia guardia), Mayank, Francesco, Sandro, Evangelos Mitsokapas (sorry mate, you have the nicest name... I had to write it all), Ye, Unai, Andrea, and Gabriele. Thanks also to the many "visitors" that immediately felt like family, Angelone (you know what I wanted to write, but I won't), Giulia, Piero, Ana and Valeria, and to the admin team for all the support, Elisa, Katie, Megan and Mike, among others. Needless to say, I had the best PhD buddy one could ask for, il Santorone.

Grazie Giovannona for every "cara","mariia", "còppo", and "Annacamilla". It has been an "emoziòne".

I am also deeply thankful to the big Turing community, Pizzo, Merve, Ayman, Nathan, Gianluca, Francesco, Charlie, Henry, Jessie, and Sanna, that shared the enrichment year with me, to Alberto and Eugenio who organized an amazing 72h workshop, to the other members of the "HIV team", Alba, Elodie, Alberto, Pau, Benja, and Sofia, that actually made it in 72h, to the CSS London crew, the triptych of the Kraken Aiello-Baronca-Perra, Lauretti, er Barucca, plus the honorary Londoners, Ross and the always present Capt. Gio (I lost track of the current title so I just stick to the original one).

Along the way, I had the fortune to find a second place to call home, which is also literally true. This is CASA, a one-of-a-kind place. Thank you Prof. Elsa Arcaute and Prof. Mike Batty for letting me in and introducing me to the family. Elsa, using her own words, has been like a mother to me. She is an extraordinary scientist and a great woman, with tons of ideas to share and a lot of advices to give, always ready to listen and dedicate me more than the very little time she has. Lots of love to all the CASA minions, Roberto, Richard, Clementine, Carlos, Antonia, Juste, Maaaaaaarten, Natalia, Bale, Bea, Dani, Fulvio, Kostas, Mateo, Meli, Nico, and Valerio.

Furthermore, collaborations are the building blocks of any scientific production. In addition to those already mentioned above, during these years I had the chance to engage and collaborate –as any aspiring network scientist should do– with bunch of extraordinary scientists (and friends!) scattered around the world: Prof. Staša Milojević, Prof. Alain Barrat, Prof. Vittorio Loreto, Dr. Benjamin Schäfer, Dr. Alice Patania, Dr. Jean-Gabriel Young, Dr. Maxime Lucas, and Dr. Enrico Ubaldi. Each one of them contributed in an essential way to the different bits composing this thesis and to my personal growth.

In addition, I would like to acknowledge the examiners, Prof. James Gleeson and Prof. Raul Mondragon. It doesn't happen often to have someone that spends so much time listening and commenting on your own research work, so I'm really glad that they agreed to examine my thesis in the first place. Many thanks for the nice discussion and all the constructive suggestions during the defence, the proofreading (which wasn't my intention to leave to you), and especially for making the viva a truly enjoyable moment!

London can be tough, moving around takes forever, and friendships are hard to build and maintain. Nevertheless, I was lucky to share these years with a number of friends that made the whole journey a lot easier. Many thanks to Fabbione, Andreas, Elif, Princess Jimena, the Corpelosi w/o Marcoilbello, Marcoilbello, Nunzia, Agnese, the Panorzi and Gherardo. Also thanks to Ali, Vale, Lilli, Yan, Stanley, Henry, Stevethebuilder, Fred&Rillo, the music of the neighbors' washing machine, the BBQ, the stained brown leather couch, the Hoxton veggie, Giuseppi, the plunger and the mould remover. The #252team has been the driving force of my days in London and will always be my family. A huge thank you to Ali for supporting me in every moment and for making each grey day of the Big Smoke way brighter.

I dedicate this thesis to Iommetta and Bibi for their blind trust and unconditional support, perché le cerque non fa' le melarance.

This work was supported by the EPSRC grant EP/N013492/1, EP/N510129/1, and EP/M023583/1.

Declaration and Publications

I hereby declare that, except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work or is the outcome of work done in collaboration with others. Ideas, results, and figures appearing in this thesis are based on the publications listed below.

[I] **Iacopini, I.**, Milojević, S., & Latora, V. (2018). Network dynamics of innovation processes. *Physical Review Letters*, 120(4), 048301 [1].

[II] **Iacopini, I.**, Petri, G., Barrat, A., & Latora, V. (2019). Simplicial models of social contagion. *Nature Communications*, 10(1), 1-9 [2].

[III] **Iacopini, I.**, Schäfer, B., Arcaute, E., Beck, C., & Latora, V. (2020). Multilayer modeling of adoption dynamics in energy demand management. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 30(1), 013153 [3].

[IV] **Iacopini, I.**, G. Di Bona, E. Ubaldi, V. Loreto & Latora, V. (2020). Interacting discovery processes on complex networks. *Physical Review Letters*, 125(24), 248301 [4].

[V] Battiston, F., Cencetti, G., **Iacopini, I.**, Latora, V., Lucas, M., Patania, A., Young, J.-G., & Petri, G. (2020). Networks beyond pairwise interactions: structure and dynamics. *Physics Reports*, 874, 1-92 [5].

[VI] **Iacopini, I.** & Latora, V. (2021). On the dual nature of adoption processes in complex networks. *Frontiers in Physics*, 9, 109 [6].

Other publications not covered in the thesis:

[V] Torrisi, V. S., Manfredi, S., **Iacopini, I.**, & Latora, V. (2019). Creative connectivity project–A network based approach to understand correlations between interdisciplinary group dynamics and creative performance. In DS 95: *Proceedings of the 21st International Conference on Engineering and Product Design Education* (E&PDE 2019), 530-535, University of Strathclyde, Glasgow. 12th-13th September 2019 [7].

[VI] Bracci, A., Casanova, P., **Iacopini, I.**, Steinegger, B., Teixeira, A.S., Antonioni, A., & Valdano, E. (2019). Evaluating the impact of PrEP on HIV and gonorrhea on a networked population of female sex workers, arXiv:1906.09085 (under submission) [8].

[VII] Vanhoof, M., Godoy-Lorite, A., Murcio, R., **Iacopini, I.**, Zdanowska, N., Raimbault, J., Milton, R., Arcaute, E., & Batty, M. (2019). Using Foursquare data to reveal spatial and temporal patterns in London. *NetMob 2019*, Jul 2019, Oxford, United Kingdom.[9].

[VIII] Batty, M., Murcio, R., **Iacopini, I.**, Vanhoof, M., & Milton, R. (2020). London in Lockdown: Mobility in the Pandemic City, arXiv:2011.07165 [10].

[IX] Martinus, K., Sigler, T., **Iacopini, I.**, & Derudder, B. (2021). The brokerage role of small states and territories in global corporate networks. *Growth and Change*, 52 (1), 12-28 [11].

[X] Sigler, T., Martinus, K., **Iacopini, I.**, & Derudder, B. (2019). The role of tax havens and offshore financial centres in shaping corporate geographies: an industry sector perspective. *Regional Studies*, 1-13 [12].

[XI] Martinus, K., Sigler, T., **Iacopini, I.**, & Derudder, B. (2019). The role of tax havens and offshore financial centers in Asia-Pacific networks: evidence from firm-subsidiary connections. *Asian Business & Management*, 18(5), 389-411 [13].

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ADM	Adoption Dynamics Model
AP	Adjacent Possible
CC	Connected Component
CR	Complex Recovery
DR	Demand Response
ER	Erdős-Rényi
ERRW	Edge-Reinforced Random Walk
HMF	Heterogeneous Mean Field
HOrS	Higher-Order System
LCC	Largest Connected Component
LSCC	Largest Strongly Connected Component
MF	Mean Field
MMCA	Microscopic Markov-Chain Approach
ODE	Ordinary Differential Equation
RFID	Radio-Frequency Identification Devices
RSC	Random Simplicial Complex
RW	Random Walk
SCC	Strongly Connected Component
SCM	Simplicial Contagion Model
SF	Scale-free
SIR	Susceptible-Infected-Recovered

SIS	Susceptible-Infected-Susceptible
SW	Small World
UM	Urn Model
UMT	Urn Model with Triggering
ZKC	Zachary Karate Club

Chapter 1

Introduction

1.1 Representing interactions

In the past decades, *network science* emerged as a cross-disciplinary field focused on taming the complexity of many real-world systems starting from their interactions [14]. Each system composed by interacting units can indeed be represented as a network, where pairs of nodes are connected by links. Simple as it is, this mathematical representation already provides an effective multi-scale description of the structural patterns behind a variety of real-world systems [15–18].

With its origins deeply rooted in graph theory [19, 20], the network representation of a system can be adopted as soon as we can decompose it into a set of interacting units. The connectivity patterns between these units, called vertices, are represented as link connecting them. In graph terms, we have a graph $G(\mathcal{V}, \mathcal{E})$ consisting in a collection of $N = |\mathcal{V}|$ nodes and $K = |\mathcal{E}|$ edges. Over the years, this formalism has allowed to investigate of complex systems in terms of (dynamic) network models and to map their dynamics into classes of dynamical processes on networks [21–23], ultimately improving our understanding of the behaviour of real-world phenomena, such as disease and rumor spreading, synchronization, and social processes. From the first seminal papers [24, 25], the field has dramatically evolved while adapting to the new interdisciplinary challenges brought by the—not unexpected—interactions with the neighbouring communities. Indeed, in these twenty years networked approaches have found their way into the most diverse disciplines [14], that span from social sciences and economics [26–30] to biology, ecology and neuroscience [31–35].

With the support of the *data revolution* and the development of the network theory, the graph-based representation continuously evolved to better capture the different levels of interactions that most systems exhibit. Considering the intrinsic non-static nature of interactions opened up further research on *temporal networks* [36]. Similarly, distinguishing between different types of interactions among the constituents led to the birth of *multilayer networks*, in which different layers encode relationships of different nature [37–39]. All together, these aspects brought brand new research questions and contributed to a better network representation of reality. Nevertheless, one can reasonably ask: are networks themselves enough to provide a complete description of a complex system?

As stressed by Butts in 2009, *"To represent an empirical phenomenon as a network is a theoretical act. It commits one to assumptions about what is interacting and the nature of that interaction"* [40]. Thus, to obtain meaningful insights from a system, choosing an appropriate representation is a crucial step that cannot be overlooked.

Indeed, when it comes to representing real-world interactions into compact and treatable mathematical objects, there are different representations that one can use, each with its own caveats [41]. In this scenario, it is clear that the network representation comes with a fundamental limit, that is, by definition, considering exclusively pairwise interactions [5]. Instead, many empirical systems are genuinely of higher-order, that means that their relationships are collective actions at the level of groups that simply cannot be factorized into pairs. This is the case of competing species in ecosystems [42], but also brain [43, 44], protein interaction [45], and gene regulatory networks [46]. We call these higher-order systems, or HOrSs [5].

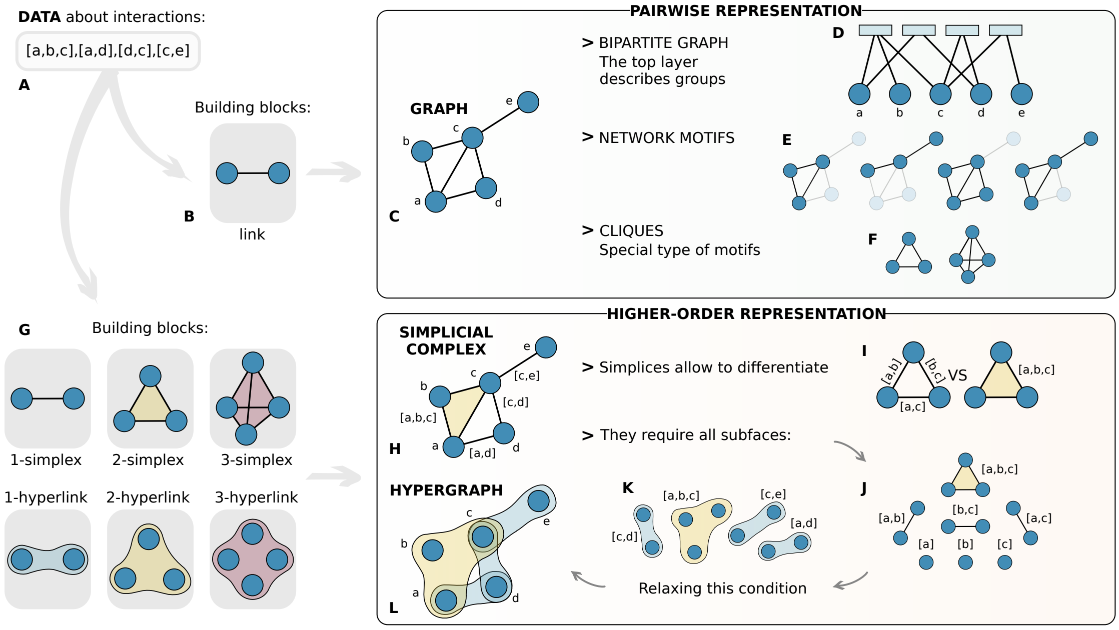


Figure 1.1: Representing interactions. A set of interaction data (**a**) can be represented as a collection of sub-interactions. These building blocks can be (**b**) pairwise interactions (links or edges) or higher order building blocks, such as simplices and hyperedges (**g**). Low order can interactions can be therefore encoded into a graph (**c**), or into a bipartite graph composed by nodes and interactions (**d**). Alternatively, network motifs (**e**) and cliques (**f**) can encode specific structures in the form of subgraphs. Similarly, higher order building blocks such as simplices and hyperedges can be assembled into higher order structures, that are, respectively, simplicial complexes (**h**) and hypergraphs (**l**). Simplices allow to distinguish genuine higher order interactions from the sum of low order ones (**i**). However, simplicial complexes also require the inclusion of all the subfaces of each simplex in the complex (**j**). This condition is relaxed in the hypergraph representation (**k**). Figure from [5].

Human social systems [28], which are at the centre of this thesis, have been also historically characterized exclusively in terms of pairwise interactions between individuals [47, 48]. Still, it is hard to imagine the fabric of social networks just in terms of pairs when many interactions involve more than two people at the same time [49]. Thus, they naturally fall into HOrSs.

How can we represent these systems when graphs are not enough? An alternative mathematical framework comes from algebraic topology [50]. Simplicial complexes (and hypergraphs) can be used to explicitly encode such many-body interactions into a new formalism [51]. Formed by simplices of different dimensions (see Fig. 1.1), these structures

are thus natural candidates to move system descriptions beyond dyadic relationships and effectively map relationships between any number of units [5]. We recall that, in its most basic definition, a k-simplex σ is a set of $k+1$ vertices $\sigma = [p_0, \dots, p_k]$. This, according to the language of graphs, a node is a 0-simplex and a link is a 1-simplex [Fig. 1.1(**b, g**)]. It is then easy to see the difference between a group interaction among three elements (a 2-simplex) and the collection of its edges (three 1-simplices). Just like a collection of edges defines a network [Fig. 1.1(**c**)], a collection of simplices defines a simplicial complex [Fig. 1.1(**h**)]. More formally, a simplicial complex $\mathcal{K}$ on a given set of vertices $\mathcal{V}$, with $|\mathcal{V}| = N$, is a collection of simplices, with the extra requirement that if simplex $\sigma \in \mathcal{K}$, then all the sub-simplices $\nu \subset \sigma$ built from subsets of σ are also contained in $\mathcal{K}$ [Fig. 1.1(**j**)]. Such a requirement, allows to distinguish between simplicial complexes and hypergraphs, with the former being a special type of the latter [Fig. 1.1(**k, l**)].
Of course, simplicial complexes are not a new idea [50], but the interest in them has been recently renewed thanks to the enhanced resolution of real-world datasets and the latest advancements in *topological data analysis* [52–55]. In particular, they recently proved to be useful in describing the architecture of complex networks [56–58] functional [59–61] and structural brain networks [44], protein interactions [62], semantic networks [63], and co-authorship networks in science [64, 65].

The latter case represents the typical (and probably easiest) example. Let us consider a paper wrote by three authors a, b, and c. Standard graph-theoretic descriptions based on links [Fig. 1.1(**b,c**)], that gave rise to a number of studies around collaboration networks in science, would consider the clique containing all the pairwise interactions between authors, namely $[a, b]$, $[b, c]$, and $[a, c]$. It is evident how this representation lacks in capturing the multi-body nature of the interaction, that can be recovered using the simplex $[a, b, c]$ formed by all the authors instead. In this setting, we are able to distinguish this paper from the hypothetical sum of three previous publications authored by two researchers at a time $\{[a, b], [b, c], [a, c]\}$ [see Fig. 1.1(**i**)].
Thus, the question is: should the sub-interactions contained be considered as well?

There is a matter of discussion whether social relationships could be better modelled by using simplicial complexes rather than hypergraphs. In the end, depending on the situation, it might be reasonable or not to assume that in a group interaction all the sub-interactions among the group members should be considered as well [66]. In this particular example, adopting the simplicial complex description would mean assuming the “ nested" nature of co-authoring, where all the sub-interactions happen [Fig. 1.1(**j**)]. This might be reasonable for certain applications involving spreading and diffusion dynamics, such as the one we will consider in Chapter 3. However, for studying the structural properties of the collaboration networks an hypergraph description might be more suitable, since it would allow to distinguish different publications authored by subsets of the authors from a collective one [65]. For a complete overview of the emergent field of *network beyond pairwise interactions* we refer the interested reader to the review in Ref. [5]. Relationships between representations are also specifically discussed in Ref. [41].

1.2 Social dynamics

Dynamical processes that emulate human behaviours have been the focus of many studies, where social relationships and interactions are typically considered as an underlying structure. Indeed, social interactions are a natural testing ground for networked approaches. Since individuals can interact in pairs or groups, the dynamics should in turn account for the effects that the structured interactions might lead to [5].

In this thesis, we focus on network models that aim at capturing the fundamental mechanisms behind the social dynamics of adoption [67, 68]. These are mostly agent-based models that describe social dynamics by relying on simple rules inspired from the physics and mathematics literature, such as spin models and interacting particle systems. We refer the reader interested in the more general world of sociophysics, statistical physics, and computational social science, to the dedicated reviews [47, 48, 69–75].

The adoption of objects and behaviours can be described both as a *contagion dynamics* on a network of individuals influencing one another, and as a *exploration dynamics* on a network of objects and behaviours that individuals can adopt [6]. Indeed, one the one hand, one can consider a single idea, or behaviour, and model its transmission from one individual to another, in an epidemic-like fashion [76, 77]. In this case, the focus is on the propagation dynamics over a social network [78–80]. On the other hand, one can start from the relationships between these ideas and shift the attention to their adoption dynamics as a sequence of items that individuals can sequentially collect [1, 81]. In this latter case, different exploration (and innovation) models have been proposed to replicate the exploration dynamics according to which one idea leads to another and a discovery can trigger further ones [82]. In this introductory section, we will present both points of view, briefly describing the different modelling approaches together with some of the key concepts and models that will be then used throughout the thesis. At the essence of these models there is the common idea to describe a social dynamics by relying on simple —yet sufficient— rules. Many efforts in this directions have been put forward, contributing to the growth of a field of physicists and computational social scientists that is often addressed as *social physics* [47, 48]. Within the broader spectrum that this field covers, ranging from the formation of norms and consensus [75, 83, 84] to the dynamics of echo chambers [85–88], we will focus on two particular dynamics. Specifically, we start introducing *spreading processes*, historically embedded within the literature of epidemics on networks [77], but recently revisited to fit the dynamics of social contagions [89]. We then continue with a wider class of models of discovery and exploration [6].

1.2.1 From epidemic spreading to social contagions

The study of spreading processes on networks is one of the branches of network science that attracted more attention among the community. Building on top of classical epidemiological compartmental models [90–92], the recent success of these models is partially due to the increasing availability of large scale data that opened up new research avenues in which researchers make use of the newly available data sources to inform the models, which on turn allow us to forecast and possibly control the disease spreading [93–97]. In light of these new advancements, network scientists have been slowly, but extensively, introducing more and more details into the modelling framework in order to increase its accuracy and ultimately its predictive power [98].

In this scenario, the typical approach is to divide the individuals of a population into a finite number of classes, or compartments. In the simplest possible case we have only two classes. Individuals or agents can either be in the susceptible (S) or in the infected (I) class, with the latter being the class of those who have an infection and are therefore potentially contagious for the rest of the population. Two of the most studied compartmental models are the Susceptible-Infected-Susceptible (SIS) and the Susceptible-Infected-Recovered (SIR). In both models, susceptible individuals (S) can get infected by means of an interaction with infectious ones (I). This SI process always leads, by construction, to the absorbing state in which all individuals are infected. The introduction of the third transition leads to richer phenomena. More specifically, in the case of the SIS, individuals can switch multiple times between the S and I states, eventually reaching a steady state in which the epidemic is sustained by a non-zero number of individuals [see Fig. 1.2(**a**)]. This SIS dynamics is indeed suitable for modelling the dynamics of those diseases that can infect an individual more than once, such as common cold. All individuals are initially assigned to the S class, with the exception of a small initial seed of infected nodes. Infected individuals can then pass the infection to the susceptible ones by means of contacts, i.e., through the links of the network. More precisely, an infectious node can pass the infection to a neighbour according to a given rate of infection β. In turn, infected individuals spontaneously recover with a rate μ and then can get infected again.

Differently from the SIS, in the SIR, individuals gain immunity to reinfections after a certain amount of time, or with a given probability per unit time. These immune individuals are then called recovered (R) and do not participate any more to the spreading dynamics [see Fig. 1.3(**a**)]. This type of models is therefore used when it comes

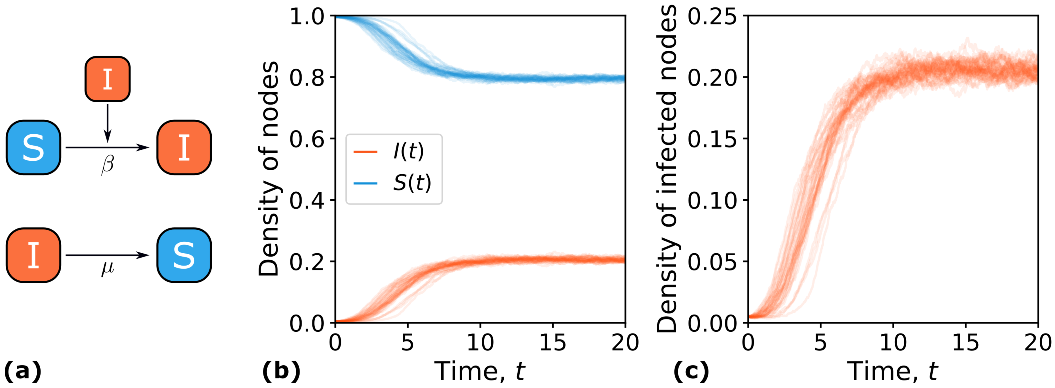


Figure 1.2: SIS model. (**a**) Transition diagram between compartments. (**b**) Evolution of the density of nodes in each compartment (I, S) as a function of time. (**c**) Zoomed temporal dynamics of the density of infected nodes. Simulations on a Barabási-Albert graph [24] ($N = 10^4$, $m = 5$) with parameters, $\beta = 0.1$ (transmission rate), $\mu = 1$ (recovery rate), $I(t = 0) = 0.005$.

to modelling infectious diseases such as Ebola, or seasonal influenza, in which individuals can acquire immunity against reinfections. As a consequence, the SIR presents also the disease-free state as an absorbing state.

Two example-scenarios for the SIS and the SIR models on a Barabási-Albert graph [24] are reported in Fig. 1.2(**b**) and Fig. 1.3(**b**) respectively. In both figures the evolution of the density of nodes in each compartment (S, I, and R) is plotted as a function of time. In Fig. 1.3(**b**), an initial seed of infectious nodes starts infecting the neighbours, causing an inevitable decrease of the susceptible ones. The epidemic reaches the peak (around $t = 5$ in this example) and then dies out, eventually leaving the system divided into susceptible and recovered nodes with acquired immunity.

In the case of the SIS shown in Fig. 1.2(**b**), the infectivity is high enough to sustain the epidemic. Indeed, the red curve, representing the evolution of the density of infected nodes $I(t) \equiv \rho(t)$, reaches a stationary state. This is a signature of the system reaching the endemic state in which, even though nodes continue to switch between S and I compartments, the total number of infected nodes fluctuates around constant value. The average number of nodes in this state gives the size of the epidemics and it is typically referred as ρ^*. The study of $\rho(t)$ as a function of the infectivity, and in particular the

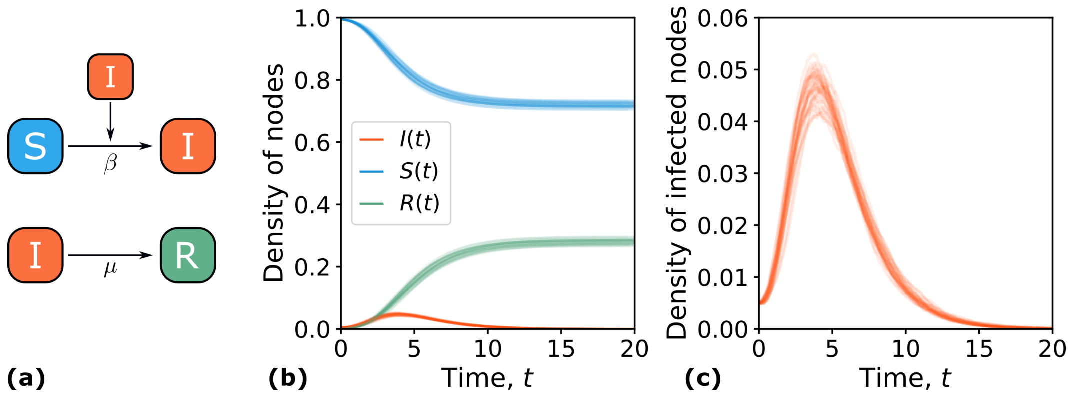


Figure 1.3: SIR model. (**a**) Transition diagram between compartments. (**b**) Evolution of the density of nodes in each compartment (I, S, R) as a function of time. (**c**) Zoomed temporal dynamics of the density of infected nodes. Simulations on a Barabási-Albert graph [24] ($N = 10^5$, $m = 5$) with parameters, $\beta = 0.1$ (transmission rate), $\mu = 1$ (recovery rate), $I(t = 0) = 0.005$.

transition between the epidemic-free and the endemic state, will be the focus of Chapter 2 and Chapter 3, in which two different extensions of the SIS model will be extensively studied.

Many theoretical approaches have been developed to analytically describe, with increasing level of complexity, the dynamics of epidemic spreading on complex networks. An accurate analytical description should include the interplay between the structure of the contact patterns and the dynamics of the spreading process on top. In this thesis, we will specifically use two of these analytical descriptions, that are probably the simplest and—one of—the most accurate ones. These are, respectively, the mean field (MF) approach and the Microscopic Markov-Chain approach (MMCA). The details of these two approaches will be discussed in Chapter 2 and 3, while we refer the reader interested in a complete overview to Ref. [77, 99–102] and references therein.

The contagion dynamics described so far goes under the name of *simple contagion*, to stress the fact that a susceptible node in contact with more than one infected neighbour can get an infection by means of independent exposures. Indeed, when modelling an epidemic spreading in a population [77], the transmission between infectious and healthy individuals is typically assumed: *(i)* to occur through pairwise interactions between

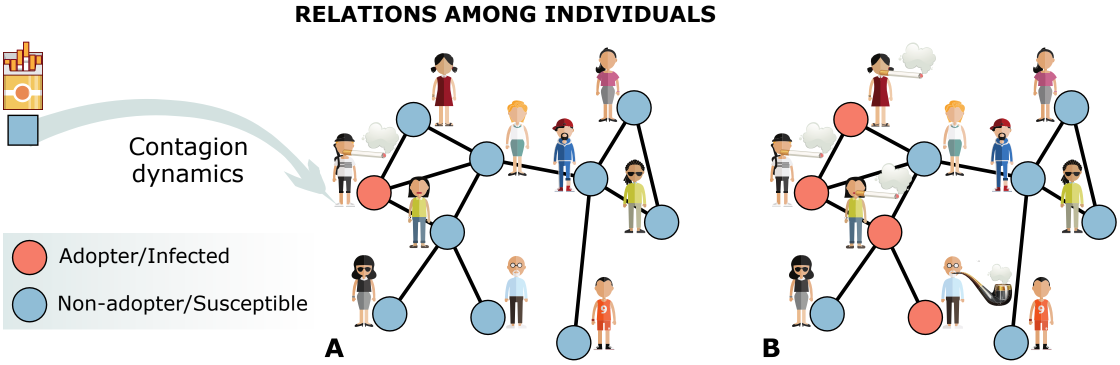


Figure 1.4: Schematic illustration of a contagion dynamics. The adoption of behaviours, ideas etc. is modelled as a spreading process on a network of social contacts. Red and blue denote adopters (or infected) and non-adopters (or susceptible) nodes respectively. For example, in (**a**) a smoker transmits the —bad— habit to his neighbours, that in turn can transmit it again (**b**). Figure adapted from [6].

infectious and healthy individuals, and *(ii)* to be caused even by a single exposure of a healthy individual to an infectious one. This last point is precisely what makes a contagion *simple*, without undermining the complexity of the epidemic model.

While the aforementioned models have been widely used to study the spread of diseases, there's a variety of other domains where they can been successfully applied, and the compartmental approach just described, initially designed to characterise the spreading of viruses, can also cover a broader class of phenomena. In fact, another equally long tradition of modellers have been using similar frameworks to characterize the spreading of social phenomena, such as the diffusion of rumours or fads, or the adoption of norms, behaviours or technological innovations [78, 103–106]. Think for example at the spreading of a behaviour, such as smoking [107], or a behavioural-related physical or mental state, like obesity [108] and happiness [109]. An illustrative example is shown in Figure 1.4, for a single selected habit: smoking. We can easily associate the addition to smoking of an individual to an infected status, as it would be for a disease. Similarly, a non-smoking behaviour denotes the possibility to become an adopter, and thus would correspond to being susceptible. Assuming now that a social relationship between two individuals is the medium in which a behaviour can propagate, we can use a network of social relationship as a structure on top of which the considered behaviour might spread, that means going for example from the configuration in Fig. 1.4(**a**) to Fig. 1.4(**b**).

We have just claimed that, under a very first approximation, the spreading dynamics of pathogens and ideas are similar. To be more precise, we introduced the idea that they can both be modelled using the same epidemic-like framework. This is, obviously, only partially true. Indeed, when dealing instead with phenomena of social contagion and human systems, the situation is more complex, and there is a variety of behavioural aspects influencing the social dynamics that cannot be overlooked [110]. In fact, often the dynamics cannot be simply explained in terms of basic disease epidemics models, which would result too reductive. Instead, the social nature of the contacts mediating these processes deserves special attention, calling for ad-hoc modelling adjustments and tailored experimental techniques to measure social effects [111, 112]. Simple epidemic-like contagions can suffice to describe some cases, such as easily convincing rumours or domino effects [79]. In other situations, however, they do not provide a satisfactory description, especially in those cases where more complex dynamics of peer influence and reinforcement mechanisms are at work [89, 110, 113]. Along this line, recent investigations have empirically shown that *simple contagion* rules (SI⟶2I) are not appropriate to describe the more complex mechanisms of social influence that are at work when humans interact [80, 107, 114, 115]. This evidence, mostly provided by digital traces, relates to different contexts, ranging from for the adoption of applications [114, 116] and technologies [117, 118] to the spreading of obesity [108], happiness [109], and music listenership [119].

These considerations gave rise to new streams of research devoted to adapt these simple models by translating theories coming from the social sciences into mechanistic models. Among these, *complex contagion* is a particularly popular theory [80, 120]. As defined by Centola & Macy [79]:

> *"a contagion is complex if its transmission requires an individual to have contact with two or more sources of activation"*, i.e. if a *"contact with a single active neighbour is*

> *not enough to trigger adoption".*

Complex contagion can hence be broadly defined as a process in which exposure to multiple sources presenting the same stimulus is needed for the contagion to occur. Current efforts in this direction have been summarized by Guilbeault et al. in Ref. [89]. The modelling of social contagion processes has been driven by these considerations in several directions. For example, threshold models assume that, in order to adopt a novel behaviour, an individual needs to be convinced by a fraction of his/her social contacts larger than a given threshold [79, 116, 121–125]. Although supported by a mounting body of empirical studies [80, 115, 116, 126, 127], complex contagion is not the only theory out there, but alternative mechanisms have been theorized. For example, Ugander et al. [126] proposed *structural diversity*, a local measure of the neighbourhood of a node, quantified in terms of number of connected components having at least one adopter. When empirically tested on data of adoption of online platforms upon invitation, this measure turned out to be a better predictor of the probability of adoption with respect to more conventional measures like the number of adopters among the peers. Complex contagion and structural diversity have also been tested against *embeddedness* and *tie strength* theories, in which friendship overlap and intensity are the key drivers of social contagion instead [128, 129]

While most of the works mentioned so far focused on the adoption mechanism from the single-sided transition leading to an adoption, in this thesis we will also put some attention to the opposite process, in which adopters abandon the new product or technology becoming "susceptible" again (I⟶S). Recently, it has been shown that differences between the recovery rates of the nodes, i.e., considering heterogeneous distributions of parameters instead of constant, can also dramatically change the epidemiological dynamics [130–132]. In Chapter 2, we will introduce a model of *complex recovery*, in which the social influence mechanism acts on the recovery rule rather than on the infection one. We will show how this change of perspective might lead to explosive adoption dynamics [3]. This behaviour will be especially pronounced in spatial systems, whose effects on the contagion dynamics have also been the focus of several other studies [106, 133–135].

Yet, in human communication, interactions can occur in groups of three or more agents, and, as discussed above, often cannot be simply factored into a collection of dyadic contacts. Thus, in Chapter 3 we will focus on the dynamics of social contagion but expanding the pairwise representation given by graphs in favour of a non-pairwise one, like simplicial complexes or hypergraphs. This is indeed a recent research direction that finds in social systems a particularly suitable playground [5]. We will introduce a model of *simplicial contagion* that shows how the inclusion of these higher-order group interactions can dramatically alter the spreading dynamics and lead to the emergence of novel phenomena, such as discontinuous transitions and bi-stability [2]. Similar results can be also found when hypergraphs are used to encode social patterns underneath the spreading process instead [136–139].

We now move the focus from the adopting individuals to the adopted items and introduce adoption processes as discovery dynamics.

1.2.2 Discovery processes and innovation

Novelties are part of our daily life. The discovery dynamics at which an individual consumes goods or listens to songs can be described, using the words by Thomas Kuhn [140], in terms of the *essential tension* between exploitation and exploration. This eternal trade-off recurs in a variety of different systems. For example, people move between different locations, mostly switching between already known places, but from time to time visit new ones [141–144]. The individual propensity towards “uncharted seas" enters in each discovery processes and enables classifications, such as the returners versus explorer dichotomy for human mobility [141]. If we think of each visit of a place, listening of a song, or, more in general, collection of an item as the addition of a symbol to a symbolic sequence, the series of actions of an individual (agent) can be represented as a sequence that grows in time, over an alphabet that represents a space

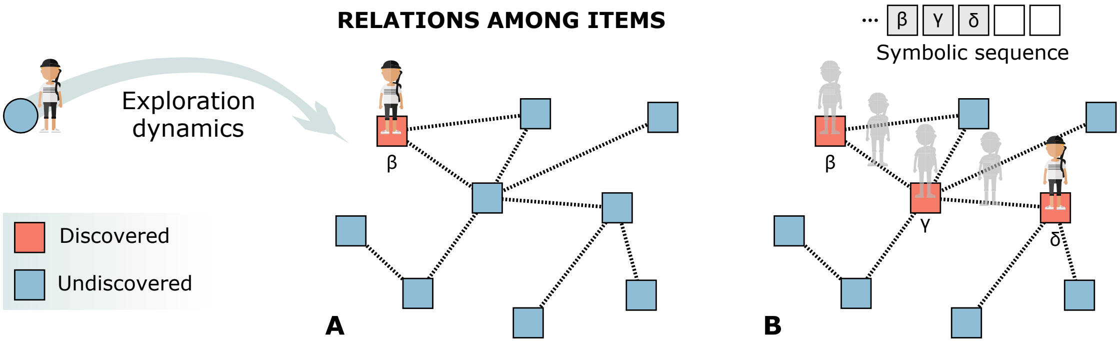


Figure 1.5: The discovery of ideas or the adoption of objects is modelled as an exploration process on a network of relationships (similarities or proximity). For example, in (**a**) an individual collects an object β and then continues the exploration by following the links of the network and sequentially collecting γ and δ. In (**b**) three objects have been discovered, and the exploration path can be seen as a symbolic sequence of discovered objects. Figure from [6]

of possibilities. Symbolic sequences have a long history in text analysis, but recently sequences of item adoptions have been used to study human behaviour, leveraging on sequences of purchases as tracked by credit card data [145] or supermarket fidelity cards [146, 147]. Any process that involves a sequence of actions involving individuals and objects can be framed in this way (see Fig. 1.5). Individuals adopt new items while often returning on their steps. Every time a new item enters the sequence it represents a novelty.

This precise mechanism of exploration and exploitation becomes particularly relevant at the collective level, where novelties can be interpreted as innovations [148]. In fact, the first discovery by any individual of a population represents a novelty for everybody. In this scenario, the essential tension between tradition and innovation has been the focus of many studies that analysed the collective action of researchers determining scientific progress [149–153]. On the same line, patent data have been largely used to explore the dynamics of technological ecosystems [154, 155], with the aim of predicting the dynamics of innovation and eventually detecting the best strategies that could influence the rate of innovation [154, 156, 157].

Researchers have been tackling the problem of the emergence of innovation from different angles [158–161]. For example, some studies have been focusing on the dynamics of substitutive systems, in which the new always replaces the old [162]. In this thesis instead, we keep the focus on the dynamics leading to the emergence of the new; we frame the problem in a cumulative way, such that the new, intentionally very broadly defined, always comes as an addition to the existing. More precisely, the existing environment is actually a necessary condition that paves the way to the emergence of the new. In fact, from Parmenides to modern evolutionary biology, “nothing comes from nothing" is a dictum at the essence of each process involving real-world systems. Thus, even if we neglect that new items might arise from the re-combination of existing ones [159, 163], there is still an essential ingredient that models should take into account, that is the structure underneath these items which determines the way in which individuals can navigate it [164, 165] and move from one item to the next. For example, knowing the bestseller of a book writer is often a condition that puts us in the position of deepening our research towards minor novels of the same author. In this setting, one can think of knowledge as an unexplored space of relationships between concepts and objects to be discovered by—more or less “innovative"—investigations and experiments [166, 167]. These could be interpreted either as an exploration processes of an abstract space of concepts, ideas, items, etc. [168], and as a knowledge acquisition process [169–171], like people acquiring information through online searches [172, 173]. An important aspect is that the structure of this space does matter, since some portions of the space are only visible from certain positions. This concept resonates with the evolutionary theory of the *adjacent possible* (AP) developed by Stuart Kauffman [174]. According to this framework, we can split the knowledge space into what has already been discovered (the actual) and what is left to explore (the possible). Notice, however, that only one tiny fraction of the possible is achievable from the actual, and this is precisely the AP, that is situated one step away from what is already known.

Recent empirical studies have investigated the emergence of novelties and innovation in a wide variety of different contexts, including science and technologies [149, 166, 175–178], knowledge and information [172, 179], goods and products [180], language [181], but also gastronomy [156] and cinema [182]. In parallel to the empirical analyses, various models have been proposed to reproduce the innovation dynamics in different domains, such as linguistics [183, 184], social systems [185], or self-organized criticality (SOC) [186]. Other approaches have modelled the emergence of innovation as an evolutionary process, such as the Schumpeterian economic dynamics proposed by Thurner *et al.* [159, 163] and the evolutionary game among innovators and developers proposed by Armano and Javarone [187]. At the core of these models, there is often a reinforcement mechanism, akin to the *rich-get-richer* paradigm [188], that accounts for self-reinforcing properties. This is an essential ingredient that allows to recover the emergent scaling laws of discovery processes in real-world systems [82, 189], like the well-known Heaps', Zipf's, and Taylor's laws [190–192], for example via sample-space-reducing mechanisms [193]. The Yule process [194] is one of the first mechanisms employed to generate the empirically observed power laws. From there, many processes with reinforcement have been developed [195]. At their root, there are the famous standard urn processes [196], like the ones of Pólya and Hoppe[197, 198]. However, these basic processes have been slowly modified and tuned with empirical data in order to better capture the observed patterns. An example is the generalization of the Yule-Simon process [199] developed to mimic the dynamics of collaborative tagging, where online users associate tags (descriptive keywords) to items, generating fat-tailed frequency distributions of tags.

Later, the urn framework has been further enriched in order to account for the dynamics of correlated novelties. In fact, empirical traces of human activities show that discoveries come in clusters, and the symbolic sequences generated by discovery processes are thus correlated [82, 189]. Models can mimic this behaviour by letting the space grow together with the process, such that novelties increase the number of possible discoveries via triggering mechanisms. A review of these models of expanding spaces can be found in Ref. [200]. Leveraging on the concept of the AP, triggering mechanisms showed good agreement with empirical data in reproducing both the scaling laws associated to the discovery processes and the correlated nature of the sequences produced. This is the case of the Urn Model with Triggering (UMT) [82], that incorporates the AP within the urn process. In what follows we briefly recap some important results for the UMT that will be useful for the rest of the thesis.

1.2.2.1 The Urn Model with Triggering

The UMT process is characterized by two parameters, ρ and ν [82, 192]. At time $t = 0$ the urn contains M_0 balls of distinct colours. At each time step t, a ball is extracted from the urn (representing a concept or an item) and is added to a sequence of items $\mathcal{S}(t)$. The ball is then put back into the urn together with ρ additional copies of it. With this mechanism, called *reinforcement*, the frequency of colours in of $\mathcal{S}(t)$ directly influences the probability of extracting colours, favouring those that have been already drawn. The first time a ball of a new colour is extracted, it represents a novelty. Every time that a novelty appears (is extracted), the space of possible discoveries increases. More precisely, according to this *triggering mechanism*, every time a novelty appears $\nu + 1$ balls of brand new colours are added to the urn. Thus, if we are interested in studying the pace of discovery of this model, we need to come up with an equation for the growth of the number of novelties $D(t)$. This can be done starting from the following considerations. $D(t)$ increases by one every time a ball is extracted for the first time, thus $D(t + 1) = D(t) + P^{\text{new}}(t)$. Here, $P^{\text{new}}(t) \in [0, 1]$ is the probability that the ball extracted at time t never appeared in the sequence $\mathcal{S}(t)$ before. In other words,

$$P^{\text{new}}(t) = \text{Prob}\left[D(t + 1) = D(t) + 1|D(t)\right]$$

and we can express it as the fraction of discoverable balls over the total number of balls available at time t. This leads to an equation for the asymptotic growth of the number of novelties that in the continuous time limit reads:

$$\frac{dD(t)}{dt} = \frac{U'(t) - D(t)}{U(t)}, \tag{1.1}$$

where the numerator is the number of unique colours available at time t minus the number of unique colours already extracted at that time, and the denominator is the total number of balls present in the urn.

Equation (1.1) can be written in terms of the parameters of the model. In particular, we can write the total number of balls in the urn up to time t, $U(t)$, as the initial number of balls M_0, plus the ρ balls added (t times) as reinforcement, plus the $(\nu+1)$ balls added ($D(t)$ times, one for each novelty) due to the triggering mechanism:

$$U(t) = M_0 + \rho t + (\nu+1)D(t). \tag{1.2}$$

Similarly, the number of unique elements in the urn at time t, $U'(t)$, can be obtained by subtracting from $U(t)$ the ρt repeated balls coming from the reinforcement, that is:

$$U'(t) - D(t) = [U(t) - \rho t] - D(t) = M_0 + \nu D(t). \tag{1.3}$$

Thus, using Eq. (1.2) and Eq. (1.3) in Eq. (1.1), we obtain:

$$\frac{dD(t)}{dt} = \frac{M_0 + \nu D(t)}{M_0 + \rho t + (\nu+1)D(t)}. \tag{1.4}$$

For great times, $t \gg M_0$, we can disregard M_0, and, after the introduction of the auxiliary variable $z(t) = \frac{D(t)}{t}$, Eq. (1.4) can be rewritten as:

$$\frac{dz(t)}{dt}t + z(t) = \frac{\nu z(t)t}{\rho t + (\nu+1)z(t)t}, \tag{1.5}$$

which can be integrated as:

$$\int_{z(t_0)}^{z(t)} \frac{\rho + (\nu+1)z(t)}{z(t)[\nu - (\nu+1)z(t) - \rho]} dz(t) = \int_{t_0}^{t} \frac{1}{t} dt. \tag{1.6}$$

The asymptotic solution ($t \to \infty$) depends on the parameters ρ and ν, and it can be shown, as detailed in the Supplemental Material of Ref. [82, 192], that is given by:

$$\begin{cases} \rho > \nu & D(t) \sim (\rho - \nu)^{\frac{\nu}{\rho}} t^{\frac{\nu}{\rho}} \\ \rho = \nu & D(t) \sim \frac{\nu}{\nu+1} \frac{t}{\ln t} \\ \rho < \nu & D(t) \sim \frac{\nu-\rho}{\nu+1} t \end{cases} \tag{1.7}$$

that is precisely the Heaps' law [190], with a sublinear growth for $\rho > \nu$ and linear for the other cases. As we will see in Chapter 4, many empirical discovery processes generate Heaps' laws [82, 190] that have sublinear behaviours. For this reason, in this thesis we will focus on the case $\rho > \nu$. In particular, in Chapter 4 we will propose a network alternative to the UMT model, the edge-reinforced random walk model [1], that encodes the AP directly into the topology of a network of concepts and ideas. Then, in Chapter 5 we will couple together different discovery processes through the links of a social network, exploring in this ways the effects of the structure of a team in processes of collective exploration [4].

1.3 Outline of the thesis

The first part of the Thesis is devoted to processes of social contagion.

In **Chapter 2**, we will devise a model of social contagion inspired by a real-world scenario of energy demand management [3]. In fact, due to the emerging of new technologies, the decentralisation of energy resources and the smart grid have forced utility services to rethink their relationships with customers. Demand response (DR) seeks to adjust the demand for power instead of adjusting the supply. However, DR business models rely on customer participation and can only be effective when large numbers of customers in close geographic vicinity, e.g. connected to the same transformer, opt in. We will introduce a model for the dynamics of service adoption on a two-layer multiplex network: the layer of social interactions among customers and the power-grid layer connecting the households. While the adoption process—based on peer-to-peer communication—runs on the social layer, the time-dependent recovery rate of the nodes depends on the states of their neighbours on the power-grid layer, making an infected node surrounded by infectious ones less keen to recover. We will perform numerical simulations of the model on synthetic and real-world networks, showing that a strong local influence of the customers' actions leads to a discontinuous transition where either none or all the nodes in the network are infected, depending on the infection rate and the social pressure to adopt. We will find that clusters of early adopters act as points of high local pressure, helping maintaining adopters, and facilitating the eventual adoption of all nodes.

In **Chapter 3**, keeping the focus on the dynamics of social contagion, we will shift the attention towards the structural aspect of the social contacts on top of which the dynamics evolves. Complex networks have been successfully used to describe the spread of diseases in populations of interacting individuals. However, pairwise interactions are often not enough to characterize social contagion processes such as opinion formation or the adoption of novelties, where complex mechanisms of influence and reinforcement are at work. We introduce a higher-order model of social contagion in which a social system is represented by a simplicial complex and contagion can occur through interactions in groups of different sizes [2, 201]. Numerical simulations of the model on both empirical and synthetic simplicial complexes will highlight the emergence of novel phenomena such as a discontinuous transition induced by higher-order interactions. We will show analytically (MF and MMCA) that the transition is discontinuous and that a bistable region appears where healthy and endemic states co-exist.

We will then shift the attention from processes of contagion to processes of discovery.

In **Chapter 4**, we will introduce a model for the emergence of the new in which cognitive processes are described as random walks on the network of links among ideas or concepts, and a discovery corresponds to the first visit of a node [1]. The transition matrix of the random walk will depends on the network weights, while in turn the weight of an edge will be reinforced by the passage of a walker. We will show how the presence of the network naturally accounts for the mechanism of the AP, and the model reproduces both the rate at which novelties emerge and the correlations among them observed empirically [82]. We will show this by using synthetic networks and by studying real data sets on the growth of knowledge in different scientific disciplines.

In **Chapter 5**, we will focus on the mechanisms of collective exploration and propose a model in which many urns, representing different explorers, are coupled through the links of a social network and exploit opportunities coming from their contacts [4]. We will study different network structures showing, both analytically and numerically, that the pace of discovery of an explorer depends on its centrality in the social network. Overall, this model of coupled urns will shed light on the role that social structures play in discovery processes.

The last Chapter summarises the main results of the thesis. Further research ideas will be also exposed based on the joint results from the two parts of the thesis.

A small **Appendix** ends the manuscript. It contains all the supplementary material and detailed analytical calculation that have been left out from the main text for readability purposes, or because redundant.

Part I

Modelling Social Contagions

Chapter 2

Multi-layered social contagions in space

2.1 Introduction

The study of dynamical processes on complex networks is a well established branch of complex systems science that aims at understanding the complex interplay between the dynamics of the process and the topology of the underlying network [22, 23]. As discussed in Sec. 1.1, networks encompass a powerful approach, in which a system can be represented by considering its connectivity patterns, encoding in this way all the interactions between the different units composing it into a compact framework [21, 202]. Systems composed by units that interact in different ways can be analogously represented by considering their multi-layered interactions [38, 39, 203–205], that means using a different network layer for each type of interaction while keeping the nodes fixed. The highly versatile essence of the network representation allows one to use it as a structure for processes of very different nature, that can ultimately be used to model real world phenomena.

2.1.1 The role of social effects in contagion dynamics

Among the most studied processes on networks, together with synchronization [206] and random walks [207], are the dynamics of spreading phenomena in a population, such as the spreading of diseases [77], norms, innovation adoption [1, 208] or knowledge diffusion [209]. Adoption dynamics becomes increasingly relevant when implementing new business models e.g. for the Internet of Things or smart grids [210].

This contagion dynamics goes typically under the name of *simple contagion*, to stress the fact that a susceptible node in contact with more than one infected neighbour can get an infection by means of independent exposures. The modelling approach just described is not only restricted to the spreading of viruses, it can also cover a broader class of phenomena, such as smart-grid technologies. However, as extensively discussed in Sec. 1.2.1, when dealing with phenomena that involve social contagion, it turns out that sometimes the simple contagion framework is not the most appropriate way to model the system under study. This is because the standard SIS model does not capture the basic dynamics of social influence and reinforcement, nor the non-linear nature of technological learning/adoption processes [110, 113]. Therefore, *complex contagion* [79, 89] has been proposed as an alternative description in which, for example, threshold mechanisms are introduced in order to account for the effects of peer pressure and social reinforcement mechanisms [125]. The fundamental difference between simple and complex contagions relies on the fact that in the latter setting multiple exposures from different sources are required for a transmission event to happen. In Chapter 3 this idea will be further extended by introducing a *simplicial contagion* model, in which a simplicial complex instead of a graph is used as the underlying structure of a social systems to encode higher-order (non-pairwise) interactions among individuals [2].

Another way of including social effects into the contagion process consists either in allowing the dynamics of infection to depend on some local properties of the node and their neighbourhood, or alternatively in letting nodes control for their connections [211–216]. Ultimately, the introduction of local effects into the contagion dynamics allows to effectively introduce mechanisms of awareness [217–221], trust [222], and risk perception [223].

All the models mentioned above focus on one of the two aspects of the dynamics of a spreading process, that is the contagion mechanism. This is generally controlled by means of the infection parameter β, which might eventually be node-dependent if local effects are considered. In the case of simple contagion the parameter β mediates two-body interactions, with a corresponding process $S + I \rightarrow 2I$, while in the case of complex and simplicial contagion one-to-many-body and group interactions are considered respectively. Conversely, less attention has been devoted to the other aspect, that is the recovery mechanism. The recovery rate parameter μ is typically considered constant for all the nodes, and it is usually absorbed into an effective infection rate β/μ. Nevertheless, recent results have shown that heterogeneity in recovery rates can have dramatic effects on the type and position of epidemic transitions, implying that heterogeneous infectious periods are as important as structural heterogeneity in the network when processes of disease spreading are considered [130, 131]. However, even when node-dependent recovery rates are endowed, the recovery remains a single-body type of process ($I \rightarrow S$).

In this Chapter, we investigate the effects of dynamical recovery rates in a model of adoption dynamics on a multiplex network[124]. The key feature of our model is the presence of a time- and node-dependent recovery mechanism that is not a spontaneous process, but depends on the states of the neighbouring nodes. Following the analogy with the processes of complex contagion, we name *complex recovery* the one-to-many-body recovery process of our model. Moreover, in the model, spreading and recovery processes are implemented on the different layers of a multiplex network [98, 224]. Within our problem of interest, that is the adoption of a new service within the smart power grid, a model considering the local effects of neighbours seems more relevant than a simple SIS model.

2.1.2 Power grid and energy demand management

In order to have a clearer perspective on smart power grids, and thus better understand the motivations behind our model, let us now spend a few words on the rapid changes that the electrical supply system is currently undergoing. To reach the ambitious climate goals set out in Paris [225], distributed generators are being installed and centrally controlled infrastructures are being replaced by decentralized ones, so that the generation of energy can be de-carbonised. New business models have then emerged to facilitate new modes of operation of the electricity supply, for example via concepts such as *smart grids* [226–228]. Within a smart grid, the different actors (agents or components) of the electrical system, ranging from fossil fuel plants and solar panels to industrial and household consumers, need to communicate and coordinate in order to allow a smooth and stable operation of the grid. One important instrument of a smart grid is the *demand response* (DR) offered from the consumer side. Instead of consuming electricity whenever the consumer wishes, they might enter a contract, guaranteeing that a certain share of consumption will be shifted to periods of low demand. Certain consumption is easily shifted, e.g. water can be heated and stored in hot water tanks for usage throughout the day, or electrical cars can be charged flexibly, given they are sufficiently charged for the next journey. DR can be offered in a static scheme with fixed low-demand periods, such as during the night, or it may be implemented as a dynamical scheme which constantly updates the consumption based on the actual available supply and demand by other customers.

Consumers are typically motivated to follow the DR scheme by price incentives. Previous studies based on game-theory and optimization approaches have shown that time-varying prices might be able to align the optimal schedule of individual power consumption with the global optimum of the system [229]. If prices are also based on consumption level, these mechanisms can be efficiently used by single companies via scheduling games in order to minimize energy costs [230]. Other studies have investigated the effects of increasing participation in DR schemes on the different market participants [231].

An important point is that, in order for DR schemes to be effective, a sufficiently large share of households is required. First, any business addressing households will not be profitable if only a very few participate. Even more importantly, large groups of consumers could act similarly to a virtual power plant [232] by providing power via demand control as a service to the grid. This is specifically profitable if many customers in a given region are part of the contract and can provide power within one distribution grid branch. Previous studies have found that consumers require positive feedback to stay within demand control contracts [233]. Hence, agents, i.e. customers opting into demand control contracts should be rewarded, e.g. by being paid a share of their contribution towards stabilizing the grid and reducing operational costs. Since large clusters of local consumers can act easily as a virtual power plant, we assume that rewards for agents geographically surrounded by other agents opted into the contract could be higher.

Here, we study the dynamics of signing contracts under DR schemes by modelling the system as a multiplex network, where the social layer of the customers and the layer of physical connections among households (as given by the power grid at the distribution level) are considered at the same time and coupled together. The adoption dynamics driving the contract signature is based on social influence mediated one-to-one social interactions. Therefore, we make use of epidemic spreading on the social network, where the contagion process is the standard simple contagion (modelling the word of mouth). Contrarily, the recovery probability depends on the local dynamics on the power-grid layer where economic incentives are implicitly included. The basic idea we want to model here, is that a power supplier will benefit from having a cluster of individuals who signed the contract within a localized geographical area, and in turn it will provide a better offer to the customers. This additional benefit, combined with the social effect of being surrounded by agents of the same type, will make the customers who signed less keen to opt out.

2.1.3 Outline

This Chapter is structured as follows:

In Sec. 2.2, we introduce the *Adoption Dynamics Model* (ADM), explaining in particular the presence of a *complex recovery* (CR) mechanism in the model, and its motivation.

In Sec. 2.3, we present analytic results of the ADM in a mean field approximation.

In Sec. 2.4, we extend the analytical formalism by introducing the more precise microscopic Markov-chain approach.

In Sec. 2.5, we discuss the results of numerical simulations of the model on two synthetic network structures, namely a duplex formed by two Erdős-Rényi random graphs, as well as another duplex consisting of a small-world network and a regular 2D lattice.

In Sec. 2.6, we focus on the application to the smart grid by using the street network as a proxy for the power grid network at the distribution level. Although not entirely representative of the real distribution of electricity, such a network encodes the geographical proximity of the households, thus it provides a more realistic representation.

Finally, in Sec. 2.6.3 and 2.6.4 we investigate the effects of the initial conditions on the temporal dynamics of the model.

Conclusions and future perspectives are summarized in Sec. 2.7.

2.2 The adoption dynamics model

Our model of adoption dynamics is formulated in terms of a multilayer network framework [38, 39, 203–205]. In particular, we consider a multiplex network $\vec{G} = \{(\mathcal{V}, \mathcal{E}_\alpha)\}_{\alpha=1,2}$ formed by two layers (a duplex network), composed by $N = |\mathcal{V}|$ nodes and $K_\alpha = |\mathcal{E}_\alpha|$ links. Every node $i = 1, ..., N$ represents a household, and it has an identical replica (i, α) at each layer α. Contrarily, the nodes interact in different ways, according to the specific layer, and have different structural patterns. In particular, the two layers represent the following two different types of interactions:

Social layer.—The top layer (layer $\alpha = 1$) represents the social network among the individuals which are living in the households -or street areas- that we consider to be

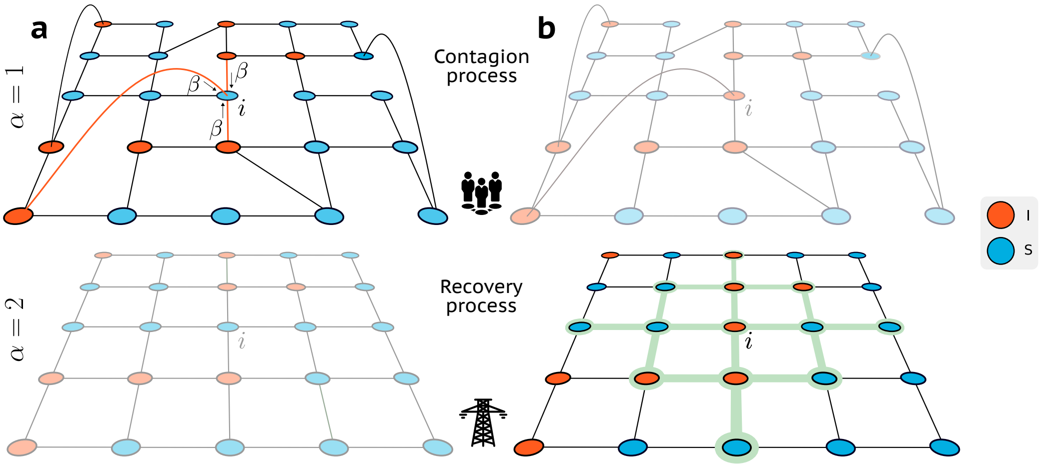


Figure 2.1: Illustration of the Adoption Dynamics Model (ADM). The two layers of the multiplex network stand for the social ties (layer 1, top) and connections in the power grid (layer 2, bottom). Susceptible and infected nodes are coloured in blue and red respectively. (**a**) The spreading dynamics takes place on the layer 1 according to a standard mechanism of simple contagion, where a susceptible node i can get infected by each one of its infected neighbours with an independent probability β. (**b**) Contrarily, an infected node recovers with a node-dependent and dynamically changing recovery probability, which depends on the states of the neighbours at layer 2. The shaded regions highlighted in green indicate the subset of nodes at distance h hops from i (the case $h = 2$ is shown here), which are considered for the computation of the dynamical recovery rate $\mu_i(t)$ [see Eq. (2.3)]. Figure from [3].

nodes. The topology of this layer is described by the binary adjacency matrix $A^{[1]} \equiv \{a_{ij}^{[1]}\}$, whose non-zero entries represent existing social links. We denote as $k_i^{[1]} = \sum_j a_{ij}^{[1]}$ the degree of the node $i \in \mathcal{V}$ at layer 1, so that $\langle k^{[1]} \rangle$ gives the average degree of this layer.

Power-grid layer.—The bottom layer (layer $\alpha = 2$) represents the physical connections among households as given by the power grid at the distribution level. While the nodes are the same as the nodes of layer 1, the connections are described by another binary adjacency matrix $A^{[2]} \equiv \{a_{ij}^{[2]}\}$. Analogously to the case of the first layer, we denote as $k_i^{[2]} = \sum_j a_{ij}^{[2]}$ and $\langle k^{[2]} \rangle$ the degree of node i and the average degree at layer 2.

Notice that as many other infrastructural networks, the power-grid layer can be represented as a spatial network [234], where the nodes (a household or a street in this case) and the links are embedded in a geographical space. The connectivity of the social layer is usually more complex. In fact, it is reasonable to assume that social ties are present at two different levels: the first one is related to physical proximity, which brings neighbours to interact more, while the second level includes long-range social links connecting people living in different areas of the city or in other cities.

Our purpose is to build a model of the dynamics of signing contracts under DR schemes, which we will name the *Adoption Dynamics Model* (ADM). Hence, at each time t, each node i of the network is characterized by a binary state variable $x_i(t) \in \{0, 1\}$. Such a variable represents the state of household i with respect to the contract at time t, with 1 indicating the user has signed a contract, and 0 indicating the user has not signed a contract yet, or has opted out. Nodes change their states according to a Susceptible-Infected-Susceptible (SIS) dynamics that takes place over the links of the first layer (see Section 1.2.1). We assume that the states 0 and 1 correspond respectively to the susceptible (S) and infected (I) states of the SIS.

In this way, each node represents a group of individuals living in a household, and each edge of the social layer stands for a social connection along which the infection can spread, i.e. a susceptible node can opt in being convinced by one of its social links. Each susceptible node has as many channels of infection per unit time as the number of infected neighbours at the social layer 1. The transition $S + I \xrightarrow{\beta} 2I$ is determined by the transmission rate β, which enters directly in the pairwise interactions between susceptible and infected nodes. In our model, the parameter β can be seen as a measure of the social or advertising pressure that convinces customers to opt into a contract. In this way, the probability $p_i(t)$ of a node i to get infected at time t reads as

$$p_i(t) = 1 - \prod_j [1 - \beta a_{ij}^{[1]} x_j(t)], \tag{2.1}$$

where the product on the right hand side gets contributions from all the infected neighbours of node i at the social layer 1, and is equal to the probability that node i is not infected by any of its infected neighbours.

The transition $I \xrightarrow{\gamma_i(t)} S$ is controlled by the parameter $\gamma_i(t)$, which represents the probability that node i recovers at time t, becoming susceptible again. Instead of the spontaneous recovery, the 1-body process typically adopted in the modelling of infectious diseases, here we consider a *complex recovery* (CR) mechanism, which is a many-body process. Namely, instead of using a constant recovery probability μ_0 equal for all nodes, here we introduce a time-dependent recovery probability $\gamma_i(t)$ which can also vary from node to node. In particular, we assume that $\gamma_i(t)$ is a function of the properties of the neighbourhood of node i at time t at the power-grid layer 2. In this way we want to model that individuals are less likely to opt out of a contract with a specific energy supplier if their neighbours, in the power grid, have signed a contract with the same company. This can be seen as an effect of a particular bonus that an energy supplier is able to offer to an individual which is part of a cluster of customers. We thus implement the CR by defining $\gamma_i(t)$ as:

$$\gamma_i(t) = (1-\theta)\mu_0 + \theta\mu_i(t), \tag{2.2}$$

where the parameter $\theta \in [0,1]$ controls for the importance of local interactions in the recovery transition with respect to a standard constant recovery parameter μ_0. Notice that, for $\theta = 0$ no local effects are considered for the recovery, and the model corresponds to the standard SIS model with a constant recovery probability μ_0. Contrarily, if $\theta = 1$, the recovery is completely determined by the dynamical term $\mu_i(t)$, which is node-dependent and that co-evolves in time together with the spreading process at layer 1.

We consider now the case in which $\mu_i(t) = \mu_{i,h}(t)$ is a function of the network hop-distance h at layer 2. Namely, we define $\mu_{i,h}(t)$ as

$$\mu_{i,h}(t) = \left(1 - \frac{|\mathcal{I}^{[2]}_{i,h}(t)|}{|\mathcal{N}^{[2]}_{i,h}|}\right)\mu_0, \tag{2.3}$$

where $\mathcal{N}^{[2]}_{i,h} \subseteq \mathcal{V}$ is the set of nodes of $\vec{G}$ which are within h hops from i on layer 2, and $\mathcal{I}^{[2]}_{i,h}(t) = \mathcal{N}^{[2]}_{i,h} \cap \{j \in \mathcal{V} : x_j(t) = 1\}$ is the subset of these nodes which are infected at time t (Fig. 2.1(**a**) bottom panel). Notice that the highest possible recovery probability in the expression above is equal to μ_0, the same as the static case, but when node i is completely surrounded by infectious neighbours $\mu_{i,h}(t)$ goes to zero. This becomes clear when inserting Eq. (2.3) into Eq. (2.2), leading to

$$\gamma_i(t) = \mu_0\left(1 - \theta\frac{|\mathcal{I}^{[2]}_{i,h}(t)|}{|\mathcal{N}^{[2]}_{i,h}|}\right). \tag{2.4}$$

In the simplest case, in which $h = 1$, we can write Eq. (2.4) directly in terms of the elements of the adjacency matrix $A^{[2]}$ as

$$\gamma_i(t) = \mu_0\left(1 - \theta\frac{\sum_j a^{[2]}_{ij} x_j(t)}{\sum_j a^{[2]}_{ij}}\right). \tag{2.5}$$

Finally, we denote the density of infected (adopters) individuals at time t as $\rho(t) = I(t)/N = \sum_{i=1}^{N} x_i(t)/N$, which represents our macroscopic order parameter. At time $t = 0$ all individuals are susceptible, with the exception of a seed $\rho_0 = \rho(t = 0) \ll 1$ of infected nodes (early adopters).

2.3 Mean field approach

The density of infected individuals and the infection threshold as function of the different control parameters of the ADM can be obtained analytically in a mean-field (MF) approximation. The MF approximation works well under the homogeneous mixing hypothesis, assuming therefore that the individuals with whom a susceptible individual has contact are chosen at random from the whole population. Furthermore, we also assume that all individuals have approximately the same number of contacts at each time, and that all contacts transmit the disease with the same probability. As a consequence, instead of considering the specific topology of the two layers, we only focus on average degree properties, so that we can write an equation for the temporal evolution of the density of infected individuals $\rho(t)$ as

$$d_t \rho(t) = -\langle \gamma_i(t) \rangle \rho(t) + \beta \langle k^{[1]} \rangle \rho(t) \left[1 - \rho(t) \right] . \tag{2.6}$$

With this approach we are assuming that each node of the social network has the same degree, equal to the average degree $\langle k^{[1]} \rangle$ of the social network at layer 1. $\langle \gamma_i(t) \rangle$ denotes the average recovery probability computed over all nodes.

For the particular case in which only the first neighbours are considered ($h = 1$) we can derive a MF expression for $\langle \gamma_i(t) \rangle$, by approximating Eq. (2.3) as

$$\langle \mu_{i,h=1}(t) \rangle \approx \left(1 - \frac{\langle k^{[2]} \rangle \rho(t)}{\langle k^{[2]} \rangle} \right) \mu_0 = \left(1 - \rho(t) \right) \mu_0. \tag{2.7}$$

Notice that if a local tree-like structure is assumed for $A^{[2]}$, the same MF approximation would hold for any h,

$$\langle \mu_{i,h}(t) \rangle \approx \left(1 - \frac{\langle k^{[2]} \rangle^h \rho(t)}{\langle k^{[2]} \rangle^h} \right) \mu_0 = \left(1 - \rho(t) \right) \mu_0. \tag{2.8}$$

Using these results for Eq. (2.2) and by substituting $\langle \gamma_{i,h}(t) \rangle$ into Eq. (2.6) we can write the final MF expression for the temporal evolution of the density $\rho(t)$ of infected nodes, which reads as

$$d_t \rho(t) = -\mu_0 [1 - \theta \rho(t)] \rho(t) + \beta \langle k^{[1]} \rangle \rho(t) [1 - \rho(t)]. \tag{2.9}$$

After defining $\lambda = \beta \langle k^{[1]} \rangle / \mu_0$ and rescaling the time as $t' = \mu_0 (\lambda - \theta) t$, we obtain the equivalent equation:

$$d_{t'} \rho(t') = \rho(t') (\rho_2^* - \rho(t')), \tag{2.10}$$

with

$$\rho_2^* = \frac{\lambda - 1}{\lambda - \theta}. \tag{2.11}$$

The associated steady state equation $d_{t'}\rho(t') = 0$ has therefore up to two acceptable solutions in the range $\rho \in [0, 1]$: a trivial solution $\rho_1^* = 0$, corresponding to the absorbing state in which there is no epidemic (no adopters) and all nodes have recovered; and a non-trivial solution ρ_2^* which depends on the parameters of the model as follows:

2.3.1 Case $\theta = 0$

Let us first consider the case $\theta = 0$ in which local dynamical effects are neglected. This case corresponds, as expected, to the standard SIS model, thus we recover the solution $\rho_2^{*[\theta=0]}$ that reads as

$$\rho_2^{*[\theta=0]} = 1 - \frac{1}{\lambda} = 1 - \frac{\mu_0}{\beta\langle k^{[1]}\rangle}. \tag{2.12}$$

The solution $\rho_2^{*[\theta=0]}$ is acceptable, i.e. non-negative, when $\lambda \geq 1$, recovering in this way the standard epidemic threshold $\lambda_c^{[\theta=0]} = 1$.

Linear stability analysis shows that the solution $\rho_1^* = 0$ is stable only when $\lambda < \lambda_c^{[\theta=0]}$. Contrarily, for values of $\lambda \geq \lambda_c^{[\theta=0]}$, the absorbing state $\rho_1^* = 0$ becomes unstable while $\rho_2^{*[\theta=0]}$ becomes stable, i.e., the epidemic takes place.

2.3.2 Case $\theta = 1$

Let us consider now the other extreme case, $\theta = 1$, in which only the local effects are considered in the CR, and therefore the recovery phase is purely dynamical. Also in this case the second solution of the stationary state equation becomes trivial, and reads $\rho_2^{*[\theta=1]} = 1$. Thus, the system presents two stationary solutions, and it is easy to show that their stability changes at the same epidemic threshold as for $\theta = 0$, so that $\lambda_c^{[\theta=1]} = \lambda_c^{[\theta=0]} = \lambda_c$. For $\lambda < \lambda_c$ ρ_1^* is stable and ρ_2^* is unstable, while for $\lambda > \lambda_c$ we have the opposite case. Therefore, contrarily from the completely non-local case ($\theta = 0$), here the system undergoes an explosive transition from the healthy to the endemic state, where all individuals are adopters.

2.3.3 General case

In the most general case, the second solution ρ_2^*, given by Eq. (2.11), depends on both the rescaled infectivity λ (therefore on the average degree $\langle k^{[1]} \rangle$) and on θ, the parameter which controls for the local effects of the CR. Notice that the solution is acceptable if $\rho_2^* \in [0, 1]$, which implies again $\lambda \in [1, \infty]$. Therefore, for any θ, the transition from the healthy to endemic state happens at the same epidemic threshold $\lambda_c = 1$, but the density of infected in the endemic state varies with θ. The stability of the fixed point ρ_2^* can be easily investigated by defining the second term in Eq. (2.9) as $F(\rho) = \lambda\rho[1 - \rho]$ and then checking the sign of the derivative of $F(\rho)$. Since $F'(\rho)|_{\rho=\rho_2^*} = 1 - \lambda$ does not depend on θ, ρ_2^* always represents a stable fixed point for the dynamics.

2.4 Microscopic Markov-chain approach

In the microscopic Markov-chain approach (MMCA), the probability of node i to be infected at time t $\text{Prob}\,[x_i(t) = 1] = \pi_i(t)$ is a random variable, and it is assumed that for different nodes these probabilities are independent [235]. The equation for the discrete-time evolution of $\pi_i(t)$ can then be written as:

$$\pi_i(t + 1) = (1 - q_i(t))(1 - \pi_i(t)) + (1 - \gamma_i(t))\pi_i(t) \tag{2.13}$$

where the two terms of the right-hand side are respectively

- the probability that node i, susceptible at time t, gets infected by a neighbour;
- the probability that node i, infected at time t, does not recover;

$q_i(t)$ represents the probability of node i not being infected by any of his neighbours at time t, and it can be written in terms of the adjacency matrix $A^{[1]}$ of layer 1, which controls the contacts between nodes i and j, as

$$q_i(t) = \prod_j 1 - \beta a_{ij}^{[1]} \pi_j(t) \tag{2.14}$$

while the recovery probability $\gamma_i(t)$, for the case $h = 1$, is given by

$$\gamma_{i,h=1}(t) = \mu_0 \left(1 - \theta \frac{\sum_j a_{ij}^{[2]} \pi_j(t)}{\sum_j a_{ij}^{[2]}} \right) \tag{2.15}$$

Finally, the stationary state $\pi_i(t+1) = \pi_i(t)$ is given by

$$\pi_i = (1 - q_i) + (q_i - \gamma_i)\pi_i \tag{2.16}$$

The system of equations given by Eq. (2.16) is then solved numerically and the density of infected is obtained by taking the average over all the nodes:

$$\rho = \frac{1}{N} \sum_i \pi_i \tag{2.17}$$

The limitations of this approach have been discussed in Ref. [236, 237]. Moreover, in our specific case, the underlying lattice-like structure of the network topologies that will be used in Sec. 2.5 can break the assumption that the probabilities $\pi_i(t)$ and $\pi_j(t)$ of two different nodes to be infected are independent.

2.5 Numerical results on synthetic networks

We present here numerical simulations of the ADM on various synthetic duplex networks. In each case the simulations are performed for different realizations of the networks, stopping each run whenever an absorbing state is reached. Alternatively, if a stationary

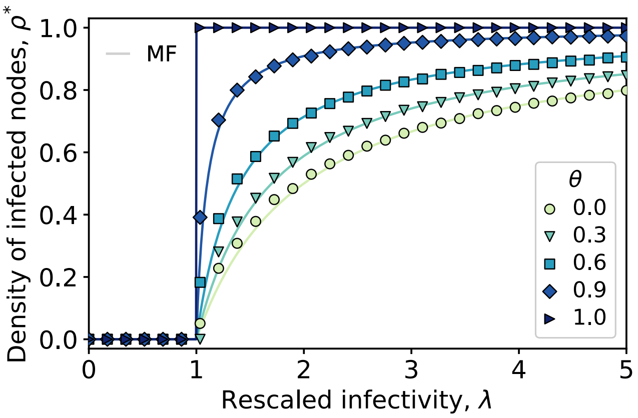


Figure 2.2: Numerical simulations of ADM ($h = 1$) on a duplex network formed by two ER networks with $N = 900$ and $\langle k \rangle = 10$. The average fraction of infected nodes is plotted against the rescaled infectivity $\lambda = \beta \langle k^{[1]} \rangle / \mu_0$. Different curves (and colours) correspond to different values of the parameter θ, which controls for the strength of the local effects in the complex recovery process, as defined in Eq (2.2). The case $\theta = 0$ corresponds to the standard SIS model. Simulations (points) are plotted together with the analytical mean-field (MF) solution of Eq. (2.9) (continuous lines). Figure from [3].

state is reached the stationary density of infected is computed by considering an average over the last 100 time-steps. Each run starts with different initial conditions, given by randomly placing a seed of ρ_0 infectious nodes (usually 1% of the nodes), and then we average the results over all the runs. Throughout all the numerical simulations presented in this paper, we restrict for simplicity to the case $h = 1$.

The first system we have considered is a duplex with $N = 900$ nodes formed by two Erdős-Rényi (ER) random graphs having average degrees $\langle k^{[1]} \rangle = \langle k^{[2]} \rangle = \langle k \rangle = 10$. Figure2.2 shows the stationary density of infected ρ^*, obtained by averaging the prevalence curves for different realizations of the numerical simulations, as a function of the rescaled infectivity $\lambda = \beta \langle k^{[1]} \rangle / \mu_0$. Different curves correspond to different values of the parameter θ, which controls for the local effects in the CR process. Indeed, the case $\theta = 0$ is equivalent to the standard SIS model with spontaneous recovery, where a non-zero density of infected nodes in the stationary state appears for values of λ larger than a critical value $\lambda_c = 1$. By increasing θ, the density of infected nodes in the endemic state $\rho^* > 0$ increases and the transition becomes steeper and steeper, until the

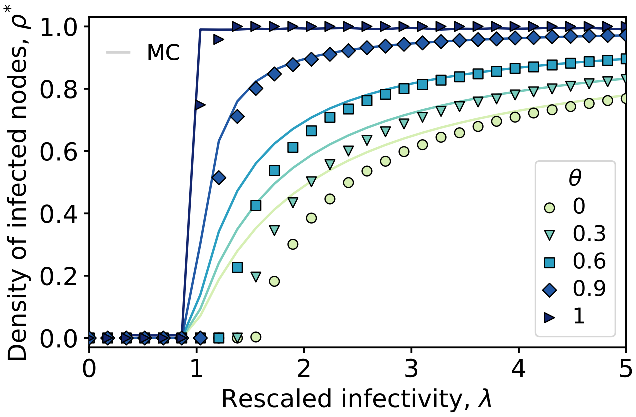


Figure 2.3: Numerical simulations of the ADM ($h = 1$) on a duplex network formed by a 2D lattice and a SW network with $N = 2500$ nodes, $\langle k^{[2]} \rangle \approx 4$, and $p = 0.01$. The average fraction of infected nodes is plotted against the rescaled infectivity $\lambda = \beta \langle k^{[1]} \rangle / \mu_0$. Different curves (and colours) correspond to different values of the parameter θ. Simulations (points) are plotted together with the curves obtained with the macroscopic Markov-chain approach (MC in the figure legend) as given by Eq. (2.17) (continuous lines). Figure from [3].

extreme case $\theta = 1$. In this latter case, i.e., when the recovery is purely dynamical, the systems undergoes a discontinuous transition from the absorbing state $\rho^* = 0$ with no adopters to the opposite state $\rho^* = 1$. Notice that the transition occurs at the same critical threshold $\lambda_c = 1$. Figure 2.2 also shows the continuous curves representing the analytical prediction in the MF approach, as given by Eq. (2.11). The match between curves and points confirms the accuracy of the MF approximation in reproducing the dynamics of the ADM in the case of random graphs, and also its ability to capture the different types of transitions the ADM exhibits when the value of θ is changed.

As a second system, we have considered a slightly more realistic synthetic duplex network. In particular, we model the power-grid layer as a 2D lattice ($N = 2500$, $\langle k^{[2]} \rangle \approx 4$) and we couple it to a social layer which is obtained from the same 2D lattice, by rewiring each of its links at random with a probability $p = 0.01$. It is worth clarifying that we will call this layer small-world (SW), given the similarity of the rewiring mechanism with the original small-world model proposed by Watts and Strogatz [25]. The rewiring mechanism, adopted only at layer 1, breaks the regularity of the lattice by introducing social connections between nodes that are not first neighbours at the level of the power grid network in layer 2. The results obtained are shown in Fig. 2.3. We notice a few differences with respect to the results reported in Fig. 2.2. In particular, we observe that the threshold λ_c slightly increases when the value of θ changes from $\theta = 0$ to $\theta = 1$, and this behaviour is not captured neither by the analytical predictions in the MF approximation, nor by the more accurate microscopic Markov-chain approach (MMCA, see Sec. 2.4), whose curves are shown as continuous lines. Such differences might be due to the limitations of the MMCA caused by the time discretization [236, 237] and, most probably, to the strong correlations between nodes induced by the underlying lattice-like structure of SW networks (which would break the assumptions of independence we used to write Eq. 2.13).

2.6 Numerical results on real-world networks

In the previous section we explored the model on two synthetic duplex networks. With the first we observed the phenomenology on two random graphs, while in the second we considered a more realistic—yet synthetic— structure composed by a lattice and a SW network. Here, we make a further step in this direction by considering street networks [238] from the real world as proxies for power grid networks at the distribution level, and a multilayer adaptation of the well-known Waxman random graph model[239] to represent the social layer. With this approach, the spatial nature of the street network is used both to embed the power grid into the physical space and to shape the connectivity patterns of the social layer. Due to the use of the Waxman model, the social connections decay exponentially with the distances on the network, which in turn are affected by the physical constraints imposed by the morphology of the territory. Details on the exact construction are given in the following section.

2.6.1 Construction of the real-world multiplex network

We construct the power-grid layer at the distribution level by taking as a proxy the street network in its primal approach [238], i.e., by representing crossroads as nodes connected by streets. According to our ADM, each node corresponds to a household, hence we approximate the households connectivity in the distribution grid by the street network formed by intersection points linked by streets.

The construction of the street network starts from the same data set used in Ref. [240]: the Ordnance Survey (OS) MasterMap [241]. This data set consists of a clean street network for the entire Britain, in which roundabouts have been replaced by single intersections, and in which each edge comes with an associated weight representing the length of the street (for more details see Ref. [240]). We first restrict our data set to the Greater London Authority by retaining only those points (nodes) falling within the boundaries of the LSOA (Lower Super Output Area)[242]. Then, we select a smaller neighbourhood in central London by following the hierarchical percolation method proposed in Ref. [240].

The method produces a clustering on the nodes based on a single parameter ϵ, which acts as a percolation threshold on the street distance between the points. More precisely, given a threshold ϵ, the graph is divided into different connected components corresponding to the sub-graphs induced by the thresholding on the nodes at a distance smaller than ϵ. For increasing values of ϵ components are then aggregated with one another, eventually collapsing into a single one (see Ref. [240] for more details). An example of the top-8 largest connected components obtained for different values of the percolation threshold ϵ (meters) for the city of London is shown in Fig. 2.4.

The method gives rise to neighbourhoods at different scales arising from the density of the street intersections. In the case of London, some scales reveal its composition in terms of historical villages, corresponding now to differentiated neighbourhoods. By selecting an appropriate scale, this method allows us to focus on a relatively small

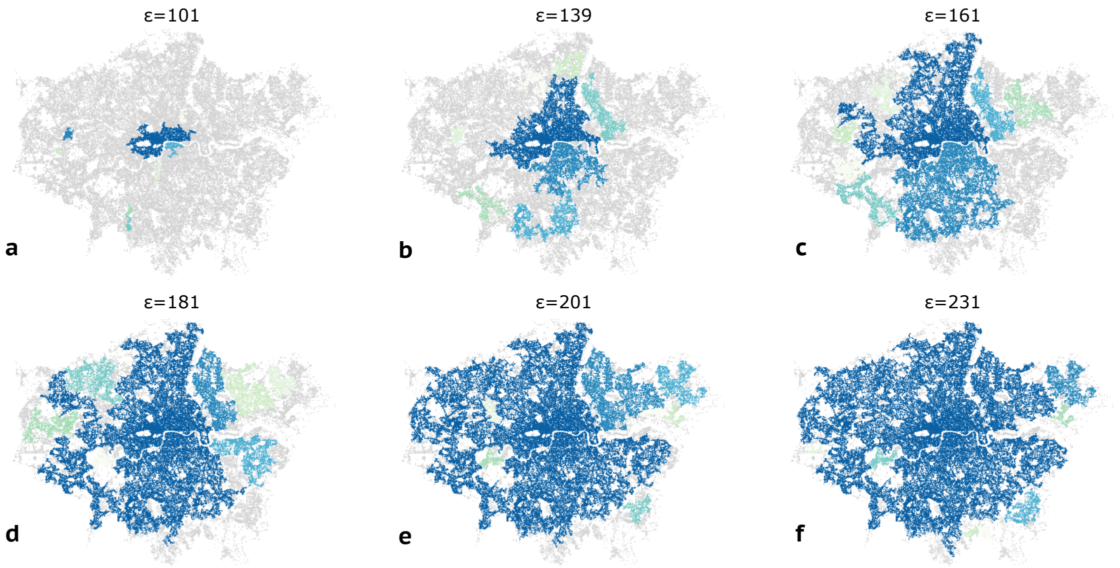


Figure 2.4: Percolation process on the street network of London. Clusters of nodes (crossroads) of the street network are depicted with different colours (only the top-8 largest clusters are shown). Each panel shows the outcome of the percolation process obtained with a different value of the percolation threshold ϵ (meters), reported on top. Notice how the areas situated south of the River Thames merge together with the one in the north when moving from (**c**) to (**d**).

portion of the city, as a targeted adoption campaign would do, while keeping at the same time the computational cost at a reasonable level. Here, we restrict our attention to a neighbourhood of the city of London [see Fig. 2.5(**a**)] corresponding to the largest connected component obtained for a threshold of $\epsilon = 89$ meters. The resulting network is composed of $N = 3379$ nodes and $K^{[2]} = 4602$ links.

We construct the geographical social network starting from the well-known Waxman random graph model [239]. In the standard model, nodes are initially placed at random over a plane and then connected in pairs with a probability that decays exponentially with their distance. Here, we modify the model in two ways: (i) nodes are not placed at random, but the geographical position of each node on the social layer corresponds to the position on the power-grid layer; (ii) instead of considering the geographical distance between the nodes, we consider the network distance. Notice that, being the network embedded in space, the network distance is already shaped by the particular spatial displacement of the nodes. More precisely, the model works as follows.

Given the set of nodes (and their position), let us call $d^{[2]}(i,j)$ their distance in layer 2 (the power-grid layer). Then, the probability that i and j are connected in layer 1 is given by:

$$P^{[1]}(i,j) = \alpha \exp\left[-d^{[2]}(i,j)/\alpha L^{[2]}\right] \tag{2.18}$$

where $L^{[2]}$ denotes the diameter of layer 2 and α is a tunable model parameter.

We construct the network for the social layer by linking the nodes of the grid layer according to the probability given by Eq. (2.18), with α tuned in order to obtained a reasonable number of influential household connections ($\alpha = 0.003$). The resulting network has $K^{[1]} = 17183$ links, corresponding to an average of $\langle k^{[1]} \rangle \approx 10$ connections per household.

2.6.2 Results

Using the procedure described in the previous section, we produce a duplex network with $N \approx 3000$ nodes, and an average degree of $\langle k^{[1]} \rangle \approx 10$ at the social layer and $\langle k^{[2]} \rangle \approx 3$ at the power-grid layer. The associated degree distributions are shown in Fig. 2.5(**b**).

As in the previous cases, we investigate the density of infected individuals in the stationary state as a function of the rescaled infectivity. The plots reported in Fig. 2.5(**c**) confirm similar results to those obtained with synthetic networks. In particular, a clear change in the nature of the transition is observed also when more realistic network structures are used both at the grid and at the social layer. Associated to the sudden transitions at large values of θ, we have also observed the appearance of hysteresis loops. An example is shown in the inset plot for the case $\theta = 0.9$. In the next section we will explore these phenomena more in details.

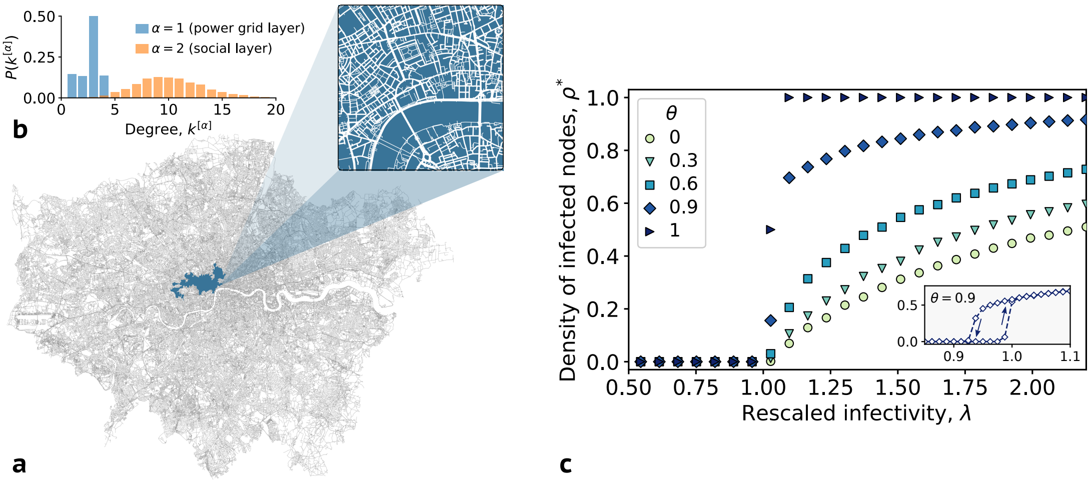


Figure 2.5: ADM ($h = 1$) on a real-world duplex network in which a street network is used as a proxy for the power grid. (**a**) A central neighbourhood in London is selected by using a hierarchical percolation approach (blue zone). The degree distributions of the street network and the coupled social network constructed from it are shown in panel (**b**). (**c**) The average fraction of infected nodes obtained by means of numerical simulations is plotted against the rescaled infectivity $\lambda = \beta \langle k^{[1]} \rangle / \mu_0$. Different curves (and colours) correspond to different values of the parameter θ, which controls for strength of the local effects in the CR process, as defined by Eq (2.2). The inset shows the hysteresis loop, which appears close to the threshold for $\theta = 0.9$. Figure from [3].

2.6.3 Varying the size of the initial seed

In what follows we briefly investigate the effects of initial conditions in the evolution of the density of infected nodes [243]. Most of the existing literature targets this problem within the domain of infectious diseases spreading, which translates into looking for optimal immunisation strategies, i.e., key nodes to vaccinate in order to limit the spread [244–246]. Here, we investigate the temporal aspect of the infection dynamics as a function of the initial conditions. This will be done in two different ways, since we can control for both the number and the position in the network of the initial adopters, i.e. of those nodes who will initiate the spreading.

We start considering a set of randomly placed infectious nodes, as before, whose size at time 0 is controlled by the density $\rho(t = 0) = \rho_0$. We then simulate the ADM with different values of the parameters (λ, θ) for different initial densities ρ_0 in the range $(0, 0.6]$. Results are shown in Fig. 2.6. Each panel corresponds to a given pair of

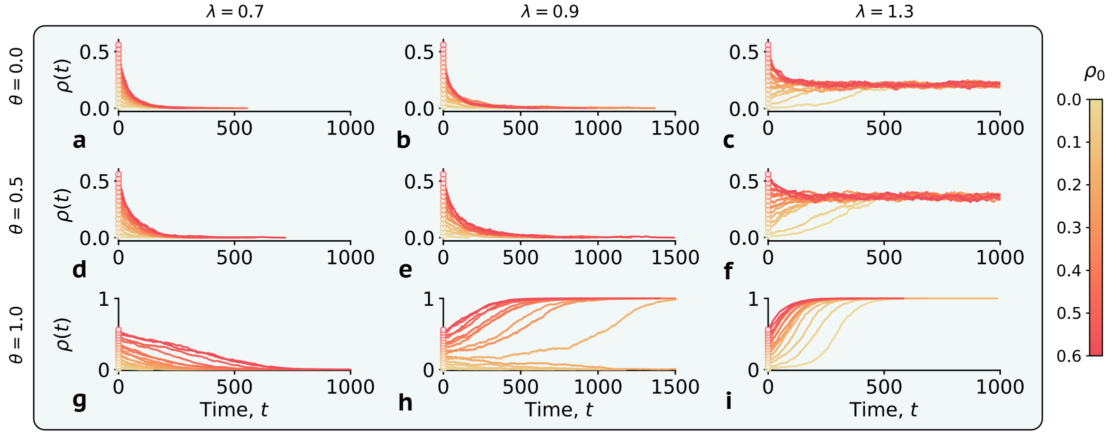


Figure 2.6: Effect of the initial density of adopters on the temporal evolution of the spreading. Each panel shows the densities of infectious nodes for different sizes of the initial seed of infectious adopters $\rho(t = 0) = \rho_0$ and for different values of (θ, λ). (**a**-**c**) refer to the standard SIS model, without local effects, while (**g**-**i**) represent the other extreme case in which the recovery process is completely controlled by the local dynamics. Different scales have been adopted for panels (**b**), (**e**), and (**h**) due to the proximity of the epidemic threshold λ_c, which makes the runs last longer. Figure from [3].

parameters (λ, θ), while different curves within the same panel display the temporal evolution of the density of infected nodes when considering different ρ_0 (see colourbar on the right-hand-side). Rows indicate different values of θ, moving from the standard SIS model with no dynamical recovery [$\theta = 0$, (**a**-**c**)], to the other extreme case in which the recovery process completely depends on the local dynamics of the neighbouring nodes ($\theta = 1$, g-i). An intermediate case with $\theta = 0.5$ is also considered (panels d-f). Similarly, we use three values of the infectivity λ: one below the epidemic threshold [$\lambda = 0.7$, (**a**,**d**,**g**)], one close to the epidemic threshold [$\lambda = 0.9$, (**b**,**e**,**h**)], and one above the epidemic threshold ($\lambda = 1.3$, c,f,i). Trivial effects are found when we are below and above the threshold. In particular, in the first case [panels (**a**,**d**,**g**)], the higher the ρ_0 the longer it takes to the system to reach the absorbing state $\rho^* = 0$. Similarly, when high values of λ are considered [panels (**c**,**f**,**i**)] what matters is the distance between the initial density of the seed ρ_0 and the final stationary state, which in turn depends on θ. However, close to the threshold, the strong dynamical effects of the CR process create a bi-stable region [panel (**h**)], in which the initial density of infectious nodes ρ_0 determines whether the systems will end up in the absorbing states without adopters ($\rho^* = 0$) or with all adopters

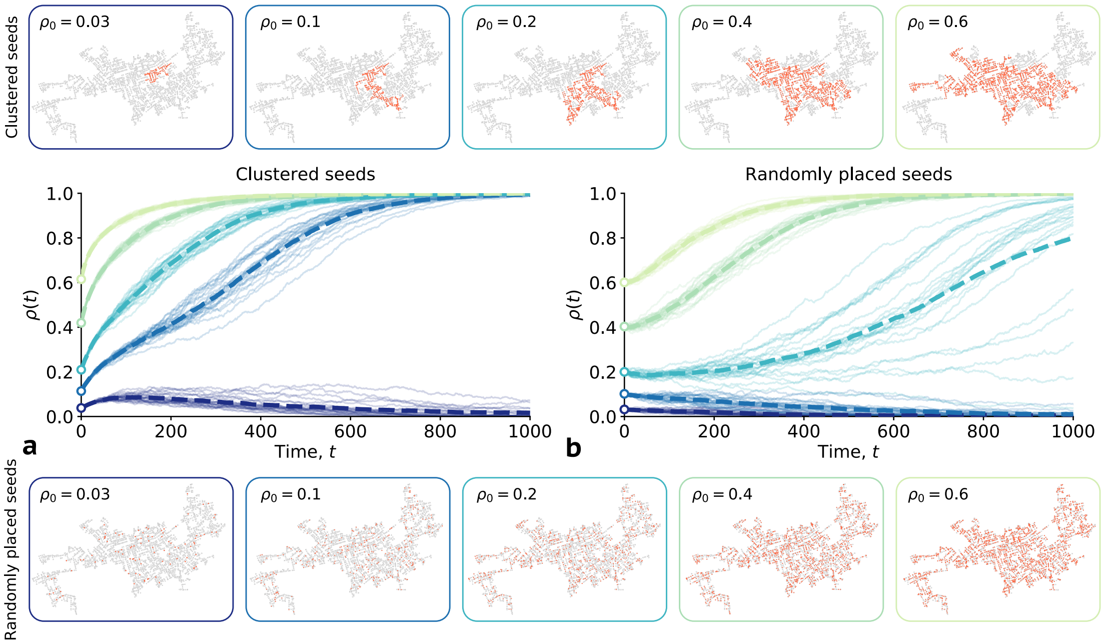


Figure 2.7: Effect of the position of the initial seed of adopters in the ADM on the real-world duplex network with parameters: $\lambda = 0.9, \theta = 1$. The temporal evolution of the densities of infectious nodes is showed for the two considered scenario: (**a**) a clustered seed of infectious nodes on the power-grid layer and (**b**) a randomly placed seed of infectious nodes. Different colours correspond to different sizes of the initial seed of infectious ρ_0 (single realizations are plotted as continuous lines, while dashed lines represent their average). The actual positions of the seeds are shown, for each ρ_0, in the top and bottom maps, representing, respectively, the clustered and the random scenario. Infectious nodes are depicted in red. Figure from [3].

($\rho^* = 1$). Notice that such bi-stability is not present in the MF formulation presented in Sec. 2.3.

2.6.4 Varying the position of the initial seed

To better understand the phenomenology of the ADM close to the threshold, we fix the parameters to the latest case ($\lambda = 0.9$, $\theta = 1$) and explore the effects of the initial position of the seed, while still varying its size. To do this, we consider a different scenario in which the initial adopters form a cluster in the power-grid layer. This cluster corresponds to a smaller neighbourhood within the considered area, and it is selected with the same hierarchical percolation approach used in the construction of the real-world multiplex network, but for smaller percolation thresholds (see Sec. 2.6 for more details on the method). We will compare this scenario with the standard case in which the initial seed of adopters is placed at random instead.

The results are shown in panel (**a**) and (**b**) of Fig. 2.7 respectively. In each case the temporal evolution for the density of infected nodes is plotted for different values of ρ_0 (different colours). Shaded curves are different realizations of the ADM, while their average is represented by the dashed thick lines. As a reference, we also add the maps showing the exact position of the seed of infectious nodes (depicted in red) for each scenario and for the different values of ρ_0 considered. Maps for the clustered seed and randomly placed seed are shown in the top and bottom row respectively. Different behaviours emerge from the two scenarios. Indeed, the prevalence curves seem to fluctuate more when the seed is placed at random (panel b). Interestingly, this phenomenon is in contrast with the results of the MF formulation and the simulations of the ADM on synthetic ER networks discussed in Sec.s 2.3 and 2.5. As a result, when a density $\rho_0 = 0.2$ of initial adopters is considered, fluctuations are so strong that, even with exactly the same (random) initial positions of adopters, the system can end up in either the absorbing state $\rho^* = 0$ or in the state $\rho^* = 1$. If clustered seeds are considered instead (panel a), the same seed size $\rho_0 = 0.2$ always drives the system to the absorbing state with all adopters ($\rho^* = 1$). In this scenario, the critical value for ρ_0 which separates the two basins of attractions seems to be better defined: see the curves for $\rho_0 = 0.03$ and $\rho_0 = 0.1$. Finally, it is worth noticing that for this last seed size, $\rho_0 = 0.1$, the initial placement completely determines the final state of the system. Indeed, only 10% of clustered adopters are enough to drive the system towards the full adoption case, while this does not happen if a seed of the same size is placed at random.

2.7 Summary and conclusions

In this Chapter, we have introduced and studied both numerically and when possible analytically a mathematical model of spreading on a network with a dynamical recovery mechanism of the nodes, which is a function of the network state. Our original purpose is to reproduce the dynamics of service adoption in demand response management [247–250], in which the behaviour of a customer is influenced by its social contacts, in addition of also depending on the specific spatial configuration of other customers in close proximity within the power grid service area. For this reason, we consider a duplex network with a social layer and a power-grid layer. The adoption process is modelled as an epidemic spreading on the social layer, with a recovery rate of the nodes that depends on the states of their neighbours on the power-grid layer. In this way the dynamics tends to preserve clusters of infected individuals by making an infected node surrounded by nodes in the same state less keen to recover.

Results suggest that the more the recovery of the nodes depends on the local influence of peers (large values of θ), the more discontinuous the transition from non-adopters (healthy) to full adoption (fully infected network) becomes. In simulations on real-world networks, such as the London network, we also noted that the final state of the system is not uniquely defined by the infection and recovery parameters β and μ, but the initial conditions can have a substantial impact on the spreading dynamics, with clustered seeds in the power-grid layer more likely leading to full network infection. We have found that a mean-field approximation captures the simulation results nicely for Erdős-Rényi networks, while more advanced analytical descriptions might be necessary to characterise our model on more realistic and complex network structures.

While we motivated our model from the electricity demand management, other applications that rely on local customer resources should follow similar dynamics [251]. This could for example include car sharing or citizen science projects. The main message we can derive for all such systems from the analysis of our model is the following. In any real application case, we would first need to determine the strength of the local influence of other customers, i.e. the magnitude of the parameter θ in our model. If such local influence is weak, a smooth transition to a non-zero density of adopters takes place when the infectivity is above a given threshold. When the local influence is strong instead, we observe a sudden transition and the appearance of an intermediate (hysteresis) region where, as soon as a critical mass is reached, (almost) everyone would adopt the new technology.

Our results also show that strong local influence is key to determine adoption, hence giving insights on how to strategically plan on the nodes to be targeted initially. Namely, reaching out to customers who are physically spread out in the power-grid layer, or explicitly targeting social clusters which are not defined spatially, is more likely to fail than starting the adoption process from clusters in the power-grid layer, which provides the positive feedback-loop for the consumers. In fact, a cluster of "adopters" on the power-grid layer is likely to stay within the contract and will also convince their neighbours to join. Therefore, this strategy will in most cases lead to a higher penetration of the new technology. Advertisement should take this into account, e.g. by explicitly advertising within local communities. Alternatively, businesses might try to re-shape the infection and recovery process itself. Already commonly adopted "hire a friend" schemes try to build a positive feedback among customers, which can directly strengthen the infection process. Our results suggest that alternative "hire a neighbour" schemes, specifically designed to target neighbours in the power-grid layer, could also positively contribute, this time by altering the recovery dynamics.

The presented approach can easily be extended into multiple directions. On the one hand, one could further improve the accuracy of the network topologies, for example by modelling the social network as a crossover between scale-free and spatial networks, as proposed in Ref. [252], or by directly using real-world friendship data. Further analytical development might also be required in this direction. On the other hand, different variations of the model are also possible. For instance, the dynamical recovery term $\mu_i(t)$ could include influences from nodes $h = 2$, or more, hops away. Alternatively, it could rely on Euclidean distances calculated on the physical space, or other distance functions more realistically mapping consumer decisions. Another natural extension might involve the introduction of other, both exogenous and endogenous, effects as additional contributions to the adoption dynamics. These could include, for example, mass media exposure [134] and the explicit introduction of economic factors such as price of energy or price incentives. Finally, in this Chapter we focused on demand control for households, specifically neglecting industrial consumers, as they follow very different rules and require independent business models which are also more likely follow different adoption schemes. Future work should extend the present framework to non-household customers.

In the next Chapter, we will take one step back and reconsider the choice of the representation for the dynamics of social contagion.

Chapter 3

Social contagions beyond pairwise interactions

3.1 Introduction

In Sec. 1.2.1 we have introduced the problem of modelling the dynamics of social contagions in a population, presenting the different frameworks and theory that are commonly used. In particular, we have focused on the differences between models of simple and complex contagion. Then, in Chapter 2 we have considered the effects of having a complex recovery mechanism in a model of multilayer adoption dynamics.
The model introduced, as well as the famous models of complex contagion present in the literature, are however still defined on networks of interactions between individuals: even when multiple interactions are needed for a contagion to take place, in both threshold and epidemic-like models, the fundamental building blocks of the system are pairwise interactions, structurally represented by the links of the network on which the process is taking place [79, 116, 121–125].

In this Chapter, we propose to go further and take into account that contagion can occur in different ways, either through pairwise interactions (the links of a network) or through group interactions, i.e., through higher-order structures [5]. Indeed, while an individual can be convinced independently by each of his/her neighbours (simple contagion), or by the successive exposure to the arguments of different neighbours (complex contagion), a fundamentally different mechanism is at work if the neighbours of an individual convince him/her in a group interaction. For example, we can adopt a new norm because of two-body processes, which means we can get convinced, separately, by each one of our first neighbours in our social network who have already adopted the norm. However, this is qualitatively different from a mechanism of contagion in which we get convinced because we are part of a social group of three individuals, and our two neighbours are both adopters. In this case the contagion is a three-body process, which mimics the simplest multiple source of reinforcement that induces adoption. The same argument can easily be generalized to larger group sizes.

To build a modelling framework based on these ideas, we formalise a social group as a simplex, and we adopt simplicial complexes as the underlying structure of the social system under consideration [see Fig. 3.1(**a**-**b**)]. This simplicial representations is indeed more suited than networks to describe the co-existence of pairwise and higher-order interactions (see Sec. 1.1). We recall that, in its most basic definition, a k-simplex σ is a set of $k+1$ vertices $\sigma = [p_0, \dots, p_k]$. It is then easy to see the difference between a group interaction among three elements, which can be represented as a 2-simplex or "full" triangle $[p_0, p_1, p_2]$, and the collection of its edges, $[p_0, p_1], [p_0, p_2], [p_1, p_2]$. Just like a collection of edges defines a network, a collection of simplices defines a simplicial complex. Formally, a simplicial complex $\mathcal{K}$ on a given set of vertices $\mathcal{V}$, with $|\mathcal{V}| = N$, is a collection of simplices, with the extra requirement that if simplex $\sigma \in \mathcal{K}$, then all the sub-simplices $\nu \subset \sigma$ built from subsets of σ are also contained in $\mathcal{K}$. Such a requirement, which makes simplicial complexes a special type of hypergraphs (see Sec. 1.1), seems appropriate in the definition of higher-dimensional groups in the context of social systems, and simplicial complexes have indeed been used to represent social aggregation in human communication [66]. Removing this extra requirement would imply, for instance, modelling a group interaction of three individuals without taking into account also the dyadic interactions among them. The same argument can be extended to interactions of four or more individuals: it is reasonable to assume that the existence of high-order interactions implies the presence of the lower-order interactions. For simplicity and coherence with the standard network nomenclature, we call nodes (or vertices) the 0-simplices and links (or edges) the 1-simplices of a simplicial complex $\mathcal{K}$, while 2-simplices correspond to the (“full") triangles, 3-simplices to the tetrahedra of $\mathcal{K}$, and so on [see Fig. 3.1(**a**)].

Here, we thus propose a new modelling framework for social contagion, namely a model of *simplicial contagion*: this epidemic-like model of social contagion on simplicial complexes takes into account the fact that contagion processes occurring through a link or through a group interaction both exist and have different rates. Our model therefore combines stochastic processes of simple contagion (pairwise interactions) and of complex contagion occurring through group interactions in which an individual is simultaneously exposed to multiple sources of contagion.

3.1.1 Outline

This Chapter is structured as follows:
In Sec. 3.2, we introduce the *Simplicial Contagion Model* (SCM), explaining in particular the role played by the higher-order interactions in the the contagion dynamics.
In Sec. 3.3, we first extract empirical high order structures from data of face-to-face interactions and then perform extensive numerical simulations on them to investigate the model behaviour.
In Sec. 3.4, we further analyse the dynamics of the model on synthetic simplicial complexes with controlled properties. This requires defining a new model for random simplicial complexes that generates simplices of different dimensions and in which the expected local connectivity can be tuned as desired.
In Sec. 3.5, we develop an analytical approach in which we derive and solve the mean

field equations describing the evolution of density of infected nodes. We analytically show, in agreement with the numerical results, that the higher-order interactions lead to the emergence of new phenomena, changing the nature of the transition at the epidemic threshold from continuous to discontinuous and leading to the appearance of a bistable region of the parameter space where both healthy and endemic asymptotic states co-exist. Indeed, the mean-field analytical approach correctly predicts the steady-state dynamics, the position and the nature of the transition and the location of the bistable region. We also show that, in the bistable region, a critical mass is needed to reach the endemic state, reminding of the recently observed minimal size of committed minorities required to initiate social changes [253].

In Sec. 3.6, we extend the analytical formalism by introducing the microscopic Markov-chain approach for simplicial contagions. This approach, differently from the mean field one, can correctly captures the behaviour of the model also on simplicial complexes with heterogeneous structures.

Conclusions and future perspectives are summarized in Sec. 3.7.

3.2 The simplicial contagion model

In order to model a simplicial contagion process, we associate a dynamical binary state variable x to each of the N vertices of $\mathcal{K}$, such that $x_i(t) \in \{0, 1\}$ represents the state of vertex i at time t. Using a standard notation, we divide the population of individuals into two classes of susceptible (S) and infectious (I) nodes, corresponding respectively to the values 0 and 1 of the state variable x. In the context of adoption processes, the state I represents individuals who have adopted a behaviour. At each time t, the macroscopic order parameter is given by the density of infectious nodes $\rho(t) = \frac{1}{N}\sum_{i=1}^{N} x_i(t)$. The model we propose here, the so-called Simplicial Contagion Model (SCM) of order D, with $D \in [1, N-1]$, is governed by a set of D control parameters $B = \{\beta_1, \beta_2, \dots, \beta_D\}$, whose elements represent the probability per unit time for a susceptible node i that participates to a simplex σ of dimension D to get the infection from each one of the

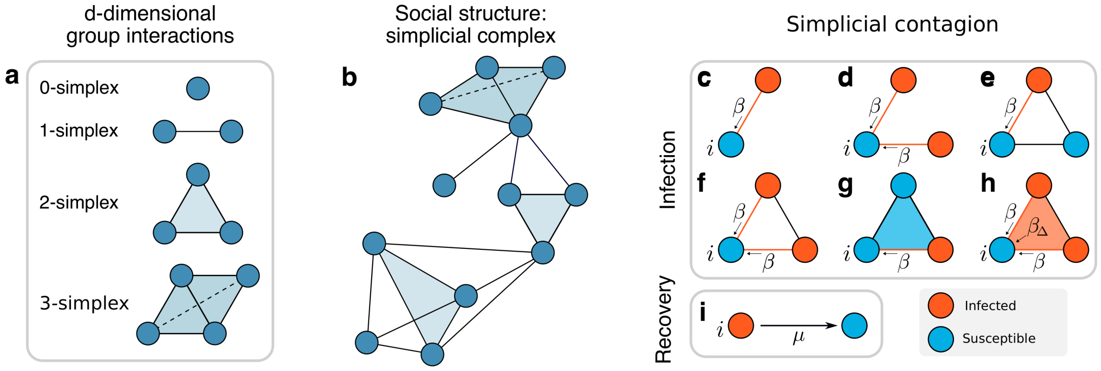


Figure 3.1: Simplicial Contagion Model (SCM). The underlying structure of a social system is made of simplices, representing d-dimensional group interactions (**a**), organized in a simplicial complex (**b**). (**c-h**) Different channels of infection for a susceptible node i in the Simplicial Contagion Model (SCM) of order $D = 2$. Susceptible and infected nodes are coloured in blue and red, respectively. Node i is in contact with one (**c**, **e**) or more (**d**, **f**) infected nodes through links (1-simplices), and it becomes infected with probability β at each time step through each of these links. (**g-h**) Node i belongs to a 2-simplex (triangle). In (**g**) one of the nodes of the 2-simplex is not infected, so i can only receive the infection from the (red) link, with probability β. In (**h**) the two other nodes of the 2-simplex are infected, so i can get the infection from each of the two 1-faces (links) of the simplex with probability β, and also from the 2-face with probability $\beta_2 = \beta_\Delta$. (**i**) Infected nodes recover with probability μ at each time step, as in the standard SIS model. Figure from [2].

subfaces composing σ, under the condition that all the other nodes of the subface are infectious. In practice, with this notation, β_1 is equal to the standard probability of infection β that a susceptible node i gets the infection from an infected neighbour j through the link (i, j) (corresponding to the process $S + I \longrightarrow 2I$). Similarly, the second parameter $\beta_2 \equiv \beta_\Delta$ corresponds to the probability per unit time that node i receives the infection from a “full" triangle (2-simplex) (i, j, k) in which both j and k are infectious, $\beta_3 = \beta_\boxtimes$ from a group of size 4 (3-simplex) to which it belongs, and so on. Such processes can be represented as $Simp(S, nI) \longrightarrow Simp((n+1)I)$: a susceptible node, part of a simplex of $n + 1$ nodes among which all other n nodes are infectious, becomes infectious with probability per unit time β_n. Thanks to the simplicial complex requirements that all sub-simplices of a simplex are included, contagion processes in a n-simplex among which $p < n$ nodes are infectious are also automatically considered, each of the $n + 1 - p$ susceptible nodes being in a simplex of size $p + 1$ with the p infectious ones. Notice, however, that this assumption can be dropped and the contagion model extended to the case of hypergraphs [254, 255], as it has been subsequently done in Ref. [137, 256]. Figure 3.1(**c**-**h**) illustrates the concrete example of the six possible ways in which a susceptible node i can undergo social contagion for an SCM of order $D = 2$ with parameters β and β_{Δ}. Finally, the recovery dynamics ($I \rightarrow S$) is controlled by the node-independent recovery probability μ [Figure 3.1(**i**)]. Notice that the SCM of order D reduces to the standard SIS model on a network when $D = 1$, since in this case the infection can only be transmitted through the links of $\mathcal{K}$.

3.3 Simplicial contagion on real-world simplicial complexes

To explore the phenomenology of the SCM, we first consider its evolution on empirical social structures. To this aim, we consider publicly available data sets describing face-to-face interactions collected by the SocioPatterns collaboration [257]. Face-to-face interactions represent indeed a typical example in which group encounters are fundamentally different from sets of binary interactions and can naturally be encoded as simplices. The time-resolved nature of the data allows us to create simplicial complexes describing the aggregated social structure.

3.3.1 Construction of simplicial complexes from face-to-face interactions data

We consider four data sets of face-to-face interactions collected in different social contexts: a workplace (InVS15) [258], a conference (SFHH) [259], a hospital (LH10) [260] and a high school (Thiers13) [261]. In each case, face-to-face interactions have been measured with a temporal resolution of 20 seconds. We first aggregated the data by using a temporal window of $\Delta t = 5$ minutes, and computed all the maximal cliques that appear (see Fig. 3.2). Since we limit our study to the case $D = 2$, we need to produce a clique complex formed by 1- and 2-simplices. Therefore, we considered all the 2- and 3-cliques and weight them according to their frequency. Note that while higher-dimensional cliques are

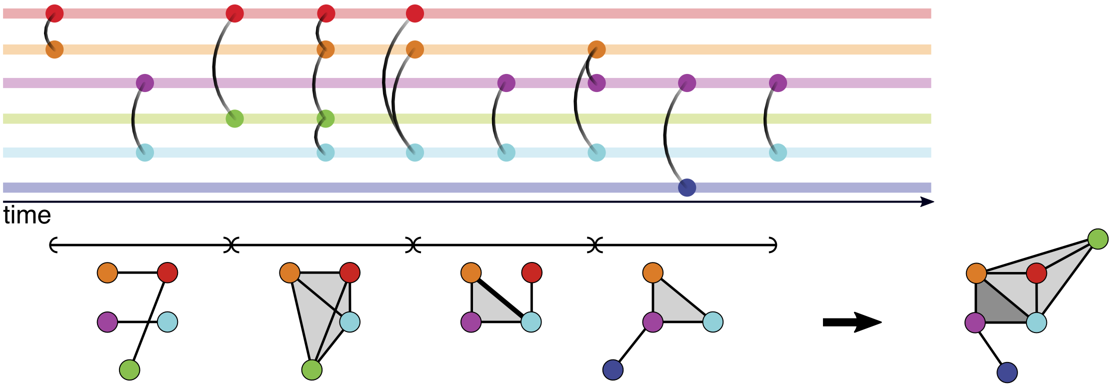


Figure 3.2: Construction simplicial complexes from face-to-face interactions data. Contacts are first aggregated by using a temporal window of $\Delta t = 5$ minutes. Maximal cliques in each window are then "promoted" to simplicial complexes. The complexes of all the snapshots are finally aggregated into a single simplicial complex keeping the frequency of each clique as the weight of each simplex. Figure inspired by [36].

Dataset	Context	$\langle k \rangle$	$\langle k_\Delta \rangle$	$\langle k \rangle^{\text{aug}}$	$\langle k_\Delta \rangle^{\text{aug}}$
InVS15	Workplace	16.9	7.0	21.0	7.0
SFHH	Conference	15.0	7.6	21.6	7.7
LH10	Hospital	19.1	17.1	25.7	17.5
Thiers13	High school	20.1	10.9	32.0	11.1

Table 3-A: Average generalized degree of the four real-world simplicial complexes constructed from the considered data sets (before and after the data augmentation).

not included in the final simplicial complex, their sub-cliques up to size 3 are considered in the counting. We then retained 20% of the simplices with the largest number of appearances. The thresholded simplicial complexes obtained in this way are those used in Fig. A.1 of Appendix A. Their connectivity properties are summarised in Table 3-A.

To reduce finite size effects, we augmented the thresholded simplicial complexes as follows: for each data set we extracted the list of sizes of the maximal simplices, also called facets, and the list of pure simplicial degrees of nodes. We then duplicated these lists five times and used the extended lists as input for the simplicial configuration model, described in Ref. [262]. The outputs of this procedure are simplicial complexes with the same statistical properties as the input complex but of significantly larger size. We used these augmented complexes as substrates for the simulations shown in Figure 3.4.

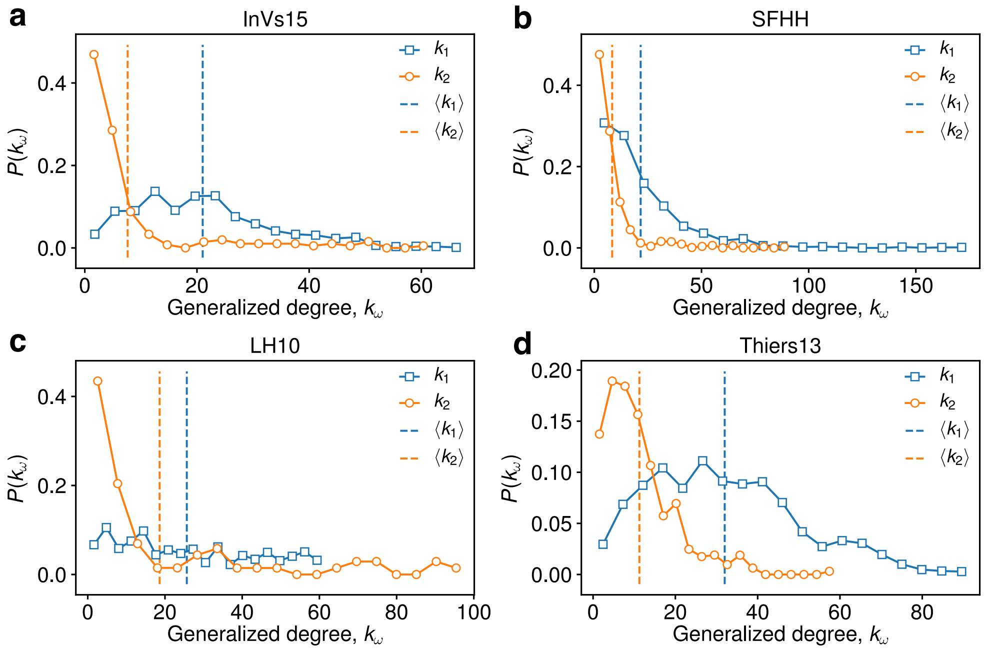


Figure 3.3: Generalised degree distributions of random simplicial complexes created from real-world data sets (see Sec. 3.3.1). The four panels correspond to different social contexts, namely (**a**) a workplace (InVS15), (**b**) a conference (SFHH), (**c**) a hospital (LH10) and (**d**) a high school (Thiers13). The generalised degrees k_1 and $k_2 = k_\Delta$ denote respectively the number of 1-simplices (blue) and 2-simplices (orange) incident in a node. The vertical dashed lines indicate the corresponding average values. Figure from [2].

3.3.2 Results on real-world simplicial complexes

We simulate the SCM over the simplicial complexes obtained from the four data sets as described in the previous Section. In particular, we start with an initial density ρ_0 of infectious nodes and we run the simulations by taking into consideration all the possible channels of infection illustrated in Figure 3.1(**c**-**h**). We stop a simulation if an absorbing state is reached, otherwise we compute the average stationary density of infectious nodes ρ^* by averaging the values measured in the last 100 time-steps after reaching a stationary state. The results are averaged over 120 runs obtained with randomly placed initial infectious nodes with the same density ρ_0. Moreover, the different data sets correspond to different densities of 1- and 2-simplices. We thus rescale the infectivity parameters β and β_Δ respectively by the average degree $\langle k \rangle$ and by the average number of 2-simplices incident on a node, $\langle k_\Delta \rangle$. We finally express all results as functions of the rescaled parameters $\lambda = \beta\langle k\rangle/\mu$ and $\lambda_\Delta = \beta_\Delta\langle k_\Delta\rangle/\mu$.

It is worth noticing that while this rescaling takes into account the different densities of simplices, it neglects the heterogeneity of contacts in the different dimensions. This is displayed in Fig. 3.3, where the generalized degree distributions show that for some data sets the the average number of contacts —per dimension— is not representative. This will have, as we will see, an impact on the position of the epidemic threshold.

Figure 3.4 shows the resulting prevalence curves for the four data sets In each panel (**b**,**d**,**f**,**h**), the average fraction of infected nodes ρ^* in the stationary state is plotted as a function of the rescaled infectivity $\lambda = \beta\langle k\rangle/\mu$ for simulations of the SCM with $\lambda_\Delta = 0.8$ (black triangles) and $\lambda_\Delta = 2$ (orange squares). For comparison, we also plot the case $\lambda_\Delta = 0$, which is equivalent to the standard SIS model with no higher-order effects (blue circles). We observe two radically different behaviours for the two values of $\lambda_\Delta \neq 0$. For $\lambda_\Delta = 0.8$, the density of infectious nodes varies as a function of λ in a very similar way to the case $\lambda_\Delta = 0$ (simple contagion), with a continuous transition. For $\lambda_\Delta = 2$ we observe instead the appearance of an endemic state with $\rho^* > 0$ at a value of λ^c well below the epidemic threshold of the other two cases. Furthermore, this transition appears to be discontinuous, and an hysteresis loop appears in a bi-stable region, where both healthy $\rho^* = 0$ and endemic $\rho^* > 0$ states can co-exist (dashed orange lines): in this parameter region, the final state depends on the initial density of infectious nodes ρ_0.

The simplicial complexes used in these simulations correspond to various social contexts and different densities of 1- and 2-simplices, and yield a similar phenomenology. These empirical structures however exhibit distributions of generalized degrees that are not well peaked around their average values (see Fig. 3.3), and do not allow us to systematically explore size effects.

To better understand the phenomenology of the simplicial contagion model, we thus now explore its behaviour on synthetic simplicial complexes with controlled properties.

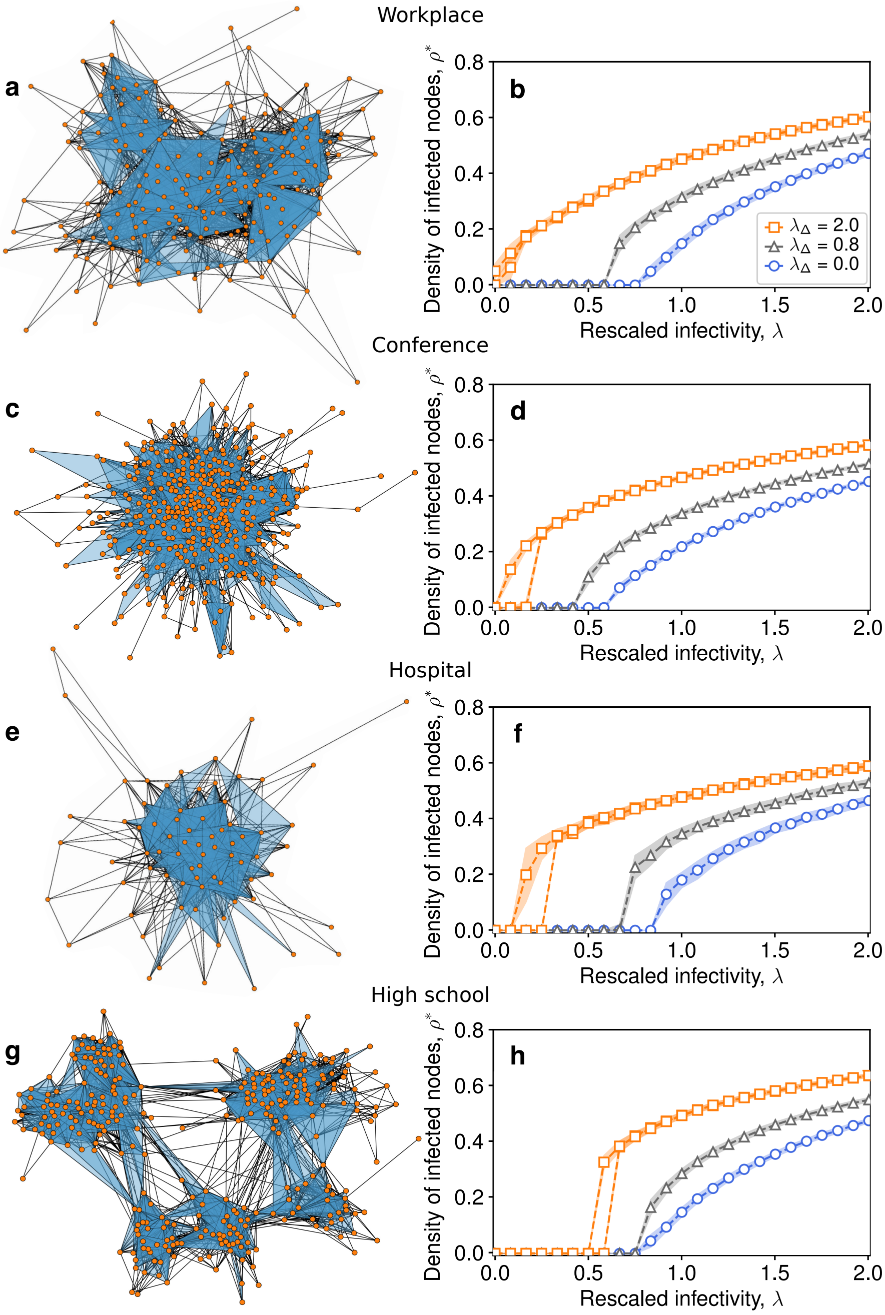


Figure 3.4: SCM of order $D = 2$ on real-world higher-order social structures. Simplicial complexes are constructed from high-resolution face-to-face contact data recorded in four different context: (**a**) a workplace, (**c**) a conference, (**e**) a hospital and (**g**) a high school. Prevalence curves are respectively reported in panels (**b**), (**d**), (**f**) and (**h**), in which the average fraction of infectious nodes obtained in the numerical simulations is plotted against the rescaled infectivity $\lambda = \beta\langle k\rangle/\mu$ for different values of the rescaled parameter $\lambda_\Delta = \beta_\Delta\langle k_\Delta\rangle/\mu$, namely $\lambda_\Delta = 0.8$ (black triangles) and $\lambda_\Delta = 2$ (orange squares). The blue circles denote the simulated curve for the equivalent standard SIS model ($\lambda_\Delta = 0$), which does not consider higher order effects. For $\lambda_\Delta = 2$ a bi-stable region appears, where healthy and endemic states co-exist. Figure from [2].

3.4 Simplicial contagion on synthetic simplicial complexes

A range of models for random simplicial complexes have been proposed so far, starting from the exponential random simplicial complex, the growing and generalized canonical ensemble [263–265] and the simplicial configuration models [262] to the more recent simplicial activity driven model [266] generalizing the activity driven temporal network model [267]. While these yield Erdös-Rényi-like models [268, 269] of arbitrary complexity, here we are interested in models generating simplicial complexes with simplices of different dimension in which we can control and tune the expected local connectivity, e.g. the number of edges and “full" triangles a node belongs to. We therefore propose, inspired by the models in Ref. [269] and Ref. [264], a new model to construct random simplicial complexes, the RSC model, which allows us to maintain the average degree of the nodes, $\langle k_1 \rangle$, fixed, while varying at the same time the expected number of “full" triangles (2-simplices) $\langle k_\Delta \rangle$ incident on a node.

3.4.1 Construction of random simplicial complexes

The Random Simplicial Complex (RSC) model of dimension D has $D + 1$ parameters, namely the number of vertices N and D probabilities $\{p_1, \dots, p_k, \dots, p_D\}$, $p_k \in [0, 1]$, which control for the creation of k-simplices up to dimension D. For the purpose of this study we limit the RSC model to $D = 2$, which restricts the set of required parameters to (N, p_1, p_2), but the procedure could easily be extended to larger D.

The model works as follows. Given a set $\mathcal{V}$ of N vertices, we first create 1-simplices (links) as in the Erdös-Rényi model [270], by connecting any two nodes $i, j \in \mathcal{V}$ of vertices with probability $p_1 \in [0, 1]$. The average degree, at this stage, is $(N - 1)p_1$. Similarly, 2-simplices are then created by connecting any triplet (i, j, k) of vertices. More precisely, we add a 2-simplex (i, j, k) with probability $p_\Delta \in [0, 1]$. Notice that simplicial complexes built in this way are radically different from the clique complexes obtained from Erdös-Rényi graphs[268], in which every subset of nodes forming a clique is automatically “promoted" to a simplex. Contrarily, in a simplicial complex generated by the RSC model proposed here, a 2-simplex (i, j, k) does not come from the promotion of an “empty" triangle composed by three 1-simplices (i, j), (j, k), (k, i) to a “full triangle" (i, j, k). This also means that the model allows for the presence of $(k + 1)$-cliques that are not considered k-simplices, therefore it is able generate simplicial complexes having both “empty" and “full" triangles, respectively encoding three 2-body interactions and one 3-body interactions [as for instance in Fig. 3.1(**b**)].

At this point each node has an average number $\langle k_\Delta \rangle = (N - 1)(N - 2)p_\Delta/2$ of incident 2-simplices that also contribute to increase the degree of the nodes. The exact contribution can be calculated by considering the different scenarios in which a 2-simplex (i, j, k) can be attached to a node i already having some links due to the first phase of the RSC construction. More precisely, the degree k_i of node i is incremented by 2 for each 2-simplex (i, j, k) such that neither the link (i, j) nor the link (i, k) are already present; this happens with probability $(1 - p_1)^2$. Analogously, if either the link (i, j) is already present but not (i, k), or vice-versa, the addition of the 2-simplex (i, j, k) increases the degree of i by 1. Since each case happens with the same probability $p_1(1 - p_1)$ the contribution is therefore $2p_1(1 - p_1)$. Overall, the degree k_i increases on average by $2(1 - p_1)$ for each 2-simplex attached to i.

Finally, for $p_1, p_\Delta \ll 1$, we can thus write the expected average degree $\langle k \rangle$ as the sum of the two contributions coming from the links and the 2-simplices, namely $\langle k \rangle \approx (N - 1)p_1 + 2\langle k_\Delta \rangle(1 - p_1)$. For any given size N, we can thus produce simplicial complexes having desired values of $\langle k \rangle$ and $\langle k_\Delta \rangle$ by fixing p_1 and p_Δ as:

$$p_1 = \frac{\langle k \rangle - 2\langle k_\Delta \rangle}{(N-1) - 2\langle k_\Delta \rangle} \tag{3.1a}$$

$$p_\Delta = \frac{2\langle k_\Delta \rangle}{(N-1)(N-2)} \, . \tag{3.1b}$$

With this procedure, for any given size N, we can produce simplicial complexes having desired values of $\langle k \rangle$ and $\langle k_\Delta \rangle$ by appropriately tuning p_1 and p_Δ.

Figure 3.5 reports the generalized degree distributions obtained with the RSC model

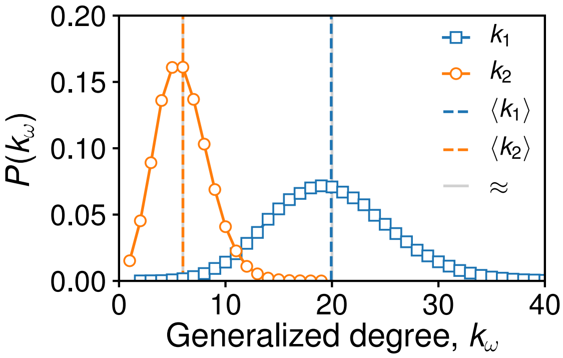


Figure 3.5: Generalised degree distributions of random simplicial complexes (RSC) generated by the model described in Sec. 3.4.1. The generalised degrees k_1 and $k_2 = k_\Delta$ denote respectively the number of 1-simplices (blue) and 2-simplices (orange) incident in a node. The vertical lines compare the average values of $\langle k_1 \rangle$ and $\langle k_2 \rangle$ obtained from multiple realizations of the model (coloured dashed lines) with the approximated values (continuous grey lines) calculated as described in the main text. Figure from [2].

just introduced. It is evident that, as for the classical Erdös-Rényi construction and in contrast to what observed for the complexes constructed from the real-world data sets (see Fig. 3.3), here both curves are well peaked around their average values. This also confirms the agreement between the expected values of $\langle k \rangle$ and $\langle k_\Delta \rangle$ (as given by Eq. (3.1) and depicted as continuous gray lines) with the empirical averages obtained from different realizations of the model (dashed coloured lines).

3.4.2 Results on random simplicial complexes

We simulate the SCM over a RSC created with the procedure described above, with $N = 2000$ nodes, $\langle k \rangle \simeq 20$ and $\langle k_\Delta \rangle \simeq 6$. As for the real-world simplicial complexes, we start with a seed of ρ_0 infectious nodes placed at random and we compute the average stationary density of infectious ρ^* by averaging over different runs, each one using a different instance of the RSC model. Results are shown in Fig.3.6(**a**), where the average fraction of infected nodes, as obtained by the simulations, is plotted as a function of the rescaled infectivity $\lambda = \beta\langle k \rangle$ for a $(D = 2)$ SCM with $\lambda_\Delta = 0.8$ (white squares), $\lambda_\Delta = 2.5$ (filled blue circles) and $\lambda_\Delta = 0$ (light blue circles).

Despite the very different properties of the underlying structure, the dynamics of

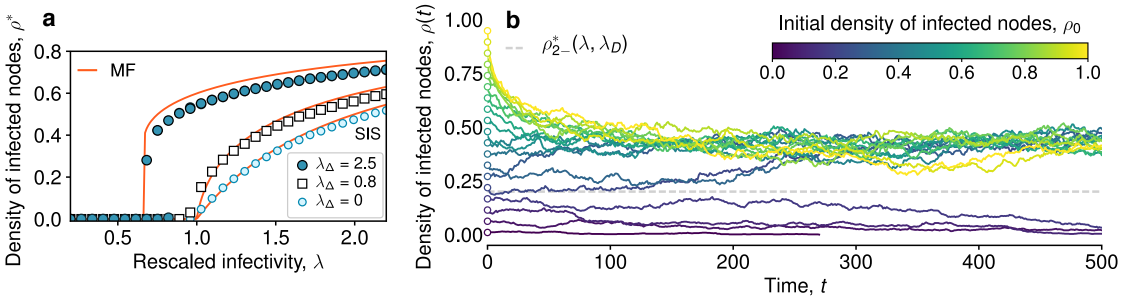


Figure 3.6: SCM of order $D = 2$ on a synthetic random simplicial complex (RSC). The RSC is generated with the procedure described in this manuscript, with parameters $N = 2000$, p_1 and p_Δ tuned in order to produce a simplicial complex with $\langle k \rangle \sim 20$ and $\langle k_\Delta \rangle \sim 6$. (**a**) The average fraction of infected obtained by means of numerical simulations is plotted against the rescaled infectivity $\lambda = \beta \langle k \rangle / \mu$ for $\lambda_\Delta = 0.8$ (white squares) and $\lambda_\Delta = 2.5$ (filled blue circles). The light blue circles give the numerical results for the standard SIS model ($\lambda_\Delta = 0$) that does not consider higher order effects. The red lines correspond to the analytical mean field solution described by Eq. (3.4). For $\lambda_\Delta = 2.5$ we observe a discontinuous transition with the formation of a bistable region where healthy and endemic states co-exist. (**b**) Effect of the initial density of infected nodes, shown by the temporal evolution of the densities of infectious nodes (a single realization is shown for each value of the initial density). The infectivity parameters are set within the range in which we observe a bistable region ($\lambda = \beta \langle k \rangle / \mu = 0.75$, $\lambda_\Delta = \beta_\Delta \langle k_\Delta \rangle / \mu = 2.5$). Different curves - and different colours - correspond to different values for the initial density of infectious nodes $\rho_0 \equiv \rho(0)$. The dashed horizontal line corresponds to the unstable branch ρ^*_{2-} of the mean field solution given by Eq. (3.5), which separates the two basins of attraction. Figure from [2].

the SCM on the RSC is very similar to the one observed on the real-world simplicial complexes. For $\lambda_\Delta = 0.8$ the model behaves similarly to a simple contagion model ($\lambda_\Delta = 0$), with a continuous transition at $\lambda^c = 1$, the well-know epidemic threshold of the standard SIS model on homogeneous networks. When a higher value of λ_Δ is considered ($\lambda_\Delta = 2.5$), the epidemic can be sustained below $\lambda^c = 1$, and both an epidemic-free and an endemic states are present in the region $\lambda^c < \lambda < 1$, with appearance of a hysteresis loop [see the filled blue circles in Fig. 3.6(**a**)]. In this region, we obtain $\rho(t \to \infty) = 0$ for $\rho(t = 0) = 0.01$, while $\rho(t \to \infty) > 0$ for $\rho(t = 0) = 0.4$.

The dependency from the initial conditions is also illustrated in Fig. 3.6(**b**), in which the temporal dynamics of single runs are shown. The various curves show how the density of infected nodes $\rho(t)$ evolves when initial seeds of infected nodes of different sizes are considered. Each colour corresponds to a different value of ρ_0, with brighter colours representing higher initial densities of infected individuals. The figure clearly shows the presence of a threshold value for ρ_0, such that $\rho(t)$ goes to the absorbing state $\rho(t) = 0$ if ρ_0 is smaller than the threshold, and to a non-trivial steady state if the initial density is above the threshold.

We now briefly investigate the size-dependence of the hysteresis loop. In particular, we check the size effects in the behaviour of the hysteresis by performing simulations of the SCM on systems of different sizes, namely $N = 500, 1000, 2000,$ and 4000, while keeping λ_Δ fixed within the region where we observe the bi-stability ($\lambda_D = 2.5$). As the Fig. 3.7 shows, we do not observe a significant variation of the dynamics when simplicial complexes of different sizes are considered, apart from a general stabilization of the incidence curves whose fluctuations tend to be smaller as the size increases. Further illustration of the finite size effects on the hysteresis loop can be found in Fig. A.2 of Appendix A.

3.5 Mean field approach

In order to study more extensively this phenomenology as λ_Δ and λ vary, and to further characterize the discontinuous transition, we consider a mean field (MF) description of the SCM, under a homogeneous mixing hypothesis [102]. Given the set of infection probabilities $B \equiv \{\beta_\omega, \omega = 1, \cdots, D\}$ and a recovery probability μ, we assume the independence between the states $x_i(t)$ and $x_j(t)$ $\forall$ $i, j \in \mathcal{V}$, and we write a MF expression for the temporal evolution of the density of infected nodes $\rho(t)$ as:

$$d_t\rho(t) = -\mu\rho(t) + \sum_{\omega=1}^{D} \beta_\omega \langle k_\omega \rangle \rho^\omega(t)\big[1 - \rho(t)\big] \tag{3.2}$$

where, for each $\omega = 1, \cdots, D$, $k_\omega(i) = k_{\omega,0}(i)$ is the generalized (simplicial) degree of a 0-dimensional face (node i), i. e., the number of ω-dimensional simplices incident to the node i [264, 265], and $\langle k_\omega \rangle$ is its average over all the nodes $i \in \mathcal{V}$. With this approximation we assume that the local connectivity of the nodes is well described by globally averaged

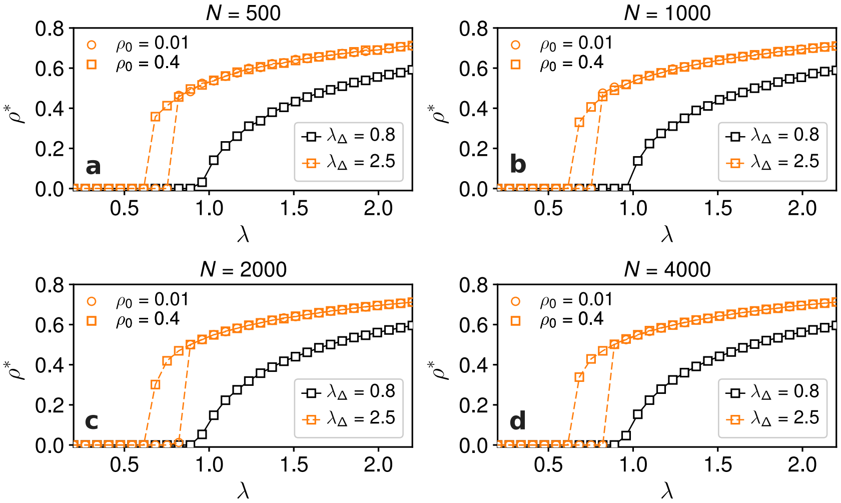


Figure 3.7: Numerical exploration of the finite size effects on the hysteresis for a SCM of order $D = 2$ on synthetic random simplicial complexes (RSC). The RSCs are generated with the procedure described in the main text, with parameters p_1 and p_Δ tuned in order to produce simplicial complexes with $\langle k \rangle \sim 20$ and $\langle k_\Delta \rangle \sim 6$. Different panels correspond to different system sizes, namely (**a**) $N = 500$, (**b**) $N = 1000$, (**c**) $N = 2000$, and (**d**) $N = 4000$. Each panel shows the average stationary fraction of infected individuals plotted against the rescaled infectivity $\lambda = \beta \langle k \rangle / \mu$. The parameter $\lambda_\Delta = \beta_\Delta \langle k_\Delta \rangle / \mu$ is set to $\lambda_\Delta = 2.5$, which corresponds to the case in which we observe a discontinuous transition, with the formation of a a bistable region where healthy and endemic states co-exist and a hysteresis appears. The two types of orange symbols correspond to two different values of the initial density of infected individuals for $\lambda_\Delta = 2.5$, namely $\rho_0 = 0.01$ (circles) and $\rho_0 = 0.4$ (squares). The case $\lambda_\Delta = 0.8$, in which we observe a continuous transition with no hysteresis, is shown for reference (black squares). Figure from [2].

properties, such as the average generalized degree. We can immediately check that in the case $D = 1$ we recover the standard MF equation for the SIS model, which leads to the well known stationary state solutions $\rho_1^{*[D=1]} = 0$ and $\rho_2^{*[D=1]} = 1 - \mu/(\beta\langle k \rangle)$. The absorbing state $\rho_1^{*[D=1]} = 0$ is the only solution for $\beta\langle k \rangle/\mu < 1$, i.e., below the epidemic threshold. When $\beta\langle k \rangle/\mu > 1$, this state becomes unstable while the solution $\rho_2^{*[D=1]}$ becomes a stable fixed point of the dynamics. The transition between these two regimes is continuous at $\beta\langle k \rangle/\mu = 1$.

3.5.1 Case $D = 2$

Let us now focus on a more interesting but still analytically tractable case in which we extend the contagion dynamics up to dimension $D = 2$, so that Eq. (3.2) reads:

$$d_t\rho(t) = -\mu\rho(t) + \beta\langle k\rangle\rho(t)\big[1 - \rho(t)\big] + \beta_\Delta\langle k_\Delta\rangle\rho^2(t)\big[1 - \rho(t)\big] \tag{3.3}$$

where $\langle k_\Delta\rangle \equiv \langle k_2\rangle$. By defining as before $\lambda = \beta\langle k\rangle/\mu$ and $\lambda_\Delta = \beta_\Delta\langle k_\Delta\rangle/\mu$, and by rescaling the time by μ, we can rewrite Eq. (3.3) as:

$$d_t\rho(t) = -\rho(t)(\rho(t) - \rho^*_{2+})(\rho(t) - \rho^*_{2-})\,, \tag{3.4}$$

where ρ^*_{2+} and ρ^*_{2-} are the solutions of the second order equation $1-\lambda(1-\rho)-\lambda_\Delta\rho(1-\rho) = 0$. We thus obtain:

$$\rho^*_{2\pm} = \frac{\lambda_\Delta - \lambda \pm \sqrt{(\lambda - \lambda_\Delta)^2 - 4\lambda_\Delta(1-\lambda)}}{2\lambda_\Delta}. \tag{3.5}$$

The steady state equation $d_t\rho(t) = 0$ has thus up to three solutions in the acceptable range $\rho \in [0, 1]$. The solution $\rho^*_1 = 0$ corresponds to the usual absorbing epidemic-free state, in which all the individuals recover and the spreading dies out. A careful analysis of the stability of this state and of the two other solutions ρ^*_{2+} and ρ^*_{2-} is however needed to fully characterize the phase diagram of the system.

Let us first consider the case $\lambda_\Delta \leq 1$. It is possible to show that ρ^*_{2-}, when it is real-valued, is always negative, i.e., it is not an acceptable solution. Moreover, ρ^*_{2+} is positive for $\lambda > 1$ and negative for $\lambda < 1$. In the regime $\lambda_\Delta \leq 1$ therefore, if $\lambda < 1$, the only acceptable solution to $d_t\rho(t) = 0$ is $\rho^*_1 = 0$; contrarily, for $\lambda > 1$, since $\rho^*_{2-} < 0$ and $\rho^*_{2+} > 0$, Eq. (3.4) shows that $d_t\rho(t)$ is positive at small $\rho(t)$: the absorbing state $\rho^*_1 = 0$ is thus unstable and the solution ρ^*_{2+} is stable. As $\rho^*_{2+} = 0$ for $\lambda = 1$, the transition at the epidemic threshold $\lambda = 1$ is continuous. In conclusion, when $\lambda_\Delta \leq 1$, the transition is similar to the one of the standard SIS model with $\lambda_\Delta = 0$.

Let us now consider the case of $\lambda_\Delta > 1$. Then, for $\lambda < \lambda^c = 2\sqrt{\lambda_\Delta} - \lambda_\Delta$, both ρ^*_{2+} and ρ^*_{2-} are outside the real domain, and the only steady state is the absorbing one $\rho^*_1 = 0$. Note that $\lambda^c < 1$, since $\lambda_\Delta > 1$. For $\lambda > \lambda^c$, we thus have two possibilities to consider. If $\lambda > 1$, we can show that $\rho^*_{2-} < 0 < \rho^*_{2+}$. Eq. (3.4) shows then that, for small $\rho(t)$, $d_t\rho(t) > 0$: as above, the absorbing state $\rho^*_1 = 0$ is unstable and the density of infectious nodes tends to ρ^*_{2+} in the large time limit; if instead $\lambda^c < \lambda < 1$, we obtain that $0 < \rho^*_{2-} < \rho^*_{2+}$. Then, still from Eq. (3.4), we obtain that $d_t\rho(t) < 0$ for $\rho(t)$ between 0 and ρ^*_{2-}, and that $d_t\rho(t) > 0$ for $\rho(t)$ between ρ^*_{2-} and ρ^*_{2+}. As a result, both $\rho^*_1 = 0$ and ρ^*_{2+} are stable steady states of the dynamics, while ρ^*_{2-} is an unstable solution. Most interestingly, the long time limit of the dynamics depends then on the initial conditions. Indeed, if the initial density of infectious nodes, $\rho(t = 0)$, is below ρ^*_{2-}, the short time derivative of $\rho(t)$ is negative, so that the density of infectious nodes decreases and the system tends to the absorbing state: $\rho(t) \xrightarrow[t\to\infty]{} 0$. On the other hand, if the initial density $\rho(t = 0)$ is large enough (namely, larger than ρ^*_{2-}), the dynamical evolution Eq. (3.4) pushes the density towards the value ρ^*_{2+}, i.e. $\rho(t) \xrightarrow[t\to\infty]{} \rho^*_{2+}$. Since $\rho^*_{2+} > 0$, the transition at λ_c is discontinuous.

We illustrate these results by showing in Fig. 3.8(**a**) the solutions ρ^*_1, ρ^*_{2+} and ρ^*_{2-} as a function of λ and for different values of λ_Δ. The vertical line corresponds to the standard epidemic threshold for the SIS model ($\lambda_\Delta = 0$). Dashed lines depict unstable branches, as given by ρ^*_{2-}.

We emphasize again two important points. First, for $\lambda_\Delta > 1$ we observe a discontinuous transition at

$$\lambda^c = 2\sqrt{\lambda_\Delta} - \lambda_\Delta, \tag{3.6}$$

instead of the usual continuous transition at the epidemic threshold. Second, for $\lambda^c < \lambda < 1$ the final state depends on the initial density of infectious nodes, as described above: the absorbing state $\rho^*_1 = 0$ is reached if the initial density $\rho(t = 0)$ is below the unstable steady state value ρ^*_{2-}; on the contrary, if $\rho(t = 0)$ is above this value, the system tends to a finite density of infectious nodes equal to ρ^*_{2+}. In other words, a critical mass

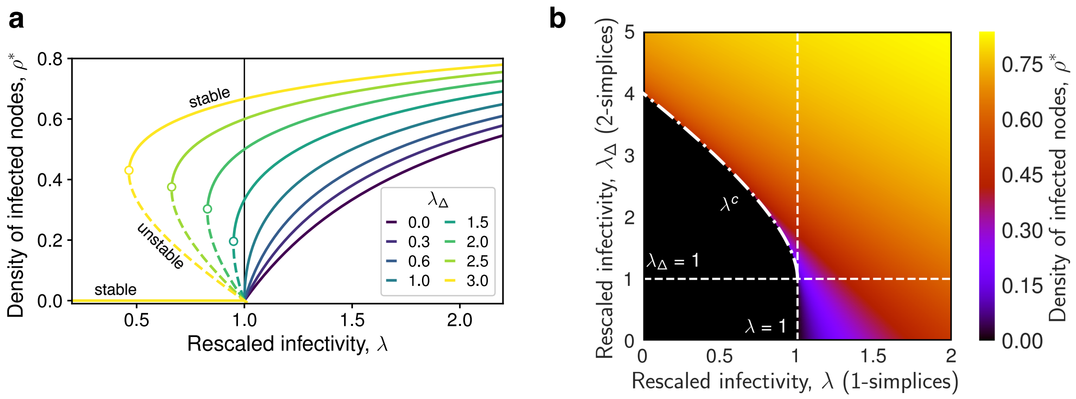


Figure 3.8: Phase diagram of the SCM of order $D = 2$ in mean field approximation. (**a**) The stationary solutions ρ^* given by Eq. (3.5) are plotted as a function of the rescaled link infectivity $\lambda = \beta\langle k\rangle/\mu$. Different curves correspond to different values of the triangle infectivity $\lambda_\Delta = \beta_\Delta\langle k_\Delta\rangle/\mu$. Continuous and dashed lines correspond to stable and unstable branches respectively, while the vertical line denotes the epidemic threshold $\lambda^c = 1$ in the standard SIS model that does not consider higher order effects. For $\lambda_\Delta \leq 1$ the higher order interactions only contribute to an increase in the density of infected individuals in the endemic state, while they leave the threshold unchanged. Conversely, when $\lambda_\Delta > 1$ we observe a shift of the epidemic threshold, and a change in the type of transition, which becomes discontinuous. (**b**) Heatmap of the stationary solution ρ^* given by Eq. (3.5) as a function of the rescaled infectivities $\lambda = \beta\langle k\rangle/\mu$ and $\lambda_\Delta = \beta_\Delta\langle k_\Delta\rangle/\mu$. The black area corresponds to the values of $(\lambda, \lambda_\Delta)$ such that the only stable solution is $\rho_1^* = 0$. The dashed vertical line corresponds to $\lambda = 1$, the epidemic threshold of the standard SIS model without higher order effects. The dash-dotted line represents the points $(\lambda^c, \lambda_\Delta)$, with $\lambda^c = 2\sqrt{\lambda_\Delta} - \lambda_\Delta$, where the system undergoes a discontinuous transition. Figure from [2].

is needed to reach the endemic state, reminding of the recently observed minimal size of committed minorities required to initiate social changes [253].

Figure 3.8(**b**) is a two-dimensional phase diagram showing ρ_{2+}^* for different values of λ and λ_Δ. Lighter colours correspond to higher values of the stationary density of infectious nodes, while the dashed vertical line corresponds to the epidemic threshold of the standard (without higher order effects) SIS model, namely $\lambda = 1$. For $\lambda_\Delta \leq 1$ (below the dashed horizontal line) the transition as λ crosses 1 is seen to be continuous, while, for $\lambda_\Delta > 1$, the transition is clearly discontinuous along the curve $\lambda^c = 2\sqrt{\lambda_\Delta} - \lambda_\Delta$ (dash-dotted line). The analytical values of ρ_{2+}^* are also reported as continuous red lines in Fig. 3.6(**a**) and compared to the results of the simulations, showing in this way the accuracy of the mean field approach just described. In addition, Fig. 3.6(**b**) shows that the unstable solution ρ^*_{2-} accurately separates the two basins of attractions for the dynamics, i.e., it defines the critical initial density of infected ρ_0 that determines whether the long term dynamics reaches the healthy state or the endemic one. Notice that the mean field approach is in fact able to correctly capture both the position of the thresholds and the discontinuous nature of the transition for the SCM with $\lambda_\Delta > 1$.

We finally note that, while a general solution for general D with arbitrary parameters $\{\beta_\omega\}$ remains out of reach, it is possible to show that the phenomenology obtained for $D = 2$ is also observed for specific cases with $D \geq 3$. We consider indeed in the next two Sections 3.5.2 and 3.5.3 two cases: $D = 3$ with $\beta_2 = 0$ and general $D > 3$ with $\beta_1 = \cdots = \beta_{D-1} = 0$. In both cases, we will show the appearance of a discontinuous transition in the regime where the simple contagion β_1 is below threshold (i.e., $\beta_1 \langle k \rangle < \mu$): similarly to the case $D = 2$, this transition occurs as β_D, which describes the rate of the high-order contagion process, increases.

3.5.2 Case $D = 3$

Let us consider here a system with maximum dimension of simplices $D = 3$. In this case the model has three spreading parameters $\beta_1, \beta_2 = \beta_\Delta$ and β_3, and the evolution equation for $\rho(t)$ reads

$$d_t\rho(t) = -\mu\rho(t) + \beta\langle k \rangle \rho(t)(1-\rho(t)) + \beta_2\langle k_2 \rangle \rho(t)^2(1-\rho(t)) + \beta_3\langle k_3 \rangle \rho(t)^3(1-\rho(t)). \quad (3.7)$$

Finding the roots of $d_t\rho(t) = 0$ yields a polynomial of degree 3, so it is possible to write these roots, corresponding to stable and unstable fixed points of the dynamics, as functions of the parameters of the model. The process is however lengthy and cumbersome, and depends moreover on three parameters, so that the representation of the whole phase diagram is not convenient.

As we want here simply to show that the phenomenology of the appearance of first order transitions obtained in the case $D = 2$, is also observed in higher dimensions, we restrict ourselves for simplicity to the case $\beta_\Delta = 0$, in which we will see that we can avoid writing the explicit solutions and resort instead to a graphical solution. This case corresponds to the hypothesis that contagion can occur only either through simple contagion or through cliques of size 4 in which 3 of the nodes are already infectious, and the evolution equation reduces to:

$$d_t\rho(t) = -\mu\rho(t) + \beta\langle k\rangle\rho(t)(1-\rho(t)) + \beta_3\langle k_3\rangle\rho(t)^3(1-\rho(t)). \tag{3.8}$$

Setting $\lambda = \beta\langle k\rangle/\mu$, $\lambda_3 = \beta_3\langle k_3\rangle/\mu$ and rescaling time by μ we obtain:

$$d_t\rho(t) = \rho(t)(1-\rho(t))\left(\lambda + \lambda_3\rho^2 - \frac{1}{1-\rho(t)}\right) \tag{3.9}$$

where we can define the functions $f_1(\rho) = \lambda + \lambda_3\rho^2$ and $f_2(\rho) = 1/(1-\rho)$. The sign of the temporal evolution of the density of infectious is thus given by the sign of the difference between $f_1 - f_2$. Note that $\rho(t)$ is by definition between 0 and 1 so we need to consider f_1 and f_2 only between these limits. In this interval, f_1 is positive and increases monotonically from λ for $\rho = 0$ to $\lambda + \lambda_3$ for $\rho = 1$. Function f_2 is also positive and strictly increasing, with $f_2(0) = 1$ and f_2 diverging towards $+\infty$ as $\rho \longrightarrow 1^-$. We also note that the equation $f_1(\rho) = f_2(\rho)$ yields a polynomial of degree 3, so it has at most 3 real roots.

Let us first consider the case $\lambda > 1$. Then at $\rho = 0$ we have $f_1 > f_2$, and as $\rho \longrightarrow 1$, f_1 becomes smaller than f_2. Therefore, at small ρ, $d_t\rho$ is positive and hence the state $\rho = 0$ is unstable. More in detail, there are two possibilities:

- either there is one single crossing point of f_1 and f_2, at ρ^*. Then, $d_t\rho(t) > 0$ if $\rho(t) < \rho^*$ and $d_t\rho(t) < 0$ if $\rho(t) > \rho^*$: for any $\rho(t = 0) > 0$, the system goes to the stationary state $\rho(t \longrightarrow \infty) = \rho^*$. This is similar to the usual SIS case with $\lambda_3 = 0$: the effect of a non-zero value of λ_3 is simply to shift the value of ρ^*.

- or there are three crossing points $\rho_1 < \rho_2 < \rho_3$. This occurs for certain combinations of values of λ and λ_3. Then for $\rho(t) < \rho_1$, $d_t\rho(t) > 0$ so the absorbing state $\rho = 0$ is again unstable. The state ρ_2 is also seen to be unstable while there are two stable fixed points ρ_1 and ρ_3: depending on the value of $\rho(t = 0)$, the system will converge to one of these values.

Hence, for $\lambda > 1$, the system always reaches a stationary state with a finite fraction of infectious nodes, which in some regions of the (λ,λ_3) phase diagram, can depend on $\rho(t = 0)$.

Let us now consider the more interesting case $\lambda < 1$. Then $f_1(\rho) < f_2(\rho)$ both for $\rho = 0$ and as $\rho \longrightarrow 1$. Hence $f_1 - f_2$ is negative both in 0 and 1, and either 0 or 2 of the roots of the equation $f_1(\rho) = f_2(\rho)$ are between 0 and 1. Hence, for $\rho \in [0, 1]$, either f_1 is always below f_2, or the two functions intersect in 2 points that we call ρ_- and ρ_+ ($\rho_- < \rho_+$):

- in the former case ($f_1(\rho) < f_2(\rho)\ \forall\rho \in [0, 1]$), $d_t\rho(t)$ is always negative so the only stationary state is the absorbing one $\rho = 0$;
- in the latter case, $d_t\rho$ is positive for $\rho(t)$ between ρ_- and ρ_+ and negative else, so that
  - if $\rho(t = 0) < \rho_-$, $d_t\rho$ is negative, hence $\rho(t)$ decreases and the system converges to $\rho = 0$
  - if $\rho(t = 0) > \rho_-$, the system converges towards $\rho(t \longrightarrow \infty) = \rho_+ > 0$.

At fixed $\lambda < 1$, the former case is obtained at small values of λ_3, while the latter is obtained for λ_3 large enough. The situation is illustrated in Fig. 3.9 for $\lambda = 0.5$. At the transition $\lambda_3 = \lambda_3^c$ between these two cases, $\rho_- = \rho_+ > 0$ (the functions f_1 and f_2 are tangent in this point): the transition from $\rho(t \longrightarrow \infty) = 0$ for $\lambda_3 < \lambda_3^c$ to $\rho(t \longrightarrow \infty) = \rho_+$ (if $\rho(t = 0) > \rho_-$) for $\lambda_3 > \lambda_3^c$ is thus a discontinuous one, in a similar way to the case $D = 2$ discussed in the main text.

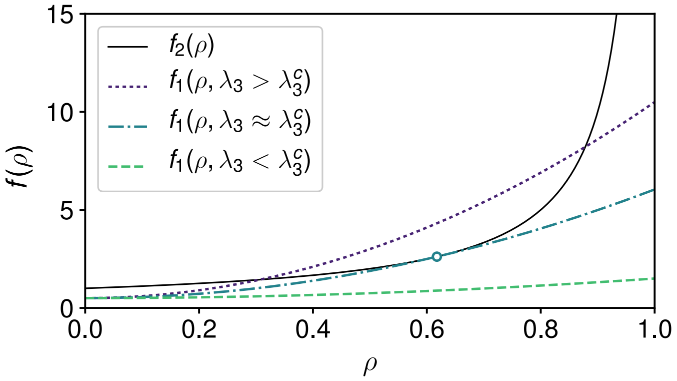


Figure 3.9: SCM of order $D = 3$, case $\lambda = 0.5$, $\lambda_2 = 0$: $f_1(\rho)$ for various λ_3 ($<$, $\approx$ and $>$ λ_3^c), and $f_2(\rho)$. f_1 is below f_2 both at $\rho = 0$ and as $\rho \longrightarrow 1$. The two curves therefore either do not cross (for $\lambda_3 < \lambda_3^c$), are tangent in $\rho_+ = \rho_-$ (for $\lambda_3 = \lambda_3^c$) or cross in two points ρ_- and ρ_+ (for $\lambda_3 > \lambda_3^c$). Figure from [2].

3.5.3 General D, with $\beta_1 = \cdots = \beta_{D-1} = 0$

For general D, there is no analytical solution for the stationary values of the density of infectious nodes. We show here however that, if we consider that contagion can occur only through cliques of size $D+1$, i.e., if all spreading rates $\beta_1, \beta_2, \ldots, \beta_{D-1}$ are null, there exists a discontinuous transition between the phase in which the spreading vanishes at low β_D and the phase in which $\rho(t \longrightarrow \infty)$ is finite at large β_D.

The evolution equation for ρ reads

$$d_t\rho(t) = -\mu\rho(t) + \beta_D\langle k_D\rangle\rho(t)^D(1-\rho(t)). \tag{3.10}$$

Defining $\lambda_D = \beta_D\langle k_D\rangle/\mu$ and rescaling time by μ we obtain

$$d_t\rho(t) = -\,\rho(t)\Big[1 - \lambda_D\rho^{D-1}(t)(1-\rho(t))\Big]. \tag{3.11}$$

Defining $F_D(\rho) = 1 - \lambda_D\rho^{D-1}(1-\rho)$, we see that the sign of $d_t\rho(t)$ is opposite to the sign of $F_D(\rho(t))$, so that we need to study the sign of the function $F_D(\rho)$ for $\rho \in [0,1]$ (as the density $\rho(t)$ is by definition between 0 and 1).

We have $F_D(0) = F_D(1) = 1$. Moreover, the derivative of F_D is

$$F'_D(\rho) = \lambda_D(D\rho^{D-1} - (D-1)\rho^{D-2}) = D\lambda_D\rho^{D-2}(\rho - (1-1/D)).$$

It is thus negative for $\rho < 1 - 1/D$ and positive for $\rho > 1 - 1/D$: F_D first decreases as ρ increases, reaches a minimum at $\rho = 1 - 1/D$ and then increases back to 1 as ρ increases to 1. We have thus two cases:

- if the minimum, $F_D(1 - 1/D)$, is positive, then $F_D(\rho) > 0$ for $\rho \in [0, 1]$: therefore, $d_t\rho(t)$ is always negative for any $\rho(t) > 0$: the density of infectious nodes can only decrease and the contagion-free state $\rho = 0$ is the only stable state.
- if instead $F_D(1 - 1/D) < 0$, then, as $F_D(0) = F_D(1) = 1$, by continuity the equation $F_D(\rho) = 0$ has two roots in $[0, 1]$, which we call ρ_- and ρ_+ ($\rho_- < \rho_+$). $F_D(\rho)$ is positive for $\rho \in [0, \rho_-)$ and $\rho \in (\rho_+, 1]$ and negative between the two roots. Therefore
  - if $\rho(t = 0) < \rho_-$, $d_t\rho(t = 0)$ is negative, hence $\rho(t)$ decreases and the system converges to $\rho = 0$
  - if $\rho(t = 0) > \rho_-$, the system converges towards $\rho(t \to \infty) = \rho_+ > 0$.

  The condition to have $F_D(1 - 1/D) < 0$ and hence a non-trivial stationary state can be written simply as

  $$1 - \lambda_D(1 - 1/D)^{D-1}(1/D) < 0$$

  i.e.,

  $$\lambda_D > \lambda_D^c = \frac{D^D}{(D-1)^{D-1}}.$$

  Note that for $\lambda_D = \lambda_D^c$, $\rho_- = \rho_+ = 1 - 1/D$ is strictly positive, showing that the transition at λ_D^c is discontinuous.

This shows therefore that for $\beta_1 = \cdots = \beta_{D-1} = 0$, we have the same phenomenology for any D as for the case $D = 2$ studied in Sec. 3.5.1: a discontinuous transition occurs at

$$\lambda_D^c = \frac{D^D}{(D-1)^{D-1}} \tag{3.12}$$

between an absorbing state $\rho = 0$ and a stationary state with a non-zero density of infectious individuals $\rho_+ > 0$.

3.6 Microscopic Markov-chain approach

Following the Microscopic Markov chain approach (MMCA) [235] we can write the probability $p_i(t) \equiv \text{Prob}\big[x_i(t) = 1\big]$ for any node i being infected at time t as a function of $p_i(t-1)$:

$$p_i(t) = \big[1 - q_i(t-1)\big]\big[1 - p_i(t-1)\big] + \big[1 - \mu\big]p_i(t-1) \tag{3.13}$$

where $q_i(t)$ denotes the probability of node $i \in \sigma$ not being infected by any of the subfaces of σ. Such a quantity can be written in terms of the parameters in $B = \{\beta_1, \beta_2, \ldots, \beta_D\}$ and of the states of the faces (links, filled triangles, etc) in which node i is involved. Considering for simplicity only contributions up to $D = 2$ we have

$$q_i(t) = \prod_{j \in \mathcal{V}} \Big[1 - \beta a_{ij} p_j(t-1)\Big] \prod_{j,l \in \mathcal{V}} \Big[1 - \beta_\Delta a_{ijl} p_j(t-1) p_l(t-1)\Big] \tag{3.14}$$

The first term of Eq. (3.14) accounts for the contagion through the links of $\mathcal{K}$. These links are fully specified by means of the standard adjacency matrix $\{a_{ij}\}$, whose elements $a_{ij} = 0, 1$ denote the absence or presence of a link (i, j). Similarly, the second term of Eq. (3.14) accounts for the contagion of i through the 2-simplices of $\mathcal{K}$ (triangles), which are analogously specified by the elements of the adjacency tensor $\{a_{ijl}\}$. This tensor is the 3-dimensional version of the adjacency matrix, in which a non-zero element (ijl) denotes the presence of a 2-simplex (i, j, l).

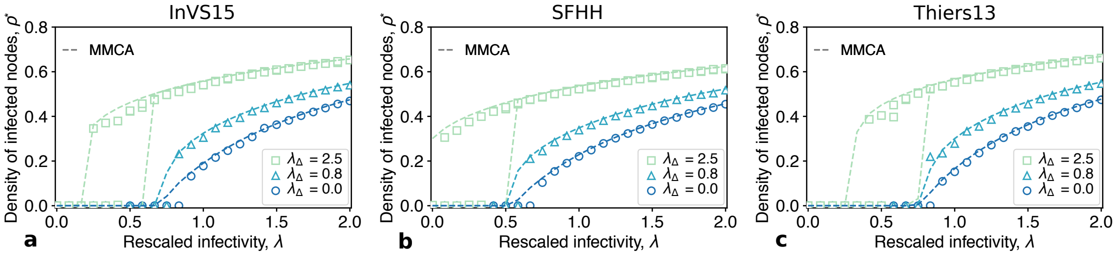


Figure 3.10: SCM of order $D = 2$ on real-world simplicial complexes constructed from high-resolution face-to-face contact data recorded in four different context: (**a**) a workplace, (**b**) a conference, and (**c**) a high school. The average densities of infectious nodes ρ^* are plotted against the rescaled infectivity $\lambda = \beta \langle k \rangle / \mu$ for different values of the rescaled parameter $\lambda_\Delta = \beta_\Delta \langle k_\Delta \rangle / \mu$. Results from numerical simulations (symbols) are compared to analytical results form the MMCA (dashed lines), as given by Eq. (3.13).

We can quickly test the accuracy of the MMCA by comparing the results of the numerical integration of Eq. (3.13) against simulations. We do this on the real-world simplicial complexes we constructed in Sec. 3.3.1 before the data augmentation. This allows us to check at the same time also the behaviour of the model and the accuracy of the analytical approach on systems of small size. Results are shown in Fig. 3.10 for the data sets referring to (**a**) a workplace (InVS15) [258], (**b**) a conference (SFHH) [259], and (**c**) a high school (Thiers13) [261]. In each panel, the average densities of infected nodes is plotted, as usual, against the rescaled infectivity λ. While symbols represent mean values coming from different simulated realizations, the dashed lines refer to the results of the MMCA. Differently from the MF, here the good match between the curves confirms the validity of the MMCA for simplicial contagions on heterogeneous structures, such as the ones considered. The results also hold in the case of the conference data set [Fig. 3.10(**b**)] in which the epidemic threshold vanishes for sufficiently high values of the 2-simplex infectivity ($\lambda_D = 2.5$).

3.7 Summary and conclusions

In summary, the simplicial model of contagion introduced here is able to capture the basic mechanisms and effects of higher-order interactions in social contagion processes. Our analytical results were derived in a mean field approximation and indeed quantitatively compared to the nondescript simplicial complexes obtained in our random simplicial complex model (akin to ER simplicial complexes [269]). However, the framework we introduced is very general and the phenomenology robust, as seen from the results obtained on empirical data sets.

In this Chapter, we focused our attention on the behaviour of the SCM on simplicial complexes. It would be interesting to further investigate the SCM on more general simplicial complexes with, for instance, heterogeneous generalized degree distribution or with community structures. Previous results on spreading dynamics on networks have already showed the impact that the presence of clusters, communities and sub-graphs might have on the epidemic threshold and on the final epidemic size [102, 271–276]. Another interesting direction would be to consider simplicial complexes with emergent properties such as hyperbolic geometry[277–279], or temporally evolving simplicial complexes [266].

Notice that some investigations are already underway. Indeed, following our suggestion [2], further efforts have been already made by other researchers in the field in order to understand the impact of the representation of the social structure on the dynamics of the model.
More precisely, Jhun et al. [137] extended the SCM to the more general case of hypergraphs. The model works exactly as the one proposed here, but this time the spreading process takes place on top of d-uniform hypergraphs in which all the hyperedges have the same size d. As for the simplicial version, a susceptible node that is part of an hyperedge α of size d can get an infection from α, with rate β_d, only if the remaining $d-1$ nodes composing α are infectious. The standard recovery probability μ is used. The authors considered the case of scale-free (SF) uniform hypergraphs. Notice that even if all the hyperedges have the same size, different nodes can belong to a different number of these hyperedges. In this sense, the heterogeneity is given by the number of hyperedges a node belongs to, which is distributed as $\sim P(k)^{\gamma}$. The heterogeneous mean-field formalism (HMF) —in which nodes of the same hyper-degree class as considered equivalent [280]—

leads to the following equation for the evolution of the stationary density of infected nodes of hyperdegree k:

$$d_t \rho_k = -\mu \rho_k + \beta_k (1 - \rho_k) k \Theta^{d-1} \tag{3.15}$$

The contagion term on the r.h.s. considers the probability that a susceptible node of hyperdegree k gets the infection from one of the hyperedges. This is, as usual, proportional to the infection rate β_k, the number of hyperedges k, and the probability Θ^{d-1} to be connected to an hyperedge having all the other nodes infected.
A comparison of Eq. (3.15) with Eq. (3.3) highlights the difference in the representation used. Indeed, differently from the simplicial case, here the contagion term does not dependent on the lower order sub-faces.

In this case, the system presents a characteristic exponent $\gamma_c = 2 + 1/(d-2)$ of the degree distribution that determines the nature of the transition. In particular, for $\gamma < \gamma_c$ the epidemic threshold vanishes ($\lambda_c = 0$). By contrast, if $\gamma = \gamma_c$ a second order transition appears, that becomes hybrid when higher values of γ are considered. The associated values of the susceptibility diverge at the transition point, as expected [281, 282]. These results are consistent with simulations on SF uniform hypergraphs, confirming the validity of the HMF approach on such topologies.

A different version of the higher-order social contagion model on hypergraph was recently proposed by de Arruda et al. [256]. Based on a similar SIS framework, the fundamental difference with respect to the other models relies in the explicit inclusion of a critical-mass dynamics into the contagion process that generalizes the SCM [2]. In the SCM, a susceptible node i part of a hyperedge α (or a simplex) of size d could get the infection from α only if all the remaining $d-1$ nodes composing it are infected. Here, the authors relax the constraints by *i*) moving from simplicial complexes to hypergraphs and *ii*) allowing an hyperedge α to be potentially infectious for $i \in \alpha$ if the number of infected nodes composing α is greater or equal to a given threshold Θ_α. The standard

SIS model is then recovered by restricting this threshold mechanism to hyperedges of size greater than two, so that a contagion through active links can always happen (no threshold). This model reveals a similar phenomenology to the one on simplices, characterized by the appearance of first and second-order transitions and hysteresis. In addition, the authors provide further analytical results on regular hypergraphs, namely an hyper-blob (a random regular network with one hyperedge containing all the nodes) and an hyper-star (a star network with one hyperedge containing all the nodes). The critical values analysis is further extended with the introduction of the concept of a "social latent" heat, interpreted as the fraction of individuals to add or remove in order to move the dynamics from one solution to the other. These findings provide a possible phenomenological explanation to some apparently contradictory results previously obtained. In fact, experimental work has showed different values of critical mass levels needed to initiate a social change, i.e., to revert an existing equilibrium to a new one by mean of a committed minority [253, 283, 284]. These threshold values, spanning from 10% to 40%, could be consistently seen as the effect of the interplay between a global critical mass and the local thresholds as given by the Θ_α, which also depend on the size of the interacting group.

Finally, further developments of the SCM based on probabilistic descriptions have showed that more complex analytical formulations, namely the microscopic Markov-chain approach [235] and the link equation [285], can improve the accuracy of the predictions [286]. Indeed, differently from the MF, these approaches can be used to analytically describe the contagion dynamics on high-order heterogeneous structures.

Part II

Modelling Discovery Dynamics

Chapter 4

Discovery processes on networks

4.1 Introduction

Creativity and innovation are the underlying forces driving the growth of our society and economy. Studying creative processes and understanding how new ideas emerge and how novelties can trigger further discoveries is therefore fundamental if we want to devise effective interventions to nurture the success and sustainable growth of our society. As already described in Chapter 1, many models have been developed in this direction. In particular, the authors of Refs. [82, 189, 199, 287] have looked at different types of temporally ordered sequences of data, such as sequences of words, songs, Wikipages and tags to study how the number $D(t)$ of novelties grows with the length of the sequence t. They have found that the Heaps' law, i.e. a power-law behaviour $D(t) \sim t^{\beta}$ originally introduced to describe the number of distinct words in a text document [190], applies to different contexts, producing different values of $\beta < 1$. Some examples are reported in Fig. 4.1 for five different real-world data sets. From the left to the right, each Heaps' law represents the growth of (i) the number of distinct songs listened by users on the online platform last.fm, (ii) the number of distinct hashtags tweeted by users on the popular social network Twitter, (iii) the number of different projects to which developers contributed to on the online hosting service GItHub, (iv) the number of distinct words

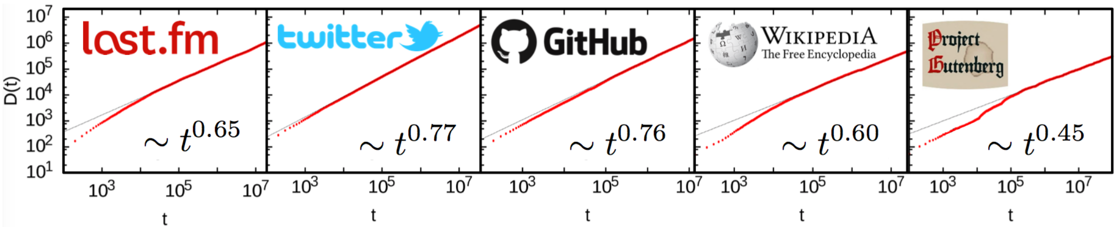


Figure 4.1: Empirical Heaps' laws from five different real-world data sets: Last.fm, Twitter, Github, Wikipedia, and Gutenberg. Red points represent the respective growth of the number of distinct elements $D(t)$ as a function of the total number of elements t. Black lines denote the associate power-law fits (exponent displayed). Figure adapted from [189].

contained on several web pages of the English Wikipedia, and (v) the number of distinct words contained in the texts of prose and poetry of the Gutenberg Project [82, 189]. Notice how these processes are radically different. Some of them are pure exploration processes, since users of last.fm are consumers of songs that are created by artists. Contrarily, hashtags are created by users and then absorbed into the system and eventually used by others.

In Section 1.2.2, we have introduced the urn model, a useful framework to study discovery and innovation processes in evolutionary biology, chemistry, sociology, economy and text analysis [288, 289]. Let us recall that in the classic Pólya urn model [197, 198], a temporal sequence of discoveries can be generated by drawing balls from an urn that contains all possible discoveries. An interesting development is the recent model by Tria and co-workers [82, 200], which adds the concept of the AP [174, 290] to the reinforcement mechanism which is already present in the original Pólya's urn framework. This model, called Urn Model with Triggering, well reproduces the empirical signatures of discovery processes, such as the Heaps' law (as it has been analytically shown in Sec. 1.2.2.1), the Zipf's law, and the semantic correlations proper of real-world systems (more details will be given in the next sections).

In this Chapter, we propose to model the dynamical mechanisms leading to discoveries and innovations as an edge-reinforced random walk (ERRW) on an underlying network of relations among concepts and ideas. It is easy to see how the network representation of the space of items naturally accounts for the AP, since paths are restricted to existing connections, and the discovery (visit) of a given node could provide access to a different set of nodes not directly accessible before.The use of networks as an underlying structure for search strategies and navigation is strictly linked to the literature in random walks and optimal foraging [207], but in lately it has been applied to a various contexts. For example, in cognitive sciences, networks have been used at length to encode the patterns behind mental representations [291–294]. Then, as for contagions, understanding the influence of these structures on the process of discovery that unfolds on top remains a fascinating problem.

Random walks on complex networks [16, 18, 21, 23, 202] have been studied at length [207]. In similar contexts, they have been used to build exploration models for social annotation [168], music album popularity [295], knowledge acquisition [170, 171], animal foraging and migration [296, 297], human mobility [298], information processing [299], human language complexity [300, 301] and evolution in research interests [302]. A special class of random walks are those with reinforcement [195, 303, 304], which have been successfully applied to biology [305] and also in human mobility [306, 307]. In particular, the concept of edge reinforcement [308–310] was introduced in the mathematical literature by Coppersmith and Diaconis [311] as a model of a person exploring a new town. Here, we will use ERWWs to mimic how different concepts are explored moving from a concept to an adjacent one in the network, with innovations being represented, in this framework, by the first discovery of nodes [312]. As supported by empirical observations, we expect indeed the walkers to move more frequently among already known concepts and, from time to time, to discover new nodes. For this reason, we introduce and study a model in which the network is co-evolving with the dynamical process taking place over it [1]. In our model, (*i*) random walkers move over a network with assigned topology and whose edge weights represent the strength of concept associations, and (*ii*) the network evolves in time through a reinforcement mechanism in which the weight of an edge is increased every time the edge is traversed by a walker, making traversed edges more likely to be traversed again.

4.1.1 Outline

This Chapter is structured as follows:

In Sec. 4.2, we introduce the *edge-reinforced random walk* (ERRW) model, explaining in particular the co-evolution of the network with the dynamics of the walker by means of the *reinforcement* mechanism.

In Sec. 4.3, we first test the model on synthetic SW networks and observe the natural emergence of a Heaps' law, characterizing the pace of innovation [190], with only two ingredients, namely the topology of the network and a parameter describing the strength of the reinforcement. We then show how, by tuning the amount of reinforcement, the model can give rise to different scaling exponents.

In Sec. 4.4 we consider a more realistic scenario, namely the growth of knowledge in modern science as tracked by a large database of scientific publications. We study the dynamics of these real innovation processes by extracting the empirical network behind them and by running our model on top of it. In such cases, the framework we propose is even simpler and easier to interpret, since the topology comes directly from the data, and the model has only one parameter.

In Sec. 4.5, we investigate the correlations in the temporal sequences of visited concepts produced by the model and compare them to appropriate null models. The correlations, in agreement with the ones of empirical trajectories, will appear as a natural consequence of the interplay between the network topology and the reinforcement mechanism that controls the exploration dynamics.

Conclusions and future perspectives are summarized in Sec. 4.6.

4.2 The edge-reinforced random walk model

Let us consider a random walker over a weighted connected graph $G(\mathcal{V},\mathcal{E})$, where $\mathcal{V}$ and $\mathcal{E}$ are, respectively, a set of $N = |\mathcal{V}|$ nodes and a set of $K = |\mathcal{E}|$ links. Each node of the graph represents a concept or an idea, and the presence of a link (i, j) denotes the existence of a direct relation between two concepts i and j. The values of N and K and the topology of the network are assumed to be fixed, while the weights of the edges can change in time according to the dynamics of the walker, which, as we will see below, is in turn influenced by the underlying network. The graph at time t, with $t = 0, 1, 2, \ldots$, is fully described by the non-negative time-dependent adjacency matrix $W^t \equiv \{w^t_{ij}\}$, where the value w^t_{ij} is different from 0 if the two concepts i and j are related, and quantifies the strength of the relationship at time t. We initialize the network assuming that at time $t = 0$ all the edges have the same weight, namely $w^0_{ij} = 1 \; \forall (i, j) \in \mathcal{E}$. The dynamics of the walkers is defined as follows: at each time step t, a walker at node i jumps to a randomly chosen neighbouring node j with a probability proportional to the weight of the connecting edge. Formally, the probability of going from node i to node j at time t is:

$$\mathrm{Prob}^t(i \longrightarrow j) = \pi^t_{ji} = \frac{w^t_{ij}}{\sum_l w^t_{il}} \tag{4.1}$$

where the time-dependent transition probability matrix $\Pi^t \equiv \{\pi^t_{ij}\}$ depends on the weights of all links at time t [313]. The transition probabilities satisfy the normalization $\sum_j \pi^t_{ji} = 1 \; \forall \; i, t$, and we assume that G has no self-loops, so that the walker changes position at each time step. On the other hand, the network co-evolves with the random walk process, since every time a walker traverses a link, it increases its weight by a quantity $\delta w > 0$, as illustrated in Fig. 4.2. This mechanism mimics the fact that the relation between two concepts is reinforced every time the two concepts are associated by a cognitive process.

Formally, the dynamics of the network is the following. Every time an edge $(i, j) \in \mathcal{E}$ is traversed at time t, the associated weight is reinforced as

$$w^{t+1}_{ij} = w^t_{ij} + \delta w \tag{4.2}$$

The quantity δw, called reinforcement, is the only tunable parameter of the model.

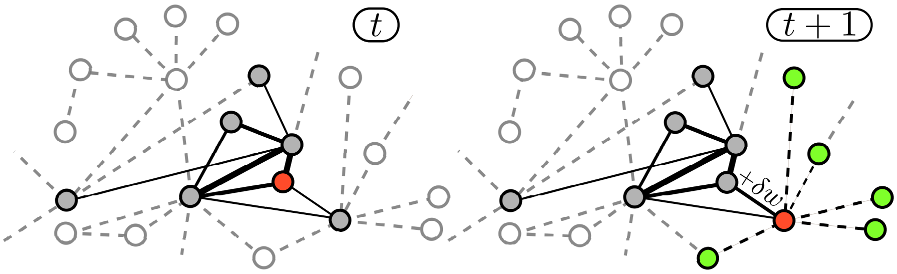


Figure 4.2: Edge-reinforced random walks (ERRWs) produce a co-evolution of the network with the dynamics of the walkers. At time t the walker is on the red node and has already visited the gray nodes, while the shaded nodes are still unexplored. The widths of edges are proportional to their weights. At time $t+1$ the walker has moved to a neighbour (red) with probability as in Eq. (4.1), and the weight of the used edge has been reinforced by δw. At this point, the walker will preferentially go back, although it can also access the set of "adjacent possible" (green). Figure from [1].

The idea of a walker preferentially returning on its steps is in line with the classical rich-get-richer paradigm, which has been extensively used in the network literature to grow scale-free graphs [24], and is here implemented in terms of reinforcement of the edges, instead of using a random walk biased on some properties of the nodes [303, 314, 315].

The co-evolution of network and walker motion induces a long-term memory in the trajectories which reproduces, as we will show below, the empirically observed correlations in the dynamics of discovery [82]. In fact, if i_t is a realization of the random variable X_t denoting the position of the walker at time t, the conditional probability Prob $[X_{t+1} = i | i_0, i_1, \dots, i_t]$ that, at time step $t+1$, the walker is at node i, after a trajectory $\mathcal{S} = (i_0, i_1, i_2, \dots, i_t)$, depends on the whole history of the visited nodes, namely on the frequency but also on the precise order in which they have been visited [307]. The strongly non-Markovian [316] nature of the random walks comes indeed from the fact that the transition matrix Π^t co-evolves with the rearrangement of the weights. This makes our approach fundamentally different from the other models based on Polya-like processes (cfr. Sec. 1.2.2.1). For instance, in the Tria *et al.* urn model [82], where an innovation corresponds to the extraction of a ball of a new colour, the probability of extracting a given colour (colours correspond to node labels in our model) at time $t+1$ only depends on the number of times each colour has been extracted up to time t, and not on the precise sequence of colours. Moreover, the use of an underlying network (see Fig. 4.2) is a natural way to include the concept of the *adjacent possible* in our model, without the need of a triggering mechanism and further parameters, which are instead necessary in the UMT (balls of new colours added into the urn whenever a colour is drawn out for the first time) and in its mapping in terms of growing graphs considered in SI of Refs. [82, 189].

Another consequence of the strongly non-Markovian nature of ERRWs is their very limited analytical tractability that makes them extremely hard to handle. Nevertheless, mathematicians managed to achieve specific analytical results by focusing on the evolution of the edges in time. This can be done by (*i*) subtracting the initial values of the weights, so that $\forall\ (i, j) \in \mathcal{E}$ and $\forall t$, $\tilde{w}^t_{ij} = w^t_{ij} - w^0_{ij}$ and (*ii*) normalizing step by step by defining a new random variable α as $\alpha^t_{ij} = \tilde{w}^t_{ij}/t\delta w$, which represents the percentage of time the walker spent on the edge (i, j) up to time t. The resulting random vector $\alpha^t := (\alpha^t_{ij})_{(i,j)\in\mathcal{E}}$ takes values in the simplex $\Delta = \{(X_{ij})_{(i,j)\in\mathcal{E}} \in (0, \infty)^{\mathcal{E}} : \sum_{ij} X_{ij} = 1\}$. It has been proved that the sequence $(\alpha^t)_{t\in\mathbb{N}}$ converges almost surely, and the associated random limit distribution can be determined analytically [309]. We refer the interested reader to a general survey of random processes with reinforcement that can be found in [195].

In the next sections we will first test our model on synthetic networks (Section 4.3), and then we will consider a real case where the underlying network of relations among concepts can be directly accessed and used (Section 4.4).

4.3 Results on synthetic networks

As a first experiment, based on the idea that concepts are organized in dense clusters connected by few long-range links, we model the relations among concepts as a small-world network (SW) [317].Our choice is supported by recent results on small-world properties of word associations[318], language networks [319] and semantic networks of creative people [320].

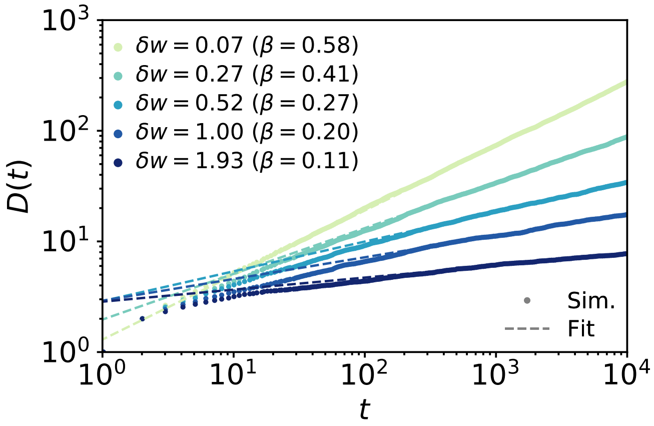


Figure 4.3: ERRW on SW networks with $N = 10^5$, $m = 1$, and $p = 0.02$. Heaps' law and associated exponents β obtained for different values of reinforcement δw. Each curve (points) represent the average of different realizations of the process. Dashed lines denote the power-law fit, with exponents given in the legend. Figure adapted from [1].

To construct SW networks, we use the procedure proposed by Newman in Ref. [321]. Namely, we start with a ring of N nodes, each connected to its $2m$ nearest neighbours, and then we add, with a tunable probability p, usually called *rewiring*, a new random edge for each of the edges of the ring. Notice that the name *rewiring* might be misleading in this case, since, differently from classic Watts-Strogatz version [25], here the new lines are not replacing the one of the ring, but are simply added.

The first thing we want to investigate is the Heaps' law for the rate at which novelties happen [82, 190]. We therefore look at how the number of distinct nodes $D(t)$ in a sequence $\mathcal{S}$ generated by a walker grows as a function of length of the sequence t. Figure 4.3 shows the curves $D(t)$ obtained by averaging over different realizations of a ERRW process with reinforcement δw on a SW network with rewiring probability $p = 0.02$. Points represent the results of these simulations, with different colours for different value of δw. All the curves can be well fitted by a power law $D(t) \sim t^{\beta}$, with an exponent β which decreases when the reinforcement δw increases. Fitted curves are plotted as dashed lines of the correspondent colour, with β exponents reported in the legend.

Finding the average number of distinct sites visited by a random walker is a well-known problem in the case of graphs without reinforcement. In particular, it has been

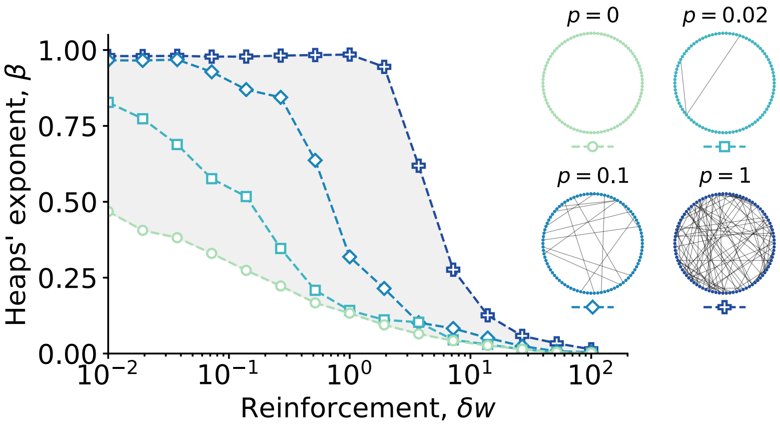


Figure 4.4: ERRW on SW networks with $N = 10^5$, $m = 1$. Heaps' exponent β as a function of the reinforcement δw for networks with different rewiring probabilities p. Figure adapted from [1].

proven that, in the absence of reinforcement, the average number of distinct sites $D(t)$ visited in t steps scales, for $t \gg 1$, as $S^{\mathrm{ring}}(t) \sim (8t/\pi)^{1/2}$ [322] in one-dimensional lattices and as $S^{\mathrm{ER}}(t) \sim t$ [323] in Erdős-Rényi random graphs [324].

While transition between these two regimes ($p = 0 \longrightarrow p = 1$) has been investigated in Refs. [325–327] for SW networks with different values of p, the effects of the reinforcement has never been explored. To this extend, we run again the process on SW networks having different values of p and extract the resulting Heaps' curves. Figure 4.4 reports the fitted values of the exponent β obtained in the case of ERRW processes under different strength of reinforcement δw. The four curves refer to SW networks with rewiring probabilities $p = 0, 0.02, 0.1$, and 1. Notice that the previously known results, $\beta^{\mathrm{ring}} = 1/2$ and $\beta^{\mathrm{ER}} = 1$, are recovered as limits of the two curves relative to $p = 0$ and $p = 1$ when $\delta w \to 0$. Furthermore, for values of p in the SW regime [328], it is possible to get values of β spanning the entire range $[0, 1]$ by tuning the amount of reinforcement δw. This means that the reinforcement mechanism we propose is able to reproduce all the Heaps' exponents empirically observed (see Fig. 4.1, [82, 189]).

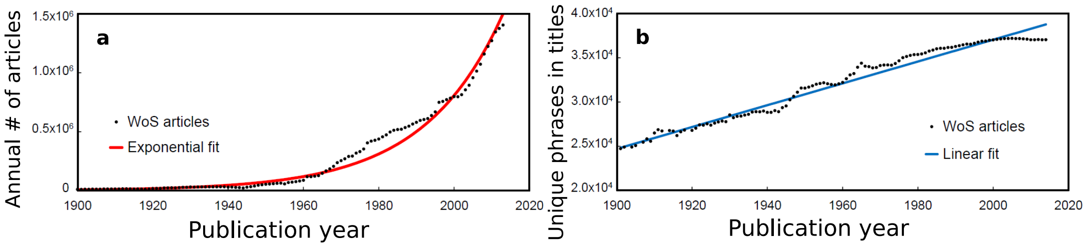


Figure 4.5: Growth of knowledge in science. (**a**) Number of scientific articles published per year. (**b**) Number of unique title phrases of scientific articles published per year. Points correspond to averages of data coming from the Web of Science database, continuous lines are fits. Figure adapted from [329].

4.4 The cognitive growth of science

To show how the model works in a real case, we need to consider a scenario in which we can access both the empirical curves $D(t)$ associated with a discovery process and an underlying network on top of which the process takes place. We therefore consider the case of the growth of knowledge in modern science.

Recently, a consistent body of scientific research has been devoted to understanding the structure and evolution of science. These efforts have been unified under the name of *science of science* [329, 330], that is generally used when performing data-driven research on citation and collaboration networks in a quantitative fashion [331–336]. Among the many interesting results of this collective investigation, it has been shown that while the number of publication per year is growing exponentially [see Fig. 4.5(**a**)], ideas grow much slower. This is reported in Fig. 4.5(**b**), in which a linear trend can be observed in the annual number of distinct title phrases of scientific articles as a function of time.

Here, however, we want to focus on the discovery novelties in science, whose growth is intuitively slower than linear. This is obviously linked to the exploration-exploitation dichotomy extensively discussed in Sec. 1.2.2 [149, 152], particularly relevant when it comes to scientific innovation [149, 153, 337, 338]. In order to do so, we move from an annual count to a publication-based time step. Furthermore, instead of considering title phrases, we look for unique concepts in abstracts, containing much more information

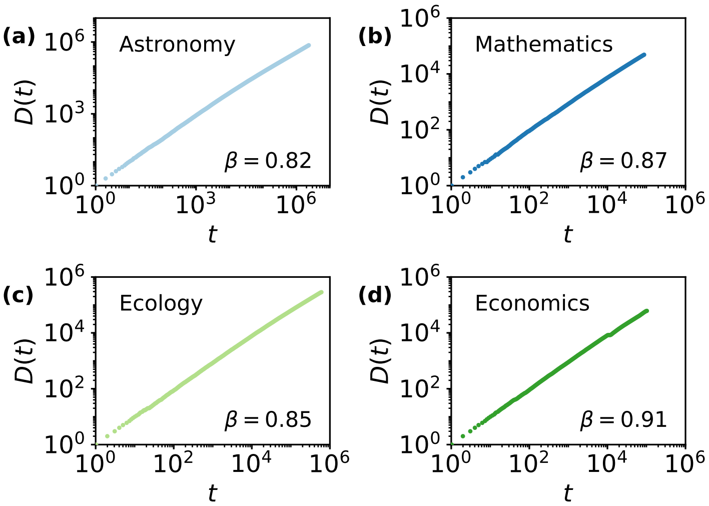


Figure 4.6: Heaps' law for the growth of knowledge in science. Number of distinct scientific novelties $D(t)$ discovered as a function of (re-scaled) time for the four considered disciplines: (**a**) astronomy, (**b**) mathematics, (**c**) ecology, and (**d**) economics. The associated fitted Heaps' exponents β are reported in the respective panels.

than the simple keywords in the titles. More specifically, we analyse 20 years (1991-2010) of scientific articles in four different disciplines, namely, astronomy, mathematics, ecology, and mathematics. Articles were taken from core journals in these four fields, and bibliographic records were downloaded from the Web of Science database. Details on data collection and the list of core journals are given in Ref. [339]. From a text analysis of each abstract, we extract relevant concepts as multi-word phrases [334] and construct, as illustrated in Fig. 4.7(**a**), the real temporal sequence $\mathcal{S}$ in each field from the publication date of the papers. By concatenating all the sub-sequences obtained from each abstract we produce a single sequence containing the evolution of knowledge in the considered field as collectively explored by the relevant community. Each sequence will contain recurrent concepts (exploitation mechanism) and novel ones (exploration). As before, the number of novelties as a function of the sequence length represents the pace of innovation of the field. Figure 4.6 shows that the number $D(t)$ of novel concepts in the considered fields grows with the length t of $\mathcal{S}$ as a power law with fitted exponents: (**a**) $\beta = 0.82$ (astronomy), (**b**) $\beta = 0.87$ (mathematics), (**c**) $\beta = 0.85$ (ecology), (**d**) and $\beta = 0.91$ (economics).

Together with the real exploration sequences we have also extracted, as illustrated

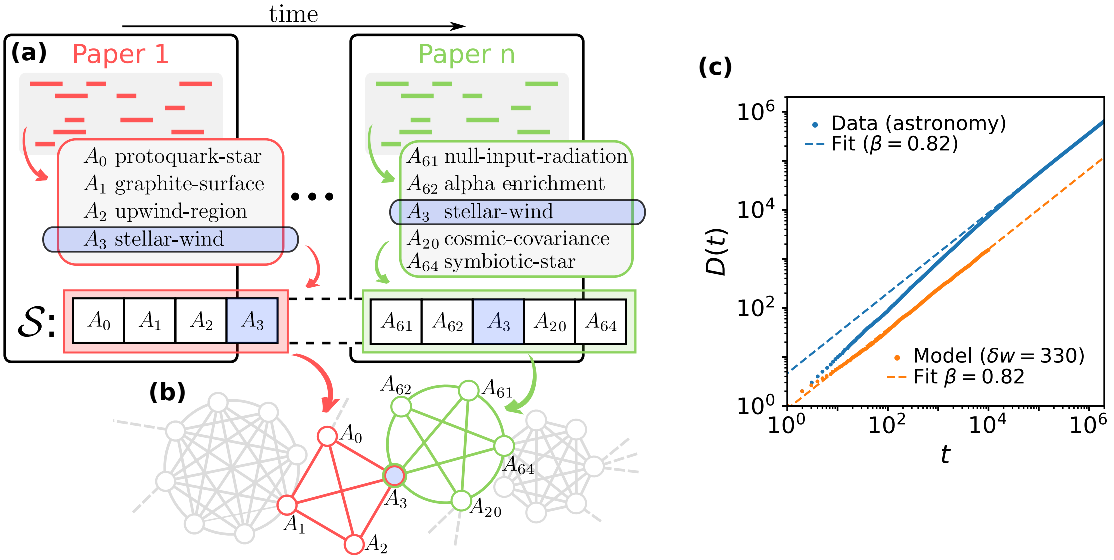


Figure 4.7: Extracting the growth of knowledge from scientific abstracts. (**a**) For each scientific field, an empirical sequence of scientific concepts $\mathcal{S}$ is extracted from the abstracts of the temporally ordered sequence of papers. (**b**) The network of relations among concepts is constructed by linking two concepts if they appear in the same abstract. The network is then used as the underlying structure for the ERRW model. (**c**) The model is tuned to the empirical data by choosing the value of the reinforcement δw that reproduces the Heaps' exponent β associated to $\mathcal{S}$. Figure adapted from [1].

in Fig. 4.7(**b**), the underlying networks of relations among concepts [293] from their co-occurrences in the abstracts, so that we do not need to rely on synthetic small-world topologies, or on the graph version of the UM (see SI of Refs. [82, 189]). Table 4-A reports basic properties, such as number of nodes N, average node degree $\langle k \rangle$, characteristic path length L and clustering coefficient C, for the largest components of the four networks we have constructed. Notice that different disciplines exhibit values of $\langle k \rangle$ ranging from 19 for mathematics to 172 for astronomy, but all of them have high values of C and low L.

Research field	Papers	N	$\langle k \rangle$	C	L	β	δw
Astronomy	97,255	103,069	172	0.41	2.48	0.82	330
Ecology	18,272	289,061	52	0.89	2.98	0.85	105
Economics	7,100	60,327	20	0.91	3.69	0.91	6
Mathematics	7,874	48,593	19	0.89	3.69	0.87	20

Table 4-A: Statistics of the network of concepts in four research fields, together with the empirical Heaps' exponent β and the value of δw that reproduces it.

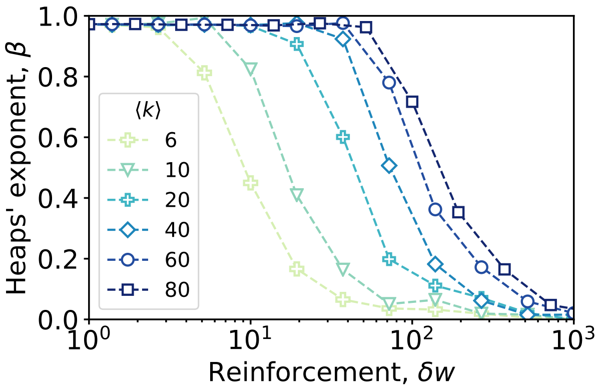


Figure 4.8: Impact of the average degree on the Heaps' exponent for ERRWs on ER networks with $N = 10^5$ and average degree $\langle k \rangle$. The Heaps' exponent β is plotted as a function of reinforcement δw for different values of $\langle k \rangle$. Figure adapted from [1].

We have then run the ERRW on each of the four networks, tuning the strength of the reinforcement δw, the only parameter of the model, so that the obtained curves for the growth of the number of distinct nodes visited by the walkers reproduce the empirical values of the exponent β. Figure 4.7(**c**) shows that, for the case of astronomy, the curve $D(t)$ of our model with $\delta w = 330$ has a power-law growth with exponent $\beta = 0.82$, equal to the one extracted from the real sequence of concepts.
The values of reinforcement obtained for the other scientific disciplines are reported in Table 4-A. Interestingly, there seems to be a clear division of the four disciplines into a class of lower reinforcement (astronomy and ecology) and a class of higher one (economics and mathematics). This difference could be due to the different research habits, approaches and techniques required by each field. For example, astronomy and ecology often rely on physical technologies, tools, field work, and experiments, while most of the research in economics and mathematics can be easily performed on a backboard or with the help of a computer. However, notice that stronger reinforcement is required to get the same β in networks with higher values of $\langle k \rangle$.

To better understand the wide range of values obtained for the reinforcement parameter from the analysis of the growth of knowledge in different scientific fields (see Table 4-A), we looked at the relation between the exponent β extracted from the

Heaps' law and the reinforcement δw in networks where we could control for a different average node degree. Results are showed in Fig. 4.8, where the fitted β exponents are plotted against the reinforcement δw. Each curve corresponds to an Erdős-Rényi random graphs with $N = 10^5$ nodes and average degrees $\langle k \rangle$ ranging from 6 to 80. As expected, the average degree significantly impacts the reinforcement. In particular, the higher the value of $\langle k \rangle$, the stronger the reinforcement δw has to be in order to produce the same Heaps' exponent. This is easily understandable if one considers the possible choices of a walker reaching a node connected to a link that has been reinforced. If the node has a high degree, the probability of selecting that specific link among all the others will be smaller, and the walker will more easily select a new link, leading to a previously undiscovered node, and therefore to a higher β. If one wants to keep a certain discovery rate in networks with higher $\langle k \rangle$, higher values of reinforcement will then need to be considered.

4.5 Correlated novelties

In addition to the Heaps' law, our model naturally captures also the correlations among novelties, which are a hallmark of real exploration sequences [82, 189]. In order to show this, we need to compare the synthetic sequences generated by the model with an appropriate null model. To this extent. in Sec. 4.5.1 we will define two null models that will then be tested in Sec. 4.5.2.

4.5.1 Null models

In order to check whether the sequences produced by our ERRW model are correlated, we need to introduce some appropriate null models, both based on a reshuffling method. More precisely, given a trajectory $\mathcal{S}$ of visited nodes (concepts), we use the the following two reshuffling procedures[82]. The simplest procedure consists in the *global reshuffling* of all the elements of $\mathcal{S}$ (indicated as *glob* in Fig. 4.9). This method destroys indeed the correlations (if there are any) in the sequence, but it also modifies the rate at which the new concepts appear, ultimately changing the exponent of the Heaps' law. Contrarily, the rate can be preserved by defining a second version of the null model, based on a *local reshuffling* (indicated as *loc* in Fig. 4.9). In this second procedure we reshuffle all the elements in $\mathcal{S}$ only after their first appearance, such that a concept cannot be randomly replaced in the sequence before the actual time it has been discovered. Algorithmically speaking, this can be achieved by following the following steps:

(i) Find all the novelties in $\mathcal{S}$ (first appearance of a symbol) and their respective indexes;

(ii) Count the number of occurrences of each symbol in $\mathcal{S}$;

(iii) Create an empty sequence $\mathcal{S}^{(loc)}$ of the same length of $\mathcal{S}$ and fill it with the novelties,keeping the same position;

(iv) Consider $\mathcal{S}$ and the last novelty, say A, that you find. If A appears $n^{(A)}$ times in $\mathcal{S}$, randomly place $n^{(A)} - 1$ symbols A in the compatible empty slots of $\mathcal{S}^{(loc)}$. These are all the available slots in $\mathcal{S}^{(loc)}$ after the appearance of A, that is already there;

(v) Repeat step (iv) for all the novelties of $\mathcal{S}$, starting from the end and going backwards until all the slots of $\mathcal{S}^{(loc)}$ are filled.

4.5.2 Results

In order to investigate the presence of correlations, we now compare the results for sequences generated by the ERRW model against their respective null models. In particular, we run the model with a reinforcement $\delta w = 0.01$ on small-world networks with rewiring probability $p = 0.02$ (results for different values of p and δw are reported in Fig. A.4 and A.5 of Appendix B). The paths produced by the walkers are symbolic sequences, that can be represented as a non-stationary process, where the alphabet is growing in time. We now show the results by considering three different quantities that are typically used to characterized symbolic sequences, such as texts.

Figure 4.9(**a**) shows that the frequency distribution $f(\Delta t)$ of inter-event times Δt between pairs of consecutive occurrences of the same concept is a power law, like the ones found

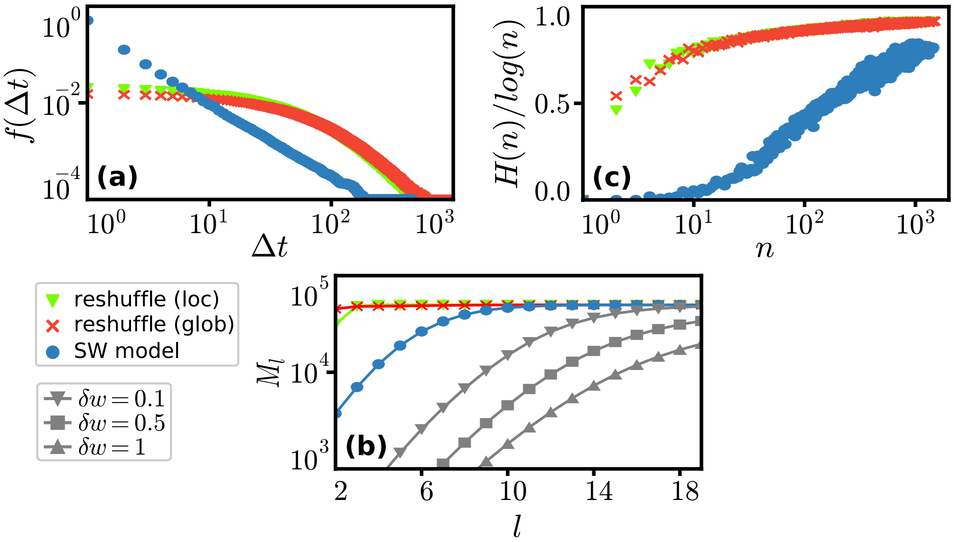


Figure 4.9: Correlations among concepts produced by an ERRW ($\delta w = 0.01$) on a SW network ($p = 0.02$). (**a**) Frequency distribution of inter-event times Δt between consecutive occurrences of the same concept (node in our model). (**b**) Number M_l of different subsequences of length l as a function of l. (**c**) Normalized entropy of the sequence of visited nodes as a function of n, the number of times the nodes have been visited. In each panel, blue circles show average values over 20 different realizations, while triangles and crosses refer to those of (globally and locally) reshuffled sequences. Figure from [1].

for novelties in Wikipedia and in other data sets in Refs. [82, 189]. Furthermore, the shape of $f(\Delta t)$ in our model significantly differs from that obtained by reshuffling the sequences locally and globally (see Sec. 4.5.1). Notice that $f(\Delta t)$ is the distribution of first return times (FRT), and it remains an interesting research question to investigate how FRT are linked to first passage times (FPT) in the case of correlated random walks [207].

We have also looked at how M_l, the number of distinct subsequences of $\mathcal{S}$ of length l, grows with l [340]. The number of different sub-sequences of length l that can be generated by an alphabet of size λ is $N_\lambda = \lambda^l$. Of course, in real data we expect to find only a small subset of the possible combinations. In particular, it has been shown that the number of allowed sub-sequences in texts scales according to a stretched exponential law [340]. In Fig. 4.9(**b**) the curve M_l generated by the ERRW model with $\delta w = 0.01$ is compared to those obtained by reshuffling the sequences. The value of M_l grows with l, until it reaches a plateau (equal to $T - l$, where $T = 5 \times 10^4$ is the number of steps of the walker in the simulation) as a consequence of the finite length of $\mathcal{S}$. Interestingly, the analogous curves for the null models immediately approach the saturation value, meaning that a process without reinforcement would generate all the possible subsequences in a sequence of length T, while with the reinforcement this number drops down because of the correlations.

In our model, the correlated sequences naturally emerge from the co-evolution of network and walker dynamics, while the UM [82] requires the introduction of an additional semantic triggering mechanism to reproduce the correlations found in the data.

To better characterize the correlations, we finally studied how homogeneously concepts occur in the sequence $\mathcal{S}$, after their first appearance. Following Tria *et al.* [82], we have divided the sequence $\mathcal{S}$ in $n^{(A)}$ subsequences of the same length, with $n^{(A)}$ being the total number of occurrence of A in $\mathcal{S}$, and we have evaluated the Shannon entropy [341]

$$H^{(A)} = -\sum_{s=1}^{n^{(A)}} p_s^{(A)} \log p_s^{(A)} \tag{4.3}$$

for every concept A, where $p_s^{(A)} = n_s^{(A)}/n^{(A)}$ denotes the probability of finding concept A in subsequence s.

Figure 4.9(**c**) shows the normalized average entropy $H(n)$ of concepts appearing n times. The maximum value is reached for a concept equally distributed along $\mathcal{S}$. Again, the large differences with respect to the null models reveal the correlated dynamics of our model. Similar results are obtained for the network of relationships among scientific concepts (see Fig. A.3 of Appendix B), confirming the validity of the choice of SW networks as underlying structures.

4.6 Summary and conclusions

In summary, the mechanism of co-evolution of network and random walks introduced in this work naturally reproduces all the properties observed in real discovery processes, including the correlated nature of empirical exploration trajectories. Notice that in our ERRW model, and differently from the original Urn Model with *semantic triggering* proposed by Tria et al., the correlations naturally appear from the network representation of the space of discoverable items and concepts. In the urn version, the correlations can only appear the introduction of *semantic labels* specifically attached to the balls (different balls and colours might share the same label), together with a mechanism named semantic triggering [82]. The semantic triggering mechanism is able to produce correlated sequences, but it also requires the addition of a third parameter, namely η, to the model. Notice, instead, that the model we propose in this paper does not need labels or additional mechanisms since correlations emerge naturally from the co-evolution of the walker dynamics and the network.

A similar model has been exploited to characterise the process of knowledge network building associated with curiosity [173]. Further applications of the ERRW framework could include the design of optimal novelty-driven exploration strategies, which could be highly relevant for scientific investigations [342]. Some preliminary in this direction, but with a different scope, include the search for edges in the space of scientific concepts where innovations might emerge [343].

In the next Chapter, we will consider the multi-agent nature of discovery processes in a model of interacting explorers [4].

Chapter 5

Interacting discovery processes

5.1 Introduction

Discoveries are essential milestones for the progress of our societies [153, 156, 158, 160, 166, 344–348]. Recently, different mathematical approaches have been proposed to investigate and model the dynamics of discovery and innovation [155, 157, 163, 168, 179, 180, 185, 186, 349–351], such as the one proposed in the previous Chapter. Among these, of particular interest are those based on random processes with reinforcement [195, 352, 353], like the ERRW and more standard Pòlya urns [197, 198]. Urns have been extensively used to study and model a variety of systems and processes, from evolutionary economics, voting and contagions [288, 354–356] to language and folksonomies [199, 357]. More recently, they have been also employed to filter information [358] and to explain the birth and evolution of social networks [351]. Interestingly, as discussed in Sec. 1.2.2, the same family of models can also be used to model discovery processes, if opportunely combined with the concept of the *adjacent possible—the set of all those things which are one step away from what is already known* [174, 290]. In fact, after the original Kauffman's formulation, the AP has been successfully translated into models. This is the case of the UMT [82] that incorporates the AP within the urn process in which the space expands together with the discovery dynamics, and the appearance of a novelty opens up the possibilities of further discoveries [82, 189, 200]. Another example is the ERRW model [1] just introduced that encodes it into the topology of a network of concepts and ideas.

UMTs have proved to be able to replicate the basic signatures of real-world discovery processes, such as the famous Heaps' and Zipf's laws [190, 191], often recurrent in complex systems [188, 193, 359–361], as well as Taylor's law [192]. It turns out that the Heaps' law, a sub-linear growth of the number of distinct elements $D(t) \sim t^{\beta}$ with the number of elements t, well describes the pace at which scientists discover concepts (see Sec. 4.4), or users collect new items [1, 82, 362], with higher values of the Heaps' exponent β denoting a faster exploration of the AP.

In Chapter 4, we have showed how reinforced random walks can be used to mimic the exploration processes leading to the appearance of the new. As we saw, the exploration dynamics considered, just like the ones previously introduced in urn models with semantic triggering, refers to a single entity, representing, for example, the collective efforts of researchers within a research field [1]. However, despite the existing models can capture some of the essential underlying mechanisms behind the discovery of novelties, little emphasis is given to the collective exploration of the space of novelties and to the benefits that social interactions could bring to the discovery process. In fact, with the exception of Ref. [351], the modelled exploration dynamics refers to a single entity, representing, for example, the collective efforts of researchers within a research field [1]. Without taking into account the multi-agent nature of the process, these models (*i*) do not capture the heterogeneity of the pace of the individual explorers and (*ii*) do not include the benefits brought by social interactions and collaborations. Indeed, empirical evidence of these mechanisms has been found in various contexts [363–365], such as music-listening, politics and voting [366, 367], health [108] and language [368].

In this Chapter, we propose a model of interacting discovery processes where an explorer is associated to each of the nodes of a social network [18, 21, 202], and its dynamics is governed by an UMT (see Sec. 1.2.2.1). Hence, the local dynamics of each node accounts for the presence of an AP, more precisely the *adjacent possible in the space of* *concepts*. The social network makes the exploration a collective one, since processes of neighboring urns are coupled. This coupling expands the notion of the AP by adding a social dimension, represented by the set of opportunities one is possibly exposed to through his/her social contacts. We call this the *adjacent possible in the social space*. Social networks have been extensively used as a substrate on top of which dynamical processes take place [22, 23]. Indeed, in Chapter 2 and 3 we have used social interactions as the underlying structure that shapes the contagion dynamics.Notice, however, that here our setting crucially differs from the typical approach in which the network mediates, for example, the diffusion of innovations or social contagions [2, 89, 105, 120, 369]. Here instead, the interactions among the many discovery processes reveals the twofold nature of the AP of each individual, highlighting the crucial role played by the social structure in determining the individual exploration dynamics [4].

5.1.1 Outline

This Chapter is structured as follows:

In Sec. 5.2, we focus on the mechanisms of collective exploration by introducing a model of many urns coupled through the links of a network. Each urn process represents a different explorer that exploits opportunities coming from his/her social contacts.

In Sec. 5.3, we restrict our attention to the pace of discovery, that is the speed at which each individual discovers new objects. We then show how it is possible to write down a system of coupled equation that governs the growth of the number of novelties of each explorer as a function of the parameters of the model, namely the reinforcement and the size of the *adjacent possible in the space of concepts*, and the topology of the social network.

In Sec. 5.4, we study different network structure, starting from small toy graphs and concluding with real-world social networks extracted from empirical data. We study the asymptotic dynamics of the model on these networks and its behaviour at transient times, finding a non-trivial dependence between the pace of discovery of each explorer and its position in the social network. Simulations are also compared to the numerical integrations of the analytical equations introduced in Sec. 5.3, thus confirming its accuracy.

In Sec. 5.5, we analytically solve the equations for the growth of the number of novelties for a generic network in the asymptotic regime. We then show how non-local centrality measures, such as the Eigenvector and α-centrality, can be used to rank the nodes according to their pace of discovery and thus predict the best innovators as a function of the social structure. These analytical results are finally tested on real-world social networks, finding an agreement with the predictions that on both undirected and directed structure.

Conclusions and future perspectives are summarized in Sec. 5.6.

5.2 The model: coupling urn processes

Let us consider an unweighted directed graph $G(\mathcal{N}, \mathcal{E})$, where $\mathcal{N}$ and $\mathcal{E}$ are respectively a set of $N = |\mathcal{N}|$ nodes and a set of $E = |\mathcal{E}|$ links. Each node of the graph represents an individual/agent, while link (i, j) denotes the existence of a directed social relation from individual i to j (such that i can benefit from j). The graph is described by its adjacency matrix $\boldsymbol{A} \equiv \{a_{ij}\}$, whose element a_{ij} is equal to 1 if link (i, j) is present, and is 0 otherwise. Each node i is equipped with an Urn Model with Triggering (UMT) that describes the discovery process of the agent i [82]. In the following we indicate as $\mathcal{U}_i(t)$ the urn i at time t, while $\mathcal{S}_i(t)$ denotes the sequence of balls generated up to time t. Notice that $\mathcal{U}_i(t)$ is an unordered multiset of size $U_i(t) = |\mathcal{U}_i(t)|$, while $\mathcal{S}_i(t)$ is an ordered multiset of size $|\mathcal{S}_i(t)| = t$. Each urn i is characterized by two parameters, ρ_i and ν_i. As in the original UMT (see Sec. 1.2.2.1), the *reinforcement* parameter ρ_i accounts for the number of balls of the same colour that are added to the urn i whenever a ball of a given colour is extracted at time t. Furthermore, the *triggering* parameter ν_i controls the size of the *adjacent possible in the space of concepts*, as $(\nu_i + 1)$ balls of new colours are added to the urn of node i whenever at time t a colour is extracted for the first time [82]. In this abstract representation, the space of concepts—made by all the colours—expands in time together with each discovery process, without relying on a predefined structure [200].

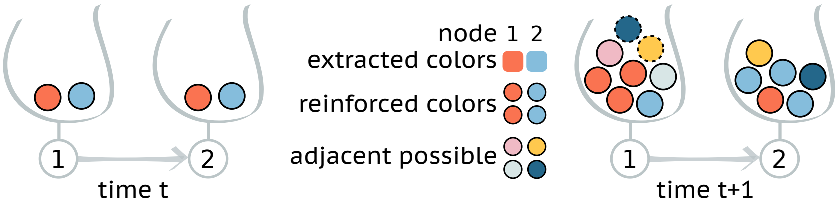


Figure 5.1:
Illustration of the model in the case of a network with two nodes. Each node is equipped with an urn obeying to the urn model with triggering (UMT) with same parameters $\rho = 1$, $\nu = 1$, and $M_0 = \nu + 1$. At the time t, the urns start with two balls, one red (R) and the other blue (B). Then, each node extracts a ball (1:R, 2:B), and therefore ρ additional balls of the same colours are added to the respective urns (reinforcement). Also, since in both cases, the extracted balls represent a novelty for the respective nodes, $\nu + 1$ balls of new colours are also added to the urns (adjacent possible). At the time $t + 1$, node 1 has access to all its balls plus two extra ones coming from the adjacent possible in the social space, i.e., the set of balls available through its neighbour, depicted with dashed borders. Figure from [4].

Discovery processes of different individuals are then coupled through the links of the network, representing social interactions. Namely, at each time t, the individual i draws a ball from an enriched urn, the so-called *social urn* of node i, $\tilde{\mathcal{U}}_i(t)$, composed by its own urn, as in the original UMT, plus the additional balls present at time t in the urns of all its neighbours, without their reinforcement. The latter wants to represent the *adjacent possible in the social space*, i.e., the AP to which we are exposed through our social contacts. The model is illustrated in Fig. 5.1 in the case of two nodes with a directed link (parameters $\rho = 1$, $\nu = 1$, and $M_0 = \nu + 1$). We thus have:

$$\tilde{\mathcal{U}}_i(t) = \mathcal{U}_i(t) + \bigcup_{j \in \mathcal{N}} a_{ij} \mathcal{U}'_j(t) \tag{5.1}$$

where $\mathcal{U}'_j(t) = \mathcal{U}_j^{[m=1]}(t) \subseteq \mathcal{U}_j(t)$ is the underlying set of the multiset $\mathcal{U}_j(t)$ (with multiplicity $m = 1$), i.e., the set of size $U'_j(t) = |\mathcal{U}'_j(t)|$ formed by its unique elements. Duplicates in the urn associated to node j at time t are indeed not considered. Thus, the “memory" of node j due to the reinforcement does not influence node i. Similarly, let us denote with $\mathcal{S}'_i(t)$ the underlying set of the sequence $\mathcal{S}_i(t)$, i.e., the sequence of all the unique elements of $\mathcal{S}_i(t)$. We consider synchronous updates for all the urn.

5.3 Pace of discovery

As previous works have shown [82], the dynamics of novelties and innovations share a number of commonalities and can, thus, be thought as two sides of the same process; a novelty refers to the discovery of something by an individual (already known to others), while innovations are novelties that are new to everybody. Here, we are interested in studying the asymptotic growth of the number of novelties—of each sequence—as a function of time (length of the sequence), representing the pace of discovery. We know, from standard results on the UMT [82], that an isolated urn i follows a Heaps' law, i.e., a power law behaviour $D_i(t) \sim t^{\beta_i}$ [190], $D_i(t) = |\mathcal{S}_i'(t)|$ being the number of different elements contained in the sequence $\mathcal{S}_i(t)$ up to time t (see analytical results in Sec. 1.2.2.1). Thus, the Heaps' exponent β_i quantifies the speed at which the urn discovers new element, and by definition it is bounded by 1. Let us consider now a node i that interacts through the network. In general, since $D_i(t)$ increases by one every time a ball is extracted for the first time, we can write

$$D_i(t+1) = D_i(t) + P_i^{\text{new}}(t) \tag{5.2}$$

where $P_i^{\text{new}}(t) \in [0,1]$ is the probability that the ball extracted at node i at time t never appeared in $\mathcal{S}_i(t)$ before. In other words, $P_i^{\text{new}}(t) = \text{Prob}\left[D_i(t+1) = D_i(t) + 1 | D_i(t)\right]$ and we can express it as the fraction of discoverable balls over the total number of balls available to node i at time t. This leads to an equation for the asymptotic Heaps' dynamics that in the continuous time limit reads:

$$\frac{dD_i(t)}{dt} = P_i^{\text{new}}(t) = \frac{|\tilde{\mathcal{U}}_i(t) \ominus \mathcal{S}_i'(t)|}{\tilde{U}_i(t)}, \tag{5.3}$$

where $\mathcal{A} \ominus \mathcal{B}$ denotes the multiset obtained by removing all the elements in set $\mathcal{B}$ from the multiset $\mathcal{A}$ (all duplicates are removed). Notice that if a node i has an out-degree $\sum_j a_{ij} = 0$, its associated Eq. (5.3) reduces to the one of an isolated urn, for which $\tilde{\mathcal{U}}_i(t) = \mathcal{U}_i(t)$. Thus, its Heaps dynamics for $\rho > \nu$ follows $D_i(t) \sim t^{\nu/\rho}$ for $t \to \infty$ [82, 192]

(see Eq. (1.7) in Sec. 1.2.2.1).

In the most general case, in which each node i is equipped with a UMT(ρ_i, ν_i), the equation for the Heaps' laws of each node $i \in \mathcal{N}$ can be written as in Eq. (5.3), by accounting for all the neighbours that are part of the social urn of node i. This can be done by using the non-zero elements of the adjacency matrix $\boldsymbol{A}$, so that the number of balls $\tilde{U}_i(t)$ in the social urn of node i at time t reads:

$$\begin{aligned}\tilde{U}_i(t) &= U_i(t) + \sum_{j\in\mathcal{N}} a_{ij}\left[M_0 + (\nu_j + 1)D_j(t)\right] \\ &= M_0 + \rho_i t + (\nu_i + 1)D_i(t) + \sum_{j\in\mathcal{N}} a_{ij}\left[M_0 + (\nu + 1)D_j(t)\right] \\ &= \rho_i t + \sum_{j\in\mathcal{N}} \left[a_{ij} + \delta_{ij}\right]\left[M_0 + (\nu_j + 1)D_j(t)\right]\end{aligned} \tag{5.4}$$

where M_0 is the initial number of balls in each urn, and δ_{ij} stands for the Kronecker delta. Notice that the term ρt does not appear on the r.h.s. of the first Eq. above, since the social urn does not account for the reinforcement of the neighbours.

Finally, from Eq. (5.3) and Eq. (5.4), the large time behaviour of the number of different elements $D_i(t)$ for each node i of our network can be written as

$$\frac{dD_i(t)}{dt} = \frac{|\tilde{\mathcal{U}}_i(t) \ominus \mathcal{S}'_i(t)|}{\tilde{U}_i(t)} = \frac{M_0 + \nu D_i(t) + \sum_j a_{ij}\left[M_0 + (\nu + 1)D_j(t)\right]}{\rho t + M_0 + (\nu + 1)D_i(t) + \sum_j a_{ij}\left[M_0 + (\nu + 1)D_j(t)\right]} \tag{5.5}$$

or, equivalently

$$\frac{dD_i(t)}{dt} = \frac{M_0 \sum_j (a_{ij} + \delta_{ij}) + \sum_j \left[\delta_{ij}\nu_j + a_{ij}(\nu_j + 1)\right]D_j(t)}{\rho_i t + \sum_j (a_{ij} + \delta_{ij})\left[M_0 + (\nu_j + 1)D_j(t)\right]}. \tag{5.6}$$

Eq.s (5.6) form a system of N coupled non-linear ordinary differential equations (ODE), with initial conditions $D_i(0) = 0\ \forall i \in \mathcal{N}$, that can be numerically integrated for any network topology $\{a_{ij}\}$.

Data set	Label	Type	N	E	$\langle k \rangle$	$\widehat{\mu}$	CCs	SCCs	S. LCC	S. LSCC
ZKC	(a)	Und.	34	78	4.6	6.7	1	1	34	34
Twitter	(b)	Dir.	4968	26875	10.8	5.2	1	4164	4968	770
NetSci	(c)	Und.	1589	2742	3.4	19.0	396	396	379	379
Jazz	(d)	Undir.	198	2742	27.7	40.0	1	1	198	198

Table 5-A: Statistics and properties of the four real-world networks considered (see Fig. 5.6): type of edges (Undirected/Directed), number of nodes N, number of edges E, average node degree $\langle k \rangle$, maximum eigenvalue $\widehat{\mu}$ of the correspondent adjacency matrix, number of (weakly) connected components (CCs), and number of strongly connected components (SCCs), size of the largest (weakly) connected component (S. LCC), and size of the largest strongly connected component (S. LSCC).

5.4 Numerical results

In this Section, we will explore the behaviour of the model in introduced in Sec. 5.2 and we will test the analytical formalism for the pace of discovery introduced in Sec. 5.3 against simulations. To do this, we will rely on small toy graphs to understand the basic mechanisms of the model and also on bigger empirical networks extracted form real-world data sets. Let us first give a brief overview of these data sets.

5.4.1 Description of the real-world social networks

We consider four data sets of real-world networks representing different types of social interactions: the Zachary Karate Club (ZKC) network [370], a network of follower relationships among Twitter users [371], a co-authorship network in Network Science [372], and a collaboration network between jazz musicians [373]. The network of Twitter from the original data set (Ref. [371]) has been reduced by performing a random walk sampling.

Some basic properties of the networks are summarized in Table 5-A, like the total number of nodes N, the total number of links E, if links are directed or undirected, the average degree $\langle k \rangle$, and the maximum eigenvalue $\widehat{\mu}$ of the related adjacency matrix. We also report the number of weakly- and strongly-connected components (CCs and SCCs respectively) together with the size of the respective largest one, since they play an

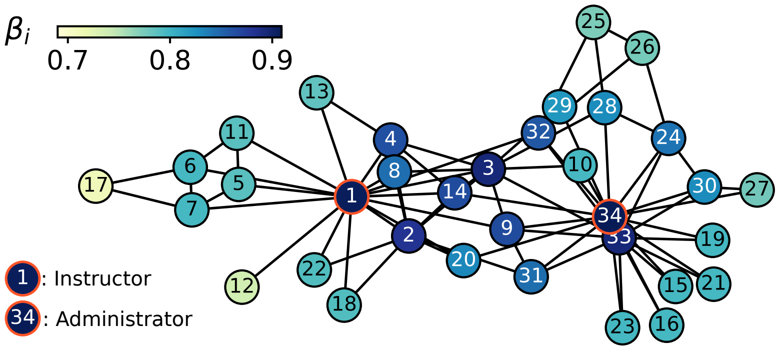


Figure 5.2: Dynamics of the interacting urns on the Zachary Karate Club network, whose nodes are coloured according to the resulting Heaps' exponent. Figure from [4].

important role in the dynamics under investigation. As we can see, the considered social networks display very different structural properties.

5.4.2 Results

We start exploring the behaviour of our model on the famous Zachary Karate Club network (ZKC) [370], a small network often used by network scientists as a benchmark for community detection algorithms [374–377] (details of the network are reported in Table 5-A).

Each node of the network is equipped with a UMT($\rho = 6, \nu = 3$) with same parameters and initial conditions. We run different simulations and observe, for each node i, the average growth of the number of distinct elements $D_i(t)$ as a function of time. We then extract the values of the Heaps' exponents of each node as $\beta_i = \beta_i(T)$, where $\beta_i(t) = \ln D_i(t)/\ln t$ and $T = 10^4$.
Figure 5.2 shows the nodes of the networks coloured accordingly. Notice the higher pace of discovery displayed by the notoriously central nodes corresponding to the instructor (node 1) and the administrator of the Club (node 34). This proves that nodes with identical UMTs can have completely different discovery dynamics, suggesting that a strategic location on the social network correlates with the discovery potential of an individual.

To further investigate this relation, we study the dynamics on five small directed networks. Figure 5.3(**a-e**) shows the temporal evolution of $D_i(t)$ for each node i of each of the networks displayed on the left. We report the simulated Heaps' laws (coloured points), whose extracted exponents β_i are shown in the legend. In addition, to assess the validity of Eq. (5.6), we also plot the curves (continuous black lines) obtained using the appropriate $\{a_{ij}\}$. It can be seen that the analytical formalism introduced in Sec. 5.3 perfectly captures the Heaps' laws, since lines are almost indistinguishable from—simulated—points. In particular, in Fig. 5.3(**a**) we observe the highest pace of discovery in the node with more outgoing links. However, the non-trivial behaviours observed in Fig. 5.3(**b-e**) for chains and graphs containing cycles indicate that the exponent of a node does not depend solely on local node properties. For instance, in Fig. 5.3(**d**) node 2 has two outgoing links, while the others have one link only. In contrast with what observed in Fig. 5.3(**a**), here the highest pace of discovery is the one of node 1, whose social urn gets the benefits of the urn of node 2. Moreover, in Fig. 5.3(**c**) and (**d**) a simple change of direction of link 4 → 2 translates into completely different dynamics. We also notice that in both Fig. 5.3(**c**) and (**e**) the presence of a cycle enhances the pace of discovery in a process of mutual exchange. However, while in (**d**) node 1 is linked to the cycle and captures the same behaviour of those in the cycle, in Fig. 5.3(**e**) node 1 behaves as an individual urn and does not affect the behaviour of the other nodes.

We have further investigated whether the extracted values of the Heaps' exponents β_i may depend on the maximum time T at which we have stopped the numerical simulations. The curves of $\beta_i(t)$ reported in Fig. 5.3(**f-j**) as function of the time t for times up to 10^8 clearly indicate that the systems, even for the small graphs considered, have not yet reached a stationary state. As we will show later, thermalisation times, that are typical of empirical trajectories of diffusion process [378], here are strongly influenced by the topology of the network. This can be seen by comparing the two $\beta_1(t)$ of Fig. 5.3(**f**) and (**g**), both approaching—as we will see later—the asymptotic value $\nu/\rho = 0.5$, but at very different timescales. Nevertheless, the ranking induced by the pace of discovery

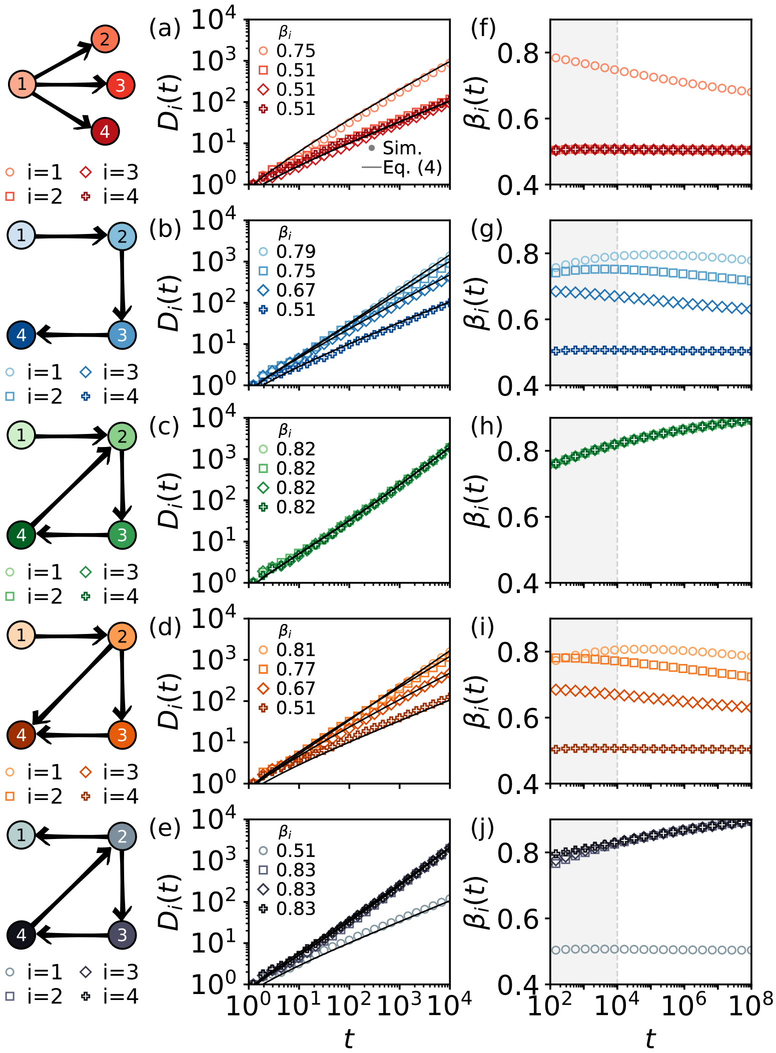


Figure 5.3: Heaps' dynamics of the interacting urns on five directed toy graphs (different symbols correspond to different nodes). Each node is equipped with a UMT with same parameters $\rho = 6$, and $\nu = 3$. (**a-e**) Temporal evolution of the number of discoveries $D_i(t)$ for each node i (associated Heaps' exponents β_i in the legend). The solutions of Eq. (5.6), shown as continuous black lines, are in perfect agreement with simulations. (**f-j**) Temporal behaviour of the associated Heaps' exponents extracted at different times. The grey area up to $T = 10^4$ corresponds to the values of (**a-e**). Figure from [4].

persists at all finite times. In the next section we will further investigate this characteristic behaviour, ultimately proving its universality for all networks.

Most of real world innovation systems operate far from equilibrium, thus we are particularly interested in the behaviour of our model at transient times. In order to numerically check the persistence of the ranking discussed in the last paragraph, we

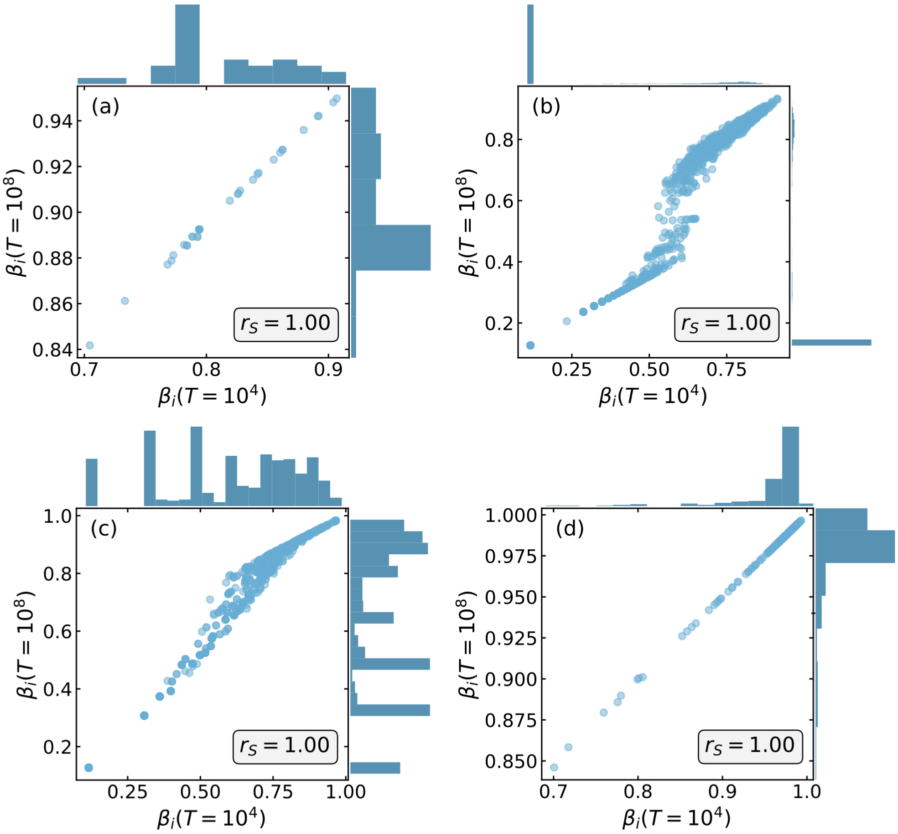


Figure 5.4: Transient behaviour and rank persistence. Scatter plot and Spearman's rank correlation coefficient r_S between fitted Heaps' exponents $\beta_i(T)$ at $T = 10^4$ and $T = 10^8$ associated to the $i = 1, \dots, N$ nodes off the four empirical networks considered: (**a**) the Zachary Karate Club network [370], (**b**) a network of follower relationships of Twitter [371], (**c**) a co-authorship network in network science [372] and (**d**) a collaboration network between jazz musicians [373]. The parameters of the model are $\rho = 10$, $\nu = 1$, $M_0 = \nu + 1$.

further run the model on each of the four empirical networks introduced in Sec. 5.4.1. We run our model connecting the urns with each one of them, keeping the same parameters, and record the Heap's exponent $\beta_i(T)$ of each node i at two different times $T = 10^4$ and $T = 10^8$. Figure 5.4 shows the scatter plot and the Spearman's rank correlation coefficient between these fitted Heaps' exponents $\beta(T = 10^4)$ and $\beta(T = 10^8)$, together with their distributions, for the node of the four real-world networks considered. In all cases, we get a Spearman's correlation of $r_S = 1$, meaning that even though the distribution of fitted exponents change (as it can be seen in the side panels of each scatter plot), the ranking is time-invariant and does not depend on the particular T at which Heaps' exponents are fitted. This is evident in the scatter plot of Fig. 5.4(**b**), where, apart from a set of nodes whose exponents span across the entire range, most of the nodes present a very low pace of discovery, with fitted exponents very close to 0. The opposite behaviour can be seen in Fig. 5.4(**d**), that displays the highest Heaps' exponents among the four networks (with all Heaps' exponents very close to 1).

These results suggest that the various paces of discovery have to depend on some structural characteristics of the networks, and the next section will focus specifically on this.

5.5 Analytical results

In this section, we derive an analytical solution to Eq. (5.6) introduced in Sec. 5.3. The system of equation fully characterises the pace of discovery of each node of the network as the temporal growth of the number of novelties. In particular, building on top of the numerical results discussed in Sec. 5.4, we are interested in finding an exact expression for the asymptotic values of the Heaps' exponents and investigating their dependence on the network topology. To achieve this, consider the system of N coupled non-linear ODEs given by Eq. (5.6) in the $t \longrightarrow \infty$ limit.

Let us suppose $\rho_i = \rho$ and $\nu_i = \nu \ \forall i \in \mathcal{N}$. For sufficiently high values of ρ we have $\lim_{t\to\infty} D_i(t)/t = 0 \ \forall i$, so that the denominator of the r.h.s. of Eq. (5.6) can be approximated by ρt, leading to:

$$\frac{dD_i(t)}{dt} \approx \frac{\sum_j \left[\delta_{ij}\nu + a_{ij}(\nu+1)\right] D_j(t)}{\rho t}, \tag{5.7}$$

which can be expressed in a more compact way as:

$$\frac{d\vec{D}(t)}{dt} \approx \frac{1}{t}\left(\frac{\nu}{\rho}\boldsymbol{I} + \frac{\nu+1}{\rho}\boldsymbol{A}\right)\vec{D}(t) = \frac{1}{t}\frac{f(\boldsymbol{A})\vec{D}(t)}{t} = \frac{1}{t}\boldsymbol{M}\vec{D}(t), \tag{5.8}$$

where $\vec{D}(t) \equiv \{D_i(t)\}_{i=1,\dots,N}$, $\boldsymbol{I}$ denotes the $N \times N$ identity matrix, and we have introduced the constant matrix

$$\boldsymbol{M} = f(\boldsymbol{A}) = (\frac{\nu}{\rho}\boldsymbol{I} + \frac{\nu+1}{\rho}\boldsymbol{A}) \tag{5.9}$$

By operating the change of variable $t = e^z$, Eq. (5.8) can be rewritten in terms of Eq. (5.9) as

$$d_z\vec{D}(z) \approx \boldsymbol{M}\vec{D}(z), \tag{5.10}$$

a standard first-order differential system, which leads to the solution

$$\vec{D}(t) \approx \sum_{\ell=1}^{r} \sum_{p=0}^{m_\ell - 1} \vec{c}_p \ln^p(t)\, t^{\lambda_\ell}, \tag{5.11}$$

where $\{\lambda_\ell\}_{\ell=1,\dots,r}$ and $\{m_\ell\}_{\ell=1,\dots,r}$ are the eigenvalues of $\boldsymbol{M}$ with their respective multiplicities, and $\vec{c}_p$ are vectors defined by the initial conditions.

The asymptotic behaviour of the number of novelties $D_i(t)$ discovered by node i at time t is then governed by the leading term in Eq. (5.11), so that we can write:

$$D_i(t) \underset{t\to\infty}{\approx} u_i \ln^{\widehat{p}(i)}(t)\, t^{\widehat{\lambda}(i)}. \tag{5.12}$$

where $\widehat{\lambda}(i)$ is the eigenvalue of $\boldsymbol{M}$ with the biggest real part such that the i-th entry of at least one of its eigenvectors $\vec{c}_p$ is different from zero. Similarly, $\widehat{p}(i)$ is the maximum value of p among these eigenvectors with non-zero i-th entries. Notice that, in general, $\widehat{\lambda}(i)$ can be smaller than the multiplicity of the eigenvalue $\widehat{\lambda}(i)$ minus one. Moreover, different nodes may have different values for these exponents. In particular, nodes in the same strongly connected component (SCC) will have the same Heaps' exponent, while these may vary across SCCs (more details, together with the full analytical derivation, are given in Appendix C.

Let us consider here a single illustrative example, that is a chain of four nodes such as the one depicted in Fig. 5.3(**b**). In this case, the asymptotic solution for a node

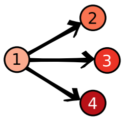
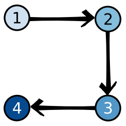
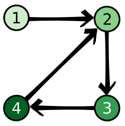
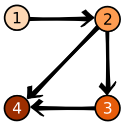
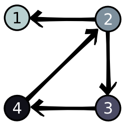
Network	 (a)	 (b)	 (c)	 (d)	 (e)
$D_1(t) \underset{t\to\infty}{\approx}$	$u_1 \ln(t)\, t^{\frac{\nu}{\rho}}$	$u_1 \ln^3(t)\, t^{\frac{\nu}{\rho}}$	$u_1\, t^{\frac{2\nu+1}{\rho}}$	$u_1 \ln^3(t)\, t^{\frac{\nu}{\rho}}$	$u_1\, t^{\frac{\nu}{\rho}}$
$D_2(t) \underset{t\to\infty}{\approx}$	$u_2\, t^{\frac{\nu}{\rho}}$	$u_2 \ln^2(t)\, t^{\frac{\nu}{\rho}}$	$u_2\, t^{\frac{2\nu+1}{\rho}}$	$u_2 \ln^2(t)\, t^{\frac{\nu}{\rho}}$	$u_2\, t^{\frac{2\nu+1}{\rho}}$
$D_3(t) \underset{t\to\infty}{\approx}$	$u_3\, t^{\frac{\nu}{\rho}}$	$u_3 \ln(t)\, t^{\frac{\nu}{\rho}}$	$u_3\, t^{\frac{2\nu+1}{\rho}}$	$u_3 \ln(t)\, t^{\frac{\nu}{\rho}}$	$u_3\, t^{\frac{2\nu+1}{\rho}}$
$D_4(t) \underset{t\to\infty}{\approx}$	$u_4\, t^{\frac{\nu}{\rho}}$	$u_4\, t^{\frac{\nu}{\rho}}$	$u_4\, t^{\frac{2\nu+1}{\rho}}$	$u_4\, t^{\frac{\nu}{\rho}}$	$u_4\, t^{\frac{2\nu+1}{\rho}}$

Table 5-B: Summary of the asymptotic Heaps' laws derived analytically for the 4 nodes composing the five directed toy networks considered in Fig. 5.3 of Sec. 5.4 (here displayed at the top). The coefficients u_i have not been reported to focus on the exponents of the power laws and the logarithms, when present.

$i = 1, \dots, N = 4$ reads

$$D_i(t) \sim u_i \ln^{N-i}(t)\, t^{\nu/\rho} \tag{5.13}$$

In this example, all the fitted exponents tend to ν/ρ at large times, while at finite times nodes with higher powers in the logarithm show higher paces of discovery, thus explaining the behaviour seen in Fig. 5.3(**g**).

Exact analytical solutions (in the asymptotic regimes) for the others directed toy graphs studied in Fig. 5.3 of Sec. 5.4 can be found in the Appendix C together with some additional toy structures, namely chains, cycles, and cliques. Here, we just report their explicit solutions in Table 5-B, where the coefficients u_i are left implicit.

In the case of strongly connected graphs, Eq. (5.12) simplifies. In particular, the logarithmic correction disappears and all the asymptotic exponents are equal to the maximum eigenvalue $\widehat{\lambda} = f(\widehat{\mu})$ of $\boldsymbol{M}$. In fact, for the Perron–Frobenius theorem [379, 380], the adjacency matrix $\boldsymbol{A}$ has a simple and positive maximum eigenvalue $\widehat{\mu}$ corresponding to an eigenvector $\vec{u}$ with all positive entries. Therefore, the approximated solution in Eq. (5.12) becomes:

$$D_i(t) \underset{t\to\infty}{\approx} u_i\, t^{\widehat{\lambda}}, \tag{5.14}$$

where u_i is proportional to the Bonacich eigenvector centrality [381] of node i, a global indicator of centrality that recursively quantifies the importance of a node from that of its neighbours, and not just from the number of neighbours (node degree).

As a consequence of Eq. (5.14), for strongly connected graphs every node has approximately the same behaviour $t^{\widehat{\lambda}}$. What makes a node different from another is precisely the multiplicative factor u_i. In cycles and cliques, nodes are all structurally equivalent ($u_i = u \; \forall i$), meaning that they all have the same curves $D_i(t)$ (see Appendix C). On the contrary, in graphs such as the ZKC (see Fig. 5.2), the different values of u_i play a very important role. Most central nodes, as the instructor and the chief administrator, are the fastest explorers (highest β_i), even if they all have the same asymptotic Heaps' exponent $\widehat{\lambda}$.

5.5.0.1 Pace of innovation and node centrality

The analytical argument just discussed showed us that for strongly connected components we expect the same asymptotic Heaps' exponents. However, the same analysis showed us that the coefficients depend on the eigenvector centrality. This factor plays a role in the transient times, when we are far from the asymptotic regime, and it is thus especially important for real-world systems.

Following the results in Eq. (5.14), we now test numerically the correlation between the eigenvector centrality and the measured Heaps' exponents at transient times for the Zachary Karate Club network (already used in Fig. 5.2). Figure 5.5(**a**) shows the scatter plot and the Spearman's rank correlation coefficient r_S of the normalised eigenvector centralities $c_i^{[E]}/c_{\max}^{[E]}$ and the fitted Heaps' exponents $\beta_i(T)$ at time $T = 10^4$ for the (largest strongly connected component of the) ZKC network. In parallel, Fig. 5.5(**b**) shows a network visualization where nodes are colour-coded accordingly to the centrality (cfr Fig. 5.2 of Sec. 5.4). Notice that the resulting Spearman's rank correlation higher than 0.98 persists changing the parameters in the simulations, even for sets of parameters in contrast with the approximations used in the analytical study, i.e. $\rho < \nu + (\nu + 1)\widehat{\mu}$. From

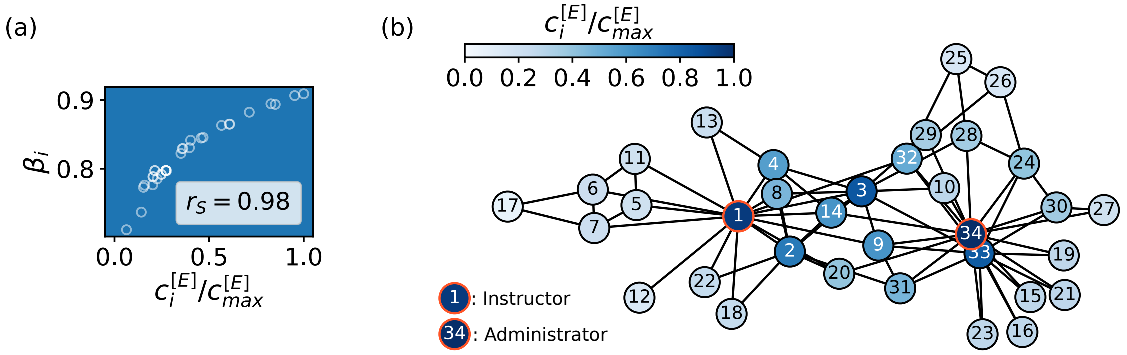


Figure 5.5: Dynamics of the interacting urns on the Zachary Karate Club network [370]. (**a**) Scatter plot and Spearman's rank correlation coefficients r_S between fitted Heaps' exponents $\beta_i(T = 10^4)$ and normalized eigenvector centrality $c_i^{[E]}/c_{\max}^{[E]}$ associated to the $i = 1, \dots, N$ nodes of the network. (**b**) Nodes are coloured according to the resulting normalized eigenvector centrality.

here, we can hence conclude that the eigenvector centrality is an optimal proxy for the distribution of Heaps' exponents in strongly connected social networks, and it can be used to give a faithful ranking of the individual expected paces of discovery.

In the more general case in which a graph is not strongly connected, Eq. (5.12) still holds, and the same argument can be applied to each of the strongly connected components in the graph to recursively find the values of u_i, $\widehat{p}(i)$, and $\widehat{\lambda}(i)$ (see Appendix C). In such cases, the eigenvector centrality needs to be replaced by its natural extension to non-strongly-connected graphs, i.e., the α-centrality [382].

The α-centrality is a measure widely used in network analysis [383–385] that has been first introduced in Ref. [382] to extend the eigenvector centrality to asymmetric graphs. The idea behind this measure is to tune the influence of the structure (adjacency matrix) by means of a parameter α, adding therefore exogenous sources to the centrality [18, 382]. Formally, it is defined as the vector $\vec{c}^{\,(\alpha)}$ such that

$$\vec{c}^{\,(\alpha)} = \alpha \boldsymbol{A} \vec{c}^{\,(\alpha)} + \vec{e}, \tag{5.15}$$

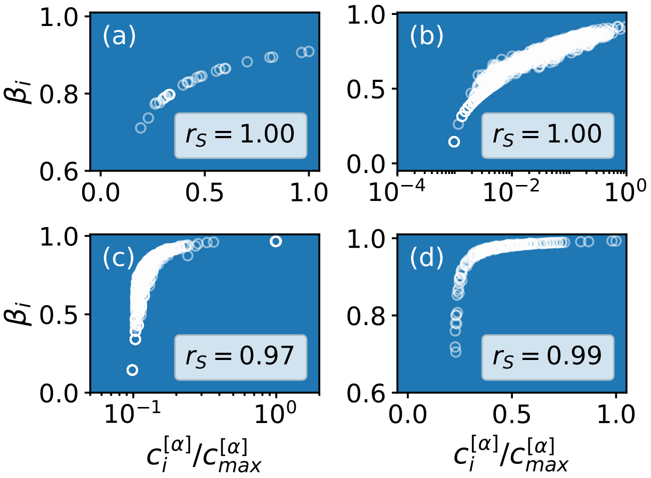


Figure 5.6: Scatter plot and Spearman's rank correlation coefficients r_S between fitted Heaps' exponents β_i and normalized α-centrality $c_i^{[\alpha]}/c_{\max}^{[\alpha]}$ associated to the $i = 1, \ldots, N$ nodes of four empirical networks: (a) the Zachary Karate Club network [370], (b) a network of follower relationships of Twitter [371], (c) a co-authorship network in network science [372] and (d) a collaboration network between jazz musicians [373]. Figure from [4].

where $\vec{e}$ is an N-dimensional vector of ones. The matricial form of Eq. (5.15) reads:

$$\vec{c}^{\,(\alpha)} = (\boldsymbol{I} - \alpha \boldsymbol{A})^{-1}\vec{e} = \left(\sum_{k=0}^{\infty} \alpha^k \boldsymbol{A}^k\right)\vec{e}, \tag{5.16}$$

where $\boldsymbol{I}$ is the N-dimensional identity matrix. It has also been shown that this centrality is equivalent to Katz-centrality [386] given by

$$\vec{c}^{\,(K)} = \left(\sum_{k=1}^{\infty} a^k \boldsymbol{A}^k\right)\vec{e}, \tag{5.17}$$

with a being an attenuation factor. In fact, it has been shown that the equality $\vec{c}^{\,(K)} = -\vec{e} + \vec{c}^{\,(\alpha)}$ holds, i.e. these two centralities differ only by a constant [382]. From Eq. (5.15) and (5.16), it is clear that the α-centrality can be both a local and global measure. In fact, for $\alpha \longrightarrow 0^+$, the relative importance of the structure as given by the adjacency matrix $\boldsymbol{A}$ decreases, in favour of the exogenous factor given by $\vec{e}$. When higher values of α are considered instead, the role of the exogenous part is damped.

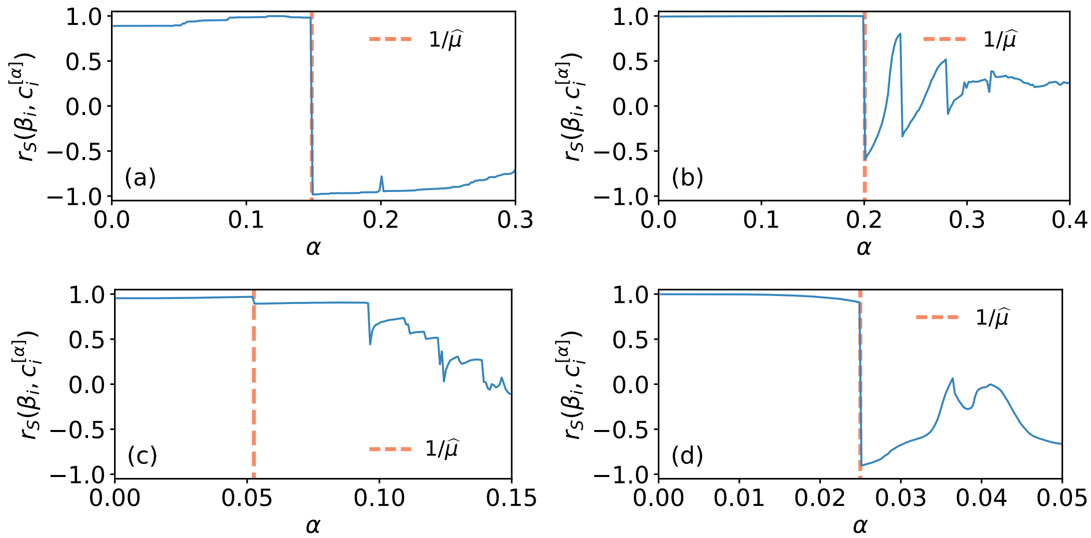


Figure 5.7: Spearman's rank correlation r_S between paces of discovery $\beta_i(10^4)$ and α-centrality $c_i^{[\alpha]}$ as a function of α for nodes $i = 1, \ldots, N$ belonging to four different real-world networks: (**a**) the Zachary Karate Club network [370], (**b**) a network of follower relationships of Twitter [371], (**c**) a co-authorship network in network science [372] and (**d**) a collaboration network between jazz musicians [373]. Each dashed vertical line corresponds the value of $1/\widehat{\mu}$, with $\widehat{\mu}$ denoting the maximum eigenvalue of the corresponding adjacency matrix. The parameters of the model are $\rho = 10$, $\nu = 1$, $M_0 = \nu + 1$.

The role of this measure in our particular setting can be explored, as before, by investigating the correlation between the α-centrality and the pace of discovery for interacting urn models connected by real-world social networks. Figure 5.6 shows the scatter plot of the number of discovered colours $D_i(T)$ against the normalized α-centrality $c_i^{[\alpha]}/c_{\max}^{[\alpha]}$ for each of the four real-world networks considered: (**a**) the ZKC, (**b**) a network of follower relationships of Twitter, a co-authorship network in network science, and (**d**) a collaboration network between jazz musicians (details in Sec. 5.4.1). The parameters of the model, which are the same for each urn, are set to $\rho = 10$, $\nu = 1$, $M_0 = \nu + 1$ and $T = 10^4$.

The high values of the Spearman's rank correlations ($r_S \geq 0.97$ in all cases) found in both undirected (**a,c,d**) and directed networks (**b**) are in agreement with our predictions. This result confirms that, together with the AP in the space of concepts, it is crucial to take into account of an AP in the social space.

Finally, we test how reliable the α-centrality is to give a ranking of the pace of discovery of the nodes when varying the value of α. This is shown in Fig. 5.7, where we plot the Spearman's rank correlation coefficient r_S between the paces of discovery $\beta_i(10^4)$ and the α-centralities $c_i^{[\alpha]}$ as a function of α for all the nodes $i = 1, \dots, N$ composing the four considered real-world networks. Although panel (**d**) displays a decrease in the correlation when approaching $1/\widehat{\mu}$, setting $\alpha < 1/\widehat{\mu}$ leads to Spearman's rank correlation coefficients $r_S > 0.89$ in all four cases.

5.6 Summary and conclusions

In conclusion, we have presented a first example in which stochastic processes are coupled over the nodes of a complex network, and analytical insights on the relations between structure and dynamics are possible. The results highlight that the structural—not just local—properties of the nodes can strongly affect their ability to discover novelties. Our networked model of social urns is not just a simple extension of UMTs. What makes it novel and different is the very same idea of coupling together many urns over a complex social network, and the concept of "social urn" we have introduced. It is such a network coupling that spontaneously produces novel behaviors, such as different exponents of the Heaps' law in a single system, and has the potential to open new areas of research and applications. This work represents only a first step toward the inclusion of structured interactions in discovery processes. Urns can, in fact, result oversimplified models for the dynamics of individual explorers. Future works could consider non-identical urns, or even explore the effects of having individuals with a finite storage capacity, or where the adoption of the new might trigger the abandoning of the old, as for substitutive systems [162]. Another natural extension would be considering discoveries and social relationships unfolding across different network layers [38] or higher-order structures [2, 5]. In addition, it would be interesting to establish a mapping between the model proposed here and models of multiple interacting random walkers on complex networks [1, 387], as well as its relationship with existing models of social spreading and meme popularity [388–390]. Finally, our results could be directly applied in studies about efficient team structures in cooperative creative tasks [7, 178, 329, 391, 392].

Conclusions and outlook

The processes of discovery and contagion are two social dynamics that been extensively studied, but have always been considered separately. In this thesis, we introduced them as the outcome of an adoption process studied from two different angles [6]. In Chapter 2 we started from the point of view of a single item (concept, or behaviour) spreading among individuals through an SIS-like dynamics whose transitions are decoupled across two different network layers, namely the social network (controlling the *simple contagion*) and the power grid (controlling the *complex recovery*) [3]. Then, with the paradigm shift proposed in Chapter 3, we studied a *simplicial contagion* model that accounts for the effects of higher-order group interactions—using the formalism of simplicial complexes instead of graphs —in processes of social contagion [2].

In the second part of the thesis we adopted the complementary point of view of a single individual sequentially adopting interlinked items. This is a good framework to model discovery and innovation processes, where novelties can indeed be seen as the first collection of an item of the first visit of a node. The network representation used in Chapter 4 enabled the development of a model that could mimic the basic dynamics and the statistical properties of real-world discovery processes by relying on a single parameter, the reinforcement [1]. Finally, in Chapter 5 we introduced a multi-agent discovery dynamics in which a social network is used to couple together discovery processes [4]. We proposed for the first time to couple urn processes on a complex network in order to explore how social interactions contribute to the collective emergence of new ideas in a team. By unveiling the crucial role played by the structure of the social network on the pace of discovery, this model opens up new research avenues towards the design of optimal group structure in teamwork.

The dual vision of adoption processes presented in this thesis and the different representations used trigger a series of open questions and research directions that could be addressed in the near future..

Coupling discovery and contagion processes. In this work, we have discussed the duality between contagion and discovery processes and presented some modelling advancements. However, a comprehensive model that bridges these complementary processes is still missing. Ideally, this model could make use of two different networks, namely the one of individuals and the one of items. Similarly to the models discussed in Chapters 2 and 3, the process of contagion could rely on social interactions and thus uses as a substrate a social network between individuals. In parallel, there could be an exploration processes with reinforcement akin to the one introduced in Chapter 4 that takes place over a different network of relationships between items. Individuals could independently perform an exploration of the space of items, that can be modelled as a random walk with reinforcement. While the topology of this network is the same for all the explorers and does not change in time, the strength of the links could vary across different explorers according to their personal history. As such, walkers explore the space according to an edge-reinforced random walk process [1], but different walkers would preferably jump towards different items according to the personal memory. Indeed, while the structure of the network of items is common to all the walkers, the strengths of the connections co-evolve with the dynamics of each walker. With the exploration mechanism just described, each walker independently performs independent searches of the space of possible discoveries [167, 200]. However, due to the reinforcement, discovering new ideas becomes more and more difficult, and walkers will often return on their steps [1]. This implies that the last novelty found will remain unchanged for some time. As for processes of individual and collective attention in social and substitutive systems [162], individuals will focus on a single novelty at the time. Thus, a

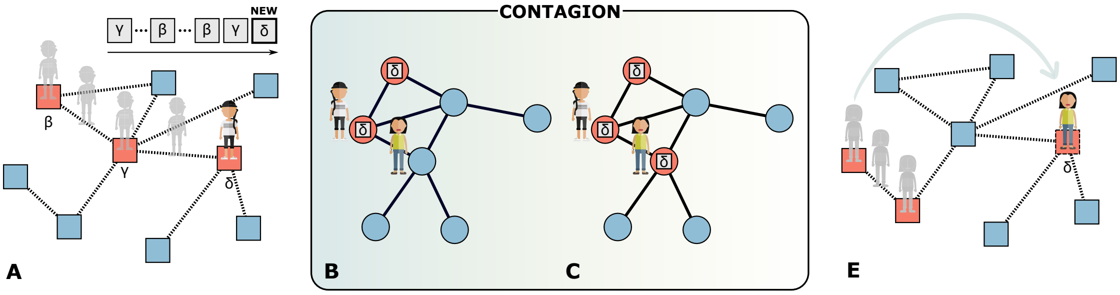


Figure 5.8: Coupling discovery and contagion processes on complex networks. (**A**) A walker explores the space of items via ERRWs. Thus, every time a node is visited the corresponding item is added to the sequence (top). Walking on δ represents the discovery of a novelty, since it never appeared in the sequence before. (**B**) The new item δ spreads from the walker to his neighbours through the link of the social network. In particular, in (**C**) a neighbour discovers δ through the social contagion dynamics. As a consequence, her position on the network of items is immediately updated by means of a flight (**D**). Figure from [6].

naive coupling could have a contagion process where each individual tries to spread this last novelty to her/his neighbours through the links of the social network (see Fig. 5.8). If it spreads enough, it will eventually become popular. This could be linked to recent works that have shown how simple mathematical models can be accurately describe processes of topics and memes that compete for collective attention, displaying bursts and decays [388, 393–395]. In a novelty-driven scenario, every time that a walker finds a novelty, she could consequently updates the item to spread to this one, select a neighbour and try to spread it. If adopted, the neighbour would immediately jump to the correspondent node on the network of items, ultimately adding flights to the random walks process [6].

The results presented in Chapter 3 highlight the importance of considering appropriate representations to model the dynamics of higher-order systems (HOrSs). The model of simplicial contagion represents only the beginning of what could be a long journey.

Dynamical processes beyond pairwise interactions. Many dynamical processes could be studied moving from pairwise to higher-order approaches [5]. This is potentially relevant to all classes of dynamical processes notoriously studied on complex structures. Developing and studying such systems might also allow to better take into account higher-order dynamical effects in real data-driven models. Recent works in this direction include higher-order extension of the Kuramoto model of synchronisation [396–399] and studies of synchronization on complex network manifolds [277, 278, 400–402]. Besides, early investigations of other dynamical processes further confirmed the key role played by HOrSs, such as for the social dynamics of consensus and cooperation [403–405], but also for other landmark dynamical processes like diffusion and random walks [406, 407].

HOrS and their representation. What is a truly genuine higher-order interaction? We already mentioned that for some systems this question is relatively easy to address. For example, data on co-authorship in publication where each paper constitutes an interaction among all authors [65] naturally come in the form of sets, and the higher-order representation can thus be adopted straightforwardly. However, there are many other systems, such as for people engaging in conversations or colleagues working in a team [408, 409], in which group interactions are not so easily identifiable and would require the design and implementation of ad-hoc data collection experiments. These novel experiments will also have to be designed to record temporal traces of social behaviour along with the patterns of interactions underneath. Having access to both the underlying—evolving—structure, and the data on the dynamical process unfolding upon it, is essential to fully understand and model the behaviour of social systems as HOrS. Behind these methodological difficulties, stands a cardinal question of an ontological nature. Once one chooses a suitable representation around a sensibly collected data set of human interactions, how does one characterize what is obtained? This is already challenging in theoretical models, but how do we tease them apart in data, and what data do we need?

HOrS and the problem with the data. Observational data, such as digital traces collected through mobile phone devices, provide an always-on picture of human behaviour that has been widely exploited by social researchers. As a matter of fact, these traces have been successfully used to find evidence of (complex) social contagions mechanisms, such as for the adoption of applications [114, 116] and the spreading of running habits [129], that complemented results obtained with more traditional methods (surveys) [107]. Data sets of this type—assuming their availability—are intrinsically ill-suited to answer HOrSs-related questions, since the patterns of interactions therein are not encoded in a higher-order fashion. With the raw data already encoded in a pairwise representation, moving to a higher-order one would require adding structural assumptions that would invalidate most of the HOrS-related research questions.

Following this path, the first option would be to leverage high-resolution data coming from already performed measurements. In fact, standard call detail records are not appropriate given their poor spatial resolution and their heterogeneous sampling biases. A number of large-scale data collection experiments have already been performed to overcome these limitations. These studies make use of modern social sensing technologies to track the proximity of individuals, a good proxy for measuring social—group—interactions. For example, physical proximity between mobile phone users—detected via Bluetooth signal strength—has been used to measure the strength of friendship ties [410] and to unveil the structural patterns in longitudinal data sets [411]. Although these data, released by the Copenhagen Networks Study [412], would represent a good testing ground for higher-order representational algorithms, they miss one essential component, that is information about social phenomena unfolding on top of the social structure.

Alternatively, one could use radio-frequency identification devices (RFID) as those deployed by the Sociopatterns collaboration [413]. In these measurements, devices are fine-tuned to capture possible routes for respiratory droplet transmission [414] via face-to-face contacts, but can also offer an adequate description of person-to-person interactions (see Fig. 3.2) from Chapter 3 where we used them to construct empirical simplicial complexes [2]). While customizing these last technologies could improve the extraction of higher-order interactions, the currently available data, designed for a different purpose, lack, again, information on any social phenomenon unfolding on the captured contacts. As such, to track both the dynamical behaviour of individuals and the many-body interactions underneath ad-hoc experiments are needed.

Computational challenges for HOrS. Finally, moving from pairwise to higher-order interactions opens up also new algorithmic and computational challenges. The first one is about efficiency in storage. For example, while matrix representations can be very convenient for extracting properties, the amount of absent relationships they contain might explode as the dimension of the considered interactions increases. Thus, just like lists of edges represent the most compact representation for pairwise interactions, it is possible to efficiently compress and sparsely store HOrS using lists of simplices. However, this representation comes trades off memory and access efficiency. Think, for example, at a simple check for the existence of a specific 2-body interaction from the list of maximal simplices representing the entire complex. This difficulty is particularly relevant when using HOrSs as a substrate for simulations of dynamical processes. In addition, many standard algorithms for networks can not be simply extended, since (i) they would result too slow, and (ii) the standard dichotomy of active/inactive links—extensively used in network algorithms—lacks a counterpart for HOrS, where a k-simplex can be "active" at k different levels. This, combined with the combinatorial necessity of adapting the dynamics to all the relevant interactions within each simplex, calls for ad-hoc algorithmic implementations of dynamical processes on higher-order structures. A way to tackle the problem with a bottom-up approach would be to extract in advance the essential structural information from a HOrS while maintaining its richness. For example, akin to Serrano et al. [415], is there a way of extracting a statistically significant backbone from HOrSs? More interestingly, one could develop order-reduction methods specifically tailored to preserve the meaningful higher-order structures that are relevant to the process under study. To achieve this, a profound rethinking of standard pairwise approaches is necessary.

References

[1] Iacopo Iacopini, Staša Milojević, and Vito Latora. Network dynamics of innovation processes. *Phys. Rev. Lett.*, 120(4):048301, 2018.

[2] Iacopo Iacopini, Giovanni Petri, Alain Barrat, and Vito Latora. Simplicial models of social contagion. *Nat. Commun.*, 10(1):2485, 2019.

[3] Iacopo Iacopini, Benjamin Schäfer, Elsa Arcaute, Christian Beck, and Vito Latora. Multilayer modeling of adoption dynamics in energy demand management. *Chaos*, 30(1):013153, 2020.

[4] Iacopo Iacopini, Gabriele Di Bona, Enrico Ubaldi, Vittorio Loreto, and Vito Latora. Interacting discovery processes on complex networks. *Phys. Rev. Lett.*, 125:248301, 2020.

[5] Federico Battiston, Giulia Cencetti, Iacopo Iacopini, Vito Latora, Maxime Lucas, Alice Patania, Jean-Gabriel Young, and Giovanni Petri. Networks beyond pairwise interactions: structure and dynamics. *Phys. Rep.*, 874:1–92, 2020.

[6] Iacopo Iacopini and Vito Latora. On the dual nature of adoption processes in complex networks. *Front. Phys.*, 9:109, 2021.

[7] Savina Torrisi, Sabato Manfredi, Iacopo Iacopini, and Vito Latora. Creative connectivity project-a network based approach to understand correlations between interdisciplinary group dynamics and creative performance. *E&PDE 2019 Towards a New Innovation Lanscape*, pages 530–535, 2019.

[8] Alba Bernini, Elodie Blouzard, Alberto Bracci, Pau Casanova, Iacopo Iacopini, Benjamin Steinegger, Andreia Sofia Teixeira, Alberto Antonioni, and Eugenio

Valdano. Evaluating the impact of prep on hiv and gonorrhea on a networked population of female sex workers. *arXiv preprint arXiv:1906.09085*, 2019.

[9] Maarten Vanhoof, Antonia Godoy-Lorite, Roberto Murcio, Iacopo Iacopini, Natalia Zdanowska, Juste Raimbault, Richard Milton, Elsa Arcaute, and Mike Batty. Using foursquare data to reveal spatial and temporal patterns in london. 2019.

[10] Michael Batty, Roberto Murcio, Iacopo Iacopini, Maarten Vanhoof, and Richard Milton. London in lockdown: Mobility in the pandemic city. *arXiv preprint arXiv:2011.07165*, 2020.

[11] Kirsten Martinus, Thomas Sigler, Iacopo Iacopini, and Ben Derudder. The brokerage role of small states and territories in global corporate networks. *Growth and Change*, 52(1):12–28, 2021.

[12] Thomas Sigler, Kirsten Martinus, Iacopo Iacopini, and Ben Derudder. The role of tax havens and offshore financial centres in shaping corporate geographies: an industry sector perspective. *Reg. Stud.*, pages 1–13, 2019.

[13] Kirsten Martinus, Thomas Sigler, Iacopo Iacopini, and Ben Derudder. The role of tax havens and offshore financial centers in asia-pacific networks: evidence from firm-subsidiary connections. *Asian Bus. Manag.*, 18(5):389–411, 2019.

[14] Alessandro Vespignani. Twenty years of network science, 2018.

[15] Mark Newman. *Networks*. Oxford university press, 2018.

[16] Stefano Boccaletti, Vito Latora, Yamir Moreno, Martin Chavez, and D-U Hwang. Complex networks: Structure and dynamics. *Phys. Rep.*, 424(4):175–308, 2006.

[17] Albert-László Barabási et al. *Network science*. Cambridge university press, 2016.

[18] Vito Latora, Vincenzo Nicosia, and Giovanni Russo. *Complex Networks: Principles, Methods and Applications*. Cambridge University Press, 2017.

[19] John Adrian Bondy, Uppaluri Siva Ramachandra Murty, et al. *Graph theory with applications*, volume 290. Macmillan London, 1976.

[20] László Lovász. *Large networks and graph limits*, volume 60. American Mathematical Soc., 2012.

[21] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74(1):47, 2002.

[22] M. A. Porter and J. P. Gleeson. *Dynamical systems on networks: A tutorial*. Springer, 2005.

[23] Alain Barrat, Marc Barthelemy, and Alessandro Vespignani. *Dynamical processes on complex networks*. Cambridge university press, 2008.

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

[25] Duncan J Watts and Steven H Strogatz. Collective dynamics of ?small-world?networks. *Nature*, 393(6684):440, 1998.

[26] Stanley Wasserman, Katherine Faust, et al. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

[27] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annu. Rev. Sociol.*, 27(1):415–444, 2001.

[28] Stephen P Borgatti, Ajay Mehra, Daniel J Brass, and Giuseppe Labianca. Network analysis in the social sciences. *Science*, 323(5916):892–895, 2009.

[29] Matthew O Jackson. *Social and economic networks*. Princeton university press, 2010.

[30] David Easley, Jon Kleinberg, et al. *Networks, crowds, and markets*, volume 8. Cambridge university press Cambridge, 2010.

[31] Uri Alon. Biological networks: the tinkerer as an engineer. *Science*, 301(5641):1866–1867, 2003.

[32] José M Montoya, Stuart L Pimm, and Ricard V Solé. Ecological networks and their fragility. *Nature*, 442(7100):259–264, 2006.

[33] Jacopo Grilli, György Barabás, Matthew J Michalska-Smith, and Stefano Allesina. Higher-order interactions stabilize dynamics in competitive network models. *Nature*, 548(7666):210–213, 2017.

[34] Ed Bullmore and Olaf Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.*, 10(3):186–198, 2009.

[35] Danielle S Bassett and Olaf Sporns. Network neuroscience. *Nat. Neurosci.*, 20(3):353, 2017.

[36] Petter Holme and Jari Saramäki. Temporal networks. *Phys. Rep.*, 519(3):97–125, 2012.

[37] Filippo Radicchi and Alex Arenas. Abrupt transition in the structural formation of interconnected networks. *Nat. Phys.*, 9(11):717, 2013.

[38] Stefano Boccaletti, Ginestra Bianconi, Regino Criado, Charo I Del Genio, Jesús Gómez-Gardenes, Miguel Romance, Irene Sendina-Nadal, Zhen Wang, and Massimiliano Zanin. The structure and dynamics of multilayer networks. *Phys. Rep.*, 544(1):1–122, 2014.

[39] Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P Gleeson, Yamir Moreno, and Mason A Porter. Multilayer networks. *J. Complex Netw.*, 2(3):203–271, 2014.

[40] Carter T Butts. Revisiting the foundations of network analysis. *Science*, 325(5939):414–416, 2009.

[41] Leo Torres, Ann S. Blevins, Danielle S. Bassett, and Tina Eliassi-Rad. The why, how, and when of representations for complex systems. *arXiv:2006.02870*, 2020.

[42] Jonathan M Levine, Jordi Bascompte, Peter B Adler, and Stefano Allesina. Beyond pairwise mechanisms of species coexistence in complex communities. *Nature*, 546(7656):56–64, 2017.

[43] Chad Giusti, Eva Pastalkova, Carina Curto, and Vladimir Itskov. Clique topology reveals intrinsic geometric structure in neural correlations. *Pro. Natl. Acad Sci. U.S.A.*, 112(44):13455–13460, 2015.

[44] Ann E Sizemore, Chad Giusti, Ari Kahn, Jean M Vettel, Richard F Betzel, and Danielle S Bassett. Cliques and cavities in the human connectome. *J. Comp. Neurosci.*, 44(1):115–145, 2018.

[45] Andreas Ruepp, Brigitte Waegele, Martin Lechner, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Gisela Fobo, Goar Frishman, Corinna Montrone, and H-Werner Mewes. Corum: the comprehensive resource of mammalian protein complexes—2009. *Nucleic acids Res.*, 38(suppl_1):D497–D501, 2010.

[46] Elena Kuzmin, Benjamin VanderSluis, Wen Wang, Guihong Tan, Raamesh Deshpande, Yiqun Chen, Matej Usaj, Attila Balint, Mojca Mattiazzi Usaj, Jolanda Van Leeuwen, et al. Systematic analysis of complex genetic interactions. *Science*, 360(6386):eaao1729, 2018.

[47] Claudio Castellano, Santo Fortunato, and Vittorio Loreto. Statistical physics of

social dynamics. *Rev. Mod. Phys.*, 81(2):591, 2009.

[48] Parongama Sen and Bikas K Chakrabarti. *Sociophysics: An Introduction*. Oxford University Press, 2014.

[49] J Miller McPherson. Hypernetwork sampling: Duality and differentiation among voluntary organizations. *Soc. Netw.*, 3(4):225–249, 1982.

[50] Pavel S Aleksandrov. *Combinatorial topology*, volume 1. Courier Corporation, 1998.

[51] Allen Hatcher. *Algebraic topology*. 2002.

[52] Vsevolod Salnikov, Daniele Cassese, and Renaud Lambiotte. Simplicial complexes and complex systems. *European Journal of Physics*, 40(1):014001, 2018.

[53] Renaud Lambiotte, Martin Rosvall, and Ingo Scholtes. From networks to optimal higher-order models of complex systems. *Nat. Phys.*, 15(4):313–320, 2019.

[54] Chad Giusti, Robert Ghrist, and Danielle S Bassett. Two's company, three (or more) is a simplex. *J. Comp. Neurosci.*, 41(1):1–14, 2016.

[55] Mason A Porter. Nonlinearity+ networks: A 2020 vision. In *Emerging Frontiers in Nonlinear Science*, pages 131–159. Springer, 2020.

[56] Giovanni Petri, Martina Scolamiero, Irene Donato, and Francesco Vaccarino. Topological strata of weighted complex networks. *PLoS One*, 8(6):e66506, 2013.

[57] Ann Sizemore, Chad Giusti, and Danielle S Bassett. Classification of weighted networks through mesoscale homological features. *J. Comp. Net.*, 5(2):245–273, 2016.

[58] Alexander P Kartun-Giles and Ginestra Bianconi. Beyond the clustering coefficient: A topological analysis of node neighbourhoods in complex networks. *Chaos, Solitons & Fract. X*, 1:100004, 2019.

[59] Giovanni Petri, Paul Expert, Federico Turkheimer, Robin Carhart-Harris, David Nutt, Peter J Hellyer, and Francesco Vaccarino. Homological scaffolds of brain functional networks. *J. Royal Soc. Interface*, 11(101):20140873, 2014.

[60] Louis-David Lord, Paul Expert, Henrique M Fernandes, Giovanni Petri, Tim J Van Hartevelt, Francesco Vaccarino, Gustavo Deco, Federico Turkheimer, and Morten L Kringelbach. Insights into brain architectures from the homological scaffolds of functional connectivity networks. *Front. Syst. Neurosci.*, 10:85, 2016.

[61] Hyekyoung Lee, Hyejin Kang, Moo K Chung, Bung-Nyun Kim, and Dong Soo Lee. Persistent brain network homology from the perspective of dendrogram. *IEEE Trans. Med. Imaging.*, 31(12):2267–2277, 2012.

[62] Ernesto Estrada and Grant J Ross. Centralities in simplicial complexes. applications to protein interaction networks. *J. Theor. Biol.*, 438:46–60, 2018.

[63] Ann E Sizemore, Elisabeth A Karuza, Chad Giusti, and Danielle S Bassett. Knowledge gaps in the early growth of semantic feature networks. *Nat. Hum. Behav.*, 2(9):682, 2018.

[64] Austin R Benson, David F Gleich, and Jure Leskovec. Higher-order organization of complex networks. *Science*, 353(6295):163–166, 2016.

[65] Alice Patania, Giovanni Petri, and Francesco Vaccarino. The shape of collaborations. *EPJ Data Sci.*, 6(1):18, 2017.

[66] Kerk F Kee, Lisa Sparks, Daniele C Struppa, and Mirco Mannucci. Social groups, social media, and higher dimensional social structures: A simplicial model of social aggregation for computational communication research. *Communication Quarterly*, 61(1):35–58, 2013.

[67] Mark S Granovetter. The strength of weak ties. In *Social networks*, pages 347–367. Elsevier, 1977.

[68] Thomas W Valente. *Network models of the diffusion of innovations*. Number 303.484 V3. Hampton Press., 1995.

[69] Andrzej Nowak, Jacek Szamrej, and Bibb Latané. From private attitude to public opinion: A dynamic theory of social impact. *Psychol. Rev.*, 97(3):362, 1990.

[70] Robert Axelrod. *The complexity of cooperation: Agent-based models of competition and collaboration*, volume 3. Princeton University Press, 1997.

[71] Serge Galam. Sociophysics: a review of galam models. *Int. J. Mod. Phys. C*, 19(03):409–440, 2008.

[72] David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. Computational social science. *Science*, 323(5915):721–723, 2009.

[73] Serge Galam. What is sociophysics about? In *Sociophysics*, pages 3–19. Springer,

2012.

[74] Rosaria Conte, Nigel Gilbert, Giulia Bonelli, Claudio Cioffi-Revilla, Guillaume Deffuant, Janos Kertesz, Vittorio Loreto, Suzy Moat, J-P Nadal, Anxo Sanchez, et al. Manifesto of computational social science. *Eur. Phys. J. Spec. Top.*, 214(1):325–346, 2012.

[75] Andrea Baronchelli. The emergence of consensus: A primer. *R. Soc Open Sci*, 5(2):172189, 2018.

[76] William Goffman and Vaun A Newill. Generalization of epidemic theory: An application to the transmission of ideas. *Nature*, 204(4955):225–228, 1964.

[77] Romualdo Pastor-Satorras, Claudio Castellano, Piet Van Mieghem, and Alessandro Vespignani. Epidemic processes in complex networks. *Rev. Mod. Phys.*, 87(3):925, 2015.

[78] Frank M Bass. A new product growth for model consumer durables. *Manag. Sci.*, 15(5):215–227, 1969.

[79] Damon Centola and Michael Macy. Complex contagions and the weakness of long ties. *Americ. J. Sociol.*, 113(3):702–734, 2007.

[80] Damon Centola. The spread of behavior in an online social network experiment. *Science*, 329(5996):1194–1197, 2010.

[81] Dale Zhou, David M Lydon-Staley, Perry Zurn, and Danielle S Bassett. The growth and form of knowledge networks by kinesthetic curiosity. *arXiv preprint arXiv:2006.02949*, 2020.

[82] Francesca Tria, Vittorio Loreto, Vito Domenico Pietro Servedio, and Steven H Strogatz. The dynamics of correlated novelties. *Sci. Rep.*, 4:5890, 2014.

[83] Cristina Bicchieri. *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press, 2005.

[84] H Peyton Young. The evolution of social norms. *Annual Review of Economics*, 7(1):359–387, 2015.

[85] R Kelly Garrett. Echo chambers online?: Politically motivated selective exposure among internet news users. *J. Comput.-Mediat. Commun.*, 14(2):265–285, 2009.

[86] Seth Flaxman, Sharad Goel, and Justin M Rao. Filter bubbles, echo chambers, and

online news consumption. *Public opinion quarterly*, 80(S1):298–320, 2016.

[87] Michele Starnini, Mattia Frasca, and Andrea Baronchelli. Emergence of metapopulations and echo chambers in mobile agents. *Sci. Rep.*, 6:31834, 2016.

[88] Fabian Baumann, Philipp Lorenz-Spreen, Igor M Sokolov, and Michele Starnini. Modeling echo chambers and polarization dynamics in social networks. *Phys. Rev. Lett.*, 124(4):048301, 2020.

[89] Douglas Guilbeault, Joshua Becker, and Damon Centola. Complex contagions: A decade in review. In *Complex Spreading Phenomena in Social Systems*, pages 3–25. Springer, 2018.

[90] William Ogilvy Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. *Proc. R. Soc. Lond. A*, 115(772):700–721, 1927.

[91] Roy M Anderson and Robert M May. *Infectious diseases of humans: dynamics and control*. Oxford university press, 1992.

[92] Herbert W Hethcote. The mathematics of infectious diseases. *SIAM Rev.*, 42(4):599–653, 2000.

[93] Qian Zhang, Kaiyuan Sun, Matteo Chinazzi, Ana Pastore y Piontti, Natalie E Dean, Diana Patricia Rojas, Stefano Merler, Dina Mistry, Piero Poletti, Luca Rossi, et al. Spread of zika virus in the americas. *Proc. Natl. Acad. Sci. U. S. A.*, 114(22):E4334–E4343, 2017.

[94] Ana Pastore y Piontti, Nicola Perra, Luca Rossi, Nicole Samay, and Alessandro Vespignani. *Charting the Next Pandemic: Modeling Infectious Disease Spreading in the Data Science Age*. Springer, 2018.

[95] Cécile Viboud and Alessandro Vespignani. The future of influenza forecasts. *Proc. Natl. Acad. Sci. U. S. A.*, 116(8):2802–2804, 2019.

[96] Adam J Kucharski, Timothy W Russell, Charlie Diamond, Yang Liu, John Edmunds, Sebastian Funk, Rosalind M Eggo, Fiona Sun, Mark Jit, James D Munday, et al. Early dynamics of transmission and control of covid-19: a mathematical modelling study. *Lancet Infect. Dis.*, 2020.

[97] Moritz U. G. Kraemer, Chia-Hung Yang, Bernardo Gutierrez, Chieh-Hsi Wu, Brennan Klein, David M. Pigott, Louis du Plessis, Nuno R. Faria, Ruoran Li,

William P. Hanage, John S. Brownstein, Maylis Layan, Alessandro Vespignani, Huaiyu Tian, Christopher Dye, Oliver G. Pybus, and Samuel V. Scarpino. The effect of human mobility and control measures on the covid-19 epidemic in china. *Science*, 2020.

[98] Manlio De Domenico, Clara Granell, Mason A Porter, and Alex Arenas. The physics of spreading processes in multilayer networks. *Nat. Phys.*, 12(10):901–906, 2016.

[99] James P Gleeson. High-accuracy approximation of binary-state dynamics on networks. *Phys. Rev. Lett.*, 107(6):068701, 2011.

[100] James P Gleeson. Binary-state dynamics on complex networks: Pair approximation and beyond. *Phys. Rev. X*, 3(2):021004, 2013.

[101] Wei Wang, Ming Tang, H Eugene Stanley, and Lidia A Braunstein. Unification of theoretical approaches for epidemic spreading on complex networks. *Rep. Prog. Phys.*, 80(3):036603, 2017.

[102] István Z Kiss, Joel C Miller, Péter L Simon, et al. Mathematics of epidemics on networks. *Cham: Springer*, 2017.

[103] Daryl J Daley and David G Kendall. Epidemics and rumours. *Nature*, 204(4963):1118–1118, 1964.

[104] Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. A theory of fads, fashion, custom, and cultural change as informational cascades. *J. Political Econ.*, 100(5):992–1026, 1992.

[105] Everett M Rogers. *Diffusion of innovations*. Simon and Schuster, 2010.

[106] Jessica T Davis, Nicola Perra, Qian Zhang, Yamir Moreno, and Alessandro Vespignani. Phase transitions in information spreading on structured populations. *Nat. Phys.*, pages 1–7, 2020.

[107] Nicholas A Christakis and James H Fowler. The collective dynamics of smoking in a large social network. *N. Engl. J. Med.*, 358(21):2249–2258, 2008.

[108] Nicholas A Christakis and James H Fowler. The spread of obesity in a large social network over 32 years. *N. Engl. J. Med.*, 357(4):370–379, 2007.

[109] James H Fowler and Nicholas A Christakis. Dynamic spread of happiness in a

large social network: longitudinal analysis over 20 years in the framingham heart study. *Bmj*, 337:a2338, 2008.

[110] Nathan O Hodas and Kristina Lerman. The simple rules of social contagion. *Sci. Rep.*, 4:4343, 2014.

[111] Sinan Aral, Lev Muchnik, and Arun Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proc. Natl. Acad. Sci. U.S.A.*, 106(51):21544–21549, 2009.

[112] Sinan Aral and Dylan Walker. Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Manage. Sci.*, 57(9):1623–1639, 2011.

[113] Flavio L Pinheiro, Marta D Santos, Francisco C Santos, and Jorge M Pacheco. Origin of peer influence in social networks. *Phys. Rev. Lett.*, 112(9):098702, 2014.

[114] Jukka-Pekka Onnela and Felix Reed-Tsochas. Spontaneous emergence of social influence in online systems. *Proceedings of the National Academy of Sciences*, 107(43):18375–18380, 2010.

[115] Bjarke Mønsted, Piotr Sapieżyński, Emilio Ferrara, and Sune Lehmann. Evidence of complex contagion of information in social media: An experiment using twitter bots. *PLoS One*, 12(9):e0184148, 2017.

[116] Márton Karsai, Gerardo Iniguez, Kimmo Kaski, and János Kertész. Complex contagion process in spreading of online innovation. *J. R. Soc. Interface*, 11(101):20140694, 2014.

[117] Oriana Bandiera and Imran Rasul. Social networks and technology adoption in northern mozambique. *Econ. J.*, 116(514):869–902, 2006.

[118] Emily Oster and Rebecca Thornton. Determinants of technology adoption: Peer effects in menstrual cup take-up. *J. Eur. Econ. Assoc.*, 10(6):1263–1293, 2012.

[119] John Ternovski and Taha Yasseri. Social complex contagion in music listenership: A natural experiment with 1.3 million participants. *Soc. Netw.*, 61:144–152, 2020.

[120] Damon Centola. *How behavior spreads: The science of complex contagions*, volume 3. Princeton University Press, 2018.

[121] Duncan J. Watts. A simple model of global cascades on random networks. *Pro.*

Natl. Acad Sci. U.S.A., 99(9):5766–5771, 2002.

[122] Sergey Melnik, Jonathan A Ward, James P Gleeson, and Mason A Porter. Multi-stage complex contagions. *Chaos*, 23(1):013124, 2013.

[123] Zhongyuan Ruan, Gerardo Iniguez, Márton Karsai, and János Kertész. Kinetics of social contagion. *Phys. Rev. Lett.*, 115(21):218702, 2015.

[124] Agnieszka Czaplicka, Raul Toral, and Maxi San Miguel. Competition of simple and complex adoption on interdependent networks. *Phys. Rev. E*, 94(6):062301, 2016.

[125] Lucas Böttcher, Jan Nagler, and Hans J Herrmann. Critical behaviors in contagion dynamics. *Phys. Rev. Lett.*, 118(8):088301, 2017.

[126] Johan Ugander, Lars Backstrom, Cameron Marlow, and Jon Kleinberg. Structural diversity in social contagion. *Pro. Natl. Acad Sci. U.S.A.*, page 201116502, 2012.

[127] Sune Lehmann and Yong-Yeol Ahn. *Complex spreading phenomena in social systems*. Springer, 2018.

[128] Sinan Aral and Dylan Walker. Tie strength, embeddedness, and social influence: A large-scale networked experiment. *Manage. Sci.*, 60(6):1352–1370, 2014.

[129] Sinan Aral and Christos Nicolaides. Exercise contagion in a global social network. *Nat. Commun.*, 8(1):1–8, 2017.

[130] Guilherme Ferraz de Arruda, Giovanni Petri, Francisco A Rodrigues, and Yamir Moreno. Impact of the distribution of recovery rates on disease spreading in complex networks. *Phys. Rev. Res.*, 2(1):013046, 2020.

[131] Alexandre Darbon, Davide Colombi, Eugenio Valdano, Lara Savini, Armando Giovannini, and Vittoria Colizza. Disease persistence on temporal contact networks accounting for heterogeneous infectious periods. *Royal Society open science*, 6(1):181404, 2019.

[132] Terry Brett, George Loukas, Yamir Moreno, and Nicola Perra. Spreading of computer viruses on time-varying networks. *Phys. Rev. E*, 99(5):050303, 2019.

[133] David Strang and Nancy Brandon Tuma. Spatial and temporal heterogeneity in diffusion. *Am. J. Sociol.*, 99(3):614–639, 1993.

[134] Jameson L Toole, Meeyoung Cha, and Marta C González. Modeling the adoption

of innovations in the presence of geographic and media influences. *PloS one*, 7(1):e29528, 2012.

[135] Balázs Lengyel, Eszter Bokányi, Riccardo Di Clemente, János Kertész, and Marta C González. The role of geography in the complex diffusion of innovations. *Sci. Rep.*, 10(1):1–11, 2020.

[136] Guilherme Ferraz de Arruda, Giovanni Petri, and Yamir Moreno. Social contagion models on hypergraphs. *Phys. Rev. Res.*, 2(2):023032, 2020.

[137] Bukyoung Jhun, Minjae Jo, and B Kahng. Simplicial sis model in scale-free uniform hypergraph. *J. Stat. Mech.: Theory Exp* ., 2019(12):123207, 2019.

[138] Guilherme Ferraz de Arruda, Michele Tizzani, and Yamir Moreno. Phase transitions and stability of dynamical processes on hypergraphs. *Commun.Phys.*, 4(1):1–9, 2021.

[139] Nicholas W Landry and Juan G Restrepo. The effect of heterogeneity on hypergraph contagion models. *Chaos*, 30(10):103117, 2020.

[140] Thomas S Kuhn and Joseph Epstein. The essential tension, 1979.

[141] Luca Pappalardo, Filippo Simini, Salvatore Rinzivillo, Dino Pedreschi, Fosca Giannotti, and Albert-Lászlό Barabási. Returners and explorers dichotomy in human mobility. *Nat. Commun.*, 6:8166, 2015.

[142] Xiao-Yong Yan, Wen-Xu Wang, Zi-You Gao, and Ying-Cheng Lai. Universal model of individual and population mobility on diverse spatial scales. *Nat. Commun.*, 8(1):1–9, 2017.

[143] Laura Alessandretti, Sune Lehmann, and Andrea Baronchelli. Understanding the interplay between social and spatial behaviour. *EPJ Data Sci.*, 7(1):36, 2018.

[144] Laura Alessandretti, Piotr Sapiezynski, Vedran Sekara, Sune Lehmann, and Andrea Baronchelli. Evidence for a conserved quantity in human mobility. *Nat. Hum. Behav.*, 2(7):485–491, 2018.

[145] Riccardo Di Clemente, Miguel Luengo-Oroz, Matias Travizano, Sharon Xu, Bapu Vaitla, and Marta C González. Sequences of purchases in credit card data reveal lifestyles in urban populations. *Nat. Commun.*, 9(1):1–8, 2018.

[146] Luca Maria Aiello, Rossano Schifanella, Daniele Quercia, and Lucia Del Prete.

Large-scale and high-resolution analysis of food purchases and health outcomes. *EPJ Data Sci.*, 8(1):14, 2019.

[147] Luca Maria Aiello, Daniele Quercia, Rossano Schifanella, and Lucia Del Prete. Tesco grocery 1.0, a large-scale dataset of grocery purchases in london. *Sci. Data*, 7(1):1–11, 2020.

[148] S. Johnson. *Where Good Ideas Come From: The Natural History of Innovation*. Penguin Books Limited, 2010.

[149] Jacob G Foster, Andrey Rzhetsky, and James A Evans. Tradition and innovation in scientists' research strategies. *Am. Sociol. Rev.*, 80(5):875–908, 2015.

[150] Aaron Clauset, Daniel B Larremore, and Roberta Sinatra. Data-driven predictions in the science of science. *Science*, 355(6324):477–480, 2017.

[151] Jaimie Murdock, Colin Allen, and Simon DeDeo. Exploration and exploitation of victorian science in darwin's reading notebooks. *Cognition*, 159:117–126, 2017.

[152] Alberto Aleta, Sandro Meloni, Nicola Perra, and Yamir Moreno. Explore with caution: mapping the evolution of scientific interest in physics. *EPJ Data Science*, 8(1):1–15, 2019.

[153] Bas Hofstra, Vivek V Kulkarni, Sebastian Munoz-Najar Galvez, Bryan He, Dan Jurafsky, and Daniel A McFarland. The diversity–innovation paradox in science. *Proc. Natl. Acad. Sci. U.S.A.*, 2020.

[154] Mario Coccia. Driving forces of technological change: the relation between population growth and technological innovation: analysis of the optimal interaction across countries. *Technol. Forecast. Soc. Change*, 82:52–65, 2014.

[155] Anton Pichler, François Lafond, and J Doyne Farmer. Technological interdependencies predict innovation dynamics. *arXiv preprint arXiv:2003.00580*, 2020.

[156] TMA Fink, M Reeves, R Palma, and RS Farr. Serendipity and strategy in rapid innovation. *Nat. Commun.*, 8(1):2002, 2017.

[157] T. M. A. Fink and M. Reeves. How much can we influence the rate of innovation? *Science Advances*, 5(1), 2019.

[158] Vishal Sood, Myléne Mathieu, Amer Shreim, Peter Grassberger, and Maya Paczuski. Interacting branching process as a simple model of innovation. *Phys. Rev. Lett.*,

105(17):178701, 2010.

[159] Hyejin Youn, Deborah Strumsky, Luis MA Bettencourt, and José Lobo. Invention as a combinatorial process: evidence from us patents. *J. R. Soc. Interface*, 12(106):20150272, 2015.

[160] Douglas H Erwin and David C Krakauer. Insights into innovation. *Science*, 304(5674):1117–1119, 2004.

[161] Andrea Tacchella, Andrea Napoletano, and Luciano Pietronero. The language of innovation. *PloS One*, 15(4):1–20, 04 2020.

[162] Ching Jin, Chaoming Song, Johannes Bjelland, Geoffrey Canright, and Dashun Wang. Emergence of scaling in complex substitutive systems. *Nat. Hum. Behav.*, 3(8):837–846, 2019.

[163] Stefan Thurner, Peter Klimek, and Rudolf Hanel. Schumpeterian economic dynamics as a quantifiable model of evolution. *New J. Phys.*, 12(7):075029, 2010.

[164] Edilson A Corrêa Jr, Vanessa Q Marinho, and Diego R Amancio. Semantic flow in language networks. *arXiv preprint arXiv:1905.07595*, 2019.

[165] Christopher W Lynn, Lia Papadopoulos, Ari E Kahn, and Danielle S Bassett. Human information processing in complex networks. *Nat.Phys.*, pages 1–9, 2020.

[166] Andrey Rzhetsky, Jacob G Foster, Ian T Foster, and James A Evans. Choosing experiments to accelerate collective discovery. *Proc. Natl. Acad. Sci. U.S.A.*, 112(47):14569–14574, 2015.

[167] Perry Zurn and Danielle S Bassett. Network architectures supporting learnability. *Philos. T. R. Soc. B*, 375(1796):20190323, 2020.

[168] Ciro Cattuto, Alain Barrat, Andrea Baldassarri, Gregory Schehr, and Vittorio Loreto. Collective dynamics of social annotation. *Proc. Natl. Acad. Sci. U.S.A.*, 106(26):10511–10515, 2009.

[169] GC Rodi, V Loreto, VDP Servedio, and F Tria. Optimal learning paths in information networks. *Sci. Rep.*, 5:10286, 2015.

[170] Henrique F de Arruda, Filipi N Silva, Luciano da F Costa, and Diego R Amancio. Knowledge acquisition: A complex networks approach. *Information Sciences*, 421:154–166, 2017.

[171] Thales S Lima, Henrique F de Arruda, Filipi N Silva, Cesar H Comin, Diego R Amancio, and Luciano da F Costa. The dynamics of knowledge acquisition via self-learning in complex networks. *Chaos*, 28(8):083106, 2018.

[172] Giovanna Chiara Rodi, Vittorio Loreto, and Francesca Tria. Search strategies of wikipedia readers. *PLoS One*, 12(2):e0170746, 2017.

[173] David M Lydon-Staley, Dale Zhou, Ann Sizemore Blevins, Perry Zurn, and Danielle S Bassett. Hunters, busybodies and the knowledge network building associated with deprivation curiosity. *Nat. Hum. Behav.*, pages 1–10, 2020.

[174] Stuart A Kauffman. Investigations on the character of autonomous agents and the worlds they mutually create. Santa Fe Institute, 1996.

[175] James McNerney, J Doyne Farmer, Sidney Redner, and Jessika E Trancik. Role of design complexity in technology improvement. *Proc. Natl. Acad. Sci. U. S. A.*, 108(22):9008–9013, 2011.

[176] Luís MA Bettencourt, Jessika E Trancik, and Jasleen Kaur. Determinants of the pace of global innovation in energy technologies. *PloS One*, 8(10), 2013.

[177] Curtis H Weiss, Julia Poncela-Casasnovas, Joshua I Glaser, Adam R Pah, Stephen D Persell, David W Baker, Richard G Wunderink, and Luís A Nunes Amaral. Adoption of a high-impact innovation in a homogeneous population. *Phys. Rev. X*, 4(4):041008, 2014.

[178] Roberta Sinatra, Dashun Wang, Pierre Deville, Chaoming Song, and Albert-László Barabási. Quantifying the evolution of individual scientific impact. *Science*, 354(6312):aaf5239, 2016.

[179] Miroslav Andjelković, Bosiljka Tadić, Marija Mitrović Dankulov, Milan Rajković, and Roderick Melnik. Topology of innovation spaces in the knowledge networks emerging through questions-and-answers. *PLoS One*, 11(5):e0154655, 2016.

[180] Fabio Saracco, Riccardo Di Clemente, Andrea Gabrielli, and Luciano Pietronero. From innovation to diversification: a simple competitive model. *PLoS One*, 10(11):e0140420, 2015.

[181] Andrea Puglisi, Andrea Baronchelli, and Vittorio Loreto. Cultural route to the emergence of linguistic categories. *Proc. Natl. Acad. Sci. U.S.A.*, 105(23):7936–7940,

2008.

[182] Sameet Sreenivasan. Quantitative analysis of the evolution of novelty in cinema through crowdsourced keywords. *Sci. Rep.*, 3:2758, 2013.

[183] Martin Gerlach and Eduardo G Altmann. Stochastic model for the vocabulary growth in natural languages. *Phys. Rev. X*, 3(2):021006, 2013.

[184] Linyuan Lü, Zi-Ke Zhang, and Tao Zhou. Deviation of zipf's and heaps' laws in human languages with limited dictionary sizes. *Sci. Rep.*, 3:1082, 2013.

[185] Marija Mitrović Dankulov, Roderick Melnik, and Bosiljka Tadić. The dynamics of meaningful social interactions and the emergence of collective knowledge. *Sci. Rep.*, 5, 2015.

[186] Bosiljka Tadić, Marija Mitrović Dankulov, and Roderick Melnik. Mechanisms of self-organized criticality in social processes of knowledge creation. *Phys. Rev. E*, 96(3):032307, 2017.

[187] Giuliano Armano and Marco Alberto Javarone. The beneficial role of mobility for the emergence of innovation. *Sci. Rep.*, 7, 2017.

[188] Matjaž Perc. The matthew effect in empirical data. *J. R. Soc. Interface*, 11(98):20140378, 2014.

[189] Bernardo Monechi, A Ruiz-Serrano, F Tria, and V Loreto. Waves of novelties in the expansion into the adjacent possible. *PLoS One*, 12(6):e0179303, 2017.

[190] Harold Stanley Heaps. *Information retrieval: Computational and theoretical aspects*. Academic Press, Inc., 1978.

[191] George Kingsley Zipf. *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio Books, 2016.

[192] Francesca Tria, Vittorio Loreto, and Vito Servedio. Zipf's, heaps' and taylor's laws are determined by the expansion into the adjacent possible. *Entropy*, 20(10):752, Sep 2018.

[193] Andrea Mazzolini, Alberto Colliva, Michele Caselle, and Matteo Osella. Heaps' law, statistics of shared components, and temporal patterns from a sample-space-reducing process. *Phys. Rev. E*, 98(5):052139, 2018.

[194] George Udny Yule. Ii.—a mathematical theory of evolution, based on the

conclusions of dr. jc willis, fr s. *Phil. Trans. R. Soc. Lond. B*, 213(402-410):21–87, 1925.

[195] Robin Pemantle et al. A survey of random processes with reinforcement. *Probab. Surv*, 4(0):1–79, 2007.

[196] Norman Lloyd Johnson and Samuel Kotz. Urn models and their application; an approach to modern discrete probability theory. 1977.

[197] George Pólya. Sur quelques points de la théorie des probabilités. In *Annales de l'institut Henri Poincaré*, volume 1, pages 117–161, 1930.

[198] Fred M Hoppe. Pólya-like urns and the ewens' sampling formula. *J. Math. Biol.*, 20(1):91–94, 1984.

[199] Ciro Cattuto, Vittorio Loreto, and Luciano Pietronero. Semiotic dynamics and collaborative tagging. *Proc. Natl. Acad. Sci. U.S.A.*, 104(5):1461–1464, 2007.

[200] Vittorio Loreto, Vito DP Servedio, Steven H Strogatz, and Francesca Tria. Dynamics on expanding spaces: modeling the emergence of novelties. In *Creativity and Universality in Language*, pages 59–83. Springer, 2016.

[201] Alain Barrat, Guilherme Ferraz de Arruda, Iacopo Iacopini, and Yamir Moreno. Social contagion on higher-order structures. *arXiv preprint arXiv:2103.03709*, 2021.

[202] Mark EJ Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[203] Manlio De Domenico, Albert Solé-Ribalta, Emanuele Cozzo, Mikko Kivelä, Yamir Moreno, Mason A Porter, Sergio Gómez, and Alex Arenas. Mathematical formulation of multilayer networks. *Phys. Rev. X*, 3(4):041022, 2013.

[204] Federico Battiston, Vincenzo Nicosia, and Vito Latora. The new challenges of multiplex networks: Measures and models. *Eur. Phys. J. Special Topics*, 226(3):401–416, 2017.

[205] Ginestra Bianconi. *Multilayer Networks: Structure and Function*. Oxford University Press, 2018.

[206] A. Arenas, A. Díaz-Guilera, J. Kurths, Y. Moreno, and C. Zhou. Synchronization in complex networks. *Phys. Rep.*, 469(3):93–153, 2008.

[207] Naoki Masuda, Mason A Porter, and Renaud Lambiotte. Random walks and

diffusion on networks. *Phys. Rep.*, 716:1–58, 2017.

[208] Thomas W Valente. Network models of the diffusion of innovations. *Comp. Math. Org. Th.*, 2(2):163–164, 1996.

[209] Robin Cowan and Nicolas Jonard. Network structure and the diffusion of knowledge. *J. Econ. Dyn. Control*, 28(8):1557–1575, 2004.

[210] Xi Fang, Satyajayant Misra, Guoliang Xue, and Dejun Yang. Smart Grids - The new and improved Power Grid: A Survey. *Communications Surveys & Tutorials, IEEE*, 14(4):944–980, 2012.

[211] Thilo Gross, Carlos J Dommar D'Lima, and Bernd Blasius. Epidemic dynamics on an adaptive network. *Phys. Rev. Lett.*, 96(20):208701, 2006.

[212] Damián H Zanette and Sebastián Risau-Gusmán. Infection spreading in a population with evolving contacts. *J. Biol. Phys.*, 34(1-2):135–148, 2008.

[213] Sebastián Risau-Gusmán and Damián H Zanette. Contact switching as a control strategy for epidemic outbreaks. *J. Theor. Biol.*, 257(1):52–60, 2009.

[214] J Gómez-Gardeñes, L Lotero, SN Taraskin, and FJ Pérez-Reche. Explosive contagion in networks. *Sci. Rep.*, 6:19767, 2016.

[215] Paula Tuzón, Juan Fernández-Gracia, and Víctor M Eguíluz. From continuous to discontinuous transitions in social diffusion. *Front. Phys.*, 6:21, 2018.

[216] Hsuan-Wei Lee, Nishant Malik, Feng Shi, and Peter J. Mucha. Social clustering in epidemic spread on coevolving networks. *Phys. Rev. E*, 99:062301, Jun 2019.

[217] Sebastian Funk, Erez Gilad, Chris Watkins, and Vincent AA Jansen. The spread of awareness and its impact on epidemic outbreaks. *Pro. Natl. Acad Sci. U.S.A.*, 106(16):6872–6877, 2009.

[218] S Funk, E Gilad, and VAA Jansen. Endemic disease, awareness, and local behavioural response. *J. Theor. Biol.*, 264(2):501–509, 2010.

[219] Qingchu Wu, Xinchu Fu, Michael Small, and Xin-Jian Xu. The impact of awareness on epidemic spreading in networks. *Chaos*, 22(1):013101, 2012.

[220] Clara Granell, Sergio Gómez, and Alex Arenas. Dynamical interplay between awareness and epidemic spreading in multiplex networks. *Phys. Rev. Lett.*, 111(12):128701, 2013.

[221] Benjamin Steinegger, Alessio Cardillo, Paolo De Los Rios, Jesús Gómez-Gardeñes, and Alex Arenas. Interplay between cost and benefits triggers nontrivial vaccination uptake. *Phys. Rev. E*, 97(3):032308, 2018.

[222] Hongrun Wu, Alex Arenas, and Sergio Gómez. Influence of trust in the spreading of information. *Phys. Rev. E*, 95(1):012301, 2017.

[223] Franco Bagnoli, Pietro Lio, and Luca Sguanci. Risk perception in epidemic modeling. *Phys. Rev. E*, 76(6):061904, 2007.

[224] Guilherme Ferraz de Arruda, Francisco A Rodrigues, and Yamir Moreno. Fundamentals of spreading processes in single and multilayer complex networks. *Phys. Rep.*, 756:1–59, 2018.

[225] Convention on Climate Change (UNFCCC) The 21st Conference of the Parties to the United Nations Framework. The Paris Agreement, 2015.

[226] Vincenzo Giordano and Gianluca Fulli. A business case for smart grid technologies: A systemic perspective. *Energy Policy*, 40:252–259, 2012.

[227] Jesús Rodríguez-Molina, Margarita Martínez-Núñez, José-Fernán Martínez, and Waldo Pérez-Aguiar. Business models in the smart grid: Challenges, opportunities and proposals for prosumer profitability. *Energies*, 7(9):6142–6171, 2014.

[228] Benjamin Schäfer, Moritz Matthiae, Marc Timme, and Dirk Witthaut. Decentral Smart Grid Control. *New J. Phys.*, 17(1):015002, 2015.

[229] Na Li, Lijun Chen, and Steven H Low. Optimal demand response based on utility maximization in power networks. In *2011 IEEE power and energy society general meeting*, pages 1–8. IEEE, 2011.

[230] Amir-Hamed Mohsenian-Rad, Vincent WS Wong, Juri Jatskevich, Robert Schober, and Alberto Leon-Garcia. Autonomous demand-side management based on game-theoretic energy consumption scheduling for the future smart grid. *IEEE transactions on Smart Grid*, 1(3):320–331, 2010.

[231] Chua-Liang Su and Daniel Kirschen. Quantifying the effect of demand response on electricity markets. *IEEE Transactions on Power Systems*, 24(3):1199–1207, 2009.

[232] Danny Pudjianto, Charlotte Ramsay, and Goran Strbac. Virtual power plant and system integration of distributed energy resources. *IET Renew. Power Gener.*,

1(1):10–16, 2007.

[233] Pierluigi Siano. Demand response and smart grids a survey. *Renew. Sustain. Energy Rev*, 30:461–478, 2014.

[234] Marc Barthélemy. Spatial networks. *Phys. Rep.*, 499(1):1–101, 2011.

[235] Sergio Gómez, Alexandre Arenas, J Borge-Holthoefer, Sandro Meloni, and Yamir Moreno. Discrete-time markov chain approach to contact-based disease spreading in complex networks. *Europhys. Lett.*, 89(3):38009, 2010.

[236] Peter G Fennell, Sergey Melnik, and James P Gleeson. Limitations of discrete-time approaches to continuous-time contagion dynamics. *Phys. Rev. E*, 94(5):052125, 2016.

[237] Eugenio Valdano, Michele Re Fiorentin, Chiara Poletto, and Vittoria Colizza. Epidemic threshold in continuous-time evolving networks. *Phys. Rev. Lett.*, 120(6):068302, 2018.

[238] Sergio Porta, Paolo Crucitti, and Vito Latora. The network analysis of urban streets: a primal approach. *Environment and Planning B: planning and design*, 33(5):705–725, 2006.

[239] Bernard M Waxman. Routing of multipoint connections. *IEEE journal on selected areas in communications*, 6(9):1617–1622, 1988.

[240] Elsa Arcaute, Carlos Molinero, Erez Hatna, Roberto Murcio, Camilo Vargas-Ruiz, A Paolo Masucci, and Michael Batty. Cities and regions in britain through hierarchical percolation. *Royal Soc. Open Sci.*, 3(4):150691, 2016.

[241] OS MasterMap Integrated Transport Network Layer [GML geospatial data], Coverage: Great Britain, Updated Jan 2010, Ordnance Survey, GB. Using: EDINA Digimap Ordnance Survey Service. Â© Crown Copyright and Database Right (February 2016). Ordnance Survey (Digimap Licence).

[242] LSOA boundaries, 2011.

[243] Panagiotis D Karampourniotis, Sameet Sreenivasan, Boleslaw K Szymanski, and Gyorgy Korniss. The impact of heterogeneous thresholds on social contagion with multiple initiators. *PLoS One*, 10(11):e0143020, 2015.

[244] Romualdo Pastor-Satorras and Alessandro Vespignani. Immunization of complex

networks. *Phys. Rev. E*, 65(3):036104, 2002.

[245] Maksim Kitsak, Lazaros K Gallos, Shlomo Havlin, Fredrik Liljeros, Lev Muchnik, H Eugene Stanley, and Hernán A Makse. Identification of influential spreaders in complex networks. *Nat. Phys.*, 6(11):888, 2010.

[246] Flaviano Morone and Hernán A Makse. Influence maximization in complex networks through optimal percolation. *Nature*, 524(7563):65, 2015.

[247] Catherine SE Bale, Nicholas J McCullen, Timothy J Foxon, Alastair M Rucklidge, and William F Gale. Harnessing social networks for promoting adoption of energy technologies in the domestic sector. *Energy Policy*, 63:833–844, 2013.

[248] Varun Rai and Scott A Robinson. Agent-based modeling of energy technology adoption: Empirical integration of social, behavioral, economic, and environmental factors. *Environmental Modelling & Software*, 70:163–177, 2015.

[249] Varun Rai and Adam Douglas Henry. Agent-based modelling of consumer energy choices. *Nature Climate Change*, 6(6):556, 2016.

[250] Laurens XW Hesselink and Emile JL Chappin. Adoption of energy efficient technologies by households–barriers, policies and agent-based modelling studies. *Renewable and Sustainable Energy Reviews*, 99:29–41, 2019.

[251] Karine Nyborg, John M Anderies, Astrid Dannenberg, Therese Lindahl, Caroline Schill, Maja Schlüter, W Neil Adger, Kenneth J Arrow, Scott Barrett, Stephen Carpenter, et al. Social norms as solutions. *Science*, 354(6308):42–43, 2016.

[252] Marc Barthélemy. Crossover from scale-free to spatial networks. *EPL (Europhysics Letters)*, 63(6):915, 2003.

[253] Damon Centola, Joshua Becker, Devon Brackbill, and Andrea Baronchelli. Experimental evidence for tipping points in social convention. *Science*, 360(6393):1116–1119, 2018.

[254] Claude Berge. *Hypergraphs: combinatorics of finite sets*, volume 45. Elsevier, 1984.

[255] Gourab Ghoshal, Vinko Zlatić, Guido Caldarelli, and MEJ Newman. Random hypergraphs and their applications. *Phys. Rev. E*, 79(6):066118, 2009.

[256] Guilherme Ferraz de Arruda, Giovanni Petri, and Yamir Moreno. Social contagion models on hypergraphs. *Phys. Rev. Res.*, 2(2):023032, 2020.

[257] SocioPatterns Collaboration. `http://www.sociopatterns.org/`, Accessed Dec 2018.

[258] Mathieu Génois and Alain Barrat. Can co-location be used as a proxy for face-to-face contacts? *EPJ Data Science*, 7(1):11, 2018.

[259] Lorenzo Isella, Juliette Stehlé, Alain Barrat, Ciro Cattuto, Jean-François Pinton, and Wouter Van den Broeck. What's in a crowd? analysis of face-to-face behavioral networks. *J. Theor. Biol*, 271(1):166–180, 2011.

[260] Philippe Vanhems, Alain Barrat, Ciro Cattuto, Jean-François Pinton, Nagham Khanafer, Corinne Régis, Byeul-a Kim, Brigitte Comte, and Nicolas Voirin. Estimating potential infection transmission routes in hospital wards using wearable proximity sensors. *PLoS One*, 8(9):e73970, 2013.

[261] Rossana Mastrandrea, Julie Fournet, and Alain Barrat. Contact patterns in a high school: a comparison between data collected using wearable sensors, contact diaries and friendship surveys. *PLoS One*, 10(9):e0136497, 2015.

[262] Jean-Gabriel Young, Giovanni Petri, Francesco Vaccarino, and Alice Patania. Construction of and efficient sampling from the simplicial configuration model. *Phys. Rev. E*, 96(3):032312, 2017.

[263] Konstantin Zuev, Or Eisenberg, and Dmitri Krioukov. Exponential random simplicial complexes. *J. Phys. A*, 48(46):465002, 2015.

[264] Owen T Courtney and Ginestra Bianconi. Generalized network structures: The configuration model and the canonical ensemble of simplicial complexes. *Phys. Rev. E*, 93(6):062311, 2016.

[265] Owen T Courtney and Ginestra Bianconi. Weighted growing simplicial complexes. *Phys. Rev. E*, 95(6):062301, 2017.

[266] Giovanni Petri and Alain Barrat. Simplicial activity driven model. *Phys. Rev. Lett.*, 121:228301, 2018.

[267] Nicola Perra, Bruno Gonçalves, Romualdo Pastor-Satorras, and Alessandro Vespignani. Activity driven modeling of time varying networks. *Sci. Rep.*, 2:469, 2012.

[268] Matthew Kahle. Topology of random clique complexes. *Discrete Mathematics*,

309(6):1658–1671, 2009.

[269] Armindo Costa and Michael Farber. Random simplicial complexes. In *Configuration Spaces*, pages 129–153. Springer, 2016.

[270] Paul Erdos and Alfréd Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60, 1960.

[271] Joel C Miller. Percolation and epidemics in random clustered networks. *Phys. Rev. E*, 80(2):020901, 2009.

[272] Joel C Miller. Spread of infectious disease through clustered populations. *J. Royal Soc. Interface*, 6(41):1121–1134, 2009.

[273] Laurent Hébert-Dufresne, Pierre-André Noël, Vincent Marceau, Antoine Allard, and Louis J Dubé. Propagation dynamics on networks featuring complex topologies. *Phys. Rev. E*, 82(3):036115, 2010.

[274] Brian Karrer and M. E. J Newman. Random graphs containing arbitrary distributions of subgraphs. *Phys. Rev. E*, 82(6):066118, 2010.

[275] Martin Ritchie, Luc Berthouze, Thomas House, and Istvan Z Kiss. Higher-order structure and epidemic dynamics in clustered networks. *J. Theor. Biol.*, 348:21–32, 2014.

[276] David JP O'Sullivan, Gary James O'Keeffe, Peter G Fennell, and James P Gleeson. Mathematical modeling of complex contagion on clustered networks. *Frontiers in Physics*, 3:71, 2015.

[277] Ginestra Bianconi and Christoph Rahmede. Emergent hyperbolic network geometry. *Sci. Rep.*, 7:41974, 2017.

[278] Daan Mulder and Ginestra Bianconi. Network geometry and complexity. *Journal of Statistical Physics*, 173(3-4):783–805, 2018.

[279] Ginestra Bianconi and Robert M Ziff. Topological percolation on hyperbolic simplicial complexes. *Physical Review E*, 98(5):052308, 2018.

[280] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.*, 86(14):3200, 2001.

[281] Sven Lübeck. Universal scaling behavior of non-equilibrium phase transitions. *Int. J. Mod. Phys. B*, 18(31n32):3977–4118, 2004.

[282] Silvio C Ferreira, Claudio Castellano, and Romualdo Pastor-Satorras. Epidemic thresholds of the susceptible-infected-susceptible model on networks: A comparison of numerical and theoretical results. *Phys. Rev. E*, 86(4):041125, 2012.

[283] Drude Dahlerup. From a small to a large minority: women in scandinavian politics. *Scand. Political Stud.*, 11(4):275–298, 1988.

[284] Sandra Grey. Numbers and beyond: The relevance of critical mass in gender research. *Politics & Gender*, 2(4):492–502, 2006.

[285] Joan T Matamalas, Alex Arenas, and Sergio Gómez. Effective approach to epidemic containment using link equations in complex networks. *Sci. Adv.*, 4(12):eaau4212, 2018.

[286] Joan T. Matamalas, Sergio Gómez, and Alex Arenas. Abrupt phase transition of epidemic spreading in simplicial complexes. *Phys. Rev. Research*, 2:012049, Feb 2020.

[287] Ciro Cattuto, Andrea Baldassarri, Vito DP Servedio, and Vittorio Loreto. Vocabulary growth in collaborative tagging systems. *arXiv preprint arXiv:0704.3316*, 2007.

[288] Mikhail V Simkin and Vwani P Roychowdhury. Re-inventing willis. *Phys. Rep.*, 502(1):1–35, 2011.

[289] Luigi Marengo and Paolo Zeppini. The arrival of the new. *J. Evol. Econ.*, 26(1):171–194, 2016.

[290] P Gravino, B Monechi, VDP Servedio, F Tria, and V Loreto. Crossing the horizon: exploring the adjacent possible in a cultural system. In *Proceedings of the Seventh International Conference on Computational Creativity*, 2016.

[291] Paul Thagard. *Mind: Introduction to cognitive science*, volume 17. MIT press Cambridge, MA, 2005.

[292] Javier Borge-Holthoefer and Alex Arenas. Semantic networks: Structure and dynamics. *Entropy*, 12(5):1264–1302, 2010.

[293] Andrea Baronchelli, Ramon Ferrer-i Cancho, Romualdo Pastor-Satorras, Nick Chater, and Morten H Christiansen. Networks in cognitive science. *Trends Cogn. Sci.*, 17(7):348–360, 2013.

[294] Nichol Castro and Cynthia SQ Siew. Contributions of modern network science to the cognitive sciences: Revisiting research spirals of representation and process. *Proc. Royal Soc. A*, 476(2238):20190825, 2020.

[295] Bernardo Monechi, Pietro Gravino, Vito D. P. Servedio, Francesca Tria, and Vittorio Loreto. Significance and popularity in music production. *R. Soc. Open Sci.*, 4(7), 2017.

[296] Gandhimohan M Viswanathan, V Afanasyev, SV Buldyrev, EJ Murphy, PA Prince, and H Eugene Stanley. Lévy flight search patterns of wandering albatrosses. *Nature*, 381(6581):413–415, 1996.

[297] Gandhimohan M Viswanathan, Marcos GE Da Luz, Ernesto P Raposo, and H Eugene Stanley. *The physics of foraging: an introduction to random searches and biological encounters*. Cambridge University Press, 2011.

[298] Riccardo Gallotti, Armando Bazzani, Sandro Rambaldi, and Marc Barthelemy. A stochastic model of randomly accelerated walkers for human mobility. *Nat. Commun.*, 7:12600, 2016.

[299] Christopher W Lynn, Lia Papadopoulos, Ari E Kahn, and Danielle S Bassett. Human information processing in complex networks. *Nat. Phys.*, 16(9):965–973, 2020.

[300] Paolo Allegrini, Paolo Grigolini, and Luigi Palatella. Intermittency and scale-free networks: a dynamical model for human language complexity. *Chaos, Solitons & Fractals*, 20(1):95–105, 2004.

[301] Enrique Alvarez-Lacalle, Beate Dorow, J-P Eckmann, and Elisha Moses. Hierarchical structures induce long-range dynamical correlations in written texts. *Natl. Acad. Sci. U. S. A.*, 103(21):7956–7961, 2006.

[302] Tao Jia, Dashun Wang, and Boleslaw K Szymanski. Quantifying patterns of research-interest evolution. *Nat. Hum. Behav.*, 1:0078, 2017.

[303] Jesús Gómez-Gardeñes and Vito Latora. Entropy rate of diffusion processes on complex networks. *Phys. Rev. E*, 78(6):065102, 2008.

[304] Elena Agliari, Raffaella Burioni, and Guido Uguzzoni. The true reinforced random walk with bias. *New J. Phys.*, 14(6):063027, 2012.

[305] Denis Boyer and Citlali Solis-Salas. Random walks with preferential relocations to places visited in the past and their application to biology. *Phys. Rev. Lett.*, 112(24):240601, 2014.

[306] Jeehye Choi, Jang-Il Sohn, K-I Goh, and I-M Kim. Modeling the mobility with memory. *Europhys. Lett.)*, 99(5):50001, 2012.

[307] Michael Szell, Roberta Sinatra, Giovanni Petri, Stefan Thurner, and Vito Latora. Understanding mobility in a social petri dish. *Sci. Rep.*, 2:457, 2012.

[308] Franz Merkl and Silke WW Rolles. Linearly edge-reinforced random walks. *Lect. Notes Monograph Ser.*, pages 66–77, 2006.

[309] Michael S Keane, Silke WW Rolles, et al. Edge-reinforced random walks on finite graphs. *Verhandelingen KNAW*, 52, 2000.

[310] Jacob G Foster, Peter Grassberger, and Maya Paczuski. Reinforced walks in two and three dimensions. *New J. Phys.*, 11(2):023009, 2009.

[311] P Coppersmith and D Diaconis. Random walk with reinforcement. *Unpublished*, 1987.

[312] Brenden M Lake, Neil D Lawrence, and Joshua B Tenenbaum. The emergence of organizing structure in conceptual representation. *Cogn. Sci.*, 42:809–832, 2018.

[313] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[314] Moreno Bonaventura, Vincenzo Nicosia, and Vito Latora. Characteristic times of biased random walks on complex networks. *Phys. Rev. E*, 89(1):012803, 2014.

[315] Vishal Sood and Peter Grassberger. Localization transition of biased random walks on random networks. *Phys. Rev. Lett.*, 99(9):098701, 2007.

[316] Crispin W Gardiner. *Stochastic methods*. Springer-Verlag, Berlin–Heidelberg–New York–Tokyo, 1985.

[317] Duncan J Watts. *Small worlds: the dynamics of networks between order and randomness*. Princeton university press, 1999.

[318] Pietro Gravino, Vito DP Servedio, Alain Barrat, and Vittorio Loreto. Complex structures and semantics in free word association. *Adv. Complex Syst.*, 15(03n04):1250054, 2012.

[319] Adilson E Motter, Alessandro PS De Moura, Ying-Cheng Lai, and Partha Dasgupta. Topology of the conceptual network of language. *Phys. Rev. E*, 65(6):065102, 2002.

[320] Mathias Benedek, Yoed N Kenett, Konstantin Umdasch, David Anaki, Miriam Faust, and Aljoscha C Neubauer. How semantic memory structure and intelligence contribute to creative thought: a network science approach. *Think. Reas.*, 23(2):158–183, 2017.

[321] Mark EJ Newman and Duncan J Watts. Scaling and percolation in the small-world network model. *Phys. Rev. E*, 60(6):7332, 1999.

[322] A Dvoretzky and P Erdös. 2nd berkeley sympos. math. stat. and prob, 1951.

[323] Caterina De Bacco, Satya N Majumdar, and Peter Sollich. The average number of distinct sites visited by a random walker on random graphs. *J. Phys. A*, 48(20):205004, 2015.

[324] Paul Erdös and Alfréd Rényi. On random graphs, i. *Publicationes Mathematicae (Debrecen)*, 6:290–297, 1959.

[325] F Jasch and A Blumen. Target problem on small-world networks. *Phys. Rev. E*, 63(4):041108, 2001.

[326] Jani Lahtinen, János Kertész, and Kimmo Kaski. Scaling of random spreading in small world networks. *Phys. Rev. E*, 64(5):057105, 2001.

[327] E Almaas, RV Kulkarni, and D Stroud. Scaling properties of random walks on small-world networks. *Phys. Rev. E*, 68(5):056105, 2003.

[328] Alain Barrat and Martin Weigt. On the properties of small-world network models. *Eur. Phys. J. B*, 13(3):547–560, 2000.

[329] Santo Fortunato, Carl T Bergstrom, Katy Börner, James A Evans, Dirk Helbing, Staša Milojević, Alexander M Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, et al. Science of science. *Science*, 359(6379):eaao0185, 2018.

[330] Federico Battiston, Federico Musciotto, Dashun Wang, Albert-László Barabási, Michael Szell, and Roberta Sinatra. Taking census of physics. *Nat. Rev. Phys.*, 1(1):89–97, 2019.

[331] Luís MA Bettencourt, David I Kaiser, and Jasleen Kaur. Scientific discovery and topological transitions in collaboration networks. *J. Inform.*, 3(3):210–221, 2009.

[332] Mark EJ Newman. The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci. U.S.A.*, 98(2):404–409, 2001.

[333] Caroline S Wagner, J David Roessner, Kamau Bobb, Julie Thompson Klein, Kevin W Boyack, Joann Keyton, Ismael Rafols, and Katy Börner. Approaches to understanding and measuring interdisciplinary scientific research (idr): A review of the literature. *J. Inform.*, 5(1):14–26, 2011.

[334] Staša Milojević. Quantifying the cognitive extent of science. *J. Inform.*, 9(4):962–973, 2015.

[335] Feng Shi, Jacob G Foster, and James A Evans. Weaving the fabric of science: Dynamic network models of science's unfolding structure. *Soc. Net.*, 43:73–85, 2015.

[336] Lindell Bromham, Russell Dinnage, and Xia Hua. Interdisciplinary research has consistently lower funding success. *Nature*, 534(7609):684–687, 2016.

[337] Brian Uzzi, Satyam Mukherjee, Michael Stringer, and Ben Jones. Atypical combinations and scientific impact. *Science*, 342(6157):468–472, 2013.

[338] Dashun Wang, Chaoming Song, and Albert-László Barabási. Quantifying long-term scientific impact. *Science*, 342(6154):127–132, 2013.

[339] Staša Milojević. How are academic age, productivity and collaboration related to citing behavior of researchers? *PLoS One*, 7(11):e49176, 2012.

[340] Werner Ebeling and Gregoire Nicolis. Word frequency and entropy of symbolic sequences: a dynamical perspective. *Chaos, Solitons & Fractals*, 2(6):635–650, 1992.

[341] Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

[342] M Coccia. How to determine the probability of discovery in research fields. Technical report, Working Paper CocciaLab n. 49B/2020, CNR–National Research Council of Italy, 2020.

[343] Serhii Brodiuk, Vasyl Palchykov, and Yurij Holovatch. Embedding technique and network analysis of scientific innovations emergence in an arxiv-based concept network. *arXiv preprint arXiv:2003.10289*, 2020.

[344] Jürgen Drews. Drug discovery: a historical perspective. *Science*, 287(5460):1960–

1964, 2000.

[345] Fang Wu and Bernardo A Huberman. Novelty and collective attention. *Proc. Natl. Acad. Sci. U.S.A.*, 104(45):17599–17601, 2007.

[346] Alexander TJ Barron, Jenny Huang, Rebecca L Spang, and Simon DeDeo. Individuals, institutions, and innovation in the debates of the french revolution. *Proc. Natl. Acad. Sci. U.S.A.*, 115(18):4607–4612, 2018.

[347] Mario Coccia. Why do nations produce science advances and new technology? *Technol. in Soc.*, 59:101124, 2019.

[348] Doheum Park, Juhan Nam, and Juyong Park. Novelty and influence of creative works, and quantifying patterns of advances based on probabilistic references networks. *EPJ Data Science*, 9(1):2, 2020.

[349] Thomas M. A. Fink and Ali Teimouri. The mathematical structure of innovation. *arXiv preprint arXiv:1912.03281*, 2019.

[350] Mario Coccia. The theory of technological parasitism for the measurement of the evolution of technology and technological forecasting. *Technol. Forecast. Soc. Change*, 141:289–304, 2019.

[351] Enrico Ubaldi, Raffaella Burioni, Vittorio Loreto, and Francesca Tria. The exploration of the adjacent possible explains the emergence and evolution of social networks. *arXiv preprint arXiv:2003.00989*, 2020.

[352] Mickaël Launay and Vlada Limic. Generalized interacting urn models. *arXiv preprint arXiv:1207.5635*, 2012.

[353] Giacomo Aletti, Irene Crimaldi, and Andrea Ghiglietti. Interacting reinforced stochastic processes: statistical inference based on the weighted empirical means. *arXiv preprint arXiv:1811.10255*, 2018.

[354] Mikhail Hayhoe, Fady Alajaji, and Bahman Gharesifard. A polya urn-based model for epidemics on networks. In *American Control Conference (ACC), 2017*, pages 358–363. IEEE, 2017.

[355] Mikhail Hayhoe, Fady Alajaji, and Bahman Gharesifard. Curing with the network polya contagion model. In *2018 Annual American Control Conference (ACC)*, pages 2644–2650. IEEE, 2018.

[356] Sven Berg. Paradox of voting under an urn model: The effect of homogeneity. *Public Choice*, 47(2):377–387, 1985.

[357] Tao Gong, Lan Shuai, Mónica Tamariz, and Gerhard Jäger. Studying language change using price equation and pólya-urn dynamics. *PLoS One*, 7(3):e33171, 2012.

[358] Riccardo Marcaccioli and Giacomo Livan. A pólya urn approach to information filtering in complex networks. *Nat. Comm.*, 10(1):745, 2019.

[359] Francesc Font-Clos, Gemma Boleda, and Alvaro Corral. A scaling law beyond zipf's law and its relation to heaps' law. *New J. Phys.*, 15(9):093033, 2013.

[360] Andrea Mazzolini, Marco Gherardi, Michele Caselle, Marco Cosentino Lagomarsino, and Matteo Osella. Statistics of shared components in complex component systems. *Phys. Rev. X*, 8(2):021023, 2018.

[361] Andrea Mazzolini, Jacopo Grilli, Eleonora De Lazzari, Matteo Osella, Marco Cosentino Lagomarsino, and Marco Gherardi. Zipf and heaps laws from dependency structures in component systems. *Phys. Rev. E*, 98(1):012315, 2018.

[362] Alessandro Mastrototaro. *A mathematical model for the emergence of innovations*. PhD thesis, Politecnico di Torino, 2018.

[363] Matthew J Salganik, Peter Sheridan Dodds, and Duncan J Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311(5762):854–856, 2006.

[364] Róbert Pálovics and András A Benczúr. Temporal influence over the last. fm social network. *Social Network Analysis and Mining*, 5(1):4, 2015.

[365] John Ternovski and Taha Yasseri. Social complex contagion in music listenership: A natural experiment with 1.3 million participants. *Soc. Netw*, 2019.

[366] Paul Felix Lazarsfeld, Bernard Berelson, and Hazel Gaudet. *The people's choice.* Duell, Sloan & Pearce, 1944.

[367] Robert M Bond, Christopher J Fariss, Jason J Jones, Adam DI Kramer, Cameron Marlow, Jaime E Settle, and James H Fowler. A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415):295, 2012.

[368] John Bryden, Shaun P Wright, and Vincent AA Jansen. How humans transmit language: horizontal transmission matches word frequencies among peers on

twitter. *J. Royal Soc. Interface*, 15(139):20170738, 2018.

[369] Mark EJ Newman. Spread of epidemic disease on networks. *Phys. Rev. E*, 66(1):016128, 2002.

[370] Wayne W Zachary. An information flow model for conflict and fission in small groups. *J. Anthropol. Res.*, 33(4):452–473, 1977.

[371] Munmun De Choudhury, Yu-Ru Lin, Hari Sundaram, Kasim Selcuk Candan, Lexing Xie, and Aisling Kelliher. How does the data sampling strategy impact the discovery of information diffusion in social media? In *Fourth International AAAI Conference on Weblogs and Social Media*, 2010.

[372] Mark EJ Newman. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E*, 74(3):036104, 2006.

[373] Pablo M Gleiser and Leon Danon. Community structure in jazz. *Adv. Complex Syst.*, 6(04):565–573, 2003.

[374] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U. S. A.*, 99(12):7821–7826, 2002.

[375] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *J. Stat. Mech.: Theory Exp*, 2008(10):P10008, 2008.

[376] Santo Fortunato. Community detection in graphs. *Phys. Rep.*, 486(3-5):75–174, 2010.

[377] Brian Karrer and Mark EJ Newman. Stochastic blockmodels and community structure in networks. *Phys. Rev. E*, 83(1):016107, 2011.

[378] Giovanni Dosi, Alessio Moneta, and Elena Stepanova. Dynamic increasing returns and innovation diffusion: bringing polya urn processes to the empirical data. *Industry and Innovation*, 26(4):461–478, 2019.

[379] Oskar Perron. "Uber matrizen. *Math. Ann.*, 64:248–263, 1907.

[380] Georg Frobenius. "Uber matrizen aus nicht negativen elementen. In *S.-B.Deutsch.Akad. Wiss. Berlin. Math-Nat. Kl.*, pages 456–477, 1912.

[381] Phillip Bonacich. Factoring and weighting approaches to status scores and clique identification. *J. Math. Sociol.*, 2(1):113–120, 1972.

[382] Phillip Bonacich and Paulette Lloyd. Eigenvector-like measures of centrality for asymmetric relations. *Soc. Netw*, 23(3):191–201, 2001.

[383] Kiyotaka Ide, Akira Namatame, Loganathan Ponnambalam, Fu Xiuju, and Rick Siow Mong Goh. A new centrality measure for probabilistic diffusion in network. *Adv. Comput. Sci.*, 3(5):115–121, 2014.

[384] Kristina Ghosh, Rumi; Lerman. Parameterized centrality metric for network analysis. *Phys. Rev. E*, 83, 6 2011.

[385] Rumi Ghosh and Kristina Lerman. Rethinking centrality: The role of dynamical processes in social network analysis. *Discrete and Continuous Dynamical Systems - Series B*, 19, 09 2012.

[386] Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.

[387] Giulia Cencetti, Federico Battiston, Duccio Fanelli, and Vito Latora. Reactive random walkers on complex networks. *Phys. Rev. E*, 98(5):052302, 2018.

[388] James P Gleeson, Jonathan A Ward, Kevin P O'sullivan, and William T Lee. Competition-induced criticality in a model of meme popularity. *Phys. Rev. Lett.*, 112(4):048701, 2014.

[389] James P Gleeson, Kevin P O'Sullivan, Raquel A Baños, and Yamir Moreno. Effects of network structure, competition and memory time on social spreading phenomena. *Phys. Rev. X*, 6(2):021019, 2016.

[390] Joseph D O'Brien, Ioannis K Dassios, and James P Gleeson. Spreading of memes on multiplex networks. *New J. Phys.*, 21(2):025001, 2019.

[391] Aaron Schecter, Andrew Pilny, Alice Leung, Marshall Scott Poole, and Noshir Contractor. Step by step: Capturing the dynamics of work team process through relational event sequences. *J. Organ. Behav.*, 39(9):1163–1181, 2018.

[392] Bernardo Monechi, Giulia Pullano, and Vittorio Loreto. Efficient team structures in an open-ended cooperative creativity experiment. *Proc. Natl. Acad. Sci. U.S.A.*, 116(44):22088–22093, 2019.

[393] Lillian Weng, Alessandro Flammini, Alessandro Vespignani, and Fillipo Menczer. Competition among memes in a world with limited attention. *Sci. Rep.*, 2:335,

2012.

[394] Philipp Lorenz-Spreen, Bjarke Mørch Mønsted, Philipp Hövel, and Sune Lehmann. Accelerating dynamics of collective attention. *Nat. Commun.*, 10(1):1–9, 2019.

[395] Cristian Candia, C Jara-Figueroa, Carlos Rodriguez-Sickert, Albert-László Barabási, and César A Hidalgo. The universal decay of collective memory and attention. *Nat. Hum. Behav.*, 3(1):82–91, 2019.

[396] Per Sebastian Skardal and Alex Arenas. Abrupt desynchronization and extensive multistability in globally coupled oscillator simplexes. *Phys. Rev. Lett.*, 122(24):248301, 2019.

[397] Per Sebastian Skardal and Alex Arenas. Higher order interactions in complex networks of phase oscillators promote abrupt synchronization switching. *Comm. Phys.*, 3(1):1–6, 2020.

[398] Ana P Millán, Joaquín J Torres, and Ginestra Bianconi. Explosive higher-order kuramoto dynamics on simplicial complexes. *Phys. Rev. Lett.*, 124(21):218301, 2020.

[399] LV Gambuzza, F Di Patti, L Gallo, S Lepri, M Romance, R Criado, M Frasca, V Latora, and S Boccaletti. The master stability function for synchronization in simplicial complexes. *arXiv preprint arXiv:2004.03913*, 2020.

[400] Ginestra Bianconi and Christoph Rahmede. Complex quantum network manifolds in dimension d> 2 are scale-free. *Sci. Rep.*, 5(1):1–10, 2015.

[401] Ginestra Bianconi and Christoph Rahmede. Network geometry with flavor: from complexity to quantum geometry. *Phys. Rev. E*, 93(3):032315, 2016.

[402] Ana P Millán, Joaquín J Torres, and Ginestra Bianconi. Synchronization in network geometries with finite spectral dimension. *Phys. Rev. E*, 99(2):022307, 2019.

[403] Leonhard Horstmeyer and Christian Kuehn. Adaptive voter model on simplicial complexes. *Phys. Rev. E*, 101(2):022305, 2020.

[404] Leonie Neuhäuser, Andrew Mellor, and Renaud Lambiotte. Multibody interactions and nonlinear consensus dynamics on networked systems. *Phys. Rev. E*, 101(3):032310, 2020.

[405] Unai Alvarez-Rodriguez, Federico Battiston, Guilherme Ferraz de Arruda, Yamir

Moreno, Matjaz Perc, and Vito Latora. Evolutionary dynamics of higher-order interactions. *Nat. Hum. Behav.*, 2021.

[406] Michael T Schaub, Austin R Benson, Paul Horn, Gabor Lippner, and Ali Jadbabaie. Random walks on simplicial complexes and the normalized hodge 1-laplacian. *SIAM Review*, 62(2):353–391, 2020.

[407] Timoteo Carletti, Federico Battiston, Giulia Cencetti, and Duccio Fanelli. Random walks on hypergraphs. *Phys. Rev. E*, 101(2):022308, 2020.

[408] Roger Guimera, Brian Uzzi, Jarrett Spiro, and Luis A Nunes Amaral. Team assembly mechanisms determine collaboration network structure and team performance. *Science*, 308(5722):697–702, 2005.

[409] Brian Uzzi and Jarrett Spiro. Collaboration and creativity: The small world problem. *Am. J. Sociol.*, 111(2):447–504, 2005.

[410] Vedran Sekara and Sune Lehmann. The strength of friendship ties in proximity sensor data. *PLoS One*, 9(7), 2014.

[411] Vedran Sekara, Arkadiusz Stopczynski, and Sune Lehmann. Fundamental structures of dynamic social networks. *Proc. Natl. Acad. Sci. U. S. A.*, 113(36):9977–9982, 2016.

[412] Piotr Sapiezynski, Arkadiusz Stopczynski, David Dreyer Lassen, and Sune Lehmann. Interaction data from the copenhagen networks study. *Sci. Data*, 6(1):1–10, 2019.

[413] Ciro Cattuto, Wouter Van den Broeck, Alain Barrat, Vittoria Colizza, Jean-François Pinton, and Alessandro Vespignani. Dynamics of person-to-person interactions from distributed rfid sensor networks. *PLoS One*, 5(7):e11596, 2010.

[414] Juliette Stehlé, Nicolas Voirin, Alain Barrat, Ciro Cattuto, Lorenzo Isella, Jean-François Pinton, Marco Quaggiotto, Wouter Van den Broeck, Corinne Régis, Bruno Lina, et al. High-resolution measurements of face-to-face contact patterns in a primary school. *PLoS One*, 6(8), 2011.

[415] M Ángeles Serrano, Marián Boguná, and Alessandro Vespignani. Extracting the multiscale backbone of complex weighted networks. *Proc. Natl. Acad. Sci. U.S.A.*, 106(16):6483–6488, 2009.

[416] Abraham Berman and Robert J. Plemmons. *Nonnegative Matrices*, chapter 2, pages

26–62. Academic Press, 1979.

[417] Robert Tarjan. Depth-first search and linear graph algorithms. *SIAM journal on computing*, 1(2):146–160, 1972.

A: Simplicial contagion on empirical and synthetic simplicial complexes

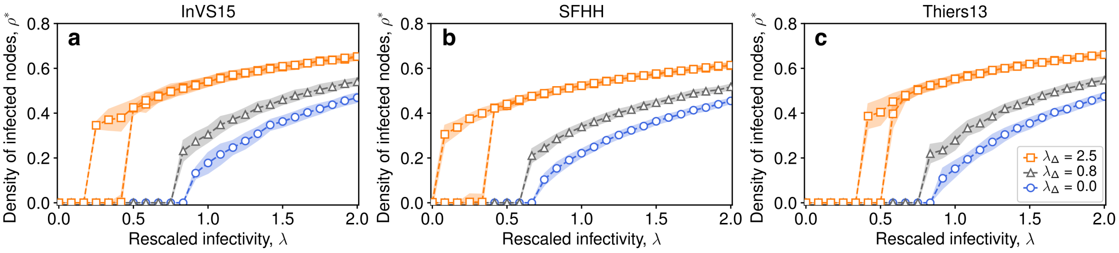


Figure A.1: SCM of order $D = 2$ on real-world higher-order social structures without data augmentation. Simplicial complexes are constructed from high-resolution face-to-face contact data recorded in a workplace (**a**), a conference (**b**), and a high school (**c**). The average fraction of infected nodes in the stationary state obtained numerically is plotted against the rescaled infectivity $\lambda = \beta\langle k\rangle/\mu$ for $\lambda_\Delta = 0.8$ (black triangles) and $\lambda_\Delta = 2.5$ (orange squares). The blue circles denote the simulated curve for the standard SIS model ($\lambda_\Delta = 0$), which does not consider higher order effects. For $\lambda_\Delta = 2.5$ a bi-stable region appears, where healthy and endemic states co-exist.

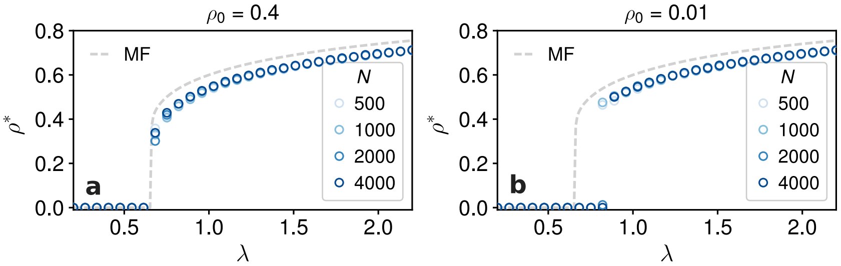


Figure A.2: Numerical exploration of the finite size effects on the hysteresis for a SCM of order $D = 2$ on synthetic random simplicial complexes (RSC). The two panels refer to two different values of the initial density of infected individuals, namely (**a**) $\rho_0 = 0.4$ and (**b**) $\rho_0 = 0.01$. The dashed line corresponds to the mean-field result. Figure from [2].

B: Correlations produced by ERRWs

Correlations produced by ERRWs on real networks

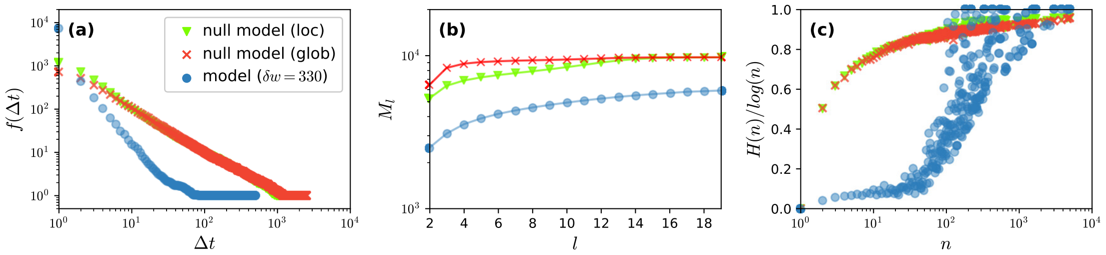


Figure A.3: Correlations among concepts for the growth of knowledge in science (Astronomy shown) produced by an ERRW model. The ERRW is tuned to the empirical data by selecting the reinforcement δw that reproduces the Heaps' exponent β obtained by fitting the associated Heaps' curve as a power law (for the Astronomy case shown $\delta w = 330$). (**a**) Frequency distribution of inter-event times Δt between consecutive occurrences of the same concept (node in our model). (**b**) Number M_l of different subsequences of length l as a function of l. (**c**) Normalized entropy of the sequence of visited nodes as a function of n, the number of times the nodes have been visited. In each panel, blue circles show average values over 20 different realizations, while triangles and crosses refer to those of (globally and locally) reshuffled sequences. Figure from [1].

In Sec. 4.5, we have shown how the ERRW model on small-world (SW) networks is able to produce correlated sequences of concepts. We have also proposed a study case of the ERRW model on real topologies extracted from empirical data. In particular, we have explored the cognitive growth of science by extracting empirical sequences of relevant concepts in different scientific fields. For each one of the considered fields, we have then tuned the reinforcement parameter of our model in order to produce sequences with the same Heaps' exponents as the empirical ones (see Fig. 4.7 and Table 4-A). Here, we investigate correlations in the sequences produced by ERRWs on real networks. Figure A.3 reports the same quantities we used to study correlations in sequences produced by ERRWs on synthetic SW networks (see Fig. 4.9), namely the average entropy of the sequence [Fig. A.3(**a**)], number M_l of different subsequences of length l as a function of l [Fig. A.3(**b**)], and frequency distribution $f(\Delta t)$ of inter-event times Δt between couples of consecutive concepts [Fig. A.3(**c**)]. In each plot, results are compared to the two null models defined in Sec. 4.5.1, confirming the correlated nature of the sequences. Furthermore, the comparison with the same statistics obtained for ERRWs on SW networks (see Fig. 4.9) confirms again that SW topologies represent a good choice for modeling the relations among concepts.

Correlations produced by ERRWs on synthetic networks

In Sec. 4.5, we have implemented the ERRW model on SW networks, which proved to be good topologies for modeling the structure of relations among concepts [318–320]). In addition to the results in Fig. 4.9, where we studied the correlations produced by an ERRW over a SW network with fixed link probability p for a fixed amount of reinforcement at $\delta w = 0.01$, here we show the curves of average entropy of sequence (Figure A.4) and frequency distribution $f(\Delta t)$ of inter-event times Δt between couples of consecutive concepts (Figure A.5) for different values of reinforcement, ranging from $\delta w = 0.001$ to $\delta w = 1$. Three different cases of SW networks with $N = 10^6$ nodes and respectively with link rewiring probability $p = 0.001$ [Fig. A.4(**a-d**) and Fig. A.5(**a-d**)], $p = 0.01$ [Fig. A.4(**e-h**) and Fig. A.5(**e-h**)] and $p = 0.1$ [Fig. A.4(**i-l**) and Fig. A.5(**i-l**)], are considered. All curves are compared to the corresponding null models as defined in Sec.4.5.1.

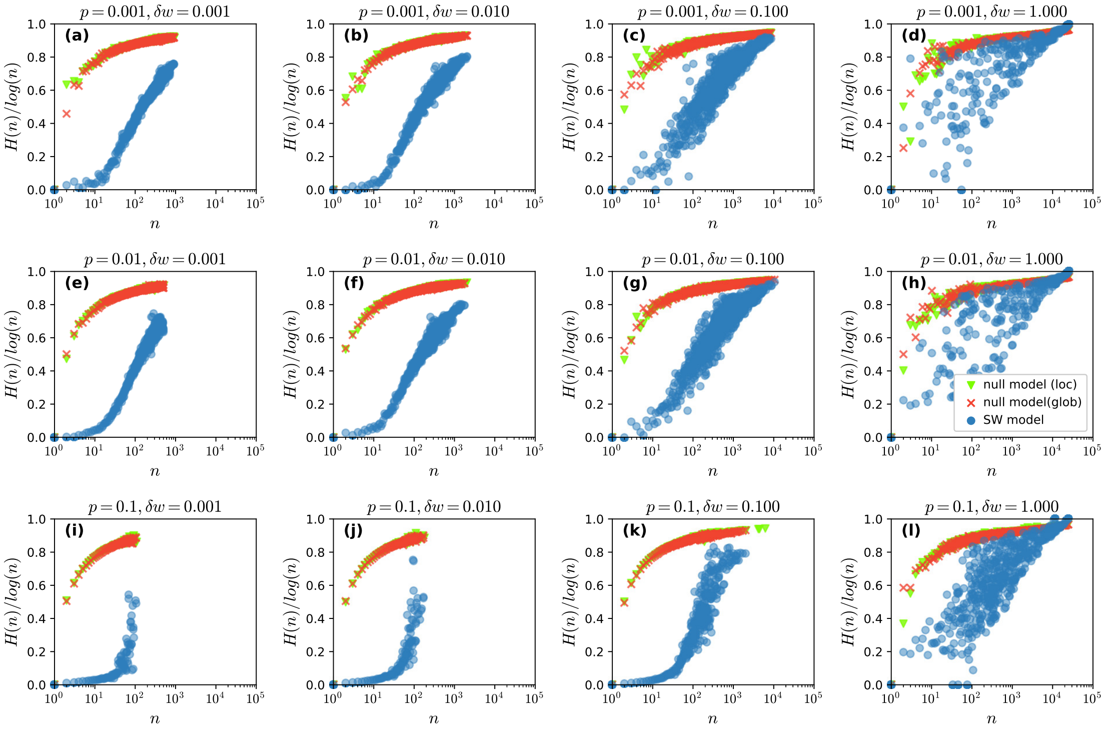


Figure A.4: Correlations among concepts produced by an edge-reinforced random walk on a SW network for different values of link probability p and reinforcement δw (see the main text for details). Normalized entropy of the sequence of visited nodes as a function of n, the number of times the nodes have been visited. In each panel, blue circles show average values over 20 different realizations, while triangles and crosses refer to those of (globally and locally) reshuffled sequences. Figure from [1].

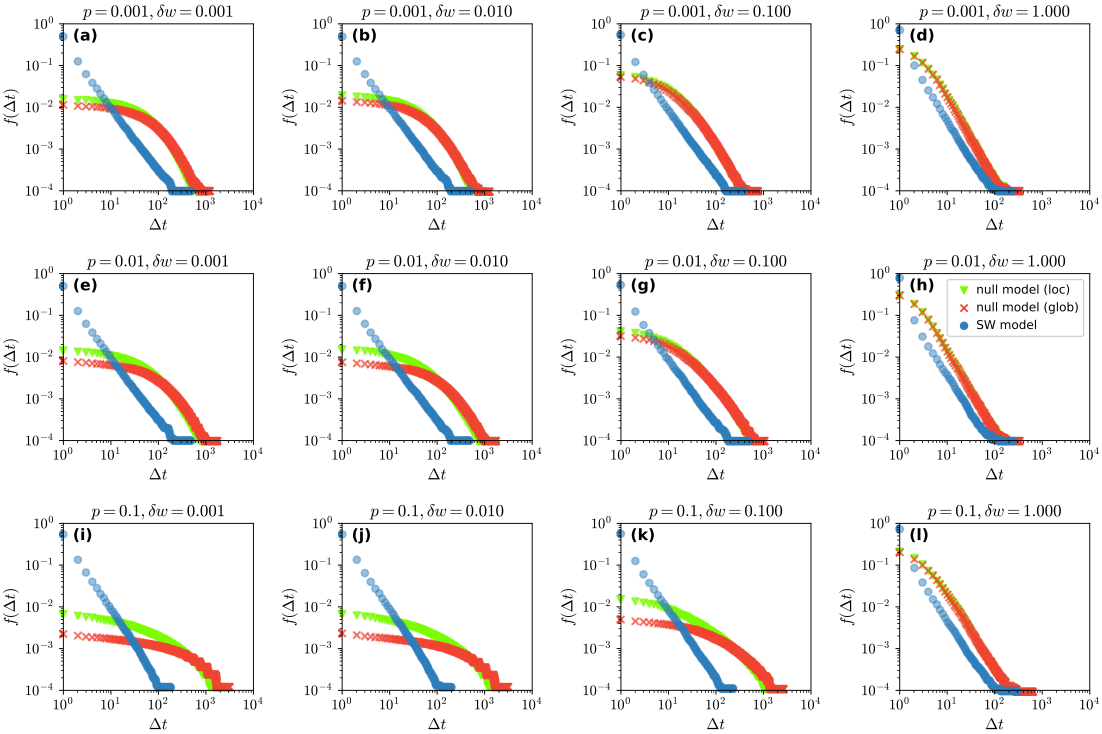


Figure A.5: Correlations among concepts produced by ERRWs on a SW network for different values of link probability p and reinforcement δw (see the main text for details). Frequency distribution of inter-event times Δt between consecutive occurrences of the same concept (node in our model). In each panel, blue circles show average values over 20 different realizations, while triangles and crosses refer to those of (globally and locally) reshuffled sequences. Figure from [1].

C: Interacting discovery processes - Analytical solutions

Here, we will study in more detail the analytical solutions associated to the model of interacting discovery processes presented in Chapter 5. In particular, we focus on the solution of the equation for the pace of innovation we derived in the main text. The case of an individual urn is equivalent to the urn model with triggering and is detailed in Sec. 1.2.2.1. We will thus consider simple cases of coupled urns, such as a pair of nodes, a chain, a cycle, a clique, ending with the formulation for a general network. Moreover, we will derive an analytical solution for each of the small networks studied in Fig. 5.3 of the main text. In every case, we will set the same parameters for each urn, so that $\rho_i = \rho$ (*reinforcement*) and $\nu_i = \nu$ (*triggering*) $\forall i = 1, \ldots, N$. Each urn will be initialized with M_0 balls of different colours. These and the other colours—added from an individual i when triggered by a discovery—will be taken from a single predefined set of discoverable balls of different colours.

Pace of discovery - Two coupled urns

Let us consider now the simplest case of two coupled urns, that is a network with only two nodes connected by a directed edge (1 → 2), as in Fig. 5.1 of the main text. This is equivalent to a directed chain of $N = 2$ nodes, that will be discussed in the next section for a general number N of nodes. The associated equations to determine the asymptotic growth of the number of novelties can be written expressing the probabilities $P_i^{\mathrm{new}}(t)$ to draw a new ball as the the fraction of discoverable balls over the total number of balls available to node i at time t:

$$\begin{cases} \dfrac{dD_1(t)}{dt} = \dfrac{|\tilde{\mathcal{U}}_1(t) \ominus \mathcal{S}_1'(t)|}{\tilde{U}_1(t)} & \text{(A1a)} \\[2ex] \dfrac{dD_2(t)}{dt} = \dfrac{|\tilde{\mathcal{U}}_2(t) \ominus \mathcal{S}_2'(t)|}{\tilde{U}_2(t)} = \dfrac{U_2'(t) - D_2(t)}{U_2(t)}. & \text{(A1b)} \end{cases}$$

Notice that the right-hand side of Eq. (A1b) is simplified since node 2 does not have any outgoing link, and therefore its dynamics is the same of an isolated urn for which

$\tilde{\mathcal{U}}_2(t) = \mathcal{U}_2(t)$. Thus, following the same procedure of the single urn (see Sec.1.2.2.1), we have, for $\rho > \nu$:

$$D_2(t) \sim (\rho - \nu)^{\frac{\nu}{\rho}} t^{\frac{\nu}{\rho}}. \tag{A2}$$

The denominator $\tilde{U}_1(t)$ of Eq. (A1a) can be expressed in terms of the two contributions coming from the two urns at time t, which reads:

$$\begin{aligned} \tilde{U}_1(t) &= \overbrace{M_0 + \rho t + (\nu + 1) D_1(t)}^{U_1(t)} + \overbrace{M_0 + (\nu + 1) D_2(t)}^{U_2'(t)} \\ &= 2M_0 + \rho t + (\nu + 1)\big[D_1(t) + D_2(t)\big]. \end{aligned} \tag{A3}$$

Similarly, the numerator of Eq. (A1a), consisting in the number of balls present in the social urn of node 1 at time t which did not appeared yet in $\mathcal{S}_1(t)$, can be written as the total number of balls in the social urn of 1 at time t, minus the number of duplicates, minus the number of balls that do not represent a novelty any more with respect to the sequence $\mathcal{S}_1(t)$, i.e.:

$$|\tilde{\mathcal{U}}_1(t) \ominus \mathcal{S}_1'(t)| = \tilde{U}_1(t) - \rho t - D_1(t). \tag{A4}$$

Then, using Eq. (A3) and Eq. (A4), the final expression for Eq. (A1a) reads:

$$\frac{dD_1(t)}{dt} = \frac{2M_0 + \nu D_1(t) + (\nu + 1) D_2(t)}{2M_0 + \rho t + (\nu + 1)\big[D_1(t) + D_2(t)\big]}. \tag{A5}$$

For large times ($t \gg M_0$) we can approximate Eq. (A5) as

$$\frac{dD_1(t)}{dt} \approx \frac{\nu D_1(t) + (\nu + 1) D_2(t)}{\rho t + (\nu + 1)\big[D_1(t) + D_2(t)\big]}. \tag{A6}$$

Let us assume now that the dynamics of node 2 relaxes before the one of node 1, so that we can solve Eq. (A6) independently from Eq. (A2). In addition, if we suppose that $\lim_{t\to\infty} D(t)/t = 0$, Eq. (A6) can be approximated as:

$$\frac{dD_1(t)}{dt} \approx \frac{\nu D_1(t)}{\rho t} + \frac{(\nu + 1) D_2(t)}{\rho t}. \tag{A7}$$

The related homogeneous equation has a similar solution of Eq. (A2), i.e.:

$$\frac{d\overline{D}_1(t)}{dt} \approx \frac{\nu \overline{D}_1(t)}{\rho t} \implies \overline{D}_1(t) \sim (\rho - \nu)^{\frac{\nu}{\rho}} t^{\frac{\nu}{\rho}}. \tag{A8}$$

We now look for a solution like $D_1(t) = \kappa(t)\overline{D}_1(t)$ that, plugged into Eq. (A7), leads to:

$$\frac{d\kappa(t)}{dt}\overline{D}_1(t) + \kappa(t)\frac{d\overline{D}_1(t)}{dt} \approx \kappa(t)\frac{d\overline{D}_1(t)}{dt} + \frac{(\nu+1)D_2(t)}{\rho t}. \tag{A9}$$

Thus, from Eq. (A2) and Eq. (A8) we get:

$$\frac{d\kappa(t)}{dt} = \frac{\nu+1}{\rho t}\frac{D_2(t)}{\overline{D}_1(t)} \approx \frac{\nu+1}{\rho t}, \tag{A10}$$

whose solution is

$$\kappa(t) \approx \frac{\nu+1}{\rho} \ln t. \tag{A11}$$

The asymptotic solution ($t \longrightarrow \infty$) of $D_1(t)$ is then approximated by:

$$D_1(t) \sim \frac{\nu+1}{\rho}(\rho - \nu)^{\frac{\nu}{\rho}} \ln(t)\, t^{\frac{\nu}{\rho}}. \tag{A12}$$

In conclusion, comparing the solutions in Eq. (A2) and Eq. (A12) the presence of an outgoing link increases the number of novelties with respect to an isolated urn dynamics. However, as we have shown here, this increase is approximately only logarithmic, meaning that we can see a slight increase at finite times which practically disappears for larger times. Le us also notice that this applies to the directed case, while in the case of an undirected link we would get identical Heaps' laws for both nodes $i = 1, 2$, without logarithmic corrections, but with higher exponents. This particular case is a cycle of two nodes, and as we will see in a dedicated section, cycles have their own behaviour.

Pace of discovery - Chain of N urns

Let us consider now a slightly more complicated case. Let us suppose that the network is composed by an open chain of N urns, where there are only directed links ($i \longrightarrow i+1$), with $i = 1, 2, \ldots, N-1$. This is the case considered in Fig. 5.3(**b,g**) of the main text, where in that case $N = 4$. Analogously to the previous case, the associated set of equations governing the growth of the number of novelties can be approximated to:

$$
\begin{cases}
\dfrac{dD_1(t)}{dt} \approx \dfrac{\nu D_1(t) + (\nu+1)D_2(t)}{\rho t + (\nu+1)\big[D_1(t) + D_2(t)\big]} & \text{(A13a)} \\
\quad\vdots & \text{(A13b)} \\
\dfrac{dD_{N-1}(t)}{dt} \approx \dfrac{\nu D_{N-1}(t) + (\nu+1)D_N(t)}{\rho t + (\nu+1)\big[D_{N-1}(t) + D_N(t)\big]} & \text{(A13c)} \\
\dfrac{dD_N(t)}{dt} \approx \dfrac{\nu D_N(t)}{\rho t + (\nu+1)D_N(t)} & \text{(A13d)}
\end{cases}
$$

We can solve the system by solving each equation, starting from the last one and recursively substituting its solution into the equation above. Indeed, since node $i = N$ does not have any outgoing link, its independent Eq. (A13d) can be immediately solved, resulting in the known asymptotic solution:

$$D_N(t) \sim (\rho - \nu)^{\frac{\nu}{\rho}} t^{\frac{\nu}{\rho}}. \tag{A14}$$

As in the previous case, in Eq. (A13c) we can consider $D_{N-1}(t)$ to be the only unknown variable. Then, following the same analytical steps presented in previous section leads to:

$$D_{N-1}(t) \approx \frac{\nu+1}{\rho}(\rho-\nu)^{\frac{\nu}{\rho}} \ln(t) t^{\nu/\rho}. \tag{A15}$$

The same reasoning can be iterated for each node i. Let us now prove by induction on i that the asymptotic solution is

$$D_i(t) = \frac{(\rho-\nu)^{\nu/\rho}}{(N-i)!} \left(\frac{\nu+1}{\rho} \ln(t)\right)^{N-i} t^{\nu/\rho}. \tag{A16}$$

We have already proved that this holds for $i = N$ and $i = N - 1$. Let us now suppose that it holds for i and let us prove it for $i - 1$, with $1 < i < N$. In the asymptotic limit, the equation for the growth of the number of novelties of node i reads

$$\frac{dD_{i-1}(t)}{dt} \approx \frac{\nu D_{i-1}(t) + (\nu+1)D_i(t)}{\rho t + (\nu+1)\left[D_{i-1}(t) + D_i(t)\right]}. \tag{A17}$$

For the induction hypothesis, in Eq. (A17) the only unknown variable is $D_i(t)$. Therefore, we can consider the homogeneous associated equation

$$\frac{d\overline{D}_{i-1}(t)}{dt} \approx \frac{\nu \overline{D}_{i-1}(t)}{\rho t}, \tag{A18}$$

which provides the approximated solution:

$$\overline{D}_{i-1}(t) \approx (\rho - \nu)^{\frac{\nu}{\rho}} t^{\frac{\nu}{\rho}}. \tag{A19}$$

As for the case of two coupled urns, we now look for a solution like $D_{i-1}(t) = \kappa(t)\overline{D}_{i-1}(t)$, that, plugged into Eq. (A17), leads to:

$$\frac{d\kappa(t)}{dt}\overline{D}_{i-1}(t) + \cancel{\kappa(t)\frac{d\overline{D}_{i-1}(t)}{dt}} \approx \cancel{\kappa(t)\frac{d\overline{D}_{i-1}(t)}{dt}} + \frac{(\nu+1)D_i(t)}{\rho t}. \tag{A20}$$

Thus, we get

$$\frac{d\kappa(t)}{dt} \approx \frac{\nu+1}{\rho t}\frac{D_i(t)}{\overline{D}_{i-1}(t)} \approx \frac{1}{(N-i)!}\frac{\nu+1}{\rho t}\left(\frac{\nu+1}{\rho}\ln(t)\right)^{N-i}, \tag{A21}$$

whose solution is

$$\kappa(t) \approx \frac{1}{(N-(i-1))!}\left(\frac{\nu+1}{\rho}\ln(t)\right)^{N-(i-1)}. \tag{A22}$$

Finally, after combining Eq. (A19) and Eq. (A22), we reach the solution for the dynamics of node $i-1$, that reads:

$$D_{i-1}(t) \approx \frac{(\rho-\nu)^{\nu/\rho}}{(N-(i-1))!}\left(\frac{\nu+1}{\rho}\ln(t)\right)^{N-(i-1)} t^{\nu/\rho}, \tag{A23}$$

which completes the proof by induction.

Finally, it is worth observing that the Heaps' laws would be very different if the links were undirected. This would indeed result, similarly to undirected cycles, in higher asymptotic Heaps' exponents.

Pace of discovery - Cycle of N urns

Directed cycle—Let us consider the case of directed cycles. As we will see, this is the simplest system leading to asymptotic Heaps' exponents that are higher than that of an individual urn. Let us hence suppose that every node i is connected just to the following one, node $i+1$, with a directed link $(i \to i+1)$, with $i = 1, 2, \dots, N$, where we identify node $N+1$ with node 1. For a generic node i, the asymptotic differential equation for the growth of the number of novelties reads:

$$\frac{dD_i(t)}{dt} \approx \frac{\nu D_i(t) + (\nu+1)D_{i+1}(t)}{\rho t + (\nu+1)\big[D_i(t) + D_{i+1}(t)\big]}. \tag{A24}$$

For symmetry reasons, the dynamics of each node is the same, implying that $D_1(t) \approx \cdots \approx D_i(t) \approx \cdots \approx D_N(t)$. Hence, Eq. (A24) becomes

$$\frac{dD_i(t)}{dt} \approx \frac{(2\nu+1)D_i(t)}{\rho t + 2(\nu+1)D_i(t)}, \tag{A25}$$

that is equal to the equation of an individual urn [see Eq. (1.4)], with $\nu' = 2\nu+1$. Therefore, from Sec. 1.2.2.1, if $\rho > \nu'$ we have the solution

$$D_i(t) \approx (\rho - 2\nu - 1)^{\frac{2\nu+1}{\rho}}\, t^{\frac{2\nu+1}{\rho}}. \tag{A26}$$

Undirected cycle—Let us now consider cycles composed by undirected links. Let us suppose that $N > 2$, considered that for $N = 1$ the network reduces to an individual urn, and for $N = 2$ it is equivalent to a directed cycle of 2 nodes. For $N > 2$, each node i is therefore connected to two different nodes $i-1$ and $i+1$, and the associated equations to be solved are:

$$\frac{dD_i(t)}{dt} \approx \frac{\nu D_i(t) + (\nu+1)D_{i-1}(t) + (\nu+1)D_{i+1}(t)}{\rho t + (\nu+1)\big[D_i(t) + (\nu+1)D_{i-1}(t) + D_{i+1}(t)\big]}. \tag{A27}$$

Again, for symmetry reasons, we can equivalently write Eq. (A27) as

$$\frac{dD_i(t)}{dt} \approx \frac{(3\nu+2)D_i(t)}{\rho t + 3(\nu+1)D_i(t)}, \tag{A28}$$

that is equal to the equation of an individual urn [see Eq. (1.4)], with $\nu'' = 3\nu + 2$. Therefore, if $\rho > \nu''$ we have the solution

$$D_i(t) \approx (\rho - 3\nu - 2)^{\frac{3\nu+2}{\rho}} t^{\frac{3\nu+2}{\rho}}. \tag{A29}$$

Undirected cycle—Let us now consider cycles composed by undirected links. Let us suppose that $N > 2$, considered that for $N = 1$ the network reduces to an individual urn, and for $N = 2$ it is equivalent to a directed cycle of 2 nodes. For $N > 2$, each node i is therefore connected to two different nodes $i - 1$ and $i + 1$, and the associated equations to be solved are:

$$\frac{dD_i(t)}{dt} \approx \frac{\nu D_i(t) + (\nu+1)D_{i-1}(t) + (\nu+1)D_{i+1}(t)}{\rho t + (\nu+1)\big[D_i(t) + (\nu+1)D_{i-1}(t) + D_{i+1}(t)\big]}. \tag{A30}$$

Again, for symmetry reasons, we can equivalently write Eq. (A30) as

$$\frac{dD_i(t)}{dt} \approx \frac{(3\nu+2)D_i(t)}{\rho t + 3(\nu+1)D_i(t)}, \tag{A31}$$

that is equal to the equation of an individual urn [see Eq. (1.4)], with $\nu'' = 3\nu + 2$. Therefore, if $\rho > \nu''$ we have the solution

$$D_i(t) \approx (\rho - 3\nu - 2)^{\frac{3\nu+2}{\rho}} t^{\frac{3\nu+2}{\rho}}. \tag{A32}$$

Notice that for undirected cycles, since all connections are mutual, the resulting paces of discovery are higher than those in the directed case. However, in both cases, directed and undirected, the dynamics of each node does not depend on the length of the cycle.

Pace of discovery - Clique of N urns

Let us consider a N-clique, that is a fully connected network of N nodes, equivalently directed or undirected. Being every node i connected to all other nodes, all nodes are equivalent, and the general equation for the growth of the number of novelties of node i reads:

$$\frac{dD_i(t)}{dt} \approx \frac{\nu D_i(t) + (\nu+1)\sum_{j\neq i} D_j(t)}{\rho t + (\nu+1)\sum_{j=1}^{N} D_j(t)}. \tag{A33}$$

For symmetry reasons, each urn follows the same dynamics and we can equivalently write Eq. (A33) as

$$\frac{dD_i(t)}{dt} \approx \frac{[N(\nu+1)-1]D_i(t)}{\rho t + N(\nu+1)D_i(t)}, \tag{A34}$$

that is equal to the equation for an individual urn [see Eq. (1.4)], with $\nu''' = N(\nu+1) - 1$. Therefore, if $\rho > \nu'''$ we have the solution

$$D_i(t) \approx (\rho - N(\nu+1) - 1)^{\frac{N(\nu+1)-1}{\rho}} t^{\frac{N(\nu+1)-1}{\rho}}. \tag{A35}$$

Let us observe that for any network with N nodes, the maximum allowed Heaps' exponent is hence $[N(\nu+1)-1]/\rho$, which occurs only in the case of a fully connected network.

Pace of discovery - The general case

Let us consider a general graph $G(\mathcal{N},\mathcal{E})$, either directed or undirected. In order to write and solve the equations for the growth of the number of novelties, we first have to calculate the probability $P_i^{\text{new}}(t)$ of drawing a new ball from the urn of each node i. This can be done by considering the number of different colours present in the social urn $\tilde{\mathcal{U}}_i(t)$ of node i at time t that have not been discovered yet by i, divided by the total number of balls $\tilde{U}_i(t)$ present in its social urn at that time. The numerator can be expressed as $|\tilde{\mathcal{U}}_i(t) \ominus \mathcal{S}'_i(t)|$, which is the length of the multiset obtained by removing from the multiset $\tilde{\mathcal{U}}_i(t)$ all the elements appeared in the sequence (taking out all duplicates). In other words, it is the number of unique colours present in the urn of node i and in the one of its neighbours (without their multiplicity) minus the number of colours already drawn (unique elements in the sequence of i). Considering that all (and only) the already discovered balls are those that have been reinforced and that the number of triggered colours added to the urn j is exactly $(\nu + 1)D_j(t)$, we can write:

$$\frac{dD_i(t)}{dt} = P_i^{\text{new}}(t) = \frac{|\tilde{\mathcal{U}}_i(t) \ominus \mathcal{S}'_i(t)|}{\tilde{U}_i(t)} = \frac{M_0 + \nu D_i(t) + \sum_{j \neq i} a_{ij}\left[M_0 + (\nu+1)D_j(t)\right]}{\rho t + M_0 + (\nu+1)D_i(t) + \sum_{j\neq i} a_{ij}\left[M_0 + (\nu+1)D_j(t)\right]}, \tag{A36}$$

or, equivalently:

$$\frac{dD_i(t)}{dt} = \frac{M_0 \sum_j (a_{ij} + \delta_{ij}) + \sum_j \left[\delta_{ij}\nu + a_{ij}(\nu+1)\right] D_j(t)}{\rho t + \sum_j (a_{ij} + \delta_{ij})\left[M_0 + (\nu+1)D_j(t)\right]}. \tag{A37}$$

For $t \gg M_0$ we can disregard the presence of M_0 in Eq. (A37). As shown above for N-cliques, in the asymptotic limit $t \to \infty$ the growth of the number of novelties obeys an Heaps' law with maximum exponent $[N(\nu+1) - 1]/\rho$. This means that if ρ is high enough, we can approximate the denominator on the r.h.s. of Eq. (A37) to ρt. After finding the approximated solution, we will estimate the set of parameters for which this approximation is valid for any topology. Therefore, in the asymptotic limit and with a proper choice of the parameters, Eq. (A37) can be rewritten as:

$$\frac{dD_i(t)}{dt} \approx \frac{\sum_j \left[\delta_{ij}\nu + a_{ij}(\nu+1)\right] D_j(t)}{\rho t}, \tag{A38}$$

which can be expressed in a more compact way as:

$$\frac{d\vec{D}(t)}{dt} \approx \frac{1}{t}\left(\frac{\nu}{\rho}\boldsymbol{I} + \frac{\nu+1}{\rho}\boldsymbol{A}\right)\vec{D}(t) = \frac{1}{t}\frac{f(\boldsymbol{A})\vec{D}(t)}{t} = \frac{1}{t}\boldsymbol{M}\vec{D}(t), \tag{A39}$$

where $\boldsymbol{I}$ is the $N \times N$ identity matrix and $\boldsymbol{M} = f(A)$, with $f(x) = \frac{\nu}{\rho} + \frac{\nu+1}{\rho}x$. By operating the change of variable $t = e^z$, Eq. (A39) can be rewritten as a standard first-order differential system, i.e. $d_z\vec{D}(z) \approx \boldsymbol{M}\vec{D}(z)$, which leads to the solution

$$\vec{D}(t) \approx \sum_{\ell=1}^{r} \sum_{p=0}^{m_\ell - 1} \vec{c}_p \ln^p(t)\, t^{\lambda_\ell}, \tag{A40}$$

where $\{\lambda_\ell\}_{\ell=1,\dots,r}$ and $\{m_\ell\}_{\ell=1,\dots,r}$ are the eigenvalues of $\boldsymbol{M}$ with their respective multiplicities, and $\vec{c}_p$ are vectors defined by the initial conditions. The asymptotic behaviour of the number of novelties $D_i(t)$ discovered by node i at time t is then governed by the leading term in Eq. (A40), so that we can write:

$$D_i(t) \underset{t\to\infty}{\approx} u_i \ln^{\widehat{p}(i)}(t)\, t^{\widehat{\lambda}(i)}. \tag{A41}$$

where $\widehat{\lambda}(i)$ is the eigenvalue of $\boldsymbol{M}$ with the biggest real part such that the i-th entry of at least one of its eigenvectors $\vec{c}_p$ is different from zero. Similarly, $\widehat{p}(i)$ is the maximum value of p among these eigenvectors with non-zero i-th entries. In general, then, $\widehat{\lambda}(i)$ might not be the maximum eigenvalue of $\boldsymbol{M}$, like $\widehat{p}(i)$ might be less than the multiplicity of the eigenvalue $\widehat{\lambda}(i)$ minus one. Moreover, different nodes may have different values for these exponents. In particular, we have the same exponents for nodes in the same strongly connected components (SCCs), while they may vary from SCC to SCC. In the following paragraphs we will investigate this aspect.

Strongly connected network—Let us suppose that the graph $G(\mathcal{N}, \mathcal{E})$ is strongly connected. In this case the solution given by Eq. (A41) simplifies. Indeed, in this case, the corresponding adjacency matrix $\boldsymbol{A} = \{a_{ij}\}$ is irreducible [416]. Let us recall that for irreducible matrices the Perron–Frobenius theorem holds [379, 380], according to which there exists a positive eigenvalue $\widehat{\mu}$ greater or equal to (in absolute value) all other eigenvalues. Such eigenvalue corresponds to a simple root of the characteristic equation and the corresponding eigenvector $\vec{u}$ has all positive entries too. The latter vector is a multiple of the Bonacich eigenvector centrality vector [381]. Widely used in network science, the Bonacich eigenvector centrality is a measure that recursively accounts for local and global properties of the network, relying on the notion that a node can be highly central either by having a high degree or by being connected to others that themselves are highly central [18]. Simple algebraic steps can prove that if μ is an eigenvalue for $\boldsymbol{A}$, then $\lambda = f(\mu)$ is an eigenvalue for $\boldsymbol{M}$. Moreover, if $\vec{u}$ is an eigenvector corresponding to the eigenvalue μ of $\boldsymbol{A}$, then $\vec{u}$ is also an eigenvector corresponding to the eigenvalue $\lambda = f(\mu)$ of $\boldsymbol{M}$. Therefore, if $\widehat{\mu}$ is the maximum eigenvalue of $\boldsymbol{A}$, then $\widehat{\lambda} = f(\widehat{\mu}) = \frac{\nu}{\rho} + \frac{\nu+1}{\rho}\widehat{\mu} > 0$ is the highest eigenvalue of $\boldsymbol{M}$, and with the same positive eigenvector $\vec{u}$. Thus, for strongly connected graphs, the approximated solution given by Eq. (A41) becomes

$$D_i(t) \underset{t\to\infty}{\approx} u_i\, t^{\widehat{\lambda}}, \tag{A42}$$

meaning that all nodes have similar Heaps' laws, and the key difference is made by their eigenvector centrality. As we saw in the main text (and we will see here more in details), these differences, more pronounced in transient times, will contribute to determine the fastest explorers in the network. Moreover, we deduce that the approximation used in Eq. (A38) is valid provided that $\widehat{\lambda} = f(\widehat{\mu}) < 1$, that is $\rho > \nu + (\nu + 1)\widehat{\mu}$, while for higher values of ρ the solution is bounded by the linear solution as seen for the individual urn in Eq. (1.7), since in the original system in Eq. (A36) we have $d_t D_i(t) \leq 1$.

Non-strongly connected network—Let us now consider the most general case, that is a directed or undirected graph with any hypotheses of connectivity. Let us construct an algorithm to determine the pace of discovery of each node, which will help us better understand analytically why some nodes have higher paces of discovery. To do this, let us partition the graphs into its strongly connected components (SCCs), i.e. maximal strongly connected subgraphs of G, which can be found in linear computational time, for example with a DFS-based algorithm [417]. Let all the SCCs be indexed as $C_1, \dots, C_p$, with $C_i \cap C_j = \varnothing\ \forall i \neq j$.

Without loss of generality, let us suppose that the graph G is weakly connected, because otherwise we can repeat the same reasoning for each weakly connected component. Let us also suppose that the number of SCCs is $p > 1$, because otherwise the graph would be strongly connected, which we already discussed in the previous paragraph. Since G is weakly connected, for each SCC C_q there must exist another component C_l, with $l \neq q$, such that there are some links from C_q to C_l or viceversa. However, there cannot be links in both directions (from C_q to C_l and viceversa), because otherwise they would be a unique SCC. It is also easy to show that there is always a SCC without any outgoing links to other SCCs. Eventually permutating the indexes of the SCCs, let us call $C_1, \ldots, C_{p_1}$ all the components with no outer links. Then, for each $1 \leq q \leq p_1$, the respective system of differential equations for D_i, $i \in C_q$, does not depend on any outer variable D_j, $j \in C_l \neq C_q$. Therefore, we can consider C_q as an independent strongly connected subgraph of G, for which the reasoning in last paragraph holds. The solution for these SCCs is then:

$$D_i(t) \underset{t\to\infty}{\approx} \gamma_i^{(q)}\, t^{\widehat{\lambda}^{(q)}}\ \forall i \in C_q, 1 \leq q \leq p_1, \tag{A43}$$

where $\widehat{\lambda}^{(q)}$ is the maximum eigenvalue of the adjacency matrix of subgraph C_q and $\gamma_i^{(q)}$ is a multiple of the eigenvector centrality for node i in C_q. Found all the Heaps' laws relative to the nodes in $C_1, \ldots, C_{p_1}$, it is possible to show that there exist SCCs $C_{p_1+1}, \ldots, C_{p_2}$ that have links only towards the previously studied SCCs $C_1, \ldots, C_{p_1}$, with $p_2 > p_1$. Then, choosing C_q one of these other SCCs, let $\overline{\lambda}^{(q)}$ be the highest eigenvalue of the adjacency matrix of C_q. Let also $\tilde{\lambda}^{(q)} = \max_{l \leq p_1}(\gamma_{ql}\widehat{\lambda}^{(l)})$ be the maximum of the Heaps' exponents in Eq. (A43) of the SCCs reachable from C_q, where $\gamma_{qr} = 1$ if there is at least a link from C_q to C_l, $\gamma_{qr} = 0$ otherwise. As we will see further in this section, the Heaps' solutions for the nodes in these SCCs is:

$$D_i(t) \underset{t\to\infty}{\approx} \begin{cases} \gamma_i^{(q)}\, t^{\overline{\lambda}^{(q)}} & \text{if } \overline{\lambda}^{(q)} > \tilde{\lambda}^{(q)} \\ \gamma_i^{(q)} \ln(t)\, t^{\tilde{\lambda}^{(q)}} & \text{if } \overline{\lambda}^{(q)} = \tilde{\lambda}^{(q)} \\ \gamma_i^{(q)}\, t^{\tilde{\lambda}^{(q)}} & \text{if } \overline{\lambda}^{(q)} < \tilde{\lambda}^{(q)} \end{cases} \quad \forall i \in C_q,\ p_1 + 1 \leq q \leq p_2, \tag{A44}$$

meaning that the Heaps' exponent $\widehat{\lambda}^{(q)}$ for node i in C_q, $p_1 + 1 \leq q \leq p_2$, is

$$\widehat{\lambda}^{(q)} = \max(\overline{\lambda}^{(q)},\ \tilde{\lambda}^{(q)}), \tag{A45}$$

that is the maximum of the highest eigenvalue $\overline{\lambda}^{(q)}$ of $\boldsymbol{M}$ relative to C_q and the highest $\tilde{\lambda}^{(q)}$ of the Heap's exponents $\widehat{\lambda}^{(l)}$ for $1 \leq l \leq p_1$. Moreover, if $\overline{\lambda}^{(q)} = \tilde{\lambda}^{(q)}$, a factor $\ln(t)$ appears in the solution. The same procedure can be repeated for all other successive SCCs C_q, keeping in mind that now a higher power $\ln^{\widehat{p}(q)}(t)$ of $\log(t)$ can appear.

In this algorithmic process, let us now consider a generic SCC, say C_q, and let us suppose we have solved inductively all the equations for the Heaps' law of the nodes in the already examined SCCs, that is $C_1, \ldots, C_{q-1}$. Let us recall that we arranged the indexes in such a way that the only outgoing links from C_q are pointed to nodes in previous SCCs, i.e. in some of the SCCs $C_1, \ldots, C_{q-1}$. For this reason, in order to solve the asymptotic differential equations responsible for the Heaps' law of the nodes in C_q, we can consider only the equations relative to the nodes in C_q in Eq. (A39), since the previous SCCs have been already solved and the following variables do not appear in these equations. We hence have to solve the following approximated equations:

$$\frac{dD_i(t)}{dt} \approx \frac{1}{t}\left(\frac{\nu}{\rho}D_i(t) + \frac{\nu+1}{\rho}\sum_{j \in C_q} a_{ij}D_j(t) + \frac{\nu+1}{\rho}\sum_{j \notin C_q} a_{ij}D_j(t)\right), \quad i \in C_q, \tag{A46}$$

where we have isolated the contributions coming from nodes outside C_q, which we have suppose to be known. Considering the general asymptotic solution for each individual Heaps' law derived for a strongly connected graph in Eq. (A41), for each $i \in C_q$ for large t we can write explicitly the functions $D_j(t)$, $j \notin C_q$, which lets us write:

$$\frac{\nu+1}{\rho}\sum_{j \notin C_q} a_{ij}D_j(t) \approx \frac{\nu+1}{\rho}\sum_{j \notin C_q} a_{ij}u_j \ln^{\widehat{p}_j}(t)\, t^{\widehat{\lambda}_j} \underset{t\to\infty}{\approx} \tilde{u}_i \ln^{\tilde{p}^{(q)}}(t)\, t^{\tilde{\lambda}^{(q)}} = f_i(t), \tag{A47}$$

where we have used the fact that $\eta_i \ln^{\tilde{p}_i}(t)\, t^{\tilde{\lambda}_i}$ is the leading term of the expression $\sum_{j \notin C_q} a_{ij}u_j \ln^{\widehat{p}_j}(t)\, t^{\widehat{\lambda}_j}$ and that we are working for large t. Then, using Eq. (A47) and calling $\vec{D}^{(q)}$ and $\boldsymbol{A}^{(q)}$ the sub-vector of $\vec{D}$ and sub-matrix of $\boldsymbol{M}$ relative to C_q, we can rewrite Eq. (A46) in a compact form as

$$\frac{d\vec{D}^{(q)}(t)}{dt} \approx \frac{\boldsymbol{M}^{(q)}\vec{D}^{(q)}(t)}{t} + \frac{\vec{f}^{(q)}(t)}{t}. \tag{A48}$$

The associated homogeneous system corresponds to the considering the sub-graph C_q without all the external links. For this system we get the same solution derived for a strongly connected graph in Eq. (A42), which is

$$\vec{\overline{D}}^{(q)}(t) \underset{t\to\infty}{\approx} \vec{\overline{u}}^{(q)}\, t^{\overline{\lambda}^{(q)}}, \tag{A49}$$

where $\overline{\lambda}^{(q)}$ is the highest eigenvalue of $\boldsymbol{M}^{(q)}$ (positive and simple for the Perron-Frobenious theorem), and $\vec{u}^{(q)}$ is a multiple of the eigenvector centrality. Let us search a solution for Eq. (A48) of the form $\vec{D}^{(q)}(t) = \vec{u}^{(q)}(t) \circ \vec{\overline{D}}^{(q)}(t)$, where $\circ$ is the Hadamard (element-wise) product, that plugged in Eq. (A48) gives:

$$\frac{d\vec{u}^{(q)}(t)}{dt} \circ \vec{\overline{D}}^{(q)}(t) + \cancel{\vec{u}^{(q)}(t) \circ \frac{d\left[\vec{\overline{D}}^{(q)}(t)\right]}{dt}} \approx \cancel{\vec{u}^{(q)}(t) \circ \frac{\boldsymbol{M}^{(q)}\vec{\overline{D}}^{(q)}(t)}{t}} + \frac{\vec{f}^{(q)}(t)}{t}, \tag{A50}$$

where the cancellation is due to the general solution in Eq. (A49) of the associated homogeneous system. Therefore, recalling Eq. (A47) and Eq. (A49) we have:

$$\frac{d\vec{u}^{(q)}(t)}{dt} \approx \vec{\tilde{u}}^{(q)} \circ \left(\vec{\overline{u}}^{(q)}\right)^{-1} \frac{\ln^{\tilde{p}^{(q)}}(t)\, t^{\tilde{\lambda}^{(q)}}}{t^{\overline{\lambda}+1}} = \vec{\gamma}\, \frac{\ln^{\tilde{p}^{(q)}}(t)\, t^{\tilde{\lambda}^{(q)}}}{t^{\overline{\lambda}+1}}, \tag{A51}$$

or equivalently, considering the i-th components:

$$\frac{du_i(t)}{dt} \approx \tilde{u}_i^{(q)} \left[\left(\vec{\overline{u}}^{(q)}\right)^{-1}\right]_i \frac{\ln^{\tilde{p}^{(q)}}(t)\, t^{\tilde{\lambda}^{(q)}}}{t^{\overline{\lambda}+1}} = \gamma_i \frac{\ln^{\tilde{p}^{(q)}}(t)\, t^{\tilde{\lambda}^{(q)}}}{t^{\overline{\lambda}+1}}, \tag{A52}$$

where we have defined $\vec{\gamma} = \vec{\tilde{u}}^{(q)} \circ \left(\vec{\overline{u}}^{(q)}\right)^{-1}$ and $\gamma_i = \tilde{u}_i^{(q)} \left[\left(\vec{\overline{u}}^{(q)}\right)^{-1}\right]_i$ its i-th component. Let us hence distinguish three cases.

1. If $\overline{\lambda}^{(q)} > \tilde{\lambda}^{(q)}$, then we have:

$$u_i(t) \approx \frac{\gamma_i}{\tilde{\lambda}^{(q)} - \overline{\lambda}^{(q)}} \ln^{\tilde{p}^{(q)}}(t)\, t^{\tilde{\lambda}^{(q)} - \overline{\lambda}^{(q)}} + u_i \underset{t\to\infty}{\approx} u_i, \tag{A53}$$

   which gives the solution:

$$D_i(t) \underset{t\to\infty}{\approx} u_i t^{\overline{\lambda}^{(q)}}. \tag{A54}$$

2. Similarly, for $\overline{\lambda}^{(q)} = \tilde{\lambda}^{(q)}$ we have:

$$u_i(t) \approx \frac{\gamma_i}{\tilde{p}^{(q)} + 1} \ln^{\tilde{p}^{(q)}+1}(t) + u_i \underset{t\to\infty}{\approx} \frac{\gamma_i}{\tilde{p}^{(q)} + 1} \ln^{\tilde{p}^{(q)}+1}(t), \tag{A55}$$

   which gives:

$$D_i(t) \underset{t\to\infty}{\approx} u_i \ln^{\tilde{p}^{(q)}+1}(t)\, t^{\tilde{\lambda}^{(q)}}. \tag{A56}$$

3. Finally, if $\overline{\lambda}^{(q)} < \tilde{\lambda}^{(q)}$ we have:

$$u_i(t) \approx \frac{\gamma_i}{\tilde{\lambda}^{(q)} - \overline{\lambda}^{(q)}} \ln^{\tilde{p}^{(q)}}(t)\, t^{\tilde{\lambda}^{(q)} - \overline{\lambda}^{(q)}} + d_1 \underset{t\to\infty}{\approx} \frac{\gamma_i}{\tilde{\lambda}^{(q)} - \overline{\lambda}^{(q)}} \ln^{\tilde{p}^{(q)}}(t)\, t^{\tilde{\lambda}^{(q)} - \overline{\lambda}^{(q)}}, \tag{A57}$$

   hence the solution:

$$D_i(t) \underset{t\to\infty}{\approx} a_i \ln^{\tilde{p}^{(q)}}(t)\, t^{\tilde{\lambda}^{(q)}}. \tag{A58}$$

To sum up, we have the following solutions:

$$D_i(t) \underset{t\to\infty}{\approx} \begin{cases} u_i t^{\overline{\lambda}^{(q)}} & \text{if } \overline{\lambda}^{(q)} > \tilde{\lambda}^{(q)} \\ u_i \ln^{\tilde{p}^{(q)}+1}(t)\, t^{\tilde{\lambda}^{(q)}} & \text{if } \overline{\lambda}^{(q)} = \tilde{\lambda}^{(q)} \quad \forall i \in C_q, q > p_1, \\ u_i \ln^{\tilde{p}^{(q)}}(t)\; t^{\tilde{\lambda}^{(q)}} & \text{if } \overline{\lambda}^{(q)} < \tilde{\lambda}^{(q)} \end{cases} \tag{A59}$$

Comparing this solution with the general one we gave in Eq. (A41), we have (a) $\widehat{\lambda}(i) = \overline{\lambda}^{(q)}$ and $\widehat{p}(i) = 0$ if $\overline{\lambda} > \tilde{\lambda}$, (b) $\widehat{\lambda}(i) = \tilde{\lambda}^{(q)}$ and $\widehat{p}(i) = \tilde{p}^{(q)} + 1$ if $\overline{\lambda} = \tilde{\lambda}$, and (c) $\widehat{\lambda}(i) = \tilde{\lambda}^{(q)}$ and $\widehat{p}(i) = \tilde{p}^{(q)}$ if $\overline{\lambda} < \tilde{\lambda}$.

In conclusion, when dealing with a network with multiple strength connected components, we solve the equations for the components that are independent from the others. Then we consider the SCCs that have links only to previous SCCs, applying the method just described. This is repeated until every SCC is studied, thus solving the whole system and describing the pace of discovery of each node of the entire network analytically, obtaining solutions of the type in Eq. (A41). In the next section this algorithmic method is applied to simple networks with $N = 4$ nodes, as we have already implicitly done above for a two nodes network and for chains.

Pace of discovery - Five small toy graphs

As an application of the analytical results of the previous sections, we study here the very same five networks depicted in Fig. 5.3 of the main text and reported again here in Table A1. In particular, we will be able to provide an explicit expression for the growth of the number of novelties at each of the four nodes of the social networks. The final results were already reported in Table 5-B of the main text.

Graph a—Let us consider a network where nodes 2, 3, and 4 do not have any outgoing links, while node 1 has the links 1 → 2, 1 → 3, and 1 → 4 to all other nodes. Let us observe that the dynamics here is very similar to the case of a couple of urns with the only link 1 → 2. Nodes 2, 3, and 4 can be considered as three individual urns, for which the Heaps' law is the same to the classic one in Eq. (A2), that is:

$$D_2(t) \underset{t\to\infty}{\approx} D_3(t) \underset{t\to\infty}{\approx} D_4(t) \underset{t\to\infty}{\approx} (\rho-\nu)^{\frac{\nu}{\rho}} t^{\frac{\nu}{\rho}}. \tag{A60}$$

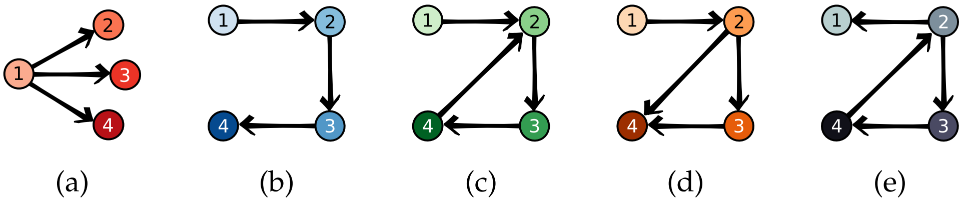


Table A1: Five small toy graphs.

As for node 1, the differential equation for the Heaps' law is approximated by:

$$\frac{dD_1(t)}{dt} \approx \frac{\nu D_1(t)}{\rho t} + \frac{(\nu+1)\big(D_2(t)+D_3(t)+D_4(t)\big)}{\rho t} \approx \frac{\nu D_1(t)}{\rho t} + \frac{3(\nu+1)(\rho-\nu)^{\frac{\nu}{\rho}} t^{\frac{\nu}{\rho}}}{\rho t}. \quad \text{(A61)}$$

The resolution of Eq. (A61) is the same as the one done for the couple of urns, with only a multiplicative factor 3. Therefore, the Heaps' solution for node 1 is:

$$D_1(t) \underset{t\to\infty}{\approx} 3\,\frac{\nu+1}{\rho}(\rho-\nu)^{\frac{\nu}{\rho}}\ln(t)\, t^{\frac{\nu}{\rho}}, \qquad \text{(A62)}$$

which means that node 1 has a higher pace of discovery than nodes 2, 3, and 4, but at asymptotic times they will show the same Heaps' exponent. Moreover, it is clear that in star-like networks adding more nodes does not increase significantly the pace of discovery.

Graph b—The next network we studied is a chain of 4 nodes, with links 1 ⟶ 2, 2 ⟶ 3, and 3 ⟶ 4. This network has already been studied above in the Appendix, and the solutions are:

$$D_i(t) \approx \frac{(\rho-\nu)^{\nu/\rho}}{(4-i)!}\left(\frac{\nu+1}{\rho}\ln(t)\right)^{4-i} t^{\nu/\rho}, \quad i = 1,2,3,4. \qquad \text{(A63)}$$

This analytical result shows us why node 1 has an higher pace of discovery than the other nodes, with lower Heaps' exponents for higher nodes. This is due to the presence of different powers of the logarithm. In the end, however, they all have the same asymptotic Heaps' exponent, meaning that the difference is visible only at finite times.

Graph c—Let us consider a network made by a directed cycle between nodes 2, 3 and 4, with links 2 ⟶ 3, 3 ⟶ 4, and 4 ⟶ 2, and another node 1 linked directly to node 2 (1 ⟶ 2). In this case, we can distinguish two SCCs, the cycle and node 1. Since there is no link going out from the cycle, we start solving the Heaps' law equations related to it. As we have seen in the dedicated section above in the Appendix, the solution is given by

Eq. (A26) with N = 3, that is:

$$D_i(t) \underset{t\to\infty}{\approx} (\rho - 2\nu - 1)^{\frac{2\nu+1}{\rho}} t^{\frac{2\nu+1}{\rho}}, \quad i = 2, 3, 4. \tag{A64}$$

Now let us consider the remaining SCC, namely node 1. Its equation is the same as Eq. (A6) for the case of two coupled urns , with the only difference that here the solution of $D_2(t)$ has a higher exponent. Then, if we search for a solution like $D_1(t) = \kappa(t)\overline{D}_1(t)$, with $\overline{D}_1(t) \approx (\rho - \nu)^{\frac{\nu}{\rho}} t^{\frac{\nu}{\rho}}$ being the solution of the associated homogeneous equation, we get:

$$\frac{d\kappa(t)}{dt} = \frac{\nu+1}{\rho t}\frac{D_2(t)}{\overline{D}_1(t)} \approx \frac{\nu+1}{\rho t}\frac{(\rho-2\nu-1)^{\frac{2\nu+1}{\rho}} t^{\frac{2\nu+1}{\rho}}}{(\rho-\nu)^{\frac{\nu}{\rho}} t^{\frac{\nu}{\rho}}} = \frac{\nu+1}{\rho}\frac{(\rho-2\nu-1)^{\frac{2\nu+1}{\rho}} t^{\frac{\nu+1}{\rho}-1}}{(\rho-\nu)^{\frac{\nu}{\rho}}}, \tag{A65}$$

whose solution is:

$$\kappa(t) \approx \frac{\nu+1}{\rho}\frac{(\rho-2\nu-1)^{\frac{2\nu+1}{\rho}} t^{\frac{\nu+1}{\rho}}}{(\rho-\nu)^{\frac{\nu}{\rho}}}, \tag{A66}$$

which gives the asymptotic solution:

$$D_1(t) \approx \frac{\nu+1}{\rho}(\rho-2\nu-1)^{\frac{2\nu+1}{\rho}} t^{\frac{2\nu+1}{\rho}}. \tag{A67}$$

We could have obtained the same result using the algorithm developed in the last section. In this case, node 1 gets the same dynamics of the nodes in the cycle, with just a scaling factor $(\nu + 1)/\rho$, since the maximum eigenvalue of its SCC (node 1 itself) is lower than the maximum eigenvalue of the SCCs he is linked to (the cycle).

Graph d—In this case we consider the same network as the last graph we just analysed swapping the direction of the link 4 → 2. Therefore, the cycle is broken (see network representation in Table 5-B), and as we are about to see, the dynamics is much more similar to a chain. We could give a detailed solution as done for the chain; instead, we are going to use directly the algorithm we developed to assess all the exponents in the Heaps' laws for every node. Let us start from node 4, which has no outgoing links. This node is hence an individual urn, with the usual solution:

$$D_4(t) \underset{t\to\infty}{\approx} (\rho-\nu)^{\frac{\nu}{\rho}} t^{\frac{\nu}{\rho}}. \tag{A68}$$

Let us move on to the SCC with outgoing links only towards previously studied SCCs, that is the SCC composed by node 3. If this SCC had no outgoing links, then it would be an isolated urn, therefore with the same exponent of the other SCC studied (node 4), meaning that the actual solution for node 3 has that exponent and a logarithmic factor. Indeed, the dynamics of node 3 is the same derived for the couple of urns in Sec. 1.2, which is:

$$D_3(t) \underset{t\to\infty}{\approx} \frac{\nu+1}{\rho}(\rho-\nu)^{\frac{\nu}{\rho}} \ln(t)\, t^{\frac{\nu}{\rho}}. \tag{A69}$$

Proceeding with node 2, we compare its exponent if it was isolated to the maximum of the exponents of node 3 and 4, which are all the same. Moreover, since node 3 has a higher power in the logarithm than node 4, in the asymptotic solution, we can disregard the presence of the link 4 → 2. Thus, the solution for node 2 has another logarithmic factor and another constant multiplicative factor than those of node 3, that is we have the solution:

$$D_2(t) \underset{t\to\infty}{\approx} \left(\frac{\nu+1}{\rho}\right)^2 (\rho-\nu)^{\frac{\nu}{\rho}} \ln^2(t)\, t^{\frac{\nu}{\rho}}. \tag{A70}$$

To complete, similarly we obtain the solution for node 1, i.e.:

$$D_1(t) \underset{t\to\infty}{\approx} \left(\frac{\nu+1}{\rho}\right)^3 (\rho-\nu)^{\frac{\nu}{\rho}} \ln^3(t)\, t^{\frac{\nu}{\rho}}. \tag{A71}$$

We can hence see that the solutions are equal to those of the chain in Sec. 1.7.b, and there are only some slight differences at finite times due to the presence of another link, but not significantly.

Graph e—The last case to be examined is again similar to Graph c, but this time we swap the direction of the link between nodes 1 and 2 (see network representation in Table 5-B). Here the order with which we study the SCCs is inverted, because now only node 1 has no outer links. Therefore, the Heaps' law for node 1 is the classic individual one in Eq. (1.7). Then we have to solve the equations for the cycle, which in this case are:

$$\begin{cases} \dfrac{dD_2(t)}{dt} \approx \dfrac{\nu D_2(t)}{\rho t} + \dfrac{(\nu+1)D_3(t)}{\rho t} + \dfrac{(\nu+1)D_1(t)}{\rho t} \\ \dfrac{dD_3(t)}{dt} \approx \dfrac{\nu D_3(t)}{\rho t} + \dfrac{(\nu+1)D_4(t)}{\rho t} \\ \dfrac{dD_4(t)}{dt} \approx \dfrac{\nu D_4(t)}{\rho t} + \dfrac{(\nu+1)D_2(t)}{\rho t}. \end{cases} \tag{A72}$$

In this system, we can consider $D_1(t)$ known, working at large time-scales. Therefore, following the algorithm described in Sec. 1.6.2, we first solve this system without the external sources (i.e. node 1), in order to find the leading solution and then compare the exponents with the external sources ones. The solution of the associated homogeneous system is the same of a directed cycle as in Eq. (A26), i.e. a power-law function with exponent $2\nu + 1/\rho$. Now, we observe that the Heaps' exponent of the cycle is higher than the exponents of outer SCCs it is linked to, that is just node 1 with exponent ν/ρ. Then, the asymptotic solution for the nodes in the cycle corresponds to the solution of the cycle as if it had no outer links. Explicit solutions are given in Table 5-B.